Reversible Gromov-Monge Sampler for Simulation-Based Inference

YoonHaeng Hur, Wenxuan Guo, and Tengyuan Liang

University of Chicago

April 18, 2022

Abstract

This paper introduces a new simulation-based inference procedure to model and sample from multi-dimensional probability distributions given access to i.i.d. samples, circumventing the usual approaches of explicitly modeling the density function or designing Markov chain Monte Carlo. Motivated by the seminal work on distance and isomorphism between metric measure spaces, we propose a new notion called the Reversible Gromov-Monge (RGM) distance and study how RGM can be used to design new transform samplers to perform simulation-based inference. Our RGM sampler can also estimate optimal alignments between two heterogeneous metric measure spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ from empirical data sets, with estimated maps that approximately push forward one measure μ to the other ν, and vice versa. Analytic properties of the RGM distance are derived; statistical rate of convergence, representation, and optimization questions regarding the induced sampler are studied. Synthetic and real-world examples showcasing the effectiveness of the RGM sampler are also demonstrated.

Keywords— Gromov-Wasserstein metric, transform sampling, simulation-based inference, generative models, isomorphism, likelihood-free inference

1 Introduction

One of the central tasks in statistics is to model and sample from a multi-dimensional probability distribution. Classic statistics approaches this problem by fitting a model to the target distribution and then sampling from a fitted model via Markov Chain Monte Carlo (MCMC) techniques. Although such model-based methods are widely used, MCMC sampling often entails several technicalities. Beyond diagnosing whether the chain mixes, obtaining i.i.d. samples from MCMC methods is complex as one has to control correlations between successive samples or run parallel chains.

An alternative approach available in statistics, reserved for the one-dimensional case, is usually referred to as the (inverse) *transform sampling*. Such an approach circumvents the calling for a parametric or nonparametric density and directly designs a sampler by transforming a simple uniform distribution. The idea is simple: one can transform a uniform measure $\mu = \mathrm{Unif}([0,1])$ to any one-dimensional target probability measure ν leveraging the following monotonic transformation $T\colon [0,1] \to \mathbb{R}$ called the inverse Cumulative Distribution Function (CDF),

$$T(x) = \inf\{y \in \mathbb{R} \ : \ \nu((-\infty, y]) \geq x\} \ . \tag{1.1}$$

Define the pushforward measure $T_{\#}\mu$ by $T_{\#}\mu(S) = \mu(\{x : T(x) \in S\})$ for any Borel set $S \subseteq \mathbb{R}$, then one can easily check that $T_{\#}\mu = \nu$; namely, with a draw from the one-dimensional uniform distribution $x \sim \mu$, the transformed sample $T(x)$ has the target probability distribution ν.

Recently, the *transform sampling* idea has been extended to the multi-dimensional setting, as seen in both machine learning (specifically, generative modeling) and computational optimal transport. Again, given a target probability measure ν supported on $\mathcal{Y}$ and a user-specified probability measure μ—that is easy to sample from such as a multivariate Gaussian—defined on $\mathcal{X}$, we aim to find a measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ such that $T_{\#}\mu = \nu$, where $T_{\#}\mu$, the pushforward measure, is defined analogously to the one-dimensional case above. Such a map T, which is called a transport map from μ to ν, transforms i.i.d. samples from μ into i.i.d. samples from ν. Therefore, with a good estimate of the transformation T, the transform sampler operates and scales more efficiently than classic MCMC approaches.

Such transform sampling ideas have been leveraged in generative modeling by designing different criteria to learn a qualified transformation T; furthermore, remarkable empirical benchmarks have been documented. The essence of these methods can be summarized as follows. A transport map is obtained by minimizing $T \mapsto \mathcal{L}(T_{\#}\widehat{\mu}, \widehat{\nu})$ over $\mathcal{F}$, where $\mathcal{F}$ is a map class that is rich enough to contain a transport map, $\widehat{\mu}$ and $\widehat{\nu}$ are empirical measures based on samples from μ and ν, respectively, and $\mathcal{L}$ measures certain discrepancy of two distributions. In summary, by properly designing a class of maps $\mathcal{F}$ and collecting sufficiently many samples, we expect a minimizer T that will satisfy $T_{\#}\mu \approx \nu$. In Generative Adversarial Networks (GAN) [23], $\mathcal{F}$ consists of neural networks, and $\mathcal{L}$ is Jensen-Shannon divergence. Moreover, different choices of $\mathcal{L}$ have led to several variants: f-divergences for f-GAN [41], Wasserstein distances for Wasserstein-GAN [2], and Maximum Mean Discrepancies (MMD) for MMD-GAN [17, 31].

The Optimal Transport (OT) theory aims to identify the optimal transformation T, quantified by the transportation cost of moving mass from μ to ν. When μ and ν both lie in the same space, say $\mathbb{R}^d$, Brenier [9] proved, under mild regularity conditions, the following remarkable result that backs up the *transform sampling* in the multi-dimensional setting. Consider the Wasserstein-p distance $W_p(\mu, \nu)$ defined as

$$W_p(\mu,\nu) \coloneqq \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_2^p \,\mathrm{d}\gamma\,(x,y)\right)^{1/p},$$

where $\Pi(\mu,\nu)$ denotes the set of all couplings between μ and ν. Brenier established that for $p = 2$, there exists a unique optimal coupling $\gamma^\star$ and a unique optimal transport map $T^\star$ such that $\gamma^\star = (\mathrm{Id}, T^\star)_{\#}\mu$; more importantly, $T^\star$ is the gradient of some convex function. As a result,

$$W_2^2(\mu,\nu) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_2^2 \,\mathrm{d}\gamma^\star\,(x,y) = \int_{\mathbb{R}^d} \|x - T^\star(x)\|_2^2 \,\mathrm{d}\mu\,(x)\,.$$

Now let's contrast this result with the one-dimensional (inverse) transform sampling: when $\mu = \mathrm{Unif}([0,1])$, it turns out the inverse CDF map $T\colon [0,1] \to \mathbb{R}$ in (1.1) minimizes the transportation cost $W_p(\mu,\nu), p \geq 1$. Brenier's result significantly enriches the one-dimensional insight to the multi-dimensional case: now the multi-dimensional map $T\colon \mathbb{R}^d \to \mathbb{R}^d$ is the gradient of a convex function, as opposed to a monotonic map from $\mathbb{R}$ to $\mathbb{R}$.

Such an OT-based approach, however, can be inefficient in practice if the target ν is a high-dimensional embedding of some low-dimensional distribution. For instance, let ν be the distribution of handwritten digit images from the MNIST data set defined on $\mathbb{R}^{28\times 28} \equiv \mathbb{R}^{784}$.[1] To use the above OT-based approach, one must choose μ on $\mathbb{R}^{784}$ and find the map $T\colon \mathbb{R}^{784} \to \mathbb{R}^{784}$ discussed above. However, the support of ν is intrinsically low-dimensional (roughly $\mathbb{R}^{15}$ as in [18]), hence other transform samplers with $\mathcal{X} = \mathbb{R}^{15}$ yielding $T\colon \mathbb{R}^{15} \to \mathbb{R}^{784}$ are more efficient than the OT-based method in terms of estimating T and computing $T(X)$ for $X \sim \mu$.

Motivated by this limitation, we propose a transform sampler having the best of both worlds: it is underpinned by OT theory, at the same time, operates when $\mathcal{X}$ and $\mathcal{Y}$ are heterogeneous spaces. The key to our approach is to utilize the Gromov-Wasserstein (GW) distance between μ and ν, well-defined for heterogeneous spaces $\mathcal{X}$ and $\mathcal{Y}$. Given two continuous functions $c_{\mathcal{X}}\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and $c_{\mathcal{Y}}\colon \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$, the

[1] Each image is properly normalized and fit into a 28×28 pixel bounding box [29].

GW distance [37, 12] is

$$\mathrm{GW}(\mu,\nu) \coloneqq \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left(c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y')\right)^2 \mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y') \right)^{1/2}. \tag{1.2}$$

A few remarks regarding the comparison between Wasserstein and Gromov-Wasserstein are as follows. First, unlike Wasserstein which solves an infinite-dimensional linear program in the coupling γ, GW formulates a Quadratic Program (QP) in γ, which is known to be computationally hard [10, 35]. Second, GW aims to match the cost functions defined on two heterogeneous spaces, intending to identify isomorphism between spaces. Despite being an elegant notion of distance between metric measure spaces [37, Definition 5.1], GW is hard to compute in practice due to its QP nature; it is also unclear how to estimate $\mathrm{GW}(\mu,\nu)$ based on finite i.i.d. samples from μ and ν, and how accurate such estimates are.

Main contributions This paper considers computational and statistical questions regarding Gromov-Wasserstein outlined above, and aims to design a new transform sampler as an approach to model and sample from multi-dimensional probability distributions given access to i.i.d. samples, circumventing the usual ways of modeling the density function or MCMC. Our transform sampler can also estimate good alignments between two heterogeneous metric measure spaces $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ from empirical data sets, with estimated maps that approximately pushforward one measure μ to the other ν, and vice versa. Towards reaching these goals, we made the following specific contributions.

- We introduce a new notion, Reversible Gromov-Monge (RGM) distance, on metric measure spaces that majorizes the usual Gromov-Wasserstein distance. Furthermore, we show several analytic properties possessed by GW naturally carry over to RGM.
- Our RGM formulation induces a transform sampler, as a relaxation of the usual GW formulation. Rather than solving a QP which is quadratic in the coupling $\gamma \in \Pi(\mu,\nu)$, we decouple the pair as $(\mathrm{Id},F)_{\#}\mu$ and $(B,\mathrm{Id})_{\#}\nu$ with $F\colon \mathcal{X}\to\mathcal{Y}$ and $B\colon \mathcal{Y}\to\mathcal{X}$, respectively, and then bind them later via the constraint $(\mathrm{Id},F)_{\#}\mu \approx (B,\mathrm{Id})_{\#}\nu$. Such a decoupling and binding idea will prove suitable for the statistical estimation problem based on finite i.i.d. samples. We will also show, from an operator viewpoint, such a decoupling and binding idea ensures that our RGM is an infinite-dimensional convex program in F, B that admits a simple representation theorem, as opposed to the otherwise intractable infinite-dimensional QP in GW.
- We derive non-asymptotic rates of convergence for the proposed RGM sampler using tools from empirical processes, for generic classes modeling the measurable maps F and B. Based on our non-asymptotic results, concrete upper bounds can be easily spelled out in the cases where F and B are parametrized by deep neural networks. As mentioned earlier, the RGM sampler also promises to identify good alignments between metric measure spaces, and to learn approximate isomorphism when possible. We demonstrate such a point using numerical experiments on MNIST.

Organization The rest of the paper is organized as follows. First, we briefly review other related studies omitted in the discussion above. Then, in Section 2, preliminary background on optimal transport and Gromov-Wasserstein distance is outlined. Next, Section 3 summarizes the primary methodology and theory regarding our proposed Reversible Gromov-Monge sampler. Synthetic and real-world examples showcasing the effectiveness of the RGM sampler are demonstrated in Section 4 as a proof of concept. The supplementary material collects details of the results in Sections 3 and 4 along with extensive discussions.

1.1 Related Literature

Inferring the underlying probability distributions from data has been a central problem in statistics and the unsupervised machine learning since the invention of histograms by Pearson a century ago. Classic mathematical statistics explicitly models the density function in a parametric or a nonparametric way [46, 54],

and studies the minimax optimality of directly estimating such density functions [50]. It is also unclear how to proceed to sample from a possibly improper[2] density estimator, even with an optimal estimator at hand. One may employ Markov Chain Monte Carlo (MCMC) techniques for sampling from specific models. However, on the computational front, it is highly non-trivial how to ensure the mixing properties of MCMC for a designed sampler [44, Chapter 7].

Recent work in unsupervised machine learning proposes to learn complex, high-dimensional distributions via (deep) generative models, either explicitly by parametrizing the sufficient statistics of the exponential families [16, 27], or implicitly by parametrizing the pushforward map transporting distributions [17, 23], with a focus on tractability in computation. Surprisingly, though lacking theoretical underpinning and optimality, the generative models' approach performs well empirically in large-scale applications where classical statistical procedures are destined to fail. There has been a growing literature on understanding distribution estimation with the implicit framework, with more general metrics and target distribution classes, to name a few, [40, 31, 17] on MMDs, [49, 32] on integral probability metrics, and [39, 3, 33, 47, 4, 55, 30, 11] on generative adversarial networks. Last but not least, we emphasize that an alternative implicit distribution estimation approach using the simulated method of moments has been formulated in the econometrics literature since [36, 42] and [24].

Originally introduced as a tool for comparing objects in computer graphics, analytic properties of the Gromov-Wasserstein distance have been studied extensively [37, 51]; the most important one is that it defines a distance between metric measure spaces, namely, metric spaces endowed with probability measures. Since many real-world data sets can be modeled as metric measure spaces, the GW distance has been utilized in various problems such as shape correspondence [48], graph matching [57], and protein comparison [21]. Certain statistical aspects of comparing metric measure spaces have been studied in [8, 56].

Computation of the GW distance amounts to a relaxation of the quadratic assignment problem [28]; both are known to be NP-hard [10, 35] in the worst case. Several approaches have been proposed for the approximate computation of the GW distance. [37] studies lower bounds on the GW distance that are easier to compute. [43] adds an entropic regularization term to the GW distance, which leads to a fast iterative algorithm; [45] further modifies this by imposing a low-rank constraint on couplings. [52] proposes the Sliced Gromov-Wasserstein distance defined by integrating GW distances over one-dimensional projections. Last but not least, recent papers [57, 6, 13] study scalable partitioning schemes to approximately compute GW distances.

2 Background

In this section, we provide background on the Optimal Transport (OT) theory and the Gromov-Wasserstein distance. First, we start with some notations. Let $\|A\|$ denote the Frobenius norm of a matrix A and $\|x\|$ denote the Euclidean norm of a vector x. Given a set $\mathcal{X}$ and a function $f\colon \mathcal{X} \to \mathbb{R}$, let $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$ denote the sup norm. For an integer $n \in \mathbb{N}$, we define $[n] = \{1, \dots, n\}$. For a metric space $\mathcal{X}$, we denote its metric as $d_\mathcal{X}$ and write $\mathcal{P}(\mathcal{X})$ to denote the collection of all Borel probability measures on $\mathcal{X}$; we call $\mathcal{X}$ a Polish space if it is complete and separable. We call a pair $(\mathcal{X}, \mu)$ a Polish probability space if $\mathcal{X}$ is a Polish space and $\mu \in \mathcal{P}(\mathcal{X})$. Given two Polish probability spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$, the collection of all transport maps from μ to ν is denoted as $\mathcal{T}(\mu, \nu) \coloneqq \{T\colon \mathcal{X} \to \mathcal{Y} \mid T_\# \mu = \nu\}$; we call $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ a coupling between μ and ν if $\gamma(A \times \mathcal{Y}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$ for all Borel subsets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$, and we denote the collection of all such couplings as $\Pi(\mu, \nu)$. For a sequence of numbers $a(n), b(n) \in \mathbb{R}$, we use $a(n) \precsim b(n)$ to denote the relationship that $a(n)/b(n) \leq C, \forall n$ with some universal constant $C > 0$.

2.1 A Brief Overview of Optimal Transport Theory

A major goal of OT is minimizing the cost associated with the transport map between two Polish probability spaces, say $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$. Consider a measurable function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$; we view $c(x, y)$ as the cost

[2] Here we mean that the estimated density is not always non-negative and integrates to one.

associated with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. For each transport map $T \in \mathcal{T}(\mu, \nu)$, we interpret $c(x, T(x))$ as a unit cost incurred by mapping each $x \in \mathcal{X}$ to $T(x) \in \mathcal{Y}$. We define the average cost incurred by the transport map T as the integration of all the unit costs with respect to μ:

$$\int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu(x) \, .$$

Minimizing the cost over $\mathcal{T}(\mu, \nu)$ is referred to as the Monge problem named after Gaspard Monge. If there exists a minimizer $T^\star$ to this problem, that is,

$$T^\star \in \operatorname*{arg\,min}_{T \in \mathcal{T}(\mu,\nu)} \int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu(x) \, ,$$

we call $T^\star$ an optimal transport map.

Another important OT problem is minimizing the cost given by couplings. We define the average cost incurred by a coupling $\gamma \in \Pi(\mu, \nu)$ as the integration of the cost $c(x, y)$ with respect to γ:

$$\int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \, .$$

Minimizing this cost over $\Pi(\mu, \nu)$ is called the Kantorovich problem credited to Leonid Kantorovich. If

$$\gamma^\star \in \operatorname*{arg\,min}_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \, ,$$

we call $\gamma^\star$ an optimal coupling.

The two OT problems are closely related: the Kantorovich problem is a relaxation of the Monge problem. To see this, for each $T \in \mathcal{T}(\mu, \nu)$, define a map $(\mathrm{Id}, T) \colon \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$ by $(\mathrm{Id}, T)(x) = (x, T(x))$. One can verify $(\mathrm{Id}, T)_{\#}\mu \in \Pi(\mu, \nu)$. Therefore, if we define $\Pi_{\mathcal{T}} \coloneqq \{(\mathrm{Id}, T)_{\#}\mu : T \in \mathcal{T}(\mu, \nu)\}$, then $\Pi_{\mathcal{T}} \subset \Pi(\mu, \nu)$ and thus

$$\inf_{T \in \mathcal{T}(\mu,\nu)} \int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu(x) = \inf_{\gamma \in \Pi_{\mathcal{T}}} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \geq \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \, ,$$

where the first equality follows from change-of-variables. In other words, two OT problems share the same objective function as a function of couplings; however, the Kantorovich problem has a larger constraint set.

Unlike the Monge problem, the Kantorovich problem has favorable properties. First, the objective function is linear in γ. Moreover, $\Pi(\mu, \nu)$ is compact in the weak topology of Borel probability measures defined on $\mathcal{X} \times \mathcal{Y}$. This suggests that we can view the Kantorovich problem as an infinite-dimensional linear program.

Besides seeking optimal transport maps or couplings, another interesting aspect of OT problems is that the least possible cost can endow a metric structure among Polish probability spaces. If $\mathcal{X} = \mathcal{Y}$ and $c = d_{\mathcal{X}}^2$, the square root of the solution of the Kantorovich problem defines a distance between μ and ν, known as the Wasserstein distance.

Definition 1. *Given a Polish space $\mathcal{X}$, the Wasserstein-2 distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined as*

$$W_2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}^2(x, y) \, \mathrm{d}\gamma(x, y) \right)^{1/2} .$$

Remark. *One can define the Wasserstein-p distance by replacing the exponent 2 above with $p \in [1, \infty]$. The Wasserstein-p distance is known to satisfy the usual metric axioms.*

2.2 Gromov-Wasserstein and Gromov-Monge Distances

Although OT problems can be defined between arbitrary Polish probability spaces, in practice, it is unclear how to design a function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ to represent meaningful cost associated with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ in two heterogeneous spaces. For instance, if $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$ with $p \neq q$, there is no simple choice for a cost function c over $\mathbb{R}^p \times \mathbb{R}^q$. As a result, classic OT theory (including Brenier's result) cannot be directly used for comparing heterogeneous Polish probability spaces.

Mémoli's pioneering work [37] resolved this issue by considering a quadratic objective function of γ:

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma\,(x,y) \Rightarrow \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2\,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')\ ,$$

where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are defined over $\mathcal{X}\times\mathcal{X}$ and $\mathcal{Y}\times\mathcal{Y}$, respectively. For instance, one can specify $c_{\mathcal{X}} = d_{\mathcal{X}}$ and $c_{\mathcal{Y}} = d_{\mathcal{Y}}$. Rather than considering a unit cost corresponding to each pair $(x,y) \in \mathcal{X}\times\mathcal{Y}$, two pairs (x,y) and (x',y') in $\mathcal{X}\times\mathcal{Y}$ are associated with the discrepancy of intra-space quantities $c_{\mathcal{X}}(x,x')$ and $c_{\mathcal{Y}}(y,y')$. In summary, by switching from the integration $\mathrm{d}\gamma$ to the double integration $\mathrm{d}\gamma\,\mathrm{d}\gamma$, we no longer need an otherwise inter-space quantity $c\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$. Therefore, we can always define this objective function whenever we have proper $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ in each individual space, which leads to the following definition.

Definition 2. *A triple $(\mathcal{X},\mu,c_{\mathcal{X}})$ is called a network space if $(\mathcal{X},\mu)$ is a Polish probability space such that $\mathrm{supp}(\mu) = \mathcal{X}$ and $c_{\mathcal{X}}\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is continuous. The Gromov-Wasserstein distance between network spaces $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ is defined as*

$$\mathrm{GW}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2\,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')\right)^{1/2}.$$

Remark. On top of the network space definition introduced in [12], we impose continuity of $c_{\mathcal{X}}$ for a cleaner analysis. A network space $(\mathcal{X},\mu,c_{\mathcal{X}})$ is called a metric measure space if $c_{\mathcal{X}} = d_{\mathcal{X}}$ as introduced in [37] and [51]. In short, a network space is a generalization of a metric measure space.

Like the Wasserstein distance, the GW distance has metric properties; it satisfies symmetry and the triangle inequality, and $\mathrm{GW}(\mu,\nu) = 0$ if $(\mathcal{X},\mu,c_{\mathcal{X}}) = (\mathcal{Y},\nu,c_{\mathcal{Y}})$. However, the converse of this last statement does not hold in general: for its validity, a suitable equivalence relation needs to be defined on the collection of network spaces.

Definition 3. *Network spaces $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ are strongly isomorphic if there exists $T \in \mathcal{T}(\mu,\nu)$ such that $T\colon \mathcal{X} \to \mathcal{Y}$ is bijective and $c_{\mathcal{X}}(x,x') = c_{\mathcal{Y}}(T(x),T(x'))$ for all $x,x' \in \mathcal{X}$. In this case, we write $(\mathcal{X},\mu,c_{\mathcal{X}}) \cong (\mathcal{Y},\nu,c_{\mathcal{Y}})$ and such a transport map T is called a strong isomorphism.*

One can easily check that $\cong$ is indeed an equivalence relation on the collection of network spaces. The following theorem states that the GW distance satisfies all metric axioms on the quotient space—under the equivalence relation $\cong$—of metric measure spaces.

Theorem 1 (Lemma 1.10 of [51])**.** *Let $\mathcal{M}$ be the collection of all network spaces $(\mathcal{X},\mu,c_{\mathcal{X}})$ such that $c_{\mathcal{X}} = d_{\mathcal{X}}$. Also, let $\mathcal{M}/_{\cong}$ be the collection of all equivalence classes of $\mathcal{M}$ induced by $\cong$. Then, GW satisfies the three metric axioms on $\mathcal{M}/_{\cong}$.*

Recall that the Monge problem is a restricted version of the Kantorovich problem with an additional constraint that couplings are given by a transport map; replacing $\Pi(\mu,\nu)$ in the Kantorovich problem with $\Pi_{\mathcal{T}}$ yields the Monge problem. Imposing the same constraint on the definition of GW leads to the Gromov-Monge distance.

Definition 4. *The Gromov-Monge distance between network spaces $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ is defined as*

$$\mathrm{GM}(\mu,\nu) = \inf_{T\in\mathcal{T}(\mu,\nu)} \left(\int_{\mathcal{X}}\int_{\mathcal{X}} (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(T(x),T(x')))^2\,\mathrm{d}\mu\,(x)\,\mathrm{d}\mu\,(x')\right)^{1/2}.$$

Loosely speaking, computing GM amounts to finding a transport map T such that $c_{\mathcal{X}}(x, x')$ best matches $c_{\mathcal{Y}}(T(x), T(x'))$ on average; we can view such a map T as a surrogate for an isomorphism. See Section 3 of [38] for more details of GM.

3 Summary of Results

Inspired by the Gromov-Wasserstein and Gromov-Monge distances, we propose a new metric—the reversible Gromov-Monge distance—between network spaces in this paper. Our formulation seeks a pair of transport maps $F \in \mathcal{T}(\mu, \nu)$ and $B \in \mathcal{T}(\nu, \mu)$ best approximating isomorphic relations between network spaces. We propose a novel transform sampling method that uses F as a push-forward map to obtain i.i.d. samples from a target distribution ν. We present two optimization formulations solving for such a pair (F, B) in order: a potentially non-convex formulation that employs the standard gradient descent method to optimize, and an infinite-dimensional convex formulation where global optima can be found efficiently. For the former, we analyze the statistical rate of convergence for generic classes $\mathcal{F} \times \mathcal{B}$ parametrizing (F, B). For the latter, we derive a new representer theorem on a suitable reproducing kernel Hilbert space (RKHS).

3.1 Metric Properties of Reversible Gromov-Monge

Our formulation is based on the following observation: for a coupling γ such that $\gamma = (\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$, which presents a binding constraint, we can simplify the objective function of GW as

$$\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \, \mathrm{d}\mu \otimes \nu \ ,$$

where $\mathrm{d}\mu \otimes \nu \coloneqq \mathrm{d}\mu\,(x)\,\mathrm{d}\nu\,(y)$ denotes the product measure of μ and ν. Imposing the binding constraint on the definition of GW leads to the following definition.

Definition 5. *For network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$, we write $(F, B) \in \mathcal{I}(\mu, \nu)$ if measurable maps $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$ satisfy the binding constraint $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$. We define the reversible Gromov-Monge (RGM) distance between $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ as*

$$\mathrm{RGM}(\mu, \nu) \coloneqq \inf_{(F,B)\in\mathcal{I}(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \, \mathrm{d}\mu \otimes \nu \right)^{1/2} . \tag{3.1}$$

Remark. A few remarks are in place for the binding constraint. If $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$, then $F_{\#}\mu = \nu$ and $B_{\#}\nu = \mu$ follow due to marginal conditions. However, the converse is not true in general. To see this, let $\mu = \nu = \mathrm{Unif}([0, 1])$, then $F_{\#}\mu = \nu$ and $B_{\#}\nu = \mu$ hold for $F(x) = B(x) = |2x - 1|$. However, $(\mathrm{Id}, F)_{\#}\mu \neq (B, \mathrm{Id})_{\#}\nu$ because $(\mathrm{Id}, F)_{\#}\mu$ is a uniform measure on $\{(x, |2x - 1|) : x \in [0, 1]\}$, whereas $(B, \mathrm{Id})_{\#}\nu$ is a uniform measure on $\{(|2y - 1|, y) : y \in [0, 1]\}$.

Roughly speaking, computing RGM consists in finding a pair $(F, B) \in \mathcal{I}(\mu, \nu)$ such that $c_{\mathcal{X}}(x, B(y))$ best matches $c_{\mathcal{Y}}(F(x), y)$ on average. Like a strong isomorphism, we can view such a pair as jointly capturing an isomorphic relation of $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$. We will use this observation later to build a transform sampling method.

We will prove that RGM possesses metric properties similar to the Gromov-Wasserstein. Motivated by Theorem 1, we derive the following result.

Theorem 2. *Let $h\colon \mathbb{R}_+ \to \mathbb{R}$ be a continuous and strictly monotone function and $\mathcal{N}^h$ be a collection of all network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ such that $c_{\mathcal{X}} = h(d_{\mathcal{X}})$. Then RGM satisfies the three metric axioms on $\mathcal{N}^h/_{\cong}$ which is the collection of all equivalence classes of $\mathcal{N}^h$ induced by $\cong$.*

Remark. Suppose $\mathcal{X}$ is a Euclidean space and $d_{\mathcal{X}}$ is the standard Euclidean distance. If $h(x) = \exp\left(-\alpha x^2\right)$ with $\alpha > 0$, then $h(d_{\mathcal{X}})$ is the radial basis function (RBF) kernel on $\mathcal{X}$; we will use this in numerical experiments.

We refer the proof of Theorem 2 and details of the properties of RGM to Section A.

3.2 Transform Sampling via RGM

With the proposed notion of RGM, we design a transform sampling method in this section. The transform sampler is based on finding a minimizing pair (F, B) of RGM, which can capture isomorphic relations between network spaces. To implement this method, we need to estimate (F, B) using only i.i.d. samples from μ and ν. Leveraging the Lagrangian form, we derive a minimization problem that can be implemented based on finite samples.

First, we rewrite the population minimization problem with the binding constraint as follows,

$$\begin{aligned}\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y}\\ B\colon \mathcal{Y}\to\mathcal{X}}} \quad & \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \,\mathrm{d}\mu\otimes\nu \\ \text{s.t.} \quad & \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu) = 0 \,.\end{aligned}$$

Here, $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ is a suitable discrepancy measure on $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ so that $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu) = 0$ is a surrogate for the original constraint $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$. In practice, we do not require that $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}} = 0$ implies $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$; in fact, the former constraint can be a relaxation of the latter. The choice of $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ will be specified later. To solve this minimization problem, we propose utilizing the Lagrangian:

$$\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y}\\ B\colon \mathcal{Y}\to\mathcal{X}}} \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \,\mathrm{d}\mu\otimes\nu + \lambda\cdot\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu) \,.$$

Given i.i.d. samples $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ from μ and ν, respectively, we replace the population objective with its empirical estimates:

$$\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y}\\ B\colon \mathcal{Y}\to\mathcal{X}}} \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n (c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 + \lambda\cdot\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) \,,$$

where $\widehat{\mu}_m$ and $\widehat{\nu}_n$ are the empirical measures based on $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$, respectively. Empirically, we find that adding the following extra terms often enhance empirical results:

$$\begin{aligned}\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y}\\ B\colon \mathcal{Y}\to\mathcal{X}}} \quad & \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n (c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 + \lambda_1\cdot\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) \\ & + \lambda_2\cdot\mathcal{L}_{\mathcal{X}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3\cdot\mathcal{L}_{\mathcal{Y}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n) \,.\end{aligned}$$

Like $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$, we utilize suitable discrepancy measures $\mathcal{L}_{\mathcal{X}}$ and $\mathcal{L}_{\mathcal{Y}}$ so that these additional terms help matching the marginals of $(\mathrm{Id}, F)_{\#}\widehat{\mu}_m$ and $(B, \mathrm{Id})_{\#}\widehat{\nu}_n$.

Lastly, we discuss the choice of $\mathcal{L}_{\mathcal{X}}, \mathcal{L}_{\mathcal{Y}}$, and $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$. We use the square of Maximum Mean Discrepancy (MMD) as the leading example.[3] For the kernel $K_{\mathcal{X}\times\mathcal{Y}}$ on the product space, we use the tensor product kernel $K_{\mathcal{X}}\otimes K_{\mathcal{Y}}$ given as

$$K_{\mathcal{X}}\otimes K_{\mathcal{Y}}((x, y), (x', y')) = K_{\mathcal{X}}(x, x')K_{\mathcal{Y}}(y, y') \,.$$

[3]This is merely a proof of concept. One may use other quantities in practice, described in Section 4. MMD between two measures is a distance between their embeddings in some reproducing kernel Hilbert space (RKHS), which is indeed a metric under mild conditions [40]. Also, MMD is representable via the reproducing kernel of the RKHS, hence one may simply choose a kernel function to define it. Concretely, for any kernel $K_{\mathcal{X}}$ on $\mathcal{X}$, the square of MMD between $\widehat{\mu}_m$ and $B_{\#}\widehat{\nu}_n$ is

$$\frac{1}{m^2}\sum_{i,i'} K_{\mathcal{X}}(x_i, x_{i'}) + \frac{1}{n^2}\sum_{j,j'} K_{\mathcal{X}}(B(y_j), B(y_{j'})) - \frac{2}{mn}\sum_{i,j} K_{\mathcal{X}}(x_i, B(y_j)) \,.$$

To utilize such a convenient closed form, we specify $\mathcal{L}_{\mathcal{X}}, \mathcal{L}_{\mathcal{Y}}, \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ as the square of corresponding MMDs by choosing kernels $K_{\mathcal{X}}, K_{\mathcal{Y}}, K_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X}, \mathcal{Y}, \mathcal{X}\times\mathcal{Y}$.

The tensor product notation is employed since the kernel on the product space inherits the feature map as the tensor product of two individual feature maps w.r.t. $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$.

Denoting the MMD associated with a kernel K as MMD_K, we obtain the following minimization problem:

$$\begin{aligned}\min_{\substack{F\colon \mathcal{X}\to\mathcal{Y}\\ B\colon \mathcal{Y}\to\mathcal{X}}} \quad & \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 \\ & + \lambda_1 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) \\ & + \lambda_2 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3 \cdot \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n)\ .\end{aligned} \tag{3.2}$$

Once we solve the problem above, the solution $\widehat{F}\colon \mathcal{X}\to\mathcal{Y}$ will serve as an approximate isomorphism and facilitate transform sampling of the target ν from a known distribution μ. The map $\widehat{B}$ possesses similar properties as $\widehat{F}$, whereas the map $\widehat{F}$ is of our primary interest for sampling purposes. The reverse map $\widehat{B}\colon \mathcal{Y}\to\mathcal{X}$ also embeds point clouds in $\mathcal{Y}$ into $\mathcal{X}$, with approximate isomorphism properties in the sense of Gromov-Monge.

3.3 Statistical Rate of Convergence

Like other transform sampling approaches for generative models, we consider (3.2) using vector-valued function classes $\mathcal{F}$ and $\mathcal{B}$ parametrized by neural networks, and then optimize using a gradient descent algorithm. We emphasize this minimization problem is much simpler than adversarial formulations as in GANs: variational problems of GANs consist of minimization over a class of generators and maximization over a class of discriminators, which requires complex saddle-point dynamics [14, 34]. In contrast, our RGM only solves a single minimization problem in network parameters. Although generally non-convex in nature, the parameter minimization problem in neural networks can often be efficiently optimized by stochastic gradient descent, and can even provably achieve the global optima if the loss satisfies certain Polyak-Łojasiewicz conditions [5].

We investigate the statistical rate of convergence for this minimization problem, assuming the empirical problem (3.2) can be solved accurately. First, define

$$\begin{aligned}C(\mu,\nu,F,B) \coloneqq & \int (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \,\mathrm{d}\mu\otimes\nu \\ & + \lambda_1 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu) \\ & + \lambda_2 \cdot \mathrm{MMD}^2_{K_{\mathcal{X}}}(\mu, B_{\#}\nu) + \lambda_3 \cdot \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\mu, \nu)\ .\end{aligned} \tag{3.3}$$

Then, the objective function of (3.2) is a plug-in estimator $C(\widehat{\mu}_m, \widehat{\nu}_n, F, B)$. We consider solving (3.2) over the transformation class $\mathcal{F}\times\mathcal{B}$ given as follows, for which we will state our non-asymptotic results in full generality. From now on, let $\mathcal{X}$ and $\mathcal{Y}$ be subsets of Euclidean spaces of dimensions $\dim(\mathcal{X})$ and $\dim(\mathcal{Y})$, respectively. $\mathcal{F}$ (resp. $\mathcal{B}$) is a collection of vector-valued measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ (resp. from $\mathcal{Y}$ to $\mathcal{X}$). For each $F\in\mathcal{F}$ and $k\in[\dim(\mathcal{Y})]$, we write $F_k(x)$ to denote the k-th coordinate of $F(x)$. Accordingly, we define $\mathcal{F}_k = \{F_k : \mathcal{X}\to\mathbb{R} \mid F\in\mathcal{F}\}$, namely, a collection of real-valued measurable functions defined on $\mathcal{X}$ that are given as the k-th coordinate of $F\in\mathcal{F}$. For $\ell\in[\dim(\mathcal{X})]$, we define B_ℓ and $\mathcal{B}_\ell = \{B_\ell : \mathcal{Y}\to\mathbb{R} \mid B\in\mathcal{B}\}$ analogously.

Then, solving (3.2) over $\mathcal{F}\times\mathcal{B}$ is written as

$$\min_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\widehat{\mu}_m, \widehat{\nu}_n, F, B)\ .$$

We prove that the empirical solution leads to an approximate infimum of $(F,B)\mapsto C(\mu,\nu,F,B)$ evaluated with the population measures μ,ν, with sufficiently large sample sizes m and n.

Overview of assumptions Before stating the next theorem, we present an overview of the assumptions. The complete statement of the assumptions and key definitions are designated to Sections B-C in the supplementary material due to space constraints. Assumptions 1 and 4 require the boundedness and Lipschitzness of the cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$. Similarly, boundedness and Lipschitzness of the kernel functions $K_{\mathcal{X}}, K_{\mathcal{Y}}$ corresponding to the MMD term are stated in Assumptions 2 and 5, respectively. The last two assumptions are imposed on the set of transformations $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$: Assumption 3 requires the transformation class is bounded, and Assumption 6 states that the classes should contain non-trivial maps. We shall employ a notion of combinatorial dimension to measure the complexity of real-valued function classes—the pseudo-dimension—formally stated in Definition 7.

Theorem 3. *Let $(\widehat{F}, \widehat{B})$ be a solution to the empirical RGM problem*

$$(\widehat{F}, \widehat{B}) \in \operatorname*{arg\,min}_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\widehat{\mu}_m, \widehat{\nu}_n, F, B) \ ,$$

with $C\colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \times \mathcal{F} \times \mathcal{B} \to \mathbb{R}$ defined in (3.3). Under Assumptions 1-6, the following inequality holds with probability $1-\delta$ on $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$

$$C(\mu, \nu, \widehat{F}, \widehat{B}) - \inf_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\mu, \nu, F, B) \precsim \mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta) \ . \tag{3.4}$$

Here, $\mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta)$ denotes a complexity measure of $(\mathcal{F}, \mathcal{B})$ given in terms of pseudo-dimensions (Pdim) of $\mathcal{F}_k$ and $\mathcal{B}_\ell$ defined in Definition 7:

$$\mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta) := \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}} + \sqrt{\frac{\log(m\vee n)}{m\wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) + \sum_{\ell=1}^{\dim(\mathcal{X})} \mathrm{Pdim}(\mathcal{B}_\ell)\right)} \ .$$

We provide required assumptions and the full proof of Theorem 3 in Section B along with the definition of the pseudo-dimension (Definition 7). When $\mathcal{F}$ and $\mathcal{B}$ are parametrized by neural network classes (the ones we will use for numerical demonstrations in Section 4), tight pseudo-dimension bounds established in [1, 25] can be plugged in Theorem 3 for concrete non-asymptotic rates.

3.4 Convex Formulation and Representer Theorem

As the last bit of our contributions, we study a convex formulation of solving (3.2) by relaxing and lifting it to an infinite-dimensional space. There are two reasons behind our convex formulation: first, as a computational alternative to the possibly non-convex optimization; second, to point out a connection with the Nadaraya-Watson estimator in classic nonparametric statistics. The crux lies in relaxing optimizing over the map $F\colon \mathcal{X} \to \mathcal{Y}$ to optimizing over its induced (dual) linear operator $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ that maps functions on $\mathcal{Y}$ to functions on $\mathcal{X}$, where $L^2_{\mathcal{X}}$ is the collection of real-valued measurable functions f defined on $\mathcal{X}$ such that $\int_{\mathcal{X}} f^2 \,\mathrm{d}\pi_{\mathcal{X}} < \infty$ given a Borel measure $\pi_{\mathcal{X}}$ on $\mathcal{X}$; similarly, define $L^2_{\mathcal{Y}}$ given a Borel measure $\pi_{\mathcal{Y}}$ on $\mathcal{Y}$. Then, for a measurable map $F\colon \mathcal{X} \to \mathcal{Y}$, we can define $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ by letting $\mathbf{F}(g) = g \circ F$ for all $g \in L^2_{\mathcal{Y}}$. Similarly, we define $\mathbf{B}\colon L^2_{\mathcal{X}} \to L^2_{\mathcal{Y}}$ for each measurable map $B\colon \mathcal{Y} \to \mathcal{X}$. We will see $\mathbf{F}$ and $\mathbf{B}$ are well-defined bounded linear operators in Section C under a mild assumption.

To state the representer theorem, consider (3.2) with $c_{\mathcal{X}} = K_{\mathcal{X}}$ and $c_{\mathcal{Y}} = K_{\mathcal{Y}}$, same as kernel functions specified in MMD terms. We show that this problem can be reduced to a finite-dimensional convex optimization by proving a representer theorem. Since finite-dimensional convex optimization can be optimized globally with provable guarantees, such a formulation can be solved numerically in an efficient way.

Let us lay out more details to state the result. Due to Mercer's theorem, let $\{\phi_k \in L^2_{\mathcal{X}}\}_{k\in\mathbb{N}}$ and $\{\psi_\ell \in L^2_{\mathcal{Y}}\}_{\ell\in\mathbb{N}}$ be countable orthonormal bases of $L^2_{\mathcal{X}}$ and $L^2_{\mathcal{Y}}$ where the kernels admit the following spectral decompositions:

$$K_{\mathcal{X}}(x, x') = \sum_k \lambda_k \phi_k(x)\phi_k(x') \ , \quad K_{\mathcal{Y}}(y, y') = \sum_\ell \gamma_\ell \psi_\ell(y)\psi_\ell(y') \ , \tag{3.5}$$

with positive eigenvalues $\lambda_k, \gamma_\ell > 0$. Since $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ defines a bounded linear operator, one can represent $\mathbf{F}$ (correspondingly $\mathbf{B}$) under the orthonormal bases

$$\mathbf{F}[\psi_\ell] = \sum_{k=1}^{\infty} \mathbf{F}_{k\ell}\phi_k \ , \quad \mathbf{B}[\phi_k] = \sum_{\ell=1}^{\infty} \mathbf{B}_{\ell k}\psi_\ell \ . \tag{3.6}$$

Here, $[\mathbf{F}_{k\ell}]$ is a semi-infinite matrix with each column describing the $L^2_{\mathcal{X}}$ representation of $\mathbf{F}[\psi_\ell]$ under the basis $\{\phi_k \in L^2_{\mathcal{X}}\}_{k\in\mathbb{N}}$. With a slight abuse of notation, we will write $\mathbf{F}$ and $\mathbf{B}$ to denote these matrices $[\mathbf{F}_{k\ell}]$ and $[\mathbf{B}_{\ell k}]$. Then, we will prove in Section C that the objective function in (3.2) with $c_{\mathcal{X}} = K_{\mathcal{X}}$ and $c_{\mathcal{Y}} = K_{\mathcal{Y}}$ is

$$\begin{aligned}
\Omega(\mathbf{F},\mathbf{B}) \coloneqq \frac{1}{mn}\sum_{i,j}(\Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} - \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j})^2 \\
+ \lambda_1 \cdot \left(\frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} \right. \\
\left. - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i}\Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \right) \\
+ \lambda_2 \cdot \left(\frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \right) \\
+ \lambda_3 \cdot \left(\frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \right) \ .
\end{aligned}$$

Here, $\mathbf{F}$ and $\mathbf{B}$ are the matrices denoting the operators induced by F and B, respectively, $\Phi_x = [\cdots, \phi_k(x), \cdots]^\top \in \mathbb{R}^\infty$ and $\Psi_y = [\cdots, \psi_\ell(y), \cdots]^\top \in \mathbb{R}^\infty$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots)$ and $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \dots)$ are diagonal matrices. Hence, (3.2) can be lifted to an infinite-dimensional optimization problem

$$\min_{(\mathbf{F},\mathbf{B})\in\mathcal{C}} \ \Omega(\mathbf{F},\mathbf{B}) \ , \tag{3.7}$$

where $\mathcal{C}$ denotes the constraint set implying that $\mathbf{F}$ and $\mathbf{B}$ are matrices corresponding to bounded linear operators induced by some maps $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$.

We will relax this problem by removing the constraint set $\mathcal{C}$, namely, by considering all matrices in $\mathbb{R}^{\infty\times\infty}$ as the decision variables,

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}} \ \Omega(\mathbf{F},\mathbf{B}) \ . \tag{3.8}$$

In other words, this relaxed problem minimizes Ω over any pair of infinite-dimensional matrices. The next result, which we refer to as the representer theorem, shows that (3.8) boils down to a finite-dimensional convex program.

Theorem 4. *Consider the optimization (3.7) under the assumptions in Proposition 10. Then, for any minimizer $(\mathbf{F}^\star, \mathbf{B}^\star)$ to the relaxed problem (3.8), we can find finite-dimensional matrices $\mathsf{F}^\star_{m,n} \in \mathbb{R}^{m\times n}$ and $\mathsf{B}^\star_{n,m} \in \mathbb{R}^{n\times m}$ such that*

$$\mathbf{F}^\star = \Lambda\Phi_m \mathsf{F}^\star_{m,n} \Psi_n^\top \ , \quad \mathbf{B}^\star = \Gamma\Psi_n \mathsf{B}^\star_{n,m} \Phi_m^\top \ ,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots)$, $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \dots)$, and $\Phi_m \in \mathbb{R}^{\infty\times m}$ and $\Psi_n \in \mathbb{R}^{\infty\times n}$ are matrices whose elements are $\phi_k(x_i)$ and $\psi_\ell(y_j)$, as defined in (3.5). In this case, $\Omega(\mathbf{F}^\star, \mathbf{B}^\star)$ can be rewritten as $\omega(\mathsf{F}^\star_{m,n}, \mathsf{B}^\star_{n,m})$

for some convex function ω defined over $\mathbb{R}^{m\times n}\times\mathbb{R}^{n\times m}$. Hence, by minimizing ω over $\mathbb{R}^{m\times n}\times\mathbb{R}^{n\times m}$, we obtain a relaxation of (3.8), that is,

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}}\Omega(\mathbf{F},\mathbf{B})\geq\min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}}\omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\ .$$

In particular, the RHS is a finite-dimensional convex optimization. Lastly, this relaxation is tight, that is,

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}}\Omega(\mathbf{F},\mathbf{B})=\min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}}\omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\ ,$$

if kernel matrices $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$ whose elements are $K_{\mathcal{X}}(x_i,x_{i'})$ and $K_{\mathcal{Y}}(y_j,y_{j'})$, are positive definite.

Remark. Looking inside the proof of Theorem 4, we know the solution to the infinite-dimensional optimization is an operator taking form of $\mathbf{F}^\star=\Lambda\Phi_m\mathsf{F}^\star_{m,n}\Psi_n^\top$, with a finite-dimensional matrix $\mathsf{F}^\star_{m,n}\in\mathbb{R}^{m\times n}$. Therefore, for any $g\in L^2_{\mathcal{Y}}$, we can deduce

$$\mathbf{F}^\star[g](x)=\underbrace{K_{\mathcal{X}}(x,X_m)}_{1\times m}\underbrace{\mathsf{F}^\star_{m,n}}_{m\times n}\underbrace{g(Y_n)}_{n\times 1}\ , \tag{3.9}$$

where $K_{\mathcal{X}}(x,X_m)$ maps each $x\in\mathcal{X}$ to a row vector whose i-th element is $K_{\mathcal{X}}(x,x_i)$ and $g(Y_n)$ denotes a column vector whose j-th element is $g(y_j)$.

Now let's draw a connection between the classic Nadaraya-Watson estimator and (3.9). For now consider a special case: (x_i,y_i)'s are paired with $m=n$. In such a case, Nadaraya-Watson estimator takes the form

$$\sum_{i,j}K_{\mathcal{X}}(x,x_i)\cdot\tfrac{1}{m}\delta_{i=j}\cdot g(y_j)\ ; \tag{3.10}$$

Namely, for a new point x, the corresponding function value $g(y)$ evaluated on its coupled $y=F(x)$ is a weighted average of $g(y_j)$'s according to the affinity $K_{\mathcal{X}}(x,x_i)$. Our solution (3.9) extends the above nonparametric smoothing idea to the decoupled data case, where the coupling weights $\mathsf{F}^\star_{m,n}$ is based on a solution to a convex program, with

$$(3.9)=\sum_{i,j}K_{\mathcal{X}}(x,x_i)\cdot\mathsf{F}^\star_{m,n}[i,j]\cdot g(y_j)\ . \tag{3.11}$$

Lastly, we draw another connection to the Monte-Carlo integration. One downstream task after learning the distribution ν is to perform numerical integration of $g\in L^2_{\mathcal{Y}}$ under the measure $\nu\in\mathcal{P}(\mathcal{Y})$. In our transform sampling framework, this amounts to evaluate $\mathbb{E}_{y\sim F^\star_\#\mu}[g(y)]=\mathbb{E}_{x\sim\mu}[g\circ F^\star(x)]$. The integration, casted in the induced operator form, has the expression

$$\mathop{\mathbb{E}}_{x\sim\mu}\big[\mathbf{F}^\star[g](x)\big]=\mathop{\mathbb{E}}_{x\sim\mu}\big[\underbrace{K_{\mathcal{X}}(x,X_m)\mathsf{F}^\star_{m,n}}_{=:W(x)\in\mathbb{R}^n}g(Y_n)\big]=\mathop{\mathbb{E}}_{x\sim\mu}\big[\sum_{j=1}^n W_j(x)g(y_j)\big] \tag{3.12}$$

where $W(x)$ can be interpreted as the importance weights in the Monte-Carlo integration. We conclude with one more remark: if plug in instead $x\sim\widehat{\mu}_m$ in (3.12), one can verify that under mild conditions,

$$\mathop{\mathbb{E}}_{x\sim\widehat{\mu}_m}\big[\mathbf{F}^\star[g](x)\big]=\frac{1}{n}\sum_{j=1}^n g(y_j)\ . \tag{3.13}$$

In other words, with the empirical measure as input, (3.12) outputs the simple sample average.

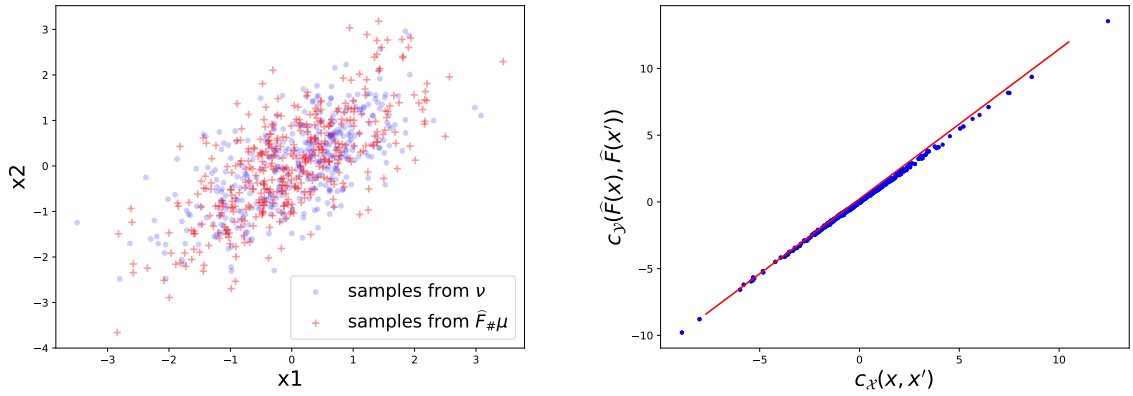
4 Experiments

This section examines the empirical performance of the reversible Gromov-Monge sampler. Following Section 3.3, we find a minimum $(\widehat{F}, \widehat{B})$ of (3.2) over a suitable class $\mathcal{F} \times \mathcal{B}$ via gradient descent; we inspect the quality of transform sampling ($\widehat{F}_{\#}\mu \approx \nu$) and space isomorphism. Complete technical details of the experiments are deferred to Section F.

Gaussian distributions Consider two strongly isomorphic Gaussian distributions on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$: the base measure $\mu = N(0, I_2)$ and the target distribution $\nu = N(0, \Sigma)$, where I_2 is the identity matrix and the entries of Σ are $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = 0.7$. We let $c_{\mathcal{X}}(x, x') = x^\top x'$ and $c_{\mathcal{Y}}(y, y') = y^\top \Sigma^{-1} y'$, then two network spaces are strongly isomorphic by design; indeed, any pair (F, B) given by $F(x) = \Sigma^{1/2} Q x$ and $B(y) = Q^\top \Sigma^{-1/2} y$ for $Q \in O(2)$, where $O(2)$ is the orthogonal group, yields $c_{\mathcal{X}}(x, B(y)) = c_{\mathcal{Y}}(F(x), y)$ for all $x, y \in \mathbb{R}^2$, hence F and B are strong isomorphisms. We aim at obtaining such a pair of (linear) isomorphisms by letting $\mathcal{F} = \mathcal{B} = \{x \mapsto Wx : W \in \mathbb{R}^{2\times 2}\}$, that is, the collection of all linear maps from $\mathbb{R}^2$ to $\mathbb{R}^2$. We set $K_{\mathcal{X}} = K_{\mathcal{Y}}$ as a degree-2 polynomial kernel that maps (x, y) to $(x^\top y + 1)^2$; the resulting MMD compares distributions by matching the first two moments, which is sufficient to distinguish Gaussian distributions. The resulting linear maps are given by $\widehat{F}(x) = \mathbf{F}x$ and $\widehat{B}(y) = \mathbf{B}y$ for some $\mathbf{F}, \mathbf{B} \in \mathbb{R}^{2\times 2}$ satisfying

$$\mathbf{F}\mathbf{F}^\top = \begin{pmatrix} 1.035 & 0.751 \\ 0.751 & 1.094 \end{pmatrix}, \quad \mathbf{B}\Sigma\mathbf{B}^\top = \begin{pmatrix} 0.940 & 0.001 \\ 0.001 & 0.944 \end{pmatrix}, \quad \mathbf{F}\mathbf{B} = \begin{pmatrix} 0.966 & 0.029 \\ -0.020 & 1.029 \end{pmatrix}.$$

Since $\mathbf{F}\mathbf{F}^\top \approx \Sigma$, $\mathbf{B}\Sigma\mathbf{B}^\top \approx I_2$, and $\mathbf{F}\mathbf{B} \approx I_2$, the pair $(\widehat{F}, \widehat{B})$ can be seen as an instance of the pair of strong isomorphisms described above. Figure 1 illustrates that $\widehat{F}$ is a strong isomorphism (Definition 3): (a) shows that $\widehat{F}_{\#}\mu \approx \nu$, that is, $\widehat{F}$ is roughly a transport map, and (b) implies that $c_{\mathcal{X}}(x, x') \approx c_{\mathcal{Y}}(\widehat{F}(x), \widehat{F}(x'))$ holds.



(a) (b)

Figure 1: Gaussian experiment: $m = n = 1000$ and $\lambda_1 = \lambda_2 = \lambda_3 = 1$. (a) shows $\{\tilde{y}_j\}_{j=1}^{400}$ versus $\{\widehat{F}(\tilde{x}_i)\}_{i=1}^{400}$, where $\{\tilde{y}_j\}_{j=1}^{400}$ and $\{\tilde{x}_i\}_{i=1}^{400}$ are i.i.d. from $\nu = N(0, \Sigma)$ and $\mu = N(0, I_2)$, respectively; they are new samples independent from $\{y_j\}_{j=1}^{1000}$ and $\{x_i\}_{i=1}^{1000}$ used in (3.2). (b) shows the points $\{(c_{\mathcal{X}}(\tilde{x}_i, \tilde{x}_{i'}), c_{\mathcal{Y}}(\widehat{F}(\tilde{x}_i), \widehat{F}(\tilde{x}_{i'})))\}_{i,i'=1}^{40}$ and a straight line $y = x$.

MNIST Next, let ν be a distribution of images corresponding to four digits (2, 4, 6, 7) from the MNIST data set, which is supported on $\mathbb{R}^{784}$. Recall from Section 1 that the support $\mathcal{Y}$ of ν is low-dimensional [18], hence choosing $\mathcal{X} = \mathbb{R}^d$ with $d \ll 784$ is reasonable. Here, we try an extreme embedding task with $d = 2$ and $\mu = N(0, I_2)$, that is, generate MNIST images by transforming two-dimensional Gaussian samples.

Unlike the Gaussian example where we design the cost functions in advance to make the two spaces strongly isomorphic, specifying them can be more complicated in general cases, which might affect the quality of the RGM sampler. Here, we briefly discuss some of the most commonly used cost functions: given a fixed exponent $p \in \mathbb{N}$ or constant $\alpha > 0$,

$$(x, y) \quad \mapsto \quad \underbrace{\|x - y\|^p}_{\text{distance-based}} \quad \text{or} \quad \underbrace{\exp\bigl(-\alpha\|x - y\|^2\bigr)}_{\text{RBF kernel}} .$$

Clearly, $\|x - y\|^p$ is the most straightforward choice in Euclidean cases; $p = 1$ and $p = 2$ are indeed widely used in the literature [43]. The RBF kernel, also referred to as the heat kernel, is a common choice in the object matching literature [48]. In this MNIST example, we have found that these cost functions provide reasonable performance once they are scaled properly. Here, we will present the results based on the RBF kernel. Concretely, first define the RBF kernel $K_d(x, y) = \exp\bigl(-\|x - y\|^2/d\bigr)$ for $d \in \mathbb{N}$ and $x, y \in \mathbb{R}^d$; here, the constant $(1/d)$ serves as a scaling factor. Then, we define the cost functions as $c_{\mathcal{X}} = (K_2 - m_{\mathcal{X}})/\mathrm{sd}_{\mathcal{X}}$ and $c_{\mathcal{Y}} = (K_{784} - m_{\mathcal{Y}})/\mathrm{sd}_{\mathcal{Y}}$, where $m_{\mathcal{X}}$ and $\mathrm{sd}_{\mathcal{X}}$ are the median and the standard error of $\{K_{\mathcal{X}}(x_i, x_{i'})\}_{i,i'=1}^m$, respectively; $m_{\mathcal{Y}}$ and $\mathrm{sd}_{\mathcal{Y}}$ are defined analogously. This additional standardization process helps aligning the cost functions.

In the same vein, $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ must be properly specified; comparing the first two moments using the degree-2 polynomial kernel is no longer sufficient as the target distribution is non-Gaussian. We suggest using RBF kernels for the MMD terms as well; let $K_{\mathcal{X}} = K_2$ and $K_{\mathcal{Y}} = K_{784}$. The MMD induced by the RBF kernel indeed defines a metric between distributions under mild assumptions [40], which allows the resulting MMD terms to represent the original constraint of the RGM distance as mentioned in Section 3.2.

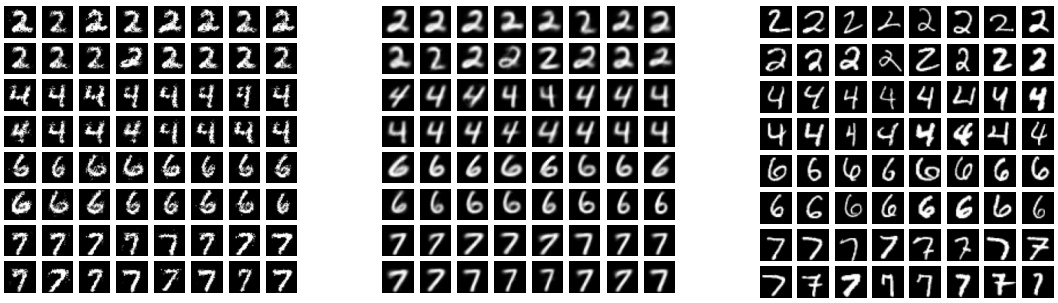
For the function classes $\mathcal{F}$ and $\mathcal{B}$, we need richer classes instead of the linear maps used in the Gaussian case. To this end, we will use fully connected neural networks with three hidden layers, each of which consists of 50 neurons. Lastly, we let $m = n = 20000$ and $\lambda_1 = \lambda_2 = \lambda_3 = 100$. Figure 2(a) shows the images generated by applying the resulting map $\widehat{F}$ to new i.i.d. samples from $\mu = N(0, I_2)$. Though not perfect, we see that recognizable images can be generated by transforming two-dimensional Gaussian samples, efficient in computation.[4] Meanwhile, the map $\widehat{B}$ shows how the MNIST images can be embedded in $\mathbb{R}^2$. Figure 3(a) shows $\{\widehat{B}(\tilde{y}_j)\}_{j=1}^{500}$, where $\{\tilde{y}_j\}_{j=1}^{500}$ are i.i.d. from ν (125×4 digits), independent from $\{y_j\}_{j=1}^{20000}$ used in (3.2). We see that each digit forms a local cluster in $\mathbb{R}^2$, each of which is roughly representable according to the range of the angular coordinate. Lastly, though not perfect as in Figure 1(b) (strongly isomorphic case), Figure 3(b) shows that $\widehat{B}$ leads to a reasonable alignment of $c_{\mathcal{X}}(\widehat{B}(y), \widehat{B}(y'))$ versus $c_{\mathcal{Y}}(y, y')$.

We clarify that the current experiment with $\mu = N(0, I_2)$ is a proof of concept. Suppose one aims to obtain images comparable to those from dedicated MNIST generators. In that case, exhaustive tests should be done for tuning each component of the RGM sampler, which is beyond the scope of this paper. Instead, we highlight that the RGM sampler with a simple modification can indeed generate significantly improved images seen in Figure 2(b). These images are generated from the following settings that are fully introduced in Section F: $(\mathcal{X}, \mu, c_{\mathcal{X}}) = (\mathbb{R}^4, N(0, I_4), K_4)$ and MMD terms in (3.2) are replaced by Sinkhorn divergences [22]. As such, the RGM sampler is amendable to other more general choices of $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ in its practical implementation.

5 Discussions

In this work, we proposed a novel distance between network spaces, called the Reversible Gromov-Monge distance, inspired by the Gromov-Wasserstein distance between metric measure spaces. Based on this,

[4] Computational cost for obtaining $\widehat{F}\colon \mathbb{R}^2 \to \mathbb{R}^{784}$ and computing $\widehat{F}(X)$ from $X \sim \mu$ is far less than that of the OT-based sampler as explained in Section 1.



(a) MMD ($\mathbb{R}^2$) (b) Sinkhorn ($\mathbb{R}^4$) (c) Original

Figure 2: (a) and (b) are generated by transforming new i.i.d. samples from μ using $\widehat{F}$: (a) from $\mu = N(0, I_2)$ with MMDs and (b) from $\mu = N(0, I_4)$ with Sinkhorn divergences. (c) shows real MNIST images.

we designed a transform sampler using empirical data that can operate between distributions defined on heterogeneous spaces. In addition, we introduced two concrete optimization methods for computing RGM given finite samples and proved their properties. The resulting RGM transform sampler is more efficient than the classic optimal transport transform sampler in terms of computation: its optimization procedure is less complicated than that of GAN-type samplers.

Lastly, we briefly mention computational aspects of the RGM distance that are not fully investigated. We have mainly relied on a practical computation of the RGM distance using the Lagrangian form instead of the constrained form (3.1), which was sufficient for the good empirical performance of the RGM sampler as witnessed in Section 4. However, one might be interested in the exact compuation instead and ask: when does minimizing the Lagrangian form lead to a close approximation to the RGM distance? Another important aspect is the comparison with the GW distance. Proposition 1 shows $\mathrm{GW}(\mu, \nu) \leq \mathrm{RGM}(\mu, \nu)$ in theory; by approximating both quantities numerically, we can see how large the gap between the two distances is. We provide our findings and insights regarding these issues in Section E for the interested readers; though not central to this paper, these can be of interest to the GW literature and we leave them as future research.

Acknowledgments

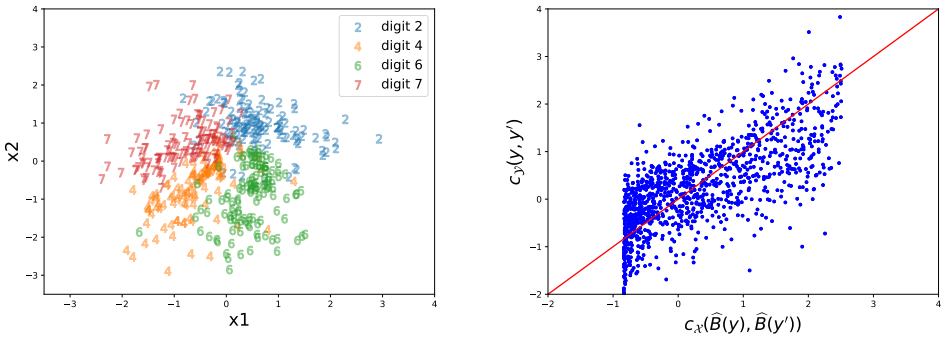
Liang acknowledges the generous support from the NSF Career award (DMS-2042473), and the William S. Fishman Faculty Research Fund at the University of Chicago Booth School of Business. Liang wishes to thank Maxim Raginsky and Chris Hansen for discussions on simulation-based inference.

References

[1] Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

[2] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR.

[3] Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. (2017). Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232. PMLR.



(a) (b)

Figure 3: (a) is generated by applying $\widehat{B}$ to 500 out-of-sample MNIST images, i.i.d. $\{\tilde{y}_j\}_{j=1}^{500}$ from ν. (b) shows the points $\{(c_{\mathcal{X}}(\widehat{B}(\tilde{y}_j), \widehat{B}(\tilde{y}_{j'})), c_{\mathcal{Y}}(\tilde{y}_j, \tilde{y}_{j'}))\}_{j,j'=1}^{50}$ and a straight line $y = x$.

[4] Bai, Y., Ma, T., and Risteski, A. (2018). Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*.

[5] Bassily, R., Belkin, M., and Ma, S. (2018). On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*.

[6] Blumberg, A. J., Carriere, M., Mandell, M. A., Rabadan, R., and Villar, S. (2020). Mrec: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. *arXiv preprint arXiv:2001.01666*.

[7] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press.

[8] Brécheteau, C. (2019). A statistical test of isomorphism between metric-measure spaces using the distance-to-a-measure signature. *Electron. J. Stat.*, 13(1):795–849.

[9] Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417.

[10] Cela, E. (1998). *The Quadratic Assignment Problem: Theory and Algorithms*. Combinatorial Optimization. Springer US.

[11] Chen, M., Liao, W., Zha, H., and Zhao, T. (2020). Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv preprint arXiv:2002.03938*.

[12] Chowdhury, S. and Mémoli, F. (2019). The Gromov–Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787.

[13] Chowdhury, S., Miller, D. W., and Needham, T. (2021). Quantized gromov-wasserstein. In *ECML/PKDD*.

[14] Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2017). Training gans with optimism. *arXiv preprint arXiv:1711.00141*.

[15] Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

[16] Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

[17] Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via Maximum Mean Discrepancy optimization. In *Proceedings of the 31st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 258–267.

[18] Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.*, 7(1):1–8.

[19] Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

[20] Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

[21] Gellert, M., Hossain, M. F., Berens, F. J. F., Bruhn, L. W., Urbainsky, C., Liebscher, V., and Lillig, C. H. (2019). Substrate specificity of thioredoxins and glutaredoxins–towards a functional classification. *Heliyon*, 5(12):e02943.

[22] Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

[23] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27.

[24] Gouriéroux, C. and Monfort, A. (1997). *Simulation-Based Econometric Methods*. Oxford University Press.

[25] Harvey, N., Liaw, C., and Mehrabian, A. (2017). Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1064–1068. PMLR.

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

[27] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[28] Koopmans, T. C. and Beckmann, M. (1957). Assignment Problems and the Location of Economic Activities. *Econometrica*, 25(1):53–76.

[29] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[30] Lei, Q., Lee, J. D., Dimakis, A. G., and Daskalakis, C. (2019). Sgd learns one-layer networks in wgans. *arXiv preprint arXiv:1910.07030*.

[31] Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR.

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

[33] Liang, T. (2021). How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41.

[34] Liang, T. and Stokes, J. (2019). Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *International Conference on Artificial Intelligence and Statistics*, volume 89, pages 907–915. PMLR.

[35] Loiola, E. M., de Abreu, N. M. M., Boaventura-Netto, P. O., Hahn, P., and Querido, T. (2007). A survey for the quadratic assignment problem. *European Journal of Operational Research*, 176(2):657–690.

[36] McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, 57(5):995–1026.

[37] Mémoli, F. (2011). Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487.

[38] Mémoli, F. and Needham, T. (2018). Distance distributions and inverse problems for metric measure spaces. *arXiv preprint arXiv:1810.09646*.

[39] Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. (2017). Sobolev gan. *arXiv preprint arXiv:1711.04894*.

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

[41] Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29.

[42] Pakes, A. and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57(5):1027–1057.

[43] Peyré, G., Cuturi, M., and Solomon, J. (2016). Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR.

[44] Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

[45] Scetbon, M., Peyré, G., and Cuturi, M. (2021). Linear-time gromov wasserstein distances using low rank couplings and costs. *arXiv preprint arXiv:2106.01128*.

[46] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

[47] Singh, S. and Póczos, B. (2018). Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*.

[48] Solomon, J., Peyré, G., Kim, V. G., and Sra, S. (2016). Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4).

[49] Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2012). On the empirical estimation of integral probability metrics. *Electron. J. Stat.*, 6:1550–1599.

[50] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053.

[51] Sturm, K.-T. (2012). The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*.

[52] Titouan, V., Flamary, R., Courty, N., Tavenard, R., and Chapel, L. (2019). Sliced gromov-wasserstein. In *Advances in Neural Information Processing Systems*, volume 32.

[53] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

[54] Wasserman, L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York.

[55] Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in wasserstein distance. *arXiv preprint arXiv:1902.01778*.

[56] Weitkamp, C. A., Proksch, K., Tameling, C., and Munk, A. (2020). Gromov-wasserstein distance based object matching: asymptotic inference. *arXiv preprint arXiv:2006.12287*.

[57] Xu, H., Luo, D., and Carin, L. (2019). Scalable gromov-wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems*, volume 32.

SUPPLEMENTARY MATERIAL

This supplementary material collects details of Sections 3 and 4 along with other technicalities.

- Section A studies analytical properties of the RGM distance.
- Section B derives the non-asymptotic rate of convergence analyzing the statistical properties of the RGM sampler.
- Section C discusses a further relaxation of the RGM into an infinite-dimensional convex program which relies on a new representer theorem.
- Section D contains the proofs of the results in Section B and auxiliary lemmas.
- Section E discusses computational aspects of the RGM distance.
- Section F provides the implementation details of the experiments in Section 4.

A Analytic Properties

In this section, we derive analytic properties of the proposed RGM distance. First, we discuss the relations among three distances: GW, GM, and RGM.

Proposition 1. *For network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ as in Definition 2,*

$$\mathrm{GW}(\mu,\nu) \leq \mathrm{GM}(\mu,\nu) \leq \mathrm{RGM}(\mu,\nu) \,.$$

Proof of Proposition 1. Define

$$Q(\gamma) = \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2 \,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')$$

for all $\gamma \in \Pi(\mu,\nu)$ so that

$$\mathrm{GW}(\mu,\nu)^2 = \inf_{\gamma\in\Pi(\mu,\nu)} Q(\gamma) \,.$$

Recall that $\Pi_{\mathcal{T}} = \{(\mathrm{Id}, T)_{\#}\mu : T \in \mathcal{T}(\mu,\nu)\} \subset \Pi(\mu,\nu)$. Hence, as noted in Section 2,

$$\mathrm{GM}(\mu,\nu)^2 = \inf_{\gamma\in\Pi_{\mathcal{T}}} Q(\gamma) \,.$$

Define $\Pi' = \{\gamma \in \Pi(\mu,\nu) : \gamma = (\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu \;\exists (F,B) \in \mathcal{I}(\nu,\mu)\}$, then one can check

$$\mathrm{RGM}(\mu,\nu)^2 = \inf_{\gamma\in\Pi'} Q(\gamma)$$

using change-of-variables. Note that Π' may be rewritten as $\{\gamma \in \Pi_{\mathcal{T}} : \gamma = (B, \mathrm{Id})_{\#}\nu \;\exists B \in \mathcal{T}(\nu,\mu)\}$. Hence, $\Pi' \subseteq \Pi_{\mathcal{T}} \subseteq \Pi(\mu,\nu)$, thus we conclude $\mathrm{GW}(\mu,\nu) \leq \mathrm{GM}(\mu,\nu) \leq \mathrm{RGM}(\mu,\nu)$. □

In short, Proposition 1 holds because our formulation can be obtained by further restricting the constraint set of couplings in GM. Note that our RGM is symmetric while the original GM is not. As a simple corollary, one can check

$$\max\{\mathrm{GM}(\mu,\nu), \mathrm{GM}(\nu,\mu)\} \leq \mathrm{RGM}(\mu,\nu) \,.$$

Now, we establish further metric properties of RGM. Symmetry of RGM is already mentioned. Next, we prove a triangle inequality using a gluing technique as in OT.

Proposition 2. *RGM satisfies the triangle inequality, that is,*

$$\mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}}) \leq \mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}) + \mathrm{RGM}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})$$

holds for three network spaces $(\mathcal{X}, \mu_{\mathcal{X}}, c_{\mathcal{X}})$, $(\mathcal{Y}, \mu_{\mathcal{Y}}, c_{\mathcal{Y}})$, and $(\mathcal{Z}, \mu_{\mathcal{Z}}, c_{\mathcal{Z}})$.

Proof of Proposition 2. Recall that

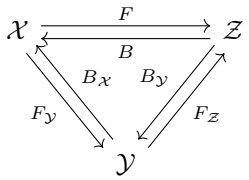
$$\mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}}) = \inf_{(F,B)\in\mathcal{I}(\mu_{\mathcal{X}},\mu_{\mathcal{Z}})} C_{\mathcal{X}\mathcal{Z}}(F, B) \ ,$$

where

$$C_{\mathcal{X}\mathcal{Z}}(F, B) = \left(\int \left(c_{\mathcal{X}}(x, B(z)) - c_{\mathcal{Z}}(F(x), z)\right)^2 \mathrm{d}\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Z}} \, (x, z) \right)^{1/2} .$$

First, $(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}})$ holds for $(F_{\mathcal{Y}}, B_{\mathcal{X}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$ and $(F_{\mathcal{Z}}, B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})$ since

$$\begin{aligned}
(\mathrm{Id}, F_{\mathcal{Z}} \circ F_{\mathcal{Y}})_{\#}\mu_{\mathcal{X}} &= (\mathrm{Id}, F_{\mathcal{Z}})_{\#}(\mathrm{Id}, F_{\mathcal{Y}})_{\#}\mu_{\mathcal{X}} \\
&= (\mathrm{Id}, F_{\mathcal{Z}})_{\#}(B_{\mathcal{X}}, \mathrm{Id})_{\#}\mu_{\mathcal{Y}} \quad (\because (F_{\mathcal{Y}}, B_{\mathcal{X}}) \in \mathcal{I}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})) \\
&= (B_{\mathcal{X}}, \mathrm{Id})_{\#}(\mathrm{Id}, F_{\mathcal{Z}})_{\#}\mu_{\mathcal{Y}} \\
&= (B_{\mathcal{X}}, \mathrm{Id})_{\#}(B_{\mathcal{Y}}, \mathrm{Id})_{\#}\mu_{\mathcal{Z}} \quad (\because (F_{\mathcal{Z}}, B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}})) \\
&= (B_{\mathcal{X}} \circ B_{\mathcal{Y}}, \mathrm{Id})_{\#}\mu_{\mathcal{Z}} \ .
\end{aligned}$$



Moreover, in this case, we have

$$C_{\mathcal{X}\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \le C_{\mathcal{X}\mathcal{Y}}(F_{\mathcal{Y}}, B_{\mathcal{X}}) + C_{\mathcal{Y}\mathcal{Z}}(F_{\mathcal{Z}}, B_{\mathcal{Y}})$$

since

$$\begin{aligned}
&C_{\mathcal{X}\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \\
&= \left(\int \left[c_{\mathcal{X}}(x, B_{\mathcal{X}} \circ B_{\mathcal{Y}}(z)) - c_{\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}(x), z)\right]^2 \mathrm{d}\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Z}} \, (x, z) \right)^{1/2} \\
&\le \left(\int \int \left[c_{\mathcal{Y}}(x, B_{\mathcal{X}} \circ B_{\mathcal{Y}}(z)) - c_{\mathcal{Y}}(F_{\mathcal{Y}}(x), B_{\mathcal{Y}}(z))\right]^2 \mathrm{d}\mu_{\mathcal{Z}} \, (z) \, \mathrm{d}\mu_{\mathcal{X}} \, (x) \right)^{1/2} \\
&\quad + \left(\int \int \left[c_{\mathcal{Y}}(F_{\mathcal{Y}}(x), B_{\mathcal{Y}}(z)) - c_{\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}(x), z)\right]^2 \mathrm{d}\mu_{\mathcal{X}} \, (x) \, \mathrm{d}\mu_{\mathcal{Z}} \, (z) \right)^{1/2} \\
&= \left(\int \int \left[c_{\mathcal{Y}}(x, B_{\mathcal{X}}(y)) - c_{\mathcal{Y}}(F_{\mathcal{Y}}(x), y)\right]^2 \mathrm{d}\mu_{\mathcal{Y}} \, (y) \, \mathrm{d}\mu_{\mathcal{X}} \, (x) \right)^{1/2} \quad (\because (B_{\mathcal{Y}})_{\#}\mu_{\mathcal{Z}} = \mu_{\mathcal{Y}}) \\
&\quad + \left(\int \int \left[c_{\mathcal{Y}}(y, B_{\mathcal{Y}}(z)) - c_{\mathcal{Z}}(F_{\mathcal{Z}}(y), z)\right]^2 \mathrm{d}\mu_{\mathcal{Y}} \, (y) \, \mathrm{d}\mu_{\mathcal{Z}} \, (z) \right)^{1/2} \quad (\because (F_{\mathcal{Y}})_{\#}\mu_{\mathcal{X}} = \mu_{\mathcal{Y}}) \\
&= C_{\mathcal{X}\mathcal{Y}}(F_{\mathcal{Y}}, B_{\mathcal{X}}) + C_{\mathcal{Y}\mathcal{Z}}(F_{\mathcal{Z}}, B_{\mathcal{Y}}) \ .
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Z}}) &= \inf_{(F,B)\in\mathcal{I}(\mu_{\mathcal{X}},\mu_{\mathcal{Z}})} C_{\mathcal{X}\mathcal{Z}}(F, B) \\
&\le \inf_{\substack{(F_{\mathcal{Y}},B_{\mathcal{X}})\in\mathcal{I}(\mu_{\mathcal{X}},\mu_{\mathcal{Y}}) \\ (F_{\mathcal{Z}},B_{\mathcal{Y}})\in\mathcal{I}(\mu_{\mathcal{Y}},\mu_{\mathcal{Z}})}} C_{\mathcal{X}\mathcal{Z}}(F_{\mathcal{Z}} \circ F_{\mathcal{Y}}, B_{\mathcal{X}} \circ B_{\mathcal{Y}}) \\
&\le \inf_{(F_{\mathcal{Y}},B_{\mathcal{X}})\in\mathcal{I}(\mu_{\mathcal{X}},\mu_{\mathcal{Y}})} C_{\mathcal{X}\mathcal{Y}}(F_{\mathcal{Y}}, B_{\mathcal{X}}) + \inf_{(F_{\mathcal{Z}},B_{\mathcal{Y}})\in\mathcal{I}(\mu_{\mathcal{Y}},\mu_{\mathcal{Z}})} C_{\mathcal{Y}\mathcal{Z}}(F_{\mathcal{Z}}, B_{\mathcal{Y}}) \\
&= \mathrm{RGM}(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}) + \mathrm{RGM}(\mu_{\mathcal{Y}}, \mu_{\mathcal{Z}}) \ .
\end{aligned}$$

□

Next, we study whether $\mathrm{RGM}(\mu,\nu)=0$ holds if and only if $(\mathcal{X},\mu,c_{\mathcal{X}})\cong(\mathcal{Y},\nu,c_{\mathcal{Y}})$. Here the equivalence relation induced by $\cong$ can be read from Definition 3. Like the Gromov-Wasserstein distance, in general, we can only assert the if part without further conditions. The following proposition states that $\mathrm{RGM}(\mu,\nu)=0$ if and only if $(\mathcal{X},\mu,c_{\mathcal{X}})\cong(\mathcal{Y},\nu,c_{\mathcal{Y}})$ under some additional conditions on $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$, thereby implying Theorem 2.

Proposition 3. *Let $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ be two network spaces. If $(\mathcal{X},\mu,c_{\mathcal{X}})\cong(\mathcal{Y},\nu,c_{\mathcal{Y}})$, then $\mathrm{RGM}(\mu,\nu)=0$. The converse is true if there exists a continuous and strictly monotone function $h\colon\mathbb{R}_+\to\mathbb{R}$ such that $c_{\mathcal{X}}=h(d_{\mathcal{X}})$ and $c_{\mathcal{Y}}=h(d_{\mathcal{Y}})$.*

Proof of Proposition 3. Suppose $\mathrm{RGM}(\mu,\nu)=0$. Due to the inequality $\mathrm{GW}(\mu,\nu)\leq\mathrm{RGM}(\mu,\nu)$, we have $\mathrm{GW}(\mu,\nu)=0$, that is,

$$\inf_{\gamma\in\Pi(\mu,\nu)}\left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}}(h(d_{\mathcal{X}}(x,x'))-h(d_{\mathcal{Y}}(y,y')))^2\,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')\right)^{1/2}=0\,.$$

Since there exists a coupling $\gamma^\star$ that achieves the minimum of GW due to Theorem 12 of [12], we conclude

$$h(d_{\mathcal{X}}(x,x'))=h(d_{\mathcal{Y}}(y,y'))$$

holds $\gamma^\star\otimes\gamma^\star$ almost surely on $(\mathcal{X}\times\mathcal{Y})^2$. Since h is strictly monotone, this means

$$d_{\mathcal{X}}(x,x')=d_{\mathcal{Y}}(y,y')$$

holds $\gamma^\star\otimes\gamma^\star$ almost surely on $(\mathcal{X}\times\mathcal{Y})^2$. Therefore,

$$\begin{aligned}&\inf_{\gamma\in\Pi(\mu,\nu)}\left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}}(d_{\mathcal{X}}(x,x')-d_{\mathcal{Y}}(y,y'))^2\,\mathrm{d}\gamma\,(x,y)\,\mathrm{d}\gamma\,(x',y')\right)^{1/2}\\&\geq\left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}}(d_{\mathcal{X}}(x,x')-d_{\mathcal{Y}}(y,y'))^2\,\mathrm{d}\gamma^\star\,(x,y)\,\mathrm{d}\gamma^\star\,(x',y')\right)^{1/2}\\&=0\,.\end{aligned}$$

Theorem 1 implies that metric measure spaces $(\mathcal{X},\mu,d_{\mathcal{X}})$ and $(\mathcal{Y},\nu,d_{\mathcal{Y}})$ are strongly isomorphic. Since $c_{\mathcal{X}}=h(d_{\mathcal{X}})$ and $c_{\mathcal{Y}}=h(d_{\mathcal{Y}})$, it follows easily that $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ are strongly isomorphic as well.

To prove the if part, suppose $(\mathcal{X},\mu,c_{\mathcal{X}})$ and $(\mathcal{Y},\nu,c_{\mathcal{Y}})$ are strongly isomorphic and consider a strong isomorphism T. Then, $(T,T^{-1})\in\mathcal{I}(\mu,\nu)$ holds since $(\mathrm{Id},T)_\#\mu=(T^{-1},\mathrm{Id})_\#T_\#\mu=(T^{-1},\mathrm{Id})_\#\nu$. Also, by definition of T, we have $c_{\mathcal{X}}(x,T^{-1}(y))=c_{\mathcal{Y}}(T(x),T\circ T^{-1}(y))=c_{\mathcal{Y}}(T(x),y)$ for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, thus

$$\mathrm{RGM}(\mu,\nu)\leq\int_{\mathcal{X}\times\mathcal{Y}}(c_{\mathcal{X}}(x,T^{-1}(y))-c_{\mathcal{Y}}(T(x),y))^2\,\mathrm{d}\mu\otimes\nu=0\,.$$

□

We conclude this section with a few more properties and examples. First, we give a sufficient condition for $(F,B)\in\mathcal{I}(\mu,\nu)$ which can be useful in practice.

Lemma 1. *Let $(F,B)\in\mathcal{T}(\mu,\nu)\times\mathcal{T}(\nu,\mu)$. If $F\circ B=\mathrm{Id}$ or $B\circ F=\mathrm{Id}$ holds, then $(F,B)\in\mathcal{I}(\mu,\nu)$.*

Proof. Without loss of generality, assume $B\circ F=\mathrm{Id}$. Then,

$$(\mathrm{Id},F)_\#\mu=(B\circ F,F)_\#\mu=(B,\mathrm{Id})_\#(F_\#\mu)=(B,\mathrm{Id})_\#\nu\,.$$

Hence, $(F,B)\in\mathcal{I}(\mu,\nu)$. □

The following example illustrates that this condition can be used to find a pair $(F, B) \in \mathcal{I}(\mu, \nu)$ when μ and ν are Gaussian distributions.

Example 1. *Given $p < q$, suppose $\mu = N(0, I_p)$ and $\nu = N(0, \Sigma)$, where $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix and $\Sigma \in \mathbb{R}^{q\times q}$ is of rank p. Then, we can find a rank-p matrix $A \in \mathbb{R}^{q\times p}$ such that $\Sigma = AA^\top$. Let $F(x) = Ax$ and $B(y) = A^\dagger y$, then one can easily check $F_\# \mu = \nu$, $B_\# \nu = \mu$, and $B \circ F = \mathrm{Id}$. Hence, $(F, B) \in \mathcal{I}(\mu, \nu)$.*

We conclude this section with a simple example that tells us that properly chosen cost functions give a strong isomorphism between two Gaussian distributions in general.

Example 2. *Consider two Gaussian distributions on $\mathbb{R}^d$, say $\mu = N(0, \Sigma_1)$ and $\nu = N(0, \Sigma_2)$. Assume Σ_1 and Σ_2 are invertible. Then two network spaces $(\mathbb{R}^d, \mu, c_\mathcal{X})$ and $(\mathbb{R}^d, \nu, c_\mathcal{Y})$ are strongly isomorphic if $c_\mathcal{X}$ and $c_\mathcal{Y}$ are Mahalanobis distances, that is,*

$$c_\mathcal{X}(x, x') = \sqrt{(x - x')^\top \Sigma_1^{-1}(x - x')}\,, \quad c_\mathcal{Y}(y, y') = \sqrt{(y - y')^\top \Sigma_2^{-1}(y - y')}\,.$$

To see this, let $T = \Sigma_2^{1/2}\Sigma_1^{-1/2}$, where $\Sigma_1^{1/2}$ and $\Sigma_2^{1/2}$ are the square roots of Σ_1 and Σ_2, respectively. Obviously, a linear map T satisfies $T \in \mathcal{T}(\mu, \nu)$ and $c_\mathcal{X}(x, x') = c_\mathcal{Y}(Tx, Tx')$ for all $x, x' \in \mathbb{R}^d$. According to Definition 3, a linear map T is a strong isomorphism. Proposition 3 implies $\mathrm{RGM}(\mu, \nu) = 0$. Notice that the same results hold for $c_\mathcal{X}(x, x') = x^\top \Sigma_1^{-1} x'$ and $c_\mathcal{Y}(y, y') = y^\top \Sigma_2^{-1} y'$ as well.

B Statistical Theory

This section serves to prove Theorem 3. Without loss of generality, we assume $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in $C(\mu, \nu, F, B)$ since the proof is essentially identical with any constants $\lambda_i, 1 \le i \le 3$. For convenience, we denote

$$\begin{aligned} C_0(F, B) &= \int (c_\mathcal{X}(x, B(y)) - c_\mathcal{Y}(F(x), y))^2 \,\mathrm{d}\mu \otimes \nu\ , \\ M(F, B) &= \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu) + \mathrm{MMD}^2_{K_\mathcal{X}}(\mu, B_\#\nu) + \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\mu, (B, \mathrm{Id})_\#\nu) \end{aligned}$$

and therefore $C(\mu, \nu, F, B) = C_0(F, B) + M(F, B)$. Similarly, define the empirical counterparts as

$$\begin{aligned} \widehat{C}_0(F, B) &= \frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n (c_\mathcal{X}(x_i, B(y_j)) - c_\mathcal{Y}(F(x_i), y_j))^2\ , \\ \widehat{M}(F, B) &= \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) + \mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) + \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n) \end{aligned}$$

and thus $C(\widehat{\mu}_m, \widehat{\nu}_n, F, B) = \widehat{C}_0(F, B) + \widehat{M}(F, B)$.

Our goal is to give an upper bound on $C(\mu, \nu, \widehat{F}, \widehat{B}) - \inf_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\mu, \nu, F, B)$. To this end, first recall that

$$C(\widehat{\mu}_m, \widehat{\nu}_n, \widehat{F}, \widehat{B}) \le C(\widehat{\mu}_m, \widehat{\nu}_n, F, B)$$

holds for any $F \in \mathcal{F}$ and $B \in \mathcal{B}$ by definition of $\widehat{F}$ and $\widehat{B}$ given in Theorem 3. Therefore,

$$C(\mu, \nu, \widehat{F}, \widehat{B}) - C(\mu, \nu, F, B) \le C(\mu, \nu, \widehat{F}, \widehat{B}) - C(\widehat{\mu}_m, \widehat{\nu}_n, \widehat{F}, \widehat{B}) + C(\widehat{\mu}_m, \widehat{\nu}_n, F, B) - C(\mu, \nu, F, B)\,.$$

The RHS can be decomposed as

$$C_0(\widehat{F}, \widehat{B}) - \widehat{C}_0(\widehat{F}, \widehat{B}) + M(\widehat{F}, \widehat{B}) - \widehat{M}(\widehat{F}, \widehat{B}) + \widehat{C}_0(F, B) - C_0(F, B) + \widehat{M}(F, B) - M(F, B)\,.$$

To further control the expression, we will first derive probabilistic bounds on $|\widehat{C}_0(F, B) - C_0(F, B)|$ and $|\widehat{M}(F, B) - M(F, B)|$ that hold for a fixed $(F, B) \in \mathcal{F}\times\mathcal{B}$ via standard concentration inequalities. Later, we will establish uniform probabilistic bounds on $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F, B) - C_0(F, B)|$ and $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F, B) - M(F, B)|$, using tools from empirical process theory.

B.1 Concentration Inequalities

We utilize the McDiarmid's inequality to derive bounds on $|\widehat{C}_0(F,B) - C_0(F,B)|$ and $|\widehat{M}(F,B) - M(F,B)|$. To give a bound on the former, we make the following boundedness assumption.

Assumption 1. *$c_{\mathcal{X}}(\cdot,\cdot), c_{\mathcal{Y}}(\cdot,\cdot)$ is uniformly bounded, that is, there exists a constant $H > 0$ such that*

$$\sup_{(x,x')\in\mathcal{X}\times\mathcal{X}} c_{\mathcal{X}}(x,x'), \sup_{(y,y')\in\mathcal{Y}\times\mathcal{Y}} c_{\mathcal{Y}}(y,y') \leq \sqrt{\frac{H}{4}} \,.$$

Proposition 4. *Under Assumption 1, for any pair $(F,B) \in \mathcal{F}\times\mathcal{B}$ and $\delta > 0$,*

$$|\widehat{C}_0(F,B) - C_0(F,B)| \precsim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}}$$

holds with probability at least $1 - 4\delta$.

To derive a similar bound on $|\widehat{M}(F,B) - M(F,B)|$, we assume that kernels are bounded.

Assumption 2. *There exists $K > 0$ such that*

$$\sup_{x\in\mathcal{X}} |K_{\mathcal{X}}(x,x)| \,, \sup_{y\in\mathcal{Y}} |K_{\mathcal{Y}}(y,y)| \leq K \,.$$

Proposition 5. *Under Assumption 2, for any pair $(F,B) \in \mathcal{F}\times\mathcal{B}$ and $\delta > 0$,*

$$|\widehat{M}(F,B) - M(F,B)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}$$

holds with probability at least $1 - 6\delta$.

B.2 Uniform Deviations

We now derive uniform deviation bounds for

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)| \,, \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F,B) - M(F,B)| \,.$$

For the former, we use the notion of uniform covering numbers defined below.

Definition 6 (Uniform Covering Number)**.** *Let $\mathcal{G}$ be a collection of real-valued functions defined on a set $\mathcal{Z}$. Given m points $z_1,\ldots,z_m \in \mathcal{Z}$ and any $\delta > 0$, we define $N_\infty(\delta,\mathcal{G},\{z_i\}_{i=1}^m)$ to be the δ-covering number of $\mathcal{G}$ under the pseudometric d induced by points $z_1,\ldots,z_m$:*

$$d(g,g') \coloneqq \max_{i\in[m]} |g(z_i) - g'(z_i)| \,.$$

Also, we define the uniform δ-covering number of $\mathcal{G}$ as follows:

$$N_\infty(\delta,\mathcal{G},m) \coloneqq \sup\left\{N_\infty(\delta,\mathcal{G},\{z_i\}_{i=1}^m) : z_1,\ldots,z_m \in \mathcal{Z}\right\} \,.$$

Here, the supremum is taken over all possible combinations of m points in $\mathcal{Z}$.

Also, we make the following assumptions.

Assumption 3. *$\mathcal{F}_k$ and $\mathcal{B}_\ell$ (see Section 3.3) consist of uniformly bounded functions, that is, there exists a constant $b > 0$ such that*

$$\max_{k\in[\dim(\mathcal{Y})]} \sup_{F_k\in\mathcal{F}_k} \|F_k\|_\infty \,, \max_{\ell\in[\dim(\mathcal{X})]} \sup_{B_\ell\in\mathcal{B}_\ell} \|B_\ell\|_\infty \leq b \,.$$

Assumption 4. *There exists a constant $L > 0$ such that*

$$|c_{\mathcal{X}}(x, x_1) - c_{\mathcal{X}}(x, x_2)| \le L\|x_1 - x_2\| \ , \ |c_{\mathcal{Y}}(y_1, y) - c_{\mathcal{Y}}(y_2, y)| \le L\|y_1 - y_2\| \ .$$

This Lipschitzness assumption ensures the smoothness of a map $(F, B) \mapsto |\widehat{C}_0(F, B) - C_0(F, B)|$ over $\mathcal{F} \times \mathcal{B}$, which allows us to utilize the uniform covering numbers.

Proposition 6. *Under Assumptions 1, 3, 4, for any $\epsilon > 0$ and $\delta > 0$,*

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F, B) - C_0(F, B)|$$
$$\precsim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m \wedge n}} + \epsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, n)}{m \wedge n}}$$

holds with probability at least $1 - 2\delta$.

Now, the remaining task is to choose ϵ carefully in Proposition 6 for a concrete upper bound. To this end, we utilize the pseudo-dimension defined below.

Definition 7 (Pseudo-Dimension)**.** *Let $\mathcal{G}$ be a collection of real-valued functions defined on a set $\mathcal{Z}$. Given a subset $S \coloneqq \{z_1, \ldots, z_m\} \subset \mathcal{Z}$, we say S is pseudo-shattered by $\mathcal{G}$ if there are $r_1, \ldots, r_m \in \mathbb{R}$ such that for each $b \in \{0, 1\}^m$ we can find $g_b \in \mathcal{G}$ satisfying $\operatorname{sign}(g_b(z_i) - r_i) = b_i$ for all $i \in [m]$. We define the pseudo-dimension of $\mathcal{G}$, denoted as $\operatorname{Pdim}(\mathcal{G})$, as the maximum cardinality of a subset $S \subset \mathcal{Z}$ that is pseudo-shattered by $\mathcal{G}$.*

Using a well-established relation of the uniform covering number and the pseudo-dimension (Lemma 4), we can simplify Proposition 6 as follows.

Corollary 1. *Under Assumptions 1, 3, 4, for any $\delta > 0$,*

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F, B) - C_0(F, B)|$$
$$\precsim \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m \wedge n}} + \sqrt{\frac{\log(m \vee n)}{m \wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})} \operatorname{Pdim}(\mathcal{F}_k) + \sum_{\ell=1}^{\dim(\mathcal{X})} \operatorname{Pdim}(\mathcal{B}_\ell)\right)}$$

holds with probability at least $1 - 2\delta$.

To derive an upper bound on $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F, B) - M(F, B)|$, we first introduce Rademacher complexities defined below.

Definition 8 (Rademacher Complexity)**.** *Let $(\mathcal{Z}, \rho)$ be a probability space and $\mathcal{G}$ be a collection of measurable functions defined on $\mathcal{Z}$. We define the Rademacher complexity of $\mathcal{G}$ with respect to m samples from ρ as follows:*

$$R_m(\mathcal{G}, \rho) = \mathop{\mathbb{E}}_{z_i \overset{\text{iid}}{\sim} \rho} \mathop{\mathbb{E}}_{\epsilon_i} \sup_{g\in\mathcal{G}} \left| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i g(z_i) \right| \ ,$$

Here, $z_1, \ldots, z_m$ are i.i.d. samples from ρ and $\epsilon_1, \ldots, \epsilon_m$ are i.i.d. Rademacher random variables such that $(z_1, \ldots, z_m)$ and $(\epsilon_1, \ldots, \epsilon_m)$ are independent.

Proposition 7. *Denote a closed unit ball of any RKHS $\mathcal{H}$ as $\mathcal{H}(1)$. Also, let $(\mathrm{Id}, \mathcal{F}) \coloneqq \{(\mathrm{Id}, F) : F \in \mathcal{F}\}$ and $(\mathcal{B}, \mathrm{Id}) \coloneqq \{(B, \mathrm{Id}) : B \in \mathcal{B}\}$; hence, they are classes of maps from $\mathcal{X}$ to $\mathcal{X} \times \mathcal{Y}$ and from $\mathcal{Y}$ to $\mathcal{X} \times \mathcal{Y}$, respectively. Under Assumption 2, for any $\delta > 0$,*

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F, B) - M(F, B)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}} + R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \mu) + R_n(\mathcal{H}_{\mathcal{X}}(1) \circ \mathcal{B}, \nu)$$
$$+ R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1) \circ (\mathrm{Id}, \mathcal{F}), \mu) + R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1) \circ (\mathcal{B}, \mathrm{Id}), \nu)$$

holds with probability at least $1-6\delta$. *Here,* $\mathcal{F}\circ\mathcal{G}=\{f\circ g : f\in\mathcal{F}, g\in\mathcal{G}\}$ *for any function classes* $\mathcal{F}$ *and* $\mathcal{G}$ *with matching input and output space.*

Now, the only remaining task is to bound four Rademacher complexities. We will derive upper bounds using the chaining technique. To illustrate the main idea, let us consider $\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}$. Recall that

$$R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\mu)=\mathop{\mathbb{E}}_{x_i\overset{\text{iid}}{\sim}\mu} R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\{x_i\}_{i=1}^m)\ ,$$

where $R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\{x_i\}_{i=1}^m)$ is the empirical Rademacher complexity of $\mathcal{H}_\mathcal{Y}(1)\circ F$ associated with $\{x_i\}_{i=1}^m$:

$$R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\{x_i\}_{i=1}^m)=\mathop{\mathbb{E}}_{\epsilon_i}\sup_{h\in\mathcal{H}_\mathcal{Y}(1),F\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m\epsilon_i h(F(x_i))\right|=\mathop{\mathbb{E}}_{\epsilon_i}\sup_{h\in\mathcal{H}_\mathcal{Y}(1),F\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\epsilon_i h(F(x_i))\ .$$

Notice that we may remove the absolute value since $\mathcal{H}_\mathcal{Y}(1)=-\mathcal{H}_\mathcal{Y}(1)$. Now, considering $\{x_i\}_{i=1}^m$ as fixed, we will first bound the empirical Rademacher complexity by replacing the Rademacher random variables with Gaussian random variables. Let g_i be i.i.d. standard Gaussian random variables, then it is well known that

$$R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\{x_i\}_{i=1}^m)\le\sqrt{\frac{\pi}{2}}\mathop{\mathbb{E}}_{g_i}\sup_{h\in\mathcal{H}_\mathcal{Y}(1),F\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m g_i h(F(x_i))=:\sqrt{\frac{\pi}{2}}\mathcal{G}_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\{x_i\}_{i=1}^m)\ .$$

Also, under the assumption that $K_\mathcal{Y}$ is bounded by K, the reproducing property and the Cauchy-Schwarz inequality imply

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_\mathcal{Y}(1),F\in\mathcal{F}}\sum_{i=1}^m g_i h(F(x_i))\\
&=\sup_{h\in\mathcal{H}_\mathcal{Y}(1),F\in\mathcal{F}}\left\langle h,\sum_{i=1}^m g_i K_\mathcal{Y}(\cdot,F(x_i))\right\rangle_{\mathcal{H}_\mathcal{Y}}\\
&\le\sup_{h\in\mathcal{H}_\mathcal{Y}(1),F\in\mathcal{F}}\|h\|_{\mathcal{H}_\mathcal{Y}}\left[\sum_{i=1}^m g_i^2K_\mathcal{Y}(F(x_i),F(x_i))+\sum_{i\ne j}g_ig_jK_\mathcal{Y}(F(x_i),F(x_j))\right]^{1/2}\\
&\le\sup_{F\in\mathcal{F}}\left[\sum_{i=1}^m g_i^2K+\sum_{i\ne j}g_ig_jK_\mathcal{Y}(F(x_i),F(x_j))\right]^{1/2}\\
&\le\left[\sum_{i=1}^m g_i^2K+\sup_{F\in\mathcal{F}}\sum_{i\ne j}g_ig_jK_\mathcal{Y}(F(x_i),F(x_j))\right]^{1/2}.
\end{aligned}$$

Here, $\langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{Y}}$ denotes the inner product on $\mathcal{H}_\mathcal{Y}$. Hence,

$$\begin{aligned}
\mathcal{G}_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\{x_i\}_{i=1}^m)&\le\frac{1}{m}\mathop{\mathbb{E}}_{g_i}\left[\sum_{i=1}^m g_i^2K+\sup_{F\in\mathcal{F}}\sum_{i\ne j}g_ig_jK_\mathcal{Y}(F(x_i),F(x_j))\right]^{1/2}\\
&\le\frac{1}{m}\left[mK+\mathop{\mathbb{E}}_{g_i}\sup_{F\in\mathcal{F}}\sum_{i\ne j}g_ig_jK_\mathcal{Y}(F(x_i),F(x_j))\right]^{1/2},
\end{aligned}$$

where the second inequality follows from the Jensen's inequality and $\mathbb{E}\,g_i^2=1$.

For any $F\colon \mathcal{X} \to \mathcal{Y}$, let $A_F \in \mathbb{R}^{m\times m}$ be a matrix whose diagonal elements are zero and (i,j)-th element is $K_{\mathcal{Y}}(F(x_i), F(x_j))$ for $i \neq j$. Then, the last term amounts to the supremum of a quadratic process

$$\mathop{\mathbb{E}}_{g} \sup_{F\in\mathcal{F}} g^\top A_F g \;,$$

where $g \coloneqq [g_1, \ldots, g_m]^\top \sim N(0, I_m)$.

We rely on the following chaining bound for the quadratic processes, derived in Section D.

Lemma 2 (Chaining Bound)**.** *Let $\mathbb{S}_0^{m\times m}$ be the collection of all symmetric matrices A whose diagonal elements are zero. Endow $\mathbb{S}_0^{m\times m}$ with a metric d given by $d(A, A') \coloneqq \|A - A'\|$. Given $\mathcal{T} \subset \mathbb{S}_0^{m\times m}$ and a fixed $A_0 \in \mathcal{T}$, define $\Delta = \sup_{A\in\mathcal{T}} d(A, A_0)$. Let $N(\delta, \mathcal{T})$ be the covering number of $\mathcal{T}$ under the metric $d(\cdot,\cdot)$, then*

$$\mathop{\mathbb{E}}_{g} \sup_{A\in\mathcal{T}} g^\top A g \le \inf_{J\in\mathbb{N}} \left\{ m\delta_J + 12 \int_{\delta_J/2}^{\Delta/2} \sqrt{2\log N(\delta, \mathcal{T})}\, \mathrm{d}\delta + 24 \int_{\delta_J/2}^{\Delta/2} \log N(\delta, \mathcal{T})\, \mathrm{d}\delta \right\} , \tag{B.1}$$

where for any integer $J \ge 0$, we define $\delta_J = 2^{-J}\Delta$.

With the above chaining bound, we can directly upper bound the Rademacher complexities of the compositional classes such as $R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \mu)$ and $R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1) \circ (\mathrm{Id}, \mathcal{F}), \mu)$. More specifically, for the former class, we will apply this chaining bound to $\mathcal{T} \coloneqq \{A_F : F \in \mathcal{F}\}$. Then, to further bound the RHS of (B.1), we make the following assumptions.

Assumption 5. *Suppose $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ are Lipschitz: there exists $L > 0$ such that*

$$|K_{\mathcal{X}}(x_1, x') - K_{\mathcal{X}}(x_2, x')| \le L\|x_1 - x_2\| \;, \quad |K_{\mathcal{Y}}(y_1, y') - K_{\mathcal{Y}}(y_2, y')| \le L\|y_1 - y_2\| \;.$$

This plays a similar role as Assumption 4: we can derive an upper bound on $d(A_F, A_{F'})$ via closeness of F and F' in $\mathcal{F}$. As a result, we will see that the covering number $N(\delta, \mathcal{T})$ can be bounded by the complexity of $\mathcal{F}$.

Assumption 6. *There exist y_0 and y_0' in $\mathcal{Y}$ with $K_{\mathcal{Y}}(y_0, y_0') \neq K_{\mathcal{Y}}(y_0, y_0)$ such that*

- *$\mathcal{F}$ contains a constant map F satisfying $F(x) = y_0$ for all $x \in \mathcal{X}$,*
- *whenever we have $x \neq x' \in \mathcal{X}$, we can find a non-constant map $F \in \mathcal{F}$ such that $F(x) = y_0$ and $F(x') = y_0'$.*

Similarly, there exist x_0 and x_0' in $\mathcal{X}$ with $K_{\mathcal{X}}(x_0, x_0') \neq K_{\mathcal{X}}(x_0, x_0)$ such that

- *$\mathcal{B}$ contains a constant map B such that $B(y) = x_0$ for all $y \in \mathcal{Y}$,*
- *whenever we have $y \neq y' \in \mathcal{Y}$, we can find a non-constant map $B \in \mathcal{B}$ such that $B(y) = x_0$ and $B(y') = x_0'$.*

The main purpose of this assumption is to exclude overly restrictive $\mathcal{F}$ and $\mathcal{B}$, and is minimal: $\mathcal{F}$ and $\mathcal{B}$ should contain constant maps, as well as non-constant maps. With these assumptions, we can derive the following result.

Proposition 8. *Under Assumptions 2, 3, 5, 6,*

$$R_m(\mathcal{H}_{\mathcal{Y}}(1) \circ \mathcal{F}, \mu) \;,\; R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1) \circ (\mathrm{Id}, \mathcal{F}), \mu) \precsim \sqrt{\frac{\log m}{m} \sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k)} \;,$$

$$R_n(\mathcal{H}_{\mathcal{X}}(1) \circ \mathcal{B}, \mu) \;,\; R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1) \circ (\mathcal{B}, \mathrm{Id}), \nu) \precsim \sqrt{\frac{\log n}{n} \sum_{k=1}^{\dim(\mathcal{X})} \mathrm{Pdim}(\mathcal{B}_k)} \;.$$

In summary, Propositions 4, 5, 7, 8 and Corollary 1 directly imply Theorem 3.

C Representer Theorem and Convex Formulation

This section provides details of the results presented in Section 3.4. Again, without loss of generality, we only consider $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in (3.2).

First, we clarify how measurable maps correspond to bounded linear operators between L^2 spaces.

Proposition 9. *Let $F\colon \mathcal{X} \to \mathcal{Y}$ be a measurable map such that $\| \mathrm{d}F_{\#}\pi_{\mathcal{X}} / \mathrm{d}\pi_{\mathcal{Y}} \|_\infty < \infty$. If we define*

$$\mathbf{F}(g) = g \circ F$$

for all $g \in L^2_{\mathcal{Y}}$, then $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ is a bounded linear operator. Similarly, a measurable map $B\colon \mathcal{Y} \to \mathcal{X}$ satisfying $\| \mathrm{d}B_{\#}\pi_{\mathcal{Y}} / \mathrm{d}\pi_{\mathcal{X}} \|_\infty < \infty$ induces a bounded linear operator $\mathbf{B}\colon L^2_{\mathcal{X}} \to L^2_{\mathcal{Y}}$ such that $\mathbf{B}(g) = g \circ B$ for all $g \in L^2_{\mathcal{X}}$.

Proof. Linearity of $\mathbf{F}$ is obvious. Since

$$\int_{\mathcal{X}} g(F(x))^2 \,\mathrm{d}\pi_{\mathcal{X}} = \int_{\mathcal{Y}} g(y)^2 \,\mathrm{d}F_{\#}\pi_{\mathcal{X}}\,(y) = \int_{\mathcal{Y}} g(y)^2 \frac{\mathrm{d}F_{\#}\pi_{\mathcal{X}}}{\mathrm{d}\pi_{\mathcal{Y}}}(y)\,\mathrm{d}\pi_{\mathcal{Y}}\,(y) \leq \|g\|^2_{L^2(\pi_{\mathcal{Y}})} \left\| \frac{\mathrm{d}F_{\#}\pi_{\mathcal{X}}}{\mathrm{d}\pi_{\mathcal{Y}}} \right\|_\infty ,$$

we can see $\mathbf{F}(g) = g \circ F \in L^2_{\mathcal{X}}$ and thus $\mathbf{F}\colon L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$. From this inequality, the operator norm of $\mathbf{F}$ is bounded by $\| \mathrm{d}F_{\#}\pi_{\mathcal{X}} / \mathrm{d}\pi_{\mathcal{Y}} \|_\infty^{1/2}$; hence boundedness of $\mathbf{F}$ follows. The same argument applies to $\mathbf{B}$. □

Next, we prove that (3.2) can be written in terms of $\mathbf{F}$ and $\mathbf{B}$ if $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ are given by the Mercer's representation:

$$K_{\mathcal{X}}(x, x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(x') \,, \tag{C.1}$$

$$K_{\mathcal{Y}}(y, y') = \sum_{\ell=1}^{\infty} \gamma_\ell \psi_\ell(y) \psi_\ell(y') \,. \tag{C.2}$$

Let $\Phi_x = [\cdots, \phi_k(x), \cdots]^\top \in \mathbb{R}^\infty$ and $\Psi_y = [\cdots, \psi_\ell(y), \cdots]^\top \in \mathbb{R}^\infty$. Then, $K_{\mathcal{X}}(x, x') = \Phi_x^\top \Lambda \Phi_{x'}$ and $K_{\mathcal{Y}}(y, y') = \Psi_y^\top \Gamma \Psi_{y'}$. Also,

$$\begin{aligned} K_{\mathcal{X}}(x, B(y)) &= \sum_k \lambda_k \phi_k(x) [\phi_k \circ B](y) \\ &= \sum_k \lambda_k \phi_k(x) \mathbf{B}[\phi_k](y) \\ &= \sum_{k,\ell} \lambda_k \phi_k(x) \mathbf{B}_{\ell k} \psi_\ell(y) \\ &= \Psi_y^\top \mathbf{B} \Lambda \Phi_x \,. \end{aligned}$$

Analogously, we can obtain

$$\begin{aligned} K_{\mathcal{Y}}(F(x), y) &= \Phi_x^\top \mathbf{F} \Gamma \Psi_y \,, \\ K_{\mathcal{X}}(B(y), B(y')) &= \Psi_y^\top \mathbf{B} \Lambda \mathbf{B}^\top \Psi_{y'} \,, \\ K_{\mathcal{Y}}(F(x), F(x')) &= \Phi_x^\top \mathbf{F} \Gamma \mathbf{F}^\top \Phi_{x'} \,. \end{aligned}$$

Using this, we have

$$\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (K_{\mathcal{X}}(x_i, B(y_j)) - K_{\mathcal{Y}}(F(x_i), y_j))^2 = \frac{1}{mn} \sum_{i,j} (\Psi_{y_j}^\top \mathbf{B} \Lambda \Phi_{x_i} - \Phi_{x_i}^\top \mathbf{F} \Gamma \Psi_{y_j})^2 \,. \tag{C.3}$$

Also,

$$\begin{aligned}\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) &= \frac{1}{m^2}\sum_{i,i'} K_\mathcal{X}(x_i, x_{i'}) + \frac{1}{n^2}\sum_{j,j'} K_\mathcal{X}(B(y_j), B(y_{j'})) - \frac{2}{mn}\sum_{i,j} K_\mathcal{X}(x_i, B(y_j)) \\ &= \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x'_i} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \,.\end{aligned} \tag{C.4}$$

Similarly, we have

$$\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) = \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \tag{C.5}$$

and

$$\begin{aligned}&\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n) \\ &= \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x'_i} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} \\ &- \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \,.\end{aligned} \tag{C.6}$$

The following proposition summarizes the discussion so far.

Proposition 10. *Given Borel measures $\pi_\mathcal{X}$ and $\pi_\mathcal{Y}$ over $\mathcal{X}$ and $\mathcal{Y}$, respectively, suppose their corresponding L^2 spaces $L^2_\mathcal{X}$ and $L^2_\mathcal{Y}$ have countable orthonormal bases: $\{\phi_k\}_{k\in\mathbb{N}}$ and $\{\psi_\ell\}_{\ell\in\mathbb{N}}$. Also, assume $K_\mathcal{X}$ and $K_\mathcal{Y}$ are given by the Mercer's representation (C.1) and (C.2). Let $\mathcal{F}_o$ and $\mathcal{B}_o$ be collections of all $F\colon \mathcal{X}\to\mathcal{Y}$ and $B\colon \mathcal{Y}\to\mathcal{X}$ such that $\|\,\mathrm{d}F_\#\pi_\mathcal{X} \,/\, \mathrm{d}\pi_\mathcal{Y}\,\|_\infty < \infty$ and $\|\,\mathrm{d}B_\#\pi_\mathcal{Y} \,/\, \mathrm{d}\pi_\mathcal{X}\,\|_\infty < \infty$, respectively. Then, solving (3.2) over $\mathcal{F}_o \times \mathcal{B}_o$ is equivalent to (3.7), where $\mathcal{C}$ denotes the collection of all pairs of matrices $(\mathbf{F}, \mathbf{B})$ that correspond to a pair of bounded linear operators induced by $(F, B) \in \mathcal{F}_o \times \mathcal{B}_o$. Also, Ω is defined as*

$$\begin{aligned}\Omega(\mathbf{F}, \mathbf{B}) \coloneqq\ & \frac{1}{mn}\sum_{i,j} (\Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} - \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j})^2 \\ &+ \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x'_i} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \\ &+ \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \\ &+ \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x'_i} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} \\ &- \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \,.\end{aligned}$$

Based on this, we now prove Theorem 4.

Proof of Theorem 4. Let $\mathbf{x}_i = \Lambda^{1/2}\Phi_{x_i}$ and $\mathbf{y}_j = \Gamma^{1/2}\Psi_{y_j}$. Notice that we can view them as elements of a Hilbert space $\ell^2_\mathbb{N}$, that is, the space of square-summable sequences:

$$\ell^2_\mathbb{N} = \{(a_k)_{k\in\mathbb{N}} : \sum_k a_k^2 < \infty\} \,.$$

Also, define $\bar{\mathbf{F}} = \Lambda^{-1/2}\mathbf{F}\Gamma^{1/2}$ and $\bar{\mathbf{B}} = \Gamma^{-1/2}\mathbf{B}\Lambda^{1/2}$ where $\Lambda, \Gamma \succ 0$. By rewriting (C.3)-(C.6) using $\mathbf{x}_i$, $\mathbf{y}_j$, $\bar{\mathbf{F}}$, and $\bar{\mathbf{B}}$, we have

$$
\begin{aligned}
&\Omega(\mathbf{F}, \mathbf{B}) \\
&= \frac{1}{mn}\sum_{i,j}(\mathbf{y}_j^\top \bar{\mathbf{B}}\mathbf{x}_i - \mathbf{x}_i^\top \bar{\mathbf{F}}\mathbf{y}_j)^2 && \text{(i)} \\
&+ \frac{1}{m^2}\sum_{i,i'}\mathbf{x}_i^\top \mathbf{x}_{i'} + \frac{1}{n^2}\sum_{j,j'}\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'} - \frac{2}{mn}\sum_{i,j}\mathbf{y}_j^\top \bar{\mathbf{B}}\mathbf{x}_i && \text{(ii)} \\
&+ \frac{1}{m^2}\sum_{i,i'}\mathbf{x}_i^\top \bar{\mathbf{F}}\bar{\mathbf{F}}^\top \mathbf{x}_{i'} + \frac{1}{n^2}\sum_{j,j'}\mathbf{y}_j^\top \mathbf{y}_{j'} - \frac{2}{mn}\sum_{i,j}\mathbf{x}_i^\top \bar{\mathbf{F}}\mathbf{y}_j && \text{(iii)} \\
&+ \frac{1}{m^2}\sum_{i,i'}(\mathbf{x}_i^\top \mathbf{x}_{i'})(\mathbf{x}_i^\top \bar{\mathbf{F}}\bar{\mathbf{F}}^\top \mathbf{x}_{i'}) + \frac{1}{n^2}\sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) - \frac{2}{mn}\sum_{i,j}(\mathbf{y}_j^\top \bar{\mathbf{B}}\mathbf{x}_i)(\mathbf{x}_i^\top \bar{\mathbf{F}}\mathbf{y}_j) && \text{(iv)} \\
&=: \bar{\Omega}(\bar{\mathbf{F}}, \bar{\mathbf{B}}) \, .
\end{aligned}
$$

As a result, (3.8) reduces to $\min_{\bar{\mathbf{F}}, \bar{\mathbf{B}} \in \mathbb{R}^{\infty \times \infty}} \bar{\Omega}(\bar{\mathbf{F}}, \bar{\mathbf{B}})$. Now, we define two finite-dimensional subspaces of $\ell^2_{\mathbb{N}}$ spanned by $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, respectively:

$$
U_m \coloneqq \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \, , \quad V_n \coloneqq \operatorname{span}\{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \, .
$$

Also, we define P_{U_m} and P_{V_n} to be matrices that correspond to the orthogonal projection operators from $\ell^2_{\mathbb{N}}$ to U_m and to V_n, respectively. Recall that P_{U_m} and P_{V_n} are symmetric and idempotent by definition.

Our goal is to prove

$$
\bar{\Omega}(\bar{\mathbf{F}}, \bar{\mathbf{B}}) \geq \bar{\Omega}(P_{U_m}\bar{\mathbf{F}}P_{V_n}, P_{V_n}\bar{\mathbf{B}}P_{U_m}) \, .
$$

More precisely, we show that four terms (i)-(iv) decrease if we replace $\bar{\mathbf{F}}$ and $\bar{\mathbf{B}}$ with $P_{U_m}\bar{\mathbf{F}}P_{V_n}$ and $P_{V_n}\bar{\mathbf{B}}P_{U_m}$, respectively. First, observe that (i) remains the same. By definition, $P_{U_m}\mathbf{x}_i = \mathbf{x}_i$ and $P_{V_n}\mathbf{y}_j = \mathbf{y}_j$, thus $\mathbf{y}_j^\top \bar{\mathbf{B}}\mathbf{x}_i = \mathbf{y}_j^\top P_{V_n}\bar{\mathbf{B}}P_{U_m}\mathbf{x}_i$ and $\mathbf{x}_i^\top \bar{\mathbf{F}}\mathbf{y}_j = \mathbf{x}_i^\top P_{U_m}\bar{\mathbf{F}}P_{V_n}\mathbf{y}_j$. Hence, (i) does not change.

To prove that (ii) decrease, it suffices to prove

$$
\sum_{j,j'}\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'} \geq \sum_{j,j'}\mathbf{y}_j^\top (P_{V_n}\bar{\mathbf{B}}P_{U_m})(P_{V_n}\bar{\mathbf{B}}P_{U_m})^\top \mathbf{y}_{j'} = \sum_{j,j'}\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'} \, .
$$

To this end, define $P_{U_m}^\perp$ to be a matrix that corresponds to the orthogonal projection from $\ell^2_{\mathbb{N}}$ to $U_m^\perp$, the orthogonal complement of U_m. By definition, $P_{U_m} + P_{U_m}^\perp$ is the identity matrix and $P_{U_m}P_{U_m}^\perp = 0$. Hence,

$$
\sum_{j,j'}\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'} = \left\|\bar{\mathbf{B}}^\top \sum_j \mathbf{y}_j\right\|^2 = \left\|P_{U_m}\bar{\mathbf{B}}^\top \sum_j \mathbf{y}_j\right\|^2 + \left\|P_{U_m}^\perp \bar{\mathbf{B}}^\top \sum_j \mathbf{y}_j\right\|^2 \geq \sum_{j,j'}\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}
$$

Here, the second equality is the Pythagorean theorem. Therefore, we can see (ii) decreases if we replace $\bar{\mathbf{B}}$ with $P_{V_n}\bar{\mathbf{B}}P_{U_m}$. Similarly, (iii) decreases.

For (iv), it suffices to prove

$$
\begin{aligned}
\sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) &\geq \sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})\left(\mathbf{y}_j^\top (P_{V_n}\bar{\mathbf{B}}P_{U_m})(P_{V_n}\bar{\mathbf{B}}P_{U_m})^\top \mathbf{y}_{j'}\right) \\
&= \sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) \, .
\end{aligned}
$$

To see this,

$$
\begin{aligned}
\sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) &= \sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) + \sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}^\perp \bar{\mathbf{B}}^\top \mathbf{y}_{j'}) \\
&\geq \sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}}P_{U_m}\bar{\mathbf{B}}^\top \mathbf{y}_{j'}) \, ,
\end{aligned}
$$

where the inequality holds since

$$\sum_{j,j'}(\mathbf{y}_j^\top \mathbf{y}_{j'})(\mathbf{y}_j^\top \bar{\mathbf{B}} P_{U_m}^\perp \bar{\mathbf{B}}^\top \mathbf{y}_{j'}) = \mathrm{Tr}\left[\left(\sum_j P_{U_m}^\perp \bar{\mathbf{B}}^\top \mathbf{y}_j \mathbf{y}_j^\top\right)\left(\sum_j P_{U_m}^\perp \bar{\mathbf{B}}^\top \mathbf{y}_j \mathbf{y}_j^\top\right)^\top\right] \geq 0 \,.$$

Similarly, we can obtain

$$\sum_{i,i'}(\mathbf{x}_i^\top \mathbf{x}_{i'})(\mathbf{x}_i^\top \bar{\mathbf{F}}\bar{\mathbf{F}}^\top \mathbf{x}_{i'}) \geq \sum_{i,i'}(\mathbf{x}_i^\top \mathbf{x}_{i'})(\mathbf{x}_i^\top \bar{\mathbf{F}} P_{V_n} \bar{\mathbf{F}}^\top \mathbf{x}_{i'}).$$

Hence, (iv) decreases.

Consequently, we have

$$(3.8) = \min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(\bar{\mathbf{F}},\bar{\mathbf{B}}) = \min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(P_{U_m}\bar{\mathbf{F}}P_{V_n}, P_{V_n}\bar{\mathbf{B}}P_{U_m}) \,.$$

By definition of a projection operator, we can find $\mathsf{U}_m \in \mathbb{R}^{m\times\infty}$ and $\mathsf{V}_n \in \mathbb{R}^{n\times\infty}$ such that

$$P_{U_m} = \Lambda^{1/2}\Phi_m \mathsf{U}_m \,, \quad P_{V_n} = \Gamma^{1/2}\Psi_n \mathsf{V}_n.$$

By letting $\mathsf{U}_m\bar{\mathbf{F}}\mathsf{V}_n^\top = \mathsf{F}_{m,n} \in \mathbb{R}^{m\times n}$ and $\mathsf{V}_n\bar{\mathbf{B}}\mathsf{U}_m^\top = \mathsf{B}_{n,m} \in \mathbb{R}^{n\times m}$, we have

$$\begin{aligned} P_{U_m}\bar{\mathbf{F}}P_{V_n} &= \Lambda^{1/2}\Phi_m \mathsf{F}_{m,n}\Psi_n^\top\Gamma^{1/2} \,, \\ P_{V_n}\bar{\mathbf{B}}P_{U_m} &= \Gamma^{1/2}\Psi_n \mathsf{B}_{n,m}\Phi_m^\top\Lambda^{1/2} \,. \end{aligned}$$

Hence,

$$\min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(P_{U_m}\bar{\mathbf{F}}P_{V_n}, P_{V_n}\bar{\mathbf{B}}P_{U_m}) = \min_{(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\in\mathsf{C}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \,,$$

where

$$\omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \coloneqq \bar{\Omega}(\Lambda^{1/2}\Phi_m\mathsf{F}_{m,n}\Psi_n^\top\Gamma^{1/2}, \Gamma^{1/2}\Psi_n\mathsf{B}_{n,m}\Phi_m^\top\Lambda^{1/2}) \,.$$

Here, C is a constraint set implying that $\mathsf{F}_{m,n}$ and $\mathsf{B}_{n,m}$ are associated with $\bar{\mathbf{F}}$ and $\bar{\mathbf{B}}$, respectively, namely,

$$\mathsf{C} = \{(\mathsf{U}_m\bar{\mathbf{F}}\mathsf{V}_n^\top, \mathsf{V}_n\bar{\mathbf{B}}\mathsf{U}_m^\top) : \bar{\mathbf{F}},\bar{\mathbf{B}} \in \mathbb{R}^{\infty\times\infty}\} \subset \mathbb{R}^{m\times n}\times\mathbb{R}^{n\times m} \,.$$

Therefore,

$$(3.8) = \min_{(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\in\mathsf{C}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \geq \min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \,.$$

Finally, note that $\mathsf{C} = \mathbb{R}^{m\times n}\times\mathbb{R}^{n\times m}$ if U_m and V_n are full rank, that is, row spaces of U_m and V_n are rank-m and rank-n, respectively. This is true if kernel matrices

$$\mathbf{K}_{\mathcal{X}} = (\Lambda^{1/2}\Phi_m)^\top(\Lambda^{1/2}\Phi_m) \,, \; \mathbf{K}_{\mathcal{Y}} = (\Gamma^{1/2}\Psi_n)^\top(\Gamma^{1/2}\Psi_n)$$

are invertible. This is equivalent to say that they are positive definite. In this case,

$$\mathsf{U}_m = \mathbf{K}_{\mathcal{X}}^{-1}(\Lambda^{1/2}\Phi_m)^\top \,, \; \mathsf{V}_n = \mathbf{K}_{\mathcal{Y}}^{-1}(\Gamma^{1/2}\Psi_n)^\top \,,$$

which are indeed full rank. Accordingly, we have

$$(3.8) = \min_{(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\in\mathsf{C}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) = \min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \,.$$

Finally, we prove ω is convex. To see this, verify

$$\begin{aligned}\omega(\mathsf{F}_{m,n}, \mathsf{B}_{n,m}) = \frac{1}{mn}\|\mathbf{K}_{\mathcal{Y}}\mathsf{B}_{n,m}\mathbf{K}_{\mathcal{X}} - \mathbf{K}_{\mathcal{Y}}\mathsf{F}_{m,n}^{\top}\mathbf{K}_{\mathcal{X}}\|^2 \\ + \left\|\mathbf{K}_{\mathcal{X}}^{1/2} \cdot \left(\frac{1}{m}\mathbf{1}_m - \mathsf{B}_{n,m}^{\top}\mathbf{K}_{\mathcal{Y}}\frac{1}{n}\mathbf{1}_n\right)\right\|^2 + \left\|\mathbf{K}_{\mathcal{Y}}^{1/2} \cdot \left(\frac{1}{n}\mathbf{1}_n - \mathsf{F}_{m,n}^{\top}\mathbf{K}_{\mathcal{X}}\frac{1}{m}\mathbf{1}_m\right)\right\|^2 \\ + \left\|\frac{1}{m}\mathbf{K}_{\mathcal{X}}^{3/2}\mathsf{F}_{m,n}\mathbf{K}_{\mathcal{Y}}^{1/2} - \frac{1}{n}\mathbf{K}_{\mathcal{X}}^{1/2}\mathsf{B}_{n,m}^{\top}\mathbf{K}_{\mathcal{Y}}^{3/2}\right\|^2,\end{aligned}$$

where $\mathbf{K}_{\mathcal{X}}^{1/2}$ and $\mathbf{K}_{\mathcal{Y}}^{1/2}$ are the square root matrices of $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_{\mathcal{Y}}$, respectively, and $\mathbf{1}_m \in \mathbb{R}^m$ and $\mathbf{1}_n \in \mathbb{R}^n$ are all-ones vectors. □

D Details of Section B

D.1 Proofs in Section B

Proof of Proposition 4. Let $h_{F,B}(x,y) \coloneqq (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2$, then

$$\begin{aligned}\widehat{C}_0(F,B) - C_0(F,B) &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathop{\mathbb{E}}_{(x,y)\sim\mu\otimes\nu} h_{F,B}(x,y) \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y)\right) \\ &\quad + \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y) - \mathop{\mathbb{E}}_{(x,y)\sim\mu\otimes\nu} h_{F,B}(x,y)\,.\end{aligned}$$

Assumption 1 implies that a function $x \mapsto \mathbb{E}_{y\sim\nu} h_{F,B}(x,y)$ is bounded in $[0,H]$. Thus, by the McDiarmid's inequality,

$$\frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y) - \mathop{\mathbb{E}}_{(x,y)\sim\mu\otimes\nu} h_{F,B}(x,y) \le \sqrt{\frac{H^2\log(1/\delta)}{2m}}$$

holds with probability at least $1-\delta$. By the same logic, for fixed x_i,

$$\frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y) \le \sqrt{\frac{H^2\log(1/\delta)}{2n}}$$

holds with probability at least $1-\delta$, where the probability is the conditional probability of $y_1,\dots,y_n$ given $x_1,\dots,x_m$. Since this is true for all x_i, the union bound implies

$$\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i,y_j) - \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y)\right) \le \sqrt{\frac{H^2\log(m/\delta)}{2n}}$$

holds with probability at least $1-\delta$. Hence,

$$\widehat{C}_0(F,B) - C_0(F,B) \precsim \sqrt{\frac{\log(m/\delta)}{m}} \le \sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}}$$

holds with probability at least $1-2\delta$. The same result holds for $C_0(F,B) - \widehat{C}_0(F,B)$, hence we complete the proof. □

Proof of Proposition 5. By the triangle inequality, $|\widehat{M}(F,B) - M(F,B)|$ is bounded above by the sum of the following three terms:

$$|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \ ,$$
$$|\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}}(\mu, B_\#\nu)| \ ,$$
$$|\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\mu, (B, \mathrm{Id})_\#\nu)| \ .$$

First, we give an upper bound on the first term. Boundedness of kernels (Assumption 2) implies

$$\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) \ , \ \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\mu, \nu) \leq 2\sqrt{K} \ .$$

Hence,

$$|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \leq 4\sqrt{K}|\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\mu, \nu)| \ .$$

Due to the triangle inequality of MMD, we have

$$|\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\mu, \nu)| \leq \mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, F_\#\mu) + \mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu) \ .$$

By Theorem 3.4 of [40],

$$\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu) \leq \sqrt{\frac{K}{n}} + \sqrt{\frac{2K\log(1/\delta)}{n}}$$

holds with probability at least $1-\delta$. Next, note that $F_\#\widehat{\mu}_m = \frac{1}{m}\sum_i \delta_{F(x_i)}$ is the empirical measure constructed from $\{F(x_i)\}_{i=1}^m$. Since they are m many i.i.d. samples from $F_\#\mu$, by the same theorem,

$$\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, F_\#\mu) \leq \sqrt{\frac{K}{m}} + \sqrt{\frac{2K\log(1/\delta)}{m}}$$

holds with probability at least $1-\delta$. Hence,

$$|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu, \nu)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}$$

holds with probability at least $1-2\delta$. Similarly, we have

$$|\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}}(\mu, B_\#\nu)| \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}} \ ,$$
$$|\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\mu, (B, \mathrm{Id})_\#\nu)|$$
$$\precsim \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}} \ ,$$

each of which holds with probability at least $1-2\delta$. Combining these three probabilistic bounds, we obtain a bound for $|\widehat{M}(F,B) - M(F,B)|$. □

Proof of Proposition 6. Without loss of generality, assume $n \geq m$. From the proof of Proposition 4,

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)| \leq \frac{1}{m}\sum_{i=1}^m \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{n}\sum_{j=1}^n h_{F,B}(x_i, y_j) - \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i, y) \right|$$
$$+ \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i, y) - \mathop{\mathbb{E}}_{x\sim\mu}\mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x, y) \right| \ .$$

Since $x \mapsto \mathbb{E}_{y\sim\nu} h_{F,B}(x,y)$ is bounded in $[0,H]$, Lemma 3 implies

$$\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y) - \mathop{\mathbb{E}}_{x\sim\mu}\mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x,y) \right| \\ \precsim \sqrt{\frac{\log(1/\delta)}{m}} + \mathop{\mathbb{E}}_{x_i}\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \epsilon_i \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y) \right|$$

holds with probability at least $1-\delta$. Since $n \geq m$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_i}\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \epsilon_i \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i,y) \right| &= \mathop{\mathbb{E}}_{x_i}\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \mathop{\mathbb{E}}_{y_1,\ldots,y_m} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i h_{F,B}(x_i,y_i) \right| \\
&\leq \mathop{\mathbb{E}}_{x_i}\mathop{\mathbb{E}}_{\epsilon_i}\mathop{\mathbb{E}}_{y_1,\ldots,y_m} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \epsilon_i h_{F,B}(x_i,y_i) \right| \\
&= \mathop{\mathbb{E}}_{x_i}\mathop{\mathbb{E}}_{y_i}\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m}\sum_{i=1}^{m} \epsilon_i h_{F,B}(x_i,y_i) \right| .
\end{aligned}$$

We first give an upper bound on

$$\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \underbrace{\left| \frac{1}{m}\sum_{i=1}^{m} \epsilon_i h_{F,B}(x_i,y_i) \right|}_{=:X_{F,B}} .$$

First, observe that Assumption 1 and Assumption 4 imply

$$\begin{aligned}
|h_{F,B}(x,y) - h_{F',B'}(x,y)| &\leq \left| \sqrt{h_{F,B}(x,y)} + \sqrt{h_{F',B'}(x,y)} \right| \left| \sqrt{h_{F,B}(x,y)} - \sqrt{h_{F',B'}(x,y)} \right| \\
&\leq 2\sqrt{H} \left(|c_{\mathcal{X}}(x,B(y)) - c_{\mathcal{X}}(x,B'(y))| + |c_{\mathcal{Y}}(F(x),y) - c_{\mathcal{Y}}(F'(x),y)| \right) \\
&\leq 2\sqrt{H}L \left(\|F(x) - F'(x)\| + \|B(y) - B'(y)\| \right) \\
&= 2\sqrt{H}L \left[\sqrt{\sum_{k=1}^{\dim(\mathcal{Y})} |F_k(x) - F'_k(x)|^2} + \sqrt{\sum_{\ell=1}^{\dim(\mathcal{X})} |B_\ell(y) - B'_\ell(y)|^2} \right] \\
&\leq 2\sqrt{H}L \left(\sum_{k=1}^{\dim(\mathcal{Y})} |F_k(x) - F'_k(x)| + \sum_{\ell=1}^{\dim(\mathcal{X})} |B_\ell(y) - B'_\ell(y)| \right) .
\end{aligned}$$

Therefore,

$$\begin{aligned}
|X_{F,B} - X_{F',B'}| &\leq \frac{1}{m}\sum_{i=1}^{m} |h_{F,B}(x_i,y_i) - h_{F',B'}(x_i,y_i)| \\
&\leq 2\sqrt{H}L \left(\sum_{k=1}^{\dim(\mathcal{Y})} \frac{1}{m}\sum_{i=1}^{m} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \frac{1}{m}\sum_{i=1}^{m} |B_\ell(y_i) - B'_\ell(y_i)| \right) \\
&\leq 2\sqrt{H}L \left(\sum_{k=1}^{\dim(\mathcal{Y})} \max_{i\in[m]} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \max_{i\in[m]} |B_\ell(y_i) - B'_\ell(y_i)| \right) \\
&=: \rho((F,B),(F',B')) .
\end{aligned}$$

For $\epsilon > 0$, let $\mathcal{N}_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)$ be the minimal ϵ-covering net of $\mathcal{F}_k$ under the pseudometric d induced by $x_1, \ldots, x_m$:

$$d(F_k, F_k') \coloneqq \max_{i\in[m]} |F_k(x_i) - F_k(x_i')| \,.$$

In other words, for any $F_k \in \mathcal{F}_k$, we can find $F_k' \in \mathcal{N}_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)$ such that $d(F_k, F_k') \le \epsilon$. Also, $|\mathcal{N}_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)| = N_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m)$. We define $\mathcal{N}_\infty(\epsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)$ in a similar fashion.

Given $\epsilon > 0$, let $T_\epsilon = \otimes_{k=1}^{\dim(\mathcal{Y})} \mathcal{N}_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m) \times \otimes_{\ell=1}^{\dim(\mathcal{X})} \mathcal{N}_\infty(\epsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)$. Then, for any $(F, B) \in \mathcal{F} \times \mathcal{B}$, we can find $(F', B') \in T_\epsilon$ such that

$$\rho((F, B), (F', B')) \le \eta\epsilon \,,$$

where $\eta = 2\sqrt{H}L(\dim(\mathcal{X}) + \dim(\mathcal{Y}))$. As a result, one can easily check

$$\begin{aligned}\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} X_{F,B} &\le \sup_{\rho((F,B),(F',B'))\le\eta\epsilon} |X_{F,B} - X_{F',B'}| + \sup_{(F,B)\in T_\epsilon} X_{F,B} \\ &\le \eta\epsilon + \sup_{(F,B)\in T_\epsilon} X_{F,B} \,.\end{aligned}$$

Note that $X_{F,B}$ is the absolute value of a sub-Gaussian random variable with parameter $H^2/(4m)$. Hence, the maximal inequality yields

$$\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} X_{F,B} \le \eta\epsilon + \mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in T_\epsilon} X_{F,B} \le \eta\epsilon + \sqrt{\frac{H^2 \log(|T_\epsilon|)}{m}} \,.$$

Using $|T_\epsilon| = \prod_{k=1}^{\dim(\mathcal{Y})} N_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m) \times \prod_{\ell=1}^{\dim(\mathcal{X})} N_\infty(\epsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)$, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left|\frac{1}{m}\sum_{i=1}^m \epsilon_i h_{F,B}(x_i, y_i)\right| \\ &\le \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)}{m}} \\ &\le \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, m)}{m}} \,.\end{aligned}$$

The second inequality is obvious from the definition of the uniform covering number. Since the last equation is independent of x_i and y_i, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{x_i}\mathop{\mathbb{E}}_{y_i}\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left|\frac{1}{m}\sum_{i=1}^m \epsilon_i h_{F,B}(x_i, y_i)\right| \\ &\le \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, m)}{m}} \,.\end{aligned}$$

As a result,

$$\begin{aligned}&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left|\frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i, y) - \mathop{\mathbb{E}}_{x\sim\mu}\mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x, y)\right| \\ &\lesssim \sqrt{\frac{\log(1/\delta)}{m}} + \epsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, m)}{m}} \\ &\le \sqrt{\frac{\log(1/\delta)}{m}} + \epsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, n)}{m}}\end{aligned} \tag{D.1}$$

holds with probability at leat $1-\delta$. Here, $\log N_\infty(\epsilon, \mathcal{B}_\ell, m) \le \log N_\infty(\epsilon, \mathcal{B}_\ell, n)$ holds since $n \ge m$, which is obvious from the definition of the uniform covering number.

Next, we give a bound on

$$\frac{1}{m}\sum_{i=1}^{m} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i, y_j) - \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i, y) \right| .$$

Considering $x_1, \ldots, x_m$ are fixed, Lemma 3 implies

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i, y_j) - \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i, y) \right| \\
&\precsim \sqrt{\frac{\log(1/\delta)}{n}} + \mathop{\mathbb{E}}_{y_j}\mathop{\mathbb{E}}_{\epsilon_j} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \underbrace{\left| \frac{1}{n}\sum_{j=1}^{n} \epsilon_j h_{F,B}(x_i, y_j) \right|}_{Y_{F,B}}
\end{aligned}$$

holds with probability at least $1-\delta$. Here, the probability should be understood as a conditional probability of $y_1, \ldots, y_n$ given $x_1, \ldots, x_m$. Again, we have

$$\begin{aligned}
|Y_{F,B} - Y_{F',B'}| &\le 2\sqrt{H}L \left(\sum_{k=1}^{\dim(\mathcal{Y})} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \frac{1}{n}\sum_{j=1}^{n} |B_\ell(y_j) - B'_\ell(y_j)| \right) \\
&\le 2\sqrt{H}L \left(\sum_{k=1}^{\dim(\mathcal{Y})} \max_{i\in[m]} |F_k(x_i) - F'_k(x_i)| + \sum_{\ell=1}^{\dim(\mathcal{X})} \max_{j\in[n]} |B_\ell(y_j) - B'_\ell(y_j)| \right) .
\end{aligned}$$

Also, $Y_{F,B}$ is the absolute value of a sub-Gaussian random variable with parameter $H^2/(4n)$. By the same argument as before,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y_j}\mathop{\mathbb{E}}_{\epsilon_j} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{n}\sum_{j=1}^{n} \epsilon_j h_{F,B}(x_i, y_j) \right| \\
&\le \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, n)}{n}} .
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{n}\sum_{j=1}^{n} h_{F,B}(x_i, y_j) - \mathop{\mathbb{E}}_{y\sim\nu} h_{F,B}(x_i, y) \right| \\
&\precsim \sqrt{\frac{\log(1/\delta)}{n}} + \epsilon + \sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, n)}{n}}
\end{aligned}$$

holds with probability (conditional probability as explained earlier) at least $1-\delta$. Since this holds for all x_i,

the union bound implies

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}h_{F,B}(x_i,y_j)-\mathop{\mathbb{E}}_{y\sim\nu}h_{F,B}(x_i,y)\right| \\
&\precsim\sqrt{\frac{\log(m/\delta)}{n}}+\epsilon+\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})}\log N_\infty(\epsilon,\mathcal{F}_k,m)+\sum_{\ell=1}^{\dim(\mathcal{X})}\log N_\infty(\epsilon,\mathcal{B}_\ell,n)}{n}} \qquad \text{(D.2)}\\
&\leq\sqrt{\frac{\log(m/\delta)}{m}}+\epsilon+\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})}\log N_\infty(\epsilon,\mathcal{F}_k,m)+\sum_{\ell=1}^{\dim(\mathcal{X})}\log N_\infty(\epsilon,\mathcal{B}_\ell,n)}{m}}
\end{aligned}$$

holds with probability at least $1-\delta$. Combining (D.1) and (D.2), for any $\epsilon>0$, we have

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}|\widehat{C}_0(F,B)-C_0(F,B)| \\
&\precsim\sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}}+\epsilon+\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})}\log N_\infty(\epsilon,\mathcal{F}_k,m)+\sum_{\ell=1}^{\dim(\mathcal{X})}\log N_\infty(\epsilon,\mathcal{B}_\ell,n)}{m\wedge n}}
\end{aligned}$$

holds with probability at least $1-2\delta$. □

Proof of Corollary 1. Combining Assumption 3 and Lemma 4, we have

$$N_\infty(\epsilon,\mathcal{F}_k,m)\leq\left(\frac{2emb}{\epsilon\cdot\mathrm{Pdim}(\mathcal{F}_k)}\right)^{\mathrm{Pdim}(\mathcal{F}_k)}.$$

Hence,

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}|\widehat{C}_0(F,B)-C_0(F,B)| \\
&\precsim\sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}}+\epsilon+\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})}\log N_\infty(\epsilon,\mathcal{F}_k,m)+\sum_{\ell=1}^{\dim(\mathcal{X})}\log N_\infty(\epsilon,\mathcal{B}_\ell,n)}{m\wedge n}} \\
&\leq\sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}}+\epsilon+\sqrt{\frac{\log\left(\frac{2eb(m\vee n)}{\epsilon}\right)}{m\wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})}\mathrm{Pdim}(\mathcal{F}_k)+\sum_{\ell=1}^{\dim(\mathcal{X})}\mathrm{Pdim}(\mathcal{B}_\ell)\right)} \\
&\precsim\sqrt{\frac{\log\left(\frac{m\vee n}{\delta}\right)}{m\wedge n}}+\sqrt{\frac{\log(m\vee n)}{m\wedge n}\left(\sum_{k=1}^{\dim(\mathcal{Y})}\mathrm{Pdim}(\mathcal{F}_k)+\sum_{\ell=1}^{\dim(\mathcal{X})}\mathrm{Pdim}(\mathcal{B}_\ell)\right)}
\end{aligned}$$

holds with probability at least $1-2\delta$, where the last bound comes from choosing $\epsilon=(m\wedge n)^{-1/2}$. □

Proof of Proposition 7. Using the triangle inequality, we bound $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}|\widehat{M}(F,B)-M(F,B)|$ by the sum of the following three terms:

$$\begin{aligned}
&\sup_{F\in\mathcal{F}}|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m,\widehat{\nu}_n)-\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu,\nu)|\,, \\
&\sup_{B\in\mathcal{B}}|\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m,B_\#\widehat{\nu}_n)-\mathrm{MMD}^2_{K_\mathcal{X}}(\mu,B_\#\nu)|\,, \\
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}|\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m,(B,\mathrm{Id})_\#\widehat{\nu}_n)-\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\mu,(B,\mathrm{Id})_\#\nu)|\,.
\end{aligned}$$

As in the proof of Proposition 5, we have

$$\begin{aligned}&\sup_{F\in\mathcal{F}}\left|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m,\widehat{\nu}_n)-\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu,\nu)\right|\\&\leq 4\sqrt{K}\left[\sup_{F\in\mathcal{F}}\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m,F_\#\mu)+\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n,\nu)\right].\end{aligned}$$

$\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n,\nu)$ has already been bounded in Proposition 5. For the first term on the RHS, observe that

$$\begin{aligned}\sup_{F\in\mathcal{F}}\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m,F_\#\mu)&=\sup_{F\in\mathcal{F}}\sup_{f\in\mathcal{H}_\mathcal{Y}(1)}\left|\int f\,\mathrm{d}F_\#\widehat{\mu}_m-\int f\,\mathrm{d}F_\#\mu\right|\\&=\sup_{F\in\mathcal{F}}\sup_{f\in\mathcal{H}_\mathcal{Y}(1)}\left|\int f\circ F\,\mathrm{d}\widehat{\mu}_m-\int f\circ F\,\mathrm{d}\mu\right|\\&=\sup_{f\in\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}}\left|\int f\,\mathrm{d}\widehat{\mu}_m-\int f\,\mathrm{d}\mu\right|,\end{aligned}$$

where the second equality follows from change-of-variables.

First, we show $\mathcal{H}_\mathcal{Y}(1)$ consists of $\sqrt{K}$-uniformly bounded functions. Let $\|\cdot\|_{\mathcal{H}_\mathcal{Y}}$ be the norm of $\mathcal{H}_\mathcal{Y}$ so that $f\in\mathcal{H}_\mathcal{Y}(1)$ is equivalent to $\|f\|_{\mathcal{H}_\mathcal{Y}}\leq 1$. Then, the reproducing property implies

$$|f(y)|\leq\|f\|_{\mathcal{H}_\mathcal{Y}}\sqrt{K_\mathcal{Y}(y,y)}\leq\sqrt{K}$$

for any $f\in\mathcal{H}_\mathcal{Y}(1)$. Accordingly, $\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}$ also consists of $\sqrt{K}$-uniformly bounded functions. Hence, Lemma 3 implies that

$$\begin{aligned}\sup_{F\in\mathcal{F}}\mathrm{MMD}_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m,F_\#\mu)&=\sup_{f\in\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F}}\left|\int f\,\mathrm{d}\widehat{\mu}_m-\int f\,\mathrm{d}\mu\right|\\&\leq 2R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\mu)+\sqrt{\frac{2K\log(1/\delta)}{m}}\end{aligned}$$

holds with probability at least $1-\delta$. Therefore, combining this with the upper bound on $\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n,\nu)$ derived in Proposition 5,

$$\sup_{F\in\mathcal{F}}\left|\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m,\widehat{\nu}_n)-\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\mu,\nu)\right|\precsim R_m(\mathcal{H}_\mathcal{Y}(1)\circ\mathcal{F},\mu)+\sqrt{\frac{\log(1/\delta)}{m}}+\sqrt{\frac{\log(1/\delta)}{n}}\tag{D.3}$$

holds with probability at least $1-2\delta$. Similarly, we can prove that

$$\sup_{B\in\mathcal{B}}\left|\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m,B_\#\widehat{\nu}_n)-\mathrm{MMD}^2_{K_\mathcal{X}}(\mu,B_\#\nu)\right|\precsim R_n(\mathcal{H}_\mathcal{X}(1)\circ\mathcal{B},\nu)+\sqrt{\frac{\log(1/\delta)}{m}}+\sqrt{\frac{\log(1/\delta)}{n}}\tag{D.4}$$

holds with probability at least $1-2\delta$.

Lastly, since $K_\mathcal{X}\otimes K_\mathcal{Y}$ is bounded by K^2, that is,

$$\sup_{(x,y),(x',y')\in\mathcal{X}\times\mathcal{Y}}K_\mathcal{X}\otimes K_\mathcal{Y}((x,y),(x',y'))\leq K^2,$$

by the same argument, we have

$$\begin{aligned}&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}}\left|\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m,(B,\mathrm{Id})_\#\widehat{\nu}_n)-\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\mu,(B,\mathrm{Id})_\#\nu)\right|\\&\leq 4K\left[\sup_{F\in\mathcal{F}}\mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m,(\mathrm{Id},F)_\#\mu)+\sup_{B\in\mathcal{B}}\mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((B,\mathrm{Id})_\#\widehat{\nu}_n,(B,\mathrm{Id})_\#\nu)\right].\end{aligned}$$

Analogously, $\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)$ consists of K-uniformly bounded functions, hence

$$\begin{aligned}
\sup_{F\in\mathcal{F}} \mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m,(\mathrm{Id},F)_\#\mu) &= \sup_{f\in\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F})}\left|\int f\,\mathrm{d}\widehat{\mu}_m - \int f\,\mathrm{d}\mu\right| \\
&\le 2R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F}),\mu) + \sqrt{\frac{2K^2\log(1/\delta)}{m}}\,, \\
\sup_{B\in\mathcal{B}} \mathrm{MMD}_{K_\mathcal{X}\otimes K_\mathcal{Y}}((B,\mathrm{Id})_\#\widehat{\nu}_n,(B,\mathrm{Id})_\#\nu) &= \sup_{f\in\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathcal{B},\mathrm{Id})}\left|\int f\,\mathrm{d}\widehat{\nu}_n - \int f\,\mathrm{d}\nu\right| \\
&\le 2R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathcal{B},\mathrm{Id}),\nu) + \sqrt{\frac{2K^2\log(1/\delta)}{n}}\,,
\end{aligned}$$

each of which holds with probability at least $1-\delta$. Therefore,

$$\begin{aligned}
&\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\widehat{\mu}_m,(B,\mathrm{Id})_\#\widehat{\nu}_n) - \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id},F)_\#\mu,(B,\mathrm{Id})_\#\nu)| \\
&\precsim R_m(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathrm{Id},\mathcal{F}),\mu) + R_n(\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}(1)\circ(\mathcal{B},\mathrm{Id}),\nu) + \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}
\end{aligned} \tag{D.5}$$

holds with probability at least $1-2\delta$. We complete the proof by combining (D.3), (D.4), (D.5). □

Proof of Lemma 2. For any integer $j\in\mathbb{N}\cup\{0\}$, define $\delta_j = 2^{-j}\Delta$, and let $\mathcal{T}_j\subset\mathbb{S}_0^{m\times m}$ be a minimal δ_j-covering net of $\mathcal{T}$; clearly, $|\mathcal{T}_j| = N(\delta_j,\mathcal{T})$. For each j, the covering set induces a mapping $\Pi_j:\mathcal{T}\to\mathcal{T}_j$ such that

$$\sup_{A\in\mathcal{T}} d(A,\Pi_j(A)) \le \delta_j\,.$$

By definition of Δ, we may assume $\mathcal{T}_0 = \{A_0\}$ so that $\Pi_0(A) = A_0$ for all $A\in\mathcal{T}$.

Note that $\mathbb{E}\, g^\top A_0 g = 0$ by definition. Using this, we write $\mathbb{E}\sup_{A\in\mathcal{T}} g^\top A g$ as a chaining sum:

$$\begin{aligned}
\mathbb{E}\sup_{A\in\mathcal{T}} g^\top A g &= \mathop{\mathbb{E}}_{g}\sup_{A\in\mathcal{T}} g^\top A g - \mathbb{E}\, g^\top A_0 g \\
&= \mathbb{E}\sup_{A\in\mathcal{T}}\left(g^\top A g - g^\top A_0 g\right) \\
&= \mathbb{E}\sup_{A\in\mathcal{T}}\left(g^\top A g - g^\top \Pi_J(A) g + \sum_{j=0}^{J-1} g^\top \Pi_{j+1}(A) g - g^\top \Pi_j(A) g\right) \\
&\le \mathbb{E}\sup_{A\in\mathcal{T}}\left(g^\top A g - g^\top \Pi_J(A) g\right) + \sum_{j=0}^{J-1}\mathbb{E}\sup_{A\in\mathcal{T}}\left(g^\top \Pi_{j+1}(A) g - g^\top \Pi_j(A) g\right)\,.
\end{aligned}$$

For the first term on RHS, using the Cauchy-Schwarz inequality and Jensen's inequality, we have

$$\mathbb{E}\sup_{A\in\mathcal{T}}\left(g^\top A g - g^\top \Pi_J(A) g\right) \le \mathbb{E}\left[\Big(\sum_{i\neq j} g_i^2 g_j^2\Big)^{1/2}\cdot\delta_J\right] \le m\delta_J\,.$$

For each summand in the second term on RHS, use Lemma 6. Note that for any j, the maximal cardinality of

$$|\{(\Pi_{j+1}(A),\Pi_j(A))\ :\ A\in\mathcal{T}\}| \le N(\delta_{j+1},\mathcal{T})\times N(\delta_j,\mathcal{T}) \le N(\delta_{j+1},\mathcal{T})^2$$

and that

$$d(\Pi_{j+1}(A),\Pi_j(A)) \le d(\Pi_{j+1}(A),A) + d(A,\Pi_j(A)) \le 3\delta_{j+1}\,.$$

Since $\Pi_{j+1}(A) - \Pi_j(A) \in \mathbb{S}_0^{m\times m}$ and $\|\Pi_{j+1}(A) - \Pi_j(A)\| \le 3\delta_{j+1}$, Lemma 6 asserts that for any j

$$\mathbb{E}\sup_{A\in\mathcal{T}}\left(g^\top \Pi_{j+1}(A)g - g^\top \Pi_j(A)g\right) \le 6\delta_{j+1}\sqrt{2\log N(\delta_{j+1},\mathcal{T})} + 12\delta_{j+1}\log N(\delta_{j+1},\mathcal{T}) \ .$$

Summing over j, we have for any J, the following inequality,

$$\mathbb{E}\sup_{A\in\mathcal{T}}\left(g^\top Ag - g^\top A_0 g\right) \le m\delta_J + 12\int_{\delta_J/2}^{\Delta/2}\sqrt{2\log N(\delta,\mathcal{T})}\,\mathrm{d}\delta + 24\int_{\delta_J/2}^{\Delta/2}\log N(\delta,\mathcal{T})\,\mathrm{d}\delta \ .$$

□

Proof of Proposition 8. To make use of the chaining inequality, we are only left to bound the covering number $N(\delta,\mathcal{T})$ with

$$\mathcal{T} \coloneqq \{A_F : F\in\mathcal{F}\} \subset \mathbb{S}_0^{m\times m} \ .$$

Lipschitzness of $K_{\mathcal{Y}}$ (Assumption 5) implies

$$|K_{\mathcal{Y}}(F(x_i),F(x_j)) - K_{\mathcal{Y}}(F'(x_i),F'(x_j))| \le \frac{L}{2}\|F(x_i)-F'(x_i)\| + \frac{L}{2}\|F(x_j)-F'(x_j)\| \ .$$

Hence,

$$\begin{aligned}
d(A_F,A_{F'}) &\le m\max_{i\neq j}|K_{\mathcal{Y}}(F(x_i),F(x_j)) - K_{\mathcal{Y}}(F'(x_i),F'(x_j))| \\
&\le \frac{mL}{2}\left(\max_{i\in[m]}\|F(x_i)-F'(x_i)\| + \max_{j\in[m]}\|F(x_j)-F'(x_j)\|\right) \\
&= mL\max_{i\in[m]}\|F(x_i)-F'(x_i)\| \\
&\le mL\max_{i\in[m]}\left(\sum_{k=1}^{\dim(\mathcal{Y})}|F_k(x_i)-F_k'(x_i)|^2\right)^{1/2} \\
&\le mL\left[\sum_{k=1}^{\dim(\mathcal{Y})}\left(\max_{i\in[m]}|F_k(x_i)-F_k'(x_i)|\right)^2\right]^{1/2} \ .
\end{aligned}$$

As in the proof of Proposition 6, for $\epsilon>0$, let $\mathcal{N}_\infty(\epsilon,\mathcal{F}_k,\{x_i\}_{i=1}^m)$ be a minimal ϵ-covering net of $\mathcal{F}_k$. Then, one can easily see that

$$\left\{A_F : F\in\otimes_{k=1}^{\dim(\mathcal{Y})}\mathcal{N}_\infty\left(\tfrac{\delta}{mL\sqrt{\dim(\mathcal{Y})}},\mathcal{F}_k,\{x_i\}_{i=1}^m\right)\right\}$$

is a δ-covering of $\mathcal{T}$. Therefore, we conclude

$$N(\delta,\mathcal{T}) \le \prod_{k=1}^{\dim(\mathcal{Y})}N_\infty\left(\tfrac{\delta}{mL\sqrt{\dim(\mathcal{Y})}},\mathcal{F}_k,\{x_i\}_{i=1}^m\right) \le \prod_{k=1}^{\dim(\mathcal{Y})}N_\infty\left(\tfrac{\delta}{mL\sqrt{\dim(\mathcal{Y})}},\mathcal{F}_k,m\right) \ .$$

Lastly, we bound $N_\infty(\delta,\mathcal{F}_k,m)$ by the pseudo-dimension of $\mathcal{F}_k$ via Lemma 4. Combining Assumption 3 and Lemma 4, we have

$$N(\delta,\mathcal{T}) \le \prod_{k=1}^{\dim(\mathcal{Y})}\left(\frac{2emb}{\frac{\delta}{mL\sqrt{\dim(\mathcal{Y})}}\cdot\mathrm{Pdim}(\mathcal{F}_k)}\right)^{\mathrm{Pdim}(\mathcal{F}_k)} \ .$$

Now, to apply Lemma 2, fix $F_0 \in \mathcal{F}$ and let $A_0 = A_{F_0}$. Then,

$$\int_{\delta_J/2}^{\Delta/2} \log N(\delta, \mathcal{T})\, \mathrm{d}\delta \leq \Delta/2 \cdot \sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) \cdot \log\left(\frac{2emb}{\frac{1}{mL\sqrt{\dim(\mathcal{Y})}} \Delta 2^{-J}/2 \cdot \mathrm{Pdim}(\mathcal{F}_k)} \right) .$$

First, we obtain an upper bound on Δ:

$$\Delta = \sup_{F \in \mathcal{F}} \|A_F - A_0\| \leq m \max_{i \neq j} |K_{\mathcal{Y}}(F(x_i), F(x_j)) - K_{\mathcal{Y}}(F_0(x_i), F_0(x_j))| \leq 2mK .$$

Next, we claim that Δ is bounded below by a universal contant; this is to upper bound Δ in the denominator. Consider y_0 and y_0' given in Assumption 6. We may assume that F_0 is the constant map explained in Assumption 6: $F_0(x) = y_0$ for all $x \in \mathcal{X}$. Without loss of generality, we assume $x_1 \neq x_2$. Then, we can find $F \in \mathcal{F}$ such that $F(x_1) = y_0$ and $F(x_2) = y_0'$ according to Assumption 6. Hence,

$$\Delta \geq |K_{\mathcal{Y}}(F(x_1), F(x_2)) - K_{\mathcal{Y}}(F_0(x_1), F_0(x_2))| = |K_{\mathcal{Y}}(y_0, y_0') - K_{\mathcal{Y}}(y_0, y_0)| > 0 .$$

Therefore, with the choice of J such that $m2^{-J} \asymp \sum_k \mathrm{Pdim}(\mathcal{F}_k)$,

$$\begin{aligned} \int_{\delta_J/2}^{\Delta/2} \log N(\delta, \mathcal{T})\, \mathrm{d}\delta &\precsim m \left[\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) \right] \cdot \log\left(\frac{m}{\min_{k \in [\dim(\mathcal{Y})]} \mathrm{Pdim}(\mathcal{F}_k)} \right) \\ &\leq m \log(m) \left[\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) \right] . \end{aligned}$$

Analogously,

$$\int_{\delta_J/2}^{\Delta/2} \sqrt{2 \log N(\delta, \mathcal{T})}\, \mathrm{d}\delta \precsim m \log(m) \left[\sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k) \right] .$$

Thus,

$$R_m(\mathcal{H}_y(1) \circ \mathcal{F}, \{x_i\}_{i=1}^m) \precsim \frac{1}{m} \left[mK + \mathop{\mathbb{E}}_{g} \sup_{F \in \mathcal{F}} g^\top A_F g \right]^{1/2} \precsim \sqrt{\frac{\log m}{m} \sum_{k=1}^{\dim(\mathcal{Y})} \mathrm{Pdim}(\mathcal{F}_k)}$$

The same argument can be applied to the other three Rademacher complexities. Hence, we have proved the proposition. □

D.2 Auxiliary Lemmas

Lemma 3 (Theorem 4.10 of [53]). *Let $(\mathcal{Z}, \rho)$ be a probability space and $\mathcal{G}$ be a class of b-uniformly bounded measurable functions defined on $\mathcal{Z}$, that is, $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq b$. Let $z_1, \ldots, z_m$ are i.i.d. samples from ρ and let $\widehat{\rho}_m$ be the empirical measure constructed from them. Then, for any $\delta > 0$,*

$$\sup_{g \in \mathcal{G}} \left| \int g\, \mathrm{d}\widehat{\rho}_m - \int g\, \mathrm{d}\rho \right| \leq 2R_m(\mathcal{G}, \rho) + \sqrt{\frac{2b^2 \log(1/\delta)}{m}}$$

holds with probability at least $1 - \delta$.

Lemma 4 (Theorem 12.2 of [1]). *Let $\mathcal{G}$ be a collection of real-valued functions defined on a set $\mathcal{Z}$. Suppose $\sup_{g \in \mathcal{G}} ||g||_\infty = b < \infty$. For $\epsilon > 0$ and $m \geq \mathrm{Pdim}(\mathcal{G})$,*

$$N_\infty(\epsilon, \mathcal{G}, m) \leq \left(\frac{2emb}{\epsilon \cdot \mathrm{Pdim}(\mathcal{G})} \right)^{\mathrm{Pdim}(\mathcal{G})} .$$

Lemma 5 (Example 2.12 of [7]). *For any $A \in \mathbb{S}_0^{m\times m}$ and $0 \le \lambda < 1/(2\|A\|_{\mathrm{op}})$,*

$$\log \mathop{\mathbb{E}}_{g} e^{\lambda g^\top A g} \le \frac{\lambda^2 \|A\|}{1 - 2\lambda\|A\|_{\mathrm{op}}} , \tag{D.6}$$

where $g \sim N(0, I_m)$. Here, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of A.

This lemma tells that $g^\top A g$ is a sub-Gamma random variable with variance factor $2\|A\|^2$ and scale parameter $2\|A\|_{\mathrm{op}}$ (see Chapter 2.4 of [7] for the definition). Using Corollary 2.6 of the same text, we can derive the following maximal inequality.

Lemma 6 (Maximal inequality). *For $A_1, \dots, A_N \in \mathbb{S}_0^{m\times m}$, suppose $\max_{i=1,\dots,N} \|A_i\| \le \delta$. Then,*

$$\mathop{\mathbb{E}}_{g} \max_{i=1,\dots,N} g^\top A_i g \le 2\delta \left(\sqrt{\log N} + \log N \right) , \tag{D.7}$$

where $g \sim N(0, I_m)$.

E Computational Aspects of the RGM Distance

Recall from Section 3.2 that we have utilized the Lagrangian form instead of the constrained form (3.1) of the RGM distance. Such a practical computation allows us to easily implement the RGM sampler and we have observed its good empirical performance in Section 4. This section shifts our focus to the exact computation of the RGM distance; we discuss conditions under which minimizing the Lagrangian form leads to a close approximation to the RGM distance. Another important computational aspect is the comparison of the RGM distance and the GW distance. $\mathrm{GW}(\mu,\nu) \le \mathrm{RGM}(\mu,\nu)$ holds in theory by Proposition 1; using a concrete example, we approximate both quantities numerically and see how large the gap between them is. Lastly, we examine the numerical performance of the convex formulation discussed in 3.4 and compare with the results from the Lagrangian form.

E.1 Approximation with the Lagrangian Form

First, we derive connections between the RGM distance and its Lagrangian formulation. As in Section B, let

$$C_0(F,B) = \int (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \, \mathrm{d}\mu \otimes \nu$$

for any $F\colon \mathcal{X} \to \mathcal{Y}$ and $B\colon \mathcal{Y} \to \mathcal{X}$ so that $\mathrm{RGM}(\mu,\nu)^2 = \inf_{(F,B)\in\mathcal{I}(\mu,\nu)} C_0(F,B)$. Define the Lagrangian form as

$$L_{\lambda_1,\lambda_2,\lambda_3}(F,B) = C_0(F,B) + \lambda_1 \cdot \underbrace{\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id},F)_{\#}\mu, (B,\mathrm{Id})_{\#}\nu)}_{\ell_1(F,B)} + \lambda_2 \cdot \underbrace{\mathcal{L}_{\mathcal{X}}(\mu, B_{\#}\nu)}_{\ell_2(F,B)} + \lambda_3 \cdot \underbrace{\mathcal{L}_{\mathcal{Y}}(F_{\#}\mu, \nu)}_{\ell_3(F,B)} ,$$

where $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$, $\mathcal{L}_{\mathcal{X}}$, and $\mathcal{L}_{\mathcal{Y}}$ are suitable nonnegative discrepancy measures as in Section 3.2; in particular, we assume $\ell_1(F,B) = \ell_2(F,B) = \ell_3(F,B) = 0$ for $(F,B) \in \mathcal{I}(\mu,\nu)$.[5]

Suppose $\lambda_1, \lambda_2, \lambda_3 \ge 0$, then

$$\inf_{\substack{F\colon \mathcal{X}\to\mathcal{Y} \\ B\colon \mathcal{Y}\to\mathcal{X}}} L_{\lambda_1,\lambda_2,\lambda_3}(F,B) \le \inf_{(F,B)\in\mathcal{I}(\mu,\nu)} L_{\lambda_1,\lambda_2,\lambda_3}(F,B) = \mathrm{RGM}(\mu,\nu)^2 = \inf_{(F,B)\in\mathcal{I}(\mu,\nu)} C_0(F,B) .$$

We seek a minimizer of $L_{\lambda_1,\lambda_2,\lambda_3}$ over $\mathcal{F}\times\mathcal{B}$, namely, the product of suitable function classes as discussed in Section 3.3. Let

$$(F^\star, B^\star) \in \mathop{\arg\min}_{(F,B)\in\mathcal{F}\times\mathcal{B}} L_{\lambda_1,\lambda_2,\lambda_3}(F,B) . \tag{E.1}$$

[5] Though we may define the Lagrangian form without $\mathcal{L}_{\mathcal{X}}$ and $\mathcal{L}_{\mathcal{Y}}$, we include them for a seamless connection with the experiment results in Section E.2.

If $\mathcal{I}(\mu,\nu) \subseteq \mathcal{F} \times \mathcal{B}$,

$$\begin{aligned} C_0(F^\star, B^\star) &\le L_{\lambda_1,\lambda_2,\lambda_3}(F^\star, B^\star) \quad (\because \lambda_1, \lambda_2, \lambda_3 \ge 0) \\ &= \inf_{(F,B)\in\mathcal{F}\times\mathcal{B}} L_{\lambda_1,\lambda_2,\lambda_3}(F, B) \quad (\because \text{(E.1)}) \\ &\le \inf_{(F,B)\in\mathcal{I}(\mu,\nu)} L_{\lambda_1,\lambda_2,\lambda_3}(F, B) \quad (\because \mathcal{I}(\mu,\nu) \subseteq \mathcal{F}\times\mathcal{B}) \\ &= \mathrm{RGM}(\mu,\nu)^2 \,. \end{aligned}$$

Roughly speaking, if the function classes are rich enough to ensure $\mathcal{I}(\mu,\nu) \subseteq \mathcal{F}\times\mathcal{B}$, the minimizer $(F^\star, B^\star)$ produces a lower bound $C_0(F^\star, B^\star)$ on $\mathrm{RGM}(\mu,\nu)^2$. On the other hand, if the minimizer satisfies the constraint $(F^\star, B^\star) \in \mathcal{I}(\mu,\nu)$, then $C_0(F^\star, B^\star)$ is an upper bound on $\mathrm{RGM}(\mu,\nu)^2$ by definition.

Therefore, a sufficient condition for $C_0(F^\star, B^\star) = \mathrm{RGM}(\mu,\nu)^2$ is that the following two hold: $\mathcal{I}(\mu,\nu) \subseteq \mathcal{F}\times\mathcal{B}$ and $(F^\star, B^\star) \in \mathcal{I}(\mu,\nu)$.

E.2 Numerical Experiments

Using a concrete example, we compute the aforementioned quantities related to the RGM distance, approximate the GW distance, and compare them; we will also discuss the results from the convex formulation in Section 3.4. Throughout, we consider two point clouds on $\mathbb{R}^2$ as in Figure 4(a), that is, μ and ν are uniform distributions supported on 30 grid points of a segment and a circle, respectively. We fix the cost functions: $c_{\mathcal{X}} = c_{\mathcal{Y}}$ is the RBF kernel that maps (x, y) to $\exp\bigl(-\|x-y\|^2\bigr)$.

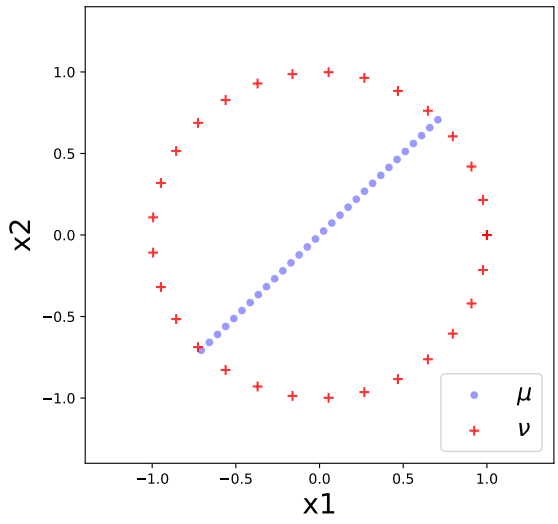


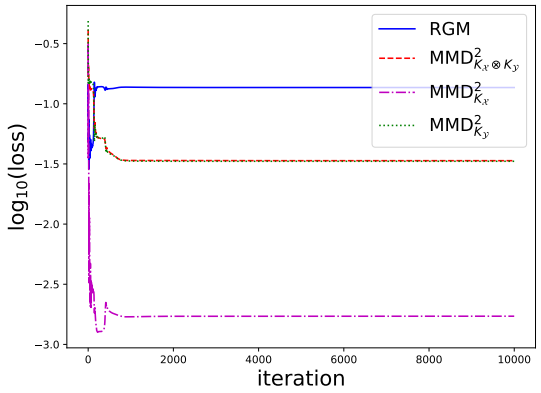
Figure 4: The supports of μ and ν are the grid points of a segment connecting $(-1,-1)$ and $(1,1)$ and a circle $x^2+y^2=1$, respectively.

First, we aim to compute the quantities discussed in Section E.1. To this end, we specify the discrepancy measures and the function classes as follows.

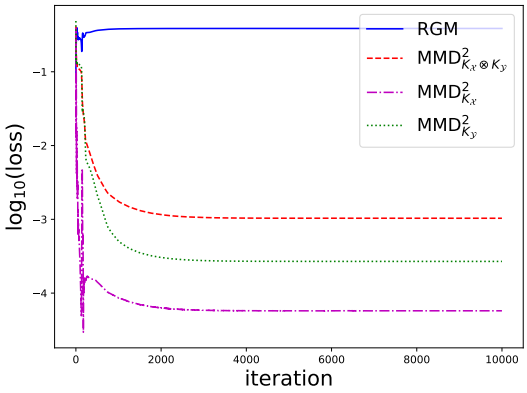
- $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}} = \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}$, $\mathcal{L}_{\mathcal{X}} = \mathrm{MMD}^2_{K_{\mathcal{X}}}$, $\mathcal{L}_{\mathcal{Y}} = \mathrm{MMD}^2_{K_{\mathcal{Y}}}$, where $K_{\mathcal{X}} = K_{\mathcal{Y}} = c_{\mathcal{X}}$.
- $\mathcal{F} = \mathcal{B}$ is the class of neural networks with two hidden layers as follows:

$$\{x \mapsto \tanh(W_2 \tanh(W_1 x + b_1)) + b_2) : W_1 \in \mathbb{R}^{2\times 30}, b_1 \in \mathbb{R}^{30}, W_2 \in \mathbb{R}^{30\times 2}, b_2 \in \mathbb{R}^2\} \,,$$

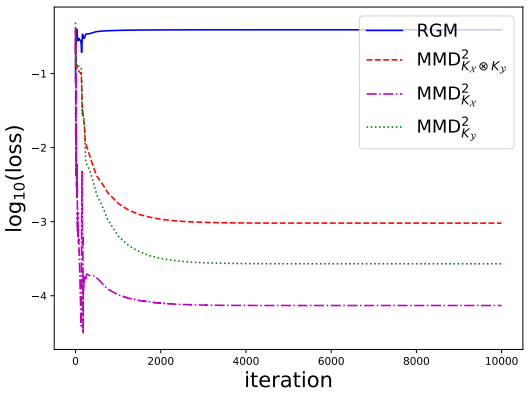
where tanh is the tangent hyperbolic function applied elementwise, that is, $\tanh(x) = (\tanh(x_1), \ldots, \tanh(x_k)) \in \mathbb{R}^k$ for $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$.



(a) $\lambda_1 = \lambda_2 = \lambda_3 = 1$



(b) $\lambda_1 = \lambda_2 = \lambda_3 = 10^2$



(c) $\lambda_1 = \lambda_2 = \lambda_3 = 10^4$

Figure 5: Training curves (10000 iterations).

We use Adam [26], a variant of stochastic gradient descent, to find a minimizer of $L_{\lambda_1,\lambda_2,\lambda_3}$ over $\mathcal{F} \times \mathcal{B}$ as in (E.1). After 10000 iterations, we can see that the loss converges as in Figure 5, indicating we have a local minimizer $(\widehat{F}^\star, \widehat{B}^\star) \in \mathcal{F} \times \mathcal{B}$. Since this optimization problem may be nonconvex, there is no guarantee that this is a global minimizer, hence

$$C_0(F^\star, B^\star) \leq C_0(\widehat{F}^\star, \widehat{B}^\star) \, .$$

To examine the experiment results in light of Section E.1, we will assume

- $\mathcal{F}$ and $\mathcal{B}$ are rich enough to ensure $\mathcal{I}(\mu, \nu) \subset \mathcal{F} \times \mathcal{B}$,
- $(F, B) \in \mathcal{I}(\mu, \nu)$ if and only if $\ell_1(F, B) = \ell_2(F, B) = \ell_3(F, B) = 0$,[6]
- $(\widehat{F}^\star, \widehat{B}^\star)$ is indeed a global minimizer: $(\widehat{F}^\star, \widehat{B}^\star) \in \arg\min_{(F,B) \in \mathcal{F} \times \mathcal{B}} L_{\lambda_1,\lambda_2,\lambda_3}(F, B)$.

[6]This is true since we are using the RBF kernel as mentioned in Section 4.

Under these assumptions, $(\widehat{F}^\star, \widehat{B}^\star) \in \mathcal{I}(\mu,\nu)$ implies $C_0(\widehat{F}^\star, \widehat{B}^\star) = \mathrm{RGM}(\mu,\nu)^2$ as discussed in Section E.1. To verify $(\widehat{F}^\star, \widehat{B}^\star) \in \mathcal{I}(\mu,\nu)$, we check the values of $\ell_1(F,B)$, $\ell_2(F,B)$, and $\ell_3(F,B)$ in Table 1. We observe that they get smaller as we increase the values of the Lagrangian multipliers. For the cases where $\lambda_1 = \lambda_2 = \lambda_3 = 10^2$ or $\lambda_1 = \lambda_2 = \lambda_3 = 10^4$, the values ℓ_1, ℓ_2, ℓ_3 are sufficiently small to conclude $(\widehat{F}^\star, \widehat{B}^\star) \in \mathcal{I}(\mu,\nu)$, hence we can roughly estimate $\mathrm{RGM}(\mu,\nu)^2 \approx C_0(\widehat{F}^\star, \widehat{B}^\star) \approx 0.39$.

	$\lambda_1 = \lambda_2 = \lambda_3 = 1$	$\lambda_1 = \lambda_2 = \lambda_3 = 10^2$	$\lambda_1 = \lambda_2 = \lambda_3 = 10^4$
$C_0(\widehat{F}^\star, \widehat{B}^\star)$	0.136	0.386	0.390
$\ell_1(\widehat{F}^\star, \widehat{B}^\star)$	3.366×10^{-2}	1.034×10^{-3}	9.550×10^{-4}
$\ell_2(\widehat{F}^\star, \widehat{B}^\star)$	1.716×10^{-3}	5.758×10^{-5}	7.379×10^{-5}
$\ell_3(\widehat{F}^\star, \widehat{B}^\star)$	3.327×10^{-2}	2.689×10^{-4}	2.698×10^{-4}
$L_{\lambda_1,\lambda_2,\lambda_3}(\widehat{F}^\star, \widehat{B}^\star)$	0.205	0.522	13.377

Table 1: Minimum values of the Lagrangian form.

Comparison with GW As discussed earlier, exact compuation of the GW distance (Definition 2) is impossible in general. Here, we estimate it using an off-the-shelf computational tool called Python Optimal Transport (POT) [20] widely used in literature, which yields $\mathrm{GW}^2(\mu,\nu) \approx 0.171$.[7] Combined with the previous computation, we can say

$$1 \le \frac{\mathrm{RGM}(\mu,\nu)}{\mathrm{GW}(\mu,\nu)} \approx \sqrt{\frac{0.390}{0.171}} = 1.515 \,,$$

indicating that the RGM distance is approximately the GW distance times 1.5. This rough computation is based on the aforementioned assumptions regarding the RGM computation and the accuracy of POT in computing the GW distance.

Instead, we may give an upper bound on the ratio of the two distances using the well-known lower bounds on the GW distance: the First Lower Bound (FLB) and the Second Lower Bound (SLB) on GW [37]. Letting $\mu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ with $m = n = 30$, these bounds are computed as follows:

$$\mathrm{FLB}^2(\mu,\nu) = W_2^2\left(\frac{1}{m}\sum_{i=1}^m \delta_{e_{\mathcal{X}}(x_i)}, \frac{1}{n}\sum_{j=1}^n \delta_{e_{\mathcal{Y}}(y_j)}\right) \approx 0.061 \,,$$

$$\mathrm{SLB}^2(\mu,\nu) = W_2^2\left(\frac{1}{m^2}\sum_{i,i'=1}^m \delta_{c_{\mathcal{X}}(x_i,x_{i'})}, \frac{1}{n^2}\sum_{j,j'=1}^n \delta_{c_{\mathcal{Y}}(y_j,y_{j'})}\right) \approx 0.135 \,,$$

where $e_{\mathcal{X}}(x_i) = \sqrt{\frac{1}{m}\sum_{i'=1}^m c_{\mathcal{X}}^2(x_i,x_{i'})}$ and $e_{\mathcal{Y}}(y_j) = \sqrt{\frac{1}{n}\sum_{j'=1}^n c_{\mathcal{Y}}^2(y_j,y_{j'})}$ are called the eccentricity; see [37] for details. These quantities, computed by using POT as well, are known to be lower bounds on $\mathrm{GW}(\mu,\nu)^2$, hence

$$1 \le \frac{\mathrm{RGM}(\mu,\nu)}{\mathrm{GW}(\mu,\nu)} \le \sqrt{\frac{0.390}{0.135}} = 1.700 \,.$$

Therefore, we can conclude that the ratio of the two distances is bounded by 1.7.

[7] Technically, this should be an upper bound on the exact value of $\mathrm{GW}^2(\mu,\nu)$ because the result of POT ought to be a local minimizer.

Convex formulation Next, we estimate the RGM distance based on the convex formulation; as in Theorem 4, we solve the convex optimization problem:

$$\min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\ , \tag{E.2}$$

where

$$\omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) = \overbrace{\frac{1}{mn}\|\mathbf{K}_{\mathcal{Y}}\mathsf{B}_{n,m}\mathbf{K}_{\mathcal{X}} - \mathbf{K}_{\mathcal{Y}}\mathsf{F}_{m,n}^{\top}\mathbf{K}_{\mathcal{X}}\|^2}^{c_0(\mathsf{F}_{m,n},\mathsf{B}_{n,m})} + \lambda_1 \cdot \overbrace{\left\|\frac{1}{m}\mathbf{K}_{\mathcal{X}}^{3/2}\mathsf{F}_{m,n}\mathbf{K}_{\mathcal{Y}}^{1/2} - \frac{1}{n}\mathbf{K}_{\mathcal{X}}^{1/2}\mathsf{B}_{n,m}^{\top}\mathbf{K}_{\mathcal{Y}}^{3/2}\right\|^2}^{m_1(\mathsf{F}_{m,n},\mathsf{B}_{n,m})}$$
$$+ \lambda_2 \cdot \underbrace{\left\|\mathbf{K}_{\mathcal{X}}^{1/2}\cdot\left(\frac{1}{m}\mathbf{1}_m - \mathsf{B}_{n,m}^{\top}\mathbf{K}_{\mathcal{Y}}\frac{1}{n}\mathbf{1}_n\right)\right\|^2}_{m_2(\mathsf{F}_{m,n},\mathsf{B}_{n,m})} + \lambda_3 \cdot \underbrace{\left\|\mathbf{K}_{\mathcal{Y}}^{1/2}\cdot\left(\frac{1}{n}\mathbf{1}_n - \mathsf{F}_{m,n}^{\top}\mathbf{K}_{\mathcal{X}}\frac{1}{m}\mathbf{1}_m\right)\right\|^2}_{m_3(\mathsf{F}_{m,n},\mathsf{B}_{n,m})}$$

as derived in Section C. It should be noted that the minimum of (E.2) is a lower bound on the minimum of the Lagrangian, that is,

$$\min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \le \min_{(F,B)\in\mathcal{F}\times\mathcal{B}} L_{\lambda_1,\lambda_2,\lambda_3}(F,B)\ .$$

To see this, first note that the RHS is exactly (3.2).[8] Then, recall from Section 3.4 that (3.2) = (3.7) is relaxed to (3.8) and is further relaxed to the convex problem (E.2) due to Theorem 4.

	$\lambda_1=\lambda_2=\lambda_3=1$	$\lambda_1=\lambda_2=\lambda_3=10^2$	$\lambda_1=\lambda_2=\lambda_3=10^4$
$c_0(\mathsf{F}^\star_{m,n},\mathsf{B}^\star_{n,m})$	0.044	0.106	0.108
$m_1(\mathsf{F}^\star_{m,n},\mathsf{B}^\star_{n,m})$	0.011	4.521×10^{-6}	6.369×10^{-10}
$m_2(\mathsf{F}^\star_{m,n},\mathsf{B}^\star_{n,m})$	0.012	2.592×10^{-6}	2.626×10^{-12}
$m_3(\mathsf{F}^\star_{m,n},\mathsf{B}^\star_{n,m})$	0.001	1.050×10^{-7}	2.461×10^{-10}
$\omega(\mathsf{F}^\star_{m,n},\mathsf{B}^\star_{n,m})$	0.068	0.107	0.108

Table 2: Mimimum values of the convex problem (E.2) obtained by CVXPY.

Table 2 shows the results obtained by a convex optimization tool called CVXPY [15]. The mimimum $\omega(\mathsf{F}^\star_{m,n},\mathsf{B}^\star_{n,m})$ of (E.2) is always smaller than $\mathrm{GW}^2(\mu,\nu)\approx 0.171$ and is between the two lower bounds: $\mathrm{FLB}^2(\mu,\nu)\approx 0.061$ and $\mathrm{SLB}^2(\mu,\nu)\approx 0.135$. Also, the MMD terms vanish if we use the large Lagrangian multipliers, indicating that the constraints (represented via the MMD terms) are met. That said, we can see that the gap between the minimum and the RGM distance can be large. Therefore, finding conditions under which this gap vanishes would be interesting future work.

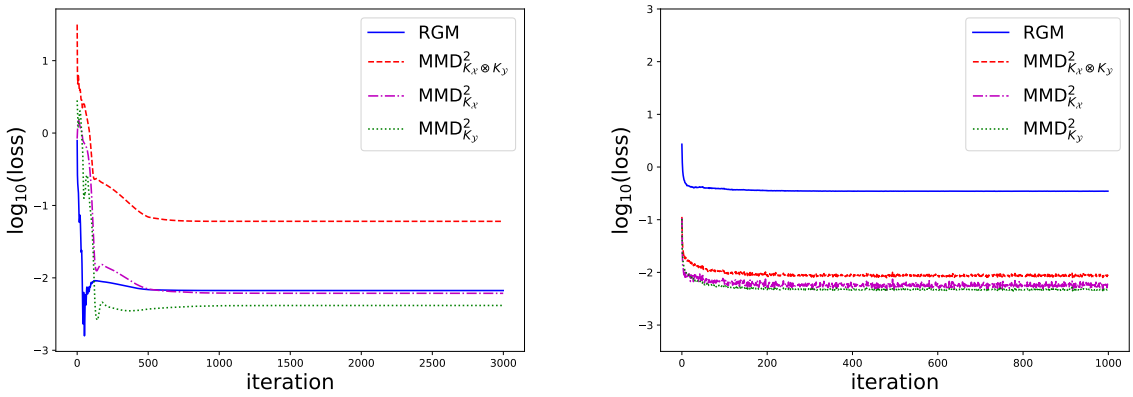
F Details of the Experiments in Section 4

Here, we provide implementation details of the experiments in Section 4.

F.1 Gaussian

In the Gaussian experiment in Section 4, we minimize (3.2) using Adam for 3000 iterations. The learning rate at the initial iteration is 0.1 and we halve it after every 500 iterations. Figure 6(a) shows the training curve.

[8] Since μ and ν are discrete, the RHS is the same as its empirical estimate (3.2).



(a) Gaussian (b) MNIST

Figure 6: Training curves for the experiments: (a) the Gaussian experiment, (b) the MNIST experiment with $\mathcal{X} = \mathbb{R}^2$.

F.2 MNIST

$\mathbb{R}^2$ and MMDs For numerical stability, we encode a variant of empirical estimate (3.2) as our loss:

$$\lambda_1 \cdot \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(c_\mathcal{X}(x_i, B(y_j)) - c_\mathcal{Y}(F(x_i), y_j))^2 + \mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n)$$
$$+ \lambda_2 \cdot \mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) + \lambda_3 \cdot \mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n)\,. \tag{F.1}$$

We choose tuning parameters $(\lambda_1, \lambda_2, \lambda_3) = (0.01, 1, 1)$. For fully connected neural networks F and B, we apply the rectified linear unit (ReLU) activation function $\sigma(x) = \max(x, 0)$ to all three hidden layers of F and B, and an additional tangent hyperbolic (tanh) function $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ to the output layer of F, both of which are elementwise activation functions. To put it explicitly, $y = F(x)$ is defined as

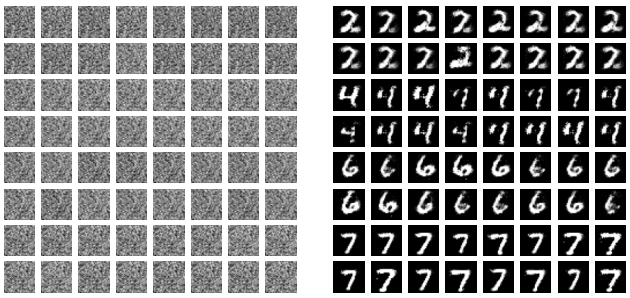
$$\begin{aligned} h_0 &= x, \quad x \in \mathbb{R}^2 \\ h_l &= \sigma(W_l h_{l-1} + b_l), \quad l = 1, 2 \\ y &= \tanh(W_3 h_2 + b_3) \end{aligned}$$

with $W_1 \in \mathbb{R}^{50\times 2}, W_2 \in \mathbb{R}^{50\times 50}, W_3 \in \mathbb{R}^{784\times 50}, b_1, b_2 \in \mathbb{R}^{50\times 1}, b_3 \in \mathbb{R}^{784\times 1}$. Similarly, $\widetilde{x} = B(\widetilde{y})$ is defined as

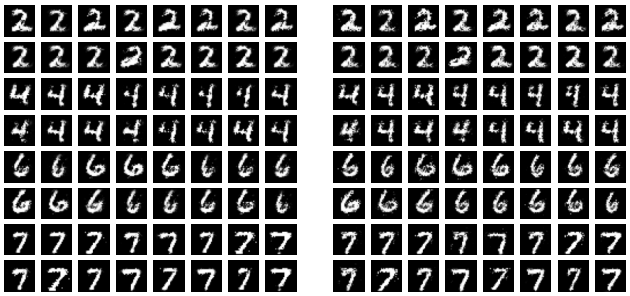
$$\begin{aligned} \widetilde{h}_0 &= \widetilde{y}, \quad \widetilde{y} \in \mathbb{R}^{784} \\ \widetilde{h}_l &= \sigma(\widetilde{W}_l \widetilde{h}_{l-1} + \widetilde{b}_l), \quad l = 1, 2 \\ \widetilde{x} &= \widetilde{W}_3 \widetilde{h}_2 + \widetilde{b}_3 \end{aligned}$$

with $\widetilde{W}_1 \in \mathbb{R}^{50\times 784}, \widetilde{W}_2 \in \mathbb{R}^{50\times 50}, \widetilde{W}_3 \in \mathbb{R}^{2\times 50}, \widetilde{b}_1, \widetilde{b}_2 \in \mathbb{R}^{50\times 1}, \widetilde{b}_3 \in \mathbb{R}^{2\times 1}$. The training set is randomly devided into minibatches of size 256, for which we run Adam again for 1000 iterations. The learning rate at the initial iteration is 0.005 and we halve it after every 500 iterations. Figure 7 shows the generated images during the training process.

$\mathbb{R}^4$ and Sinkhorn divergences We implement the Sinkhorn divergence with squared Euclidean cost by using GeomLoss [19]. Concretely, we first define the entropic regularized Kantorovich problem between



(a) Before training (b) After 20 iterations



(c) After 50 iterations (d) After 1000 iterations

Figure 7: Generated images on MNIST data for digits $2, 4, 6, 7$ during the training process, under the experimental setup with $\mathbb{R}^2$ and MMD discrepancy.

$\widehat{\mu}_m$ and $B_{\#}\widehat{\nu}_n$ on some Euclidean space $\mathcal{X}$

$$W^2_{2,\epsilon}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) = \min_{\gamma \in \Pi(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n)} \sum_{i=1}^{m}\sum_{j=1}^{n} \gamma_{ij} \left(\|x_i - B(y_j)\|^2 + \epsilon \log(\gamma_{ij}) \right)$$

where γ is a coupling matrix and γ_{ij} denotes its (i, j) element. Then the Sinkhorn divergence between two empirical measures is defined by $s_{\epsilon,\mathcal{X}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) = W^2_{2,\epsilon}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) - \frac{1}{2}W^2_{2,\epsilon}(\widehat{\mu}_m, \widehat{\mu}_m) - \frac{1}{2}W^2_{2,\epsilon}(B_{\#}\widehat{\nu}_n, B_{\#}\widehat{\nu}_n)$. We encode our loss by replacing MMDs in (F.1) with Sinkhorn divergences

$$\begin{aligned} &\lambda_1 \cdot \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2 + s_{\epsilon,\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n) \\ &+ \lambda_2 \cdot s_{\epsilon,\mathcal{X}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3 \cdot s_{\epsilon,\mathcal{Y}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n) \ , \end{aligned}$$

and choose tuning parameters $(\lambda_1, \lambda_2, \lambda_3) = (1, 1, 1)$. The Sinkhorn parameter ϵ is set to be 0.0001 for all three discrepancy measures. Again, $F\colon \mathbb{R}^4 \to \mathbb{R}^{784}$ and $B\colon \mathbb{R}^{784} \to \mathbb{R}^4$ are parametrized by fully connected neural networks with three hidden layers, whose activation functions are same as the MMD case. The rest of the setups, including the choice of optimizer, number of iterations, and batchsize, are same as the MMD case above.

Comparison between RGM and GW Lastly, let us estimate the gap between the RGM distance and the GW distance in the MNIST example ($\mathbb{R}^2$ and MMDs) based on the discussions in Section E. Recall that we have obtained a minimizer $(\widehat{F}, \widehat{B})$ of (3.2) over $\mathcal{F} \times \mathcal{B}$ using samples $\{x_i\}_{i=1}^{20000}$ and $\{y_j\}_{j=1}^{20000}$ from $\mu = N(0, I_2)$ and $\nu =$ the distribution of the four digits, respectively; though this is a local minimizer as the optimization problem may be nonconvex, we will assume that this is indeed a global minimizer as in Section E. Letting $\widehat{\mu}_m$ and $\widehat{\nu}_n$ be the empirical measures constructed by $\{x_i\}_{i=1}^{20000}$ and $\{y_j\}_{j=1}^{20000}$, respectively $(m = n = 20000)$, we obtain the following quantities:

$$\begin{aligned}
\widehat{C}_0(\widehat{F}, \widehat{B}) &\overset{\text{Section B}}{=} \int (c_{\mathcal{X}}(x, \widehat{B}(y)) - c_{\mathcal{Y}}(\widehat{F}(x), y))^2 \, \mathrm{d}\widehat{\mu}_m \otimes \widehat{\nu}_n \approx 0.348 \,, \\
m_1 &:= \mathrm{MMD}^2_{K_{\mathcal{X}} \otimes K_{\mathcal{Y}}}((\mathrm{Id}, \widehat{F})_{\#}\widehat{\mu}_m, (\widehat{B}, \mathrm{Id})_{\#}\widehat{\nu}_n) \approx 2.192 \times 10^{-3} \,, \\
m_2 &:= \mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, \widehat{B}_{\#}\widehat{\nu}_n) \approx 8.261 \times 10^{-5} \,, \\
m_3 &:= \mathrm{MMD}^2_{K_{\mathcal{Y}}}(\widehat{F}_{\#}\widehat{\mu}_m, \widehat{\nu}_n) \approx 1.437 \times 10^{-3} \,.
\end{aligned}$$

Hence, $C(\widehat{\mu}_m, \widehat{\nu}_n, \widehat{F}, \widehat{B}) = \widehat{C}_0(\widehat{F}, \widehat{B}) + \sum_{k=1}^3 \lambda_k m_k = 0.719$, where C is defined in (3.3). Recall from Section B that

$$\begin{aligned}
|C(\widehat{\mu}_m, \widehat{\nu}_n, \widehat{F}, \widehat{B}) - C(\mu, \nu, \widehat{F}, \widehat{B})| &\le \sup_{(F,B) \in \mathcal{F} \times \mathcal{B}} |C(\widehat{\mu}_m, \widehat{\nu}_n, F, B) - C(\mu, \nu, F, B)| \\
&\lesssim \mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta)
\end{aligned}$$

holds with probability at least $1 - \delta$, where $\mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta)$ is defined in Theorem 3. Assuming that this complexity measure is sufficiently small for $m = n = 20000$, we may roughly say $C(\mu, \nu, \widehat{F}, \widehat{B}) \approx C(\widehat{\mu}_m, \widehat{\nu}_n, \widehat{F}, \widehat{B}) = 0.719$. In the same vein, Theorem 3 indicates

$$\inf_{(F,B) \in \mathcal{F} \times \mathcal{B}} C(\mu, \nu, F, B) \approx C(\mu, \nu, \widehat{F}, \widehat{B}) \approx 0.719 \,.$$

Now, we combine this result with the discussion in Section E. First, by definition,

$$\inf_{(F,B) \in \mathcal{F} \times \mathcal{B}} C(\mu, \nu, F, B) = \inf_{(F,B) \in \mathcal{F} \times \mathcal{B}} L_{\lambda_1, \lambda_2, \lambda_3}(F, B) \,.$$

We have derived in Section E that

$$\mathrm{RGM}(\mu, \nu)^2 = C_0(F^\star, B^\star) = L_{\lambda_1, \lambda_2, \lambda_3}(F^\star, B^\star) = \inf_{(F,B) \in \mathcal{F} \times \mathcal{B}} L_{\lambda_1, \lambda_2, \lambda_3}(F, B)$$

if $\mathcal{I}(\mu, \nu) \subseteq \mathcal{F} \times \mathcal{B}$ and the minimizer $(F^\star, B^\star)$ defined in (E.1) satisfies $(F^\star, B^\star) \in \mathcal{I}(\mu, \nu)$. Therefore, under these assumptions, we can roughly estimate

$$\mathrm{RGM}^2(\mu, \nu) \approx 0.719.$$

Lastly, let us estimate the lower bounds, FLB and SLB, on the GW distance as in Section E.2. Due to the computational complexity, we will use subsets of the training data, say $\{x_i\}_{i=1}^{2000}$ and $\{y_j\}_{j=1}^{2000}$, to construct plug-in estimators $\mathrm{FLB}(\widehat{\mu}_{2000}, \widehat{\nu}_{2000})$ and $\mathrm{SLB}(\widehat{\mu}_{2000}, \widehat{\nu}_{2000})$; using POT again, we have

$$\begin{aligned}
\mathrm{FLB}^2(\mu, \nu) &\approx \mathrm{FLB}^2(\widehat{\mu}_{2000}, \widehat{\nu}_{2000}) \approx 0.006 \,, \\
\mathrm{SLB}^2(\mu, \nu) &\approx \mathrm{SLB}(\widehat{\mu}_{2000}, \widehat{\nu}_{2000}) \approx 0.148 \,.
\end{aligned}$$

Therefore, we can roughly conclude

$$1 \le \frac{\mathrm{RGM}(\mu, \nu)}{\mathrm{GW}(\mu, \nu)} \le \sqrt{\frac{0.719}{0.148}} = 2.204 \,.$$