Reversible Gromov-Monge Sampler for Simulation-Based Inference

YoonHaeng Hur, Wenxuan Guo, and Tengyuan Liang*

University of Chicago

September 28, 2021

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 usual approaches of explicitly modeling the density function or designing Markov chain Monte Carlo. Motivated by the seminal work of [1] and [2] 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 in order to perform simulation-based inference. Our RGM sampler can also estimate optimal alignments between two heterogenous 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 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 difficult as one has to control correlations between successive samples, or to 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 non-parametric 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 transform $T : [0,1] \to \mathbb{R}$ called the inverse Cumulative

*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.

Distribution Function (CDF),

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

Define the pushforward measure $T_{\#}\mu$ by $T_{\#}\mu(S) = \mu(\{x : T(x) \in S\})$ for any Borel set $S \subseteq \mathbb{R}$, and 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 (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) [3], $\mathcal{F}$ 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 [4], Wasserstein distances for Wasserstein-GAN [5], and Maximum Mean Discrepancies (MMD) for MMD-GAN [6, 7].

The optimal transport 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 [8] 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

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

where $\Pi(\mu, \nu)$ denotes all couplings of (μ, ν). Brenier established that for $p = 2$, there exists a unique optimal transport map $T^\star$ given as the gradient of some convex function. Also, there exists a unique optimal coupling $\gamma^\star$ and $\gamma^\star = (\mathrm{Id}, T^\star)_{\#}\mu$ holds. 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 $T : [0, 1] \to \mathbb{R}$ 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 $T : \mathbb{R}^d \to \mathbb{R}^d$ is the gradient of a convex function, as opposed to a monotonic map $\mathbb{R} \to \mathbb{R}$.

Elegant as it is, one limitation of the above optimal transport theory is that both μ and ν are supported on a common Euclidean space. To see why such a limitation is a non-trifling issue in practice, consider a simple task where one aims to transform a multivariate Gaussian sample (say $\mathcal{X} = \mathbb{R}^{10}$) to a sample of the MNIST image ($\mathcal{Y} \subset \mathbb{R}^{784}$). The probability distribution of MNIST images are supported on some image manifold $\mathcal{Y}$, which clearly differs from the usual Euclidean space $\mathcal{X}$. It turns out, when $\mathcal{X}$ and $\mathcal{Y}$ are two heterogenous spaces, Gromov-Wasserstein (GW) distance was proposed to attack the aforementioned limitation. Let $c_{\mathcal{X}}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_{\mathcal{Y}}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be two continuous cost functions on $\mathcal{X}$ and $\mathcal{Y}$,

respectively, [1] defined a notion of GW distance as

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

A few remarks regarding the comparison between Wasserstein and Gromov-Wasserstein are warranted here: First, unlike Wasserstein which solves for an infinite-dimensional linear program in the coupling γ, GW formulates a Quadratic Assignment Program (QAP) in γ, which is known to be computationally hard; Second, GW aims to match the cost functions defined on two heterogenous spaces over a pair of couplings ($\gamma(x,y)$ and $\gamma(x',y')$), and thus holds promise to identify isomorphism between spaces. Despite being an elegant notion of distance between metric measure spaces (see [1, Definition 5.1]), GW is hard to compute in practice due to its QAP 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 heterogenous 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 carry naturally over to our RGM.
- Our RGM formulation naturally induces a transform sampler, as a relaxation of the usual GW formulation. Rather than solving a QAP 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 $F:\mathcal{X}\to\mathcal{Y}$ and $\mathcal{B}:\mathcal{Y}\to\mathcal{X}$, respectively, and then later bind them via the constraints $(\mathrm{Id},F)_{\#}\mu \approx (B,\mathrm{Id})_{\#}\nu$. 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 QAP 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,B. Based on our non-asymptotic results, concrete upper bounds can be easily spelled out in the cases when F,B are parametrized by deep neural networks. As mentioned, 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 that are 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 (RGM) Sampler, with detailed derivations deferred later. To be specific, Section 4 collects analytical properties of the RGM metric, Section 5 derives the non-asymptotic rate of convergence analyzing the statistical properties of the RGM sampler, and Section 6 discusses a further relaxation of the RGM into an infinite-dimensional convex program which relies on a new representer theorem. Finally, synthetic and real-world examples showcasing the effectiveness of the RGM sampler are demonstrated in Section 7. Other proof details omitted in the main text are designated to Appendix A.

1.1 Related Literature

Modern data sets are mostly in an unstructured form. Inferring the underlying probability distributions from data has been a central problem in statistics and 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 [9, 10], and studies the minimax optimality of directly estimating such density functions [11]. It is also unclear how to proceed to sample from a possibly improper[1] density estimator, even with an optimal estimator at hand. One may employ Markov Chain Monte Carlo (MCMC) techniques for sampling from certain models. However, on the computational front, it is highly non-trivial how to ensure the mixing properties of MCMC for a designed sampler [12, 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 [13, 14], or implicitly by parametrizing the pushforward map transporting distributions [6, 3], 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, [15, 7, 6] on MMDs, [16, 17] on integral probability metrics, and [18, 19, 20, 21, 22, 23, 24, 25] on generative adversarial networks. Last but not the least, we emphasize that an alternative implicit distribution estimation approach using the simulated method of moments has been formulated in the econometrics literature since [26, 27] and [28].

The Gromov-Wasserstein distance was introduced in [29] as a relaxation of the Gromov-Hausdorff distance widely used for comparing metric spaces. Analytic properties of the Gromov-Wasserstein distance have been studied extensively [1, 2]; the most important one is that it defines a distance between metric measure spaces, namely, metric spaces endowed with probability measures. Since it is possible to model a number of real-world data sets via metric measure spaces, the Gromov-Wasserstein distance has been utilized in various problems that aim to compare two data sets such as shape correspondence [30], graph matching [31], and network matching [32].

Being a quadratic assignment program, computation of the Gromov-Wasserstein distances is intractable in general. QAP, which dates back to [33], is known to be NP-hard [34] in the worst case. As a result, several approaches have been proposed for the approximate computation of Gromov-Wasserstein distance, its upper and lower bounds, and other variants. [1] studies lower bounds on the Gromov-Wasserstein distance that are easier to compute. [35] adds an entropic regularization term to the Gromov-Wasserstein distance, which leads to a fast iterative algorithm. Recently, the Sliced Gromov-Wasserstein distance has been proposed by [36], which amounts to integrating the Gromov-Wasserstein distances over one-dimensional projections.

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. We denote the Frobenius norm of a matrix A as $\|A\|$. For any $x \in \mathbb{R}^p$ with $p \in \mathbb{N} \cup \{\infty\}$, we denote its ℓ^2 norm as $\|x\|$. Given a set $\mathcal{X}$ and a function $f\colon \mathcal{X} \to \mathbb{R}$, we define $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$ with the essential supremum. For an integer $n \in \mathbb{N}$, we define $[n] = \{1, \dots, n\}$. Given a Polish space $\mathcal{X}$, that is, a complete and separable metric space, 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 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 $\mathcal{T}(\mu, \nu) := \{T : \mathcal{X} \to \mathcal{Y} \mid T_\# \mu = \nu\}$. $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is called a coupling 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 $B \subset \mathcal{Y}$. 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 asymptotic relationship that $\limsup_{n\to\infty} a(n)/b(n) < \infty$.

[1] Here we mean that the estimated density is non-negative and integrates to 1.

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 given Polish probability spaces. To be concrete, let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be Polish probability spaces. Consider a measurable function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$; we view $c(x, y)$ as the cost 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 between $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$. 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.

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 $(\mathrm{Id}, T)\colon \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$ with $(\mathrm{Id}, T)(x) = (x, T(x))$. One can verify $(\mathrm{Id}, T)_{\#}\mu \in \Pi(\mu, \nu)$. Therefore, if we define $\Pi_{\mathcal{T}} := \{(\mathrm{Id}, T)_{\#}\mu : T \in \mathcal{T}(\mu, \nu)\}$, 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 cost 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. More precisely, if $\mathcal{X} = \mathcal{Y}$ and $c = d_{\mathcal{X}}^2$, then the minimum of Kantorovich's 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 heterogenous 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) is not suited for comparing heterogenous Polish probability spaces.

Mémoli's pioneering work [1] 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 integration $\mathrm{d}\gamma\otimes\gamma$, 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. This definition is based on Definition 1 of [32]. 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 [1] and [2]. In short, a network space is a generalization of a metric measure space.

Like the Wasserstein distance, the Gromov-Wasserstein 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 Gromov-Wasserstein distance satisfies all metric axioms on the quotient space—under the equivalence relation $\cong$—of metric measure spaces.

Theorem 1 (Lemma 1.10 of [2])**.** *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.

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 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 suitable reproducing kernel Hilbert spaces (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 $\mathrm{d}\mu \otimes \nu := \mathrm{d}\mu\,(x)\,\mathrm{d}\nu\,(y)$ 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*

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

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 $h(x) = \exp(-\alpha x^2)$ 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 on analytic properties of RGM to Section 4.

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 $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}} = 0$ 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 $\mathcal{L}_{\mathcal{X}}$, 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 MMD as the leading example for two reasons. First, MMD is a metric between Borel probability measures under mild conditions. Also, the square of MMD between discrete distributions admits a closed form. For instance, let $K_{\mathcal{X}}$ be a kernel on $\mathcal{X}$, then MMD between $\widehat{\mu}_m$ and $B_{\#}\widehat{\nu}_n$ is

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

By choosing kernels $K_{\mathcal{X}}, K_{\mathcal{Y}}, K_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X}, \mathcal{Y}, \mathcal{X}\times\mathcal{Y}$, respectively, we can specify $\mathcal{L}_{\mathcal{X}}, \mathcal{L}_{\mathcal{Y}}, \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ as the square of corresponding MMDs. 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

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

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:

$$\min_{\substack{F:\,\mathcal{X}\to\mathcal{Y}\\ B:\,\mathcal{Y}\to\mathcal{X}}} \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(c_{\mathcal{X}}(x_i,B(y_j))-c_{\mathcal{Y}}(F(x_i),y_j))^2+\lambda_1\cdot\mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id},F)_{\#}\widehat{\mu}_m,(B,\mathrm{Id})_{\#}\widehat{\nu}_n) \\ +\lambda_2\cdot\mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m,B_{\#}\widehat{\nu}_n)+\lambda_3\cdot\mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m,\widehat{\nu}_n)\,. \tag{3.1}$$

Once we solve the problem above, the solution $\widehat{F}:\mathcal{X}\to\mathcal{Y}$ 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 $\widehat{B}:\mathcal{Y}\to\mathcal{X}$ 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 (3.1) 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 [37, 38]. 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 in the overparametrized regime [39].

We investigate the statistical rate of convergence for this minimization problem, assuming the empirical problem (3.1) can be solved accurately. First, define

$$C(\mu,\nu,F,B):=\int(c_{\mathcal{X}}(x,B(y))-c_{\mathcal{Y}}(F(x),y))^2\,\mathrm{d}\mu\otimes\nu+\lambda_1\cdot\mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id},F)_{\#}\mu,(B,\mathrm{Id})_{\#}\nu) \\ +\lambda_2\cdot\mathrm{MMD}^2_{K_{\mathcal{X}}}(\mu,B_{\#}\nu)+\lambda_3\cdot\mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\mu,\nu)\,. \tag{3.2}$$

Then, the objective function of (3.1) is a plug-in estimator $C(\widehat{\mu}_m,\widehat{\nu}_n,F,B)$. Now, consider solving (3.1) over the transformation class $\mathcal{F}\times\mathcal{B}$ given as follows. Our non-asymptotic results work for generic classes $\mathcal{F}\times\mathcal{B}$.

Definition 6 (Transformation Classes)**.** *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.*

For $\mathcal{F}$ and $\mathcal{B}$ given in Definition 6, solving (3.1) 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.

Theorem 3. *For $\mathcal{F}$ and $\mathcal{B}$ given in Definition 6, 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 $C : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \times \mathcal{F} \times \mathcal{B} \to \mathbb{R}$ defined in (3.2). Under Assumptions 1-6, the following inequality holds with probability $1-\delta$ on $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$

$$C(\mu, \nu, \widehat{F}, \widehat{B}) - \inf_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\mu, \nu, F, B) \precsim \mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta). \tag{3.3}$$

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 $\mathcal{B}_\ell$:

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

We will provide required assumptions and the proof of Theorem 3 in Section 5 along with the definition of the pseudo-dimension. When $\mathcal{F}$ and $\mathcal{B}$ are parametrized by neural network classes (the ones we will use for numerical demonstrations in Section 7), tight pseudo-dimension bounds established in [40, 41] 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 (3.1) 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 $F : \mathcal{X} \to \mathcal{Y}$ to optimizing over its induced (dual) linear operator $\mathbf{F} : L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ that maps functions on $\mathcal{Y}$ to functions on $\mathcal{X}$. Let $\pi_{\mathcal{X}}$ be a Borel measure on $\mathcal{X}$ and $L^2_{\mathcal{X}}$ be the collection of real-valued measurable functions f defined on $\mathcal{X}$ such that $\int_{\mathcal{X}} f^2 \, \mathrm{d}\pi_{\mathcal{X}} < \infty$. 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 6 under a mild assumption.

To state the representer theorem, consider (3.1) 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:

$$K_{\mathcal{X}}(x, x') = \sum_k \lambda_k \phi_k(x)\phi_k(x') \, , \quad K_{\mathcal{Y}}(y, y') = \sum_\ell \gamma_\ell \psi_\ell(y)\psi_\ell(y') \, , \tag{3.4}$$

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

$$\mathbf{F}[\psi_\ell] = \sum_{k=1}^{\infty} \mathbf{F}_{k\ell}\phi_k \, , \quad \mathbf{B}[\phi_k] = \sum_{\ell=1}^{\infty} \mathbf{B}_{\ell k}\psi_\ell \, . \tag{3.5}$$

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 $[\mathbf{B}_{\ell k}]$. With these notations, (3.1) with $c_{\mathcal{X}} = K_{\mathcal{X}}$ and $c_{\mathcal{Y}} = K_{\mathcal{Y}}$ can be lifted to an infinite-dimensional optimization where the decision variables are matrices $\mathbf{F}$ and $\mathbf{B}$

$$\min_{(\mathbf{F},\mathbf{B})\in\mathcal{C}} \ \Omega(\mathbf{F}, \mathbf{B}) \, , \tag{3.6}$$

where the exact form of Ω is deferred to Section 6. Here, $\mathcal{C}$ 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,

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}} \Omega(\mathbf{F},\mathbf{B})\ . \tag{3.7}$$

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 (3.7) boils down to a finite-dimensional convex program.

Theorem 4. *Consider the optimization (3.6) under the assumptions in Proposition 10. Then, for any minimizer $(\mathbf{F}^\star, \mathbf{B}^\star)$ to the relaxed problem (3.7), 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*

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

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 (3.4). 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 $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$. In other words, by minimizing ω over $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$, we obtain a relaxation of (3.7), 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

$$\mathbf{F}^\star[g](x) = \underbrace{K_\mathcal{X}(x, X_m)}_{1\times m} \underbrace{\mathsf{F}^\star_{m,n}}_{m\times n} \underbrace{g(Y_n)}_{n\times 1}\ , \tag{3.8}$$

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 (3.8). 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

$$\sum_{i,j} K_\mathcal{X}(x, x_i) \cdot \tfrac{1}{m}\delta_{i=j} \cdot g(y_j)\ ; \tag{3.9}$$

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 (3.8) 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

$$(3.8) = \sum_{i,j} K_\mathcal{X}(x, x_i) \cdot \mathsf{F}^\star_{m,n}[i, j] \cdot g(y_j)\ . \tag{3.10}$$

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

$$\mathop{\mathbb{E}}_{x\sim\mu}\left[\mathbf{F}^\star[g](x)\right] = \mathop{\mathbb{E}}_{x\sim\mu}\Big[\underbrace{K_{\mathcal{X}}(x,X_m)\mathsf{F}^\star_{m,n}}_{=:W(x)\in\mathbb{R}^n} g(Y_n)\Big] = \mathop{\mathbb{E}}_{x\sim\mu}\Big[\sum_{j=1}^n W_j(x)g(y_j)\Big] \tag{3.11}$$

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 (3.11), one can verify that under mild conditions,

$$\mathop{\mathbb{E}}_{x\sim\widehat{\mu}_m}\left[\mathbf{F}^\star[g](x)\right] = \frac{1}{n}\sum_{j=1}^n g(y_j)\ . \tag{3.12}$$

In other words, with the empirical measure as input, (3.11) outputs the simple sample average.

4 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,*

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

The proof follows 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

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

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,*

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

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

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 $c_{\mathcal{Y}} = h(d_{\mathcal{Y}})$.*

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 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.

5 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.

5.1 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 $1 - 6\delta$.

5.2 Uniform Deviations

We now derive uniform deviation bounds for $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{C}_0(F,B) - C_0(F,B)|$ and $\sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} |\widehat{M}(F,B) - M(F,B)|$. For the former, we use the notion of uniform covering numbers defined below.

Definition 7 (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 $z_1, \ldots, z_m$:*

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

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

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

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 Definition 6) consist of uniformly bounded functions, that is, there exists a constant $b > 0$ such that*

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

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 $\delta > 0$,*

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

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 8 (Pseudo-Dimension)**.** *Let $\mathcal{G}$ be a collection of real-valued functions defined on a set $\mathcal{Z}$. Given a subset $S := \{z_1, \dots, z_m\} \subset \mathcal{Z}$, we say S is pseudo-shattered by $\mathcal{G}$ if there are $r_1, \dots, r_m \in \mathbb{R}$ 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 $\delta > 0$,*

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

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 9 (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:*

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

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

Proposition 7. *Denote a closed unit ball of any RKHS $\mathcal{H}$ as $\mathcal{H}(1)$. Also, let $(\mathrm{Id}, \mathcal{F}) := \{(\mathrm{Id}, F) : F \in \mathcal{F}\}$ and $(\mathcal{B}, \mathrm{Id}) := \{(B, \mathrm{Id}) : B \in \mathcal{B}\}$; 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

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

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

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

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

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

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 $d(A, A') := \|A - A'\|$. 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*

$$\mathop{\mathbb{E}}_{g} \sup_{A\in\mathcal{T}} g^\top A g \le \inf_{J\in\mathbb{N}} \left\{ m\delta_J + 12\int_{\delta_J/2}^{\Delta/2} \sqrt{2\log N(\delta,\mathcal{T})}\,\mathrm{d}\delta + 24\int_{\delta_J/2}^{\Delta/2} \log N(\delta,\mathcal{T})\,\mathrm{d}\delta \right\} , \tag{5.1}$$

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 $\mathcal{T} := \{A_F : F \in \mathcal{F}\}$. Then, to further bound the RHS of (5.1), 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.

6 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 (3.1).

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

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

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 (3.1) 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:

$$K_{\mathcal{X}}(x, x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(x') \,, \tag{6.1}$$

$$K_{\mathcal{Y}}(y, y') = \sum_{\ell=1}^{\infty} \gamma_\ell \psi_\ell(y) \psi_\ell(y') \,. \tag{6.2}$$

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

$$\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (K_{\mathcal{X}}(x_i, B(y_j)) - K_{\mathcal{Y}}(F(x_i), y_j))^2 = \frac{1}{mn} \sum_{i,j} (\Psi_{y_j}^\top \mathbf{B} \Lambda \Phi_{x_i} - \Phi_{x_i}^\top \mathbf{F} \Gamma \Psi_{y_j})^2 . \tag{6.3}$$

Also,

$$\begin{aligned}\mathrm{MMD}^2_{K_\mathcal{X}}(\widehat{\mu}_m, B_\#\widehat{\nu}_n) &= \frac{1}{m^2}\sum_{i,i'} K_\mathcal{X}(x_i, x_{i'}) + \frac{1}{n^2}\sum_{j,j'} K_\mathcal{X}(B(y_j), B(y_{j'})) - \frac{2}{mn}\sum_{i,j} K_\mathcal{X}(x_i, B(y_j)) \\ &= \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i} \,.\end{aligned} \tag{6.4}$$

Similarly, we have

$$\mathrm{MMD}^2_{K_\mathcal{Y}}(F_\#\widehat{\mu}_m, \widehat{\nu}_n) = \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \tag{6.5}$$

and

$$\begin{aligned}&\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}((\mathrm{Id}, F)_\#\widehat{\mu}_m, (B, \mathrm{Id})_\#\widehat{\nu}_n) \\ &= \frac{1}{m^2}\sum_{i,i'} \Phi_{x_i}^\top \Lambda \Phi_{x_i'} \Phi_{x_i}^\top \mathbf{F}\Gamma\mathbf{F}^\top \Phi_{x_{i'}} + \frac{1}{n^2}\sum_{j,j'} \Psi_{y_j}^\top \Gamma \Psi_{y_{j'}} \Psi_{y_j}^\top \mathbf{B}\Lambda\mathbf{B}^\top \Psi_{y_{j'}} - \frac{2}{mn}\sum_{i,j} \Psi_{y_j}^\top \mathbf{B}\Lambda\Phi_{x_i}\Phi_{x_i}^\top \mathbf{F}\Gamma\Psi_{y_j} \,.\end{aligned} \tag{6.6}$$

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 (6.1) and (6.2). 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 (3.1) over $\mathcal{F}_o \times \mathcal{B}_o$ is equivalent to (3.6), 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*

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

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 (6.3)-(6.6) 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, (3.7) 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 $(\mathbf{x}_1,\dots,\mathbf{x}_m)$ and $(\mathbf{y}_1,\dots,\mathbf{y}_n)$, respectively:

$$U_m := \operatorname{span}\{\mathbf{x}_1,\dots,\mathbf{x}_m\}\ , \quad V_n := \operatorname{span}\{\mathbf{y}_1,\dots,\mathbf{y}_n\}\ .$$

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

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

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

$$(3.7) = \min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(\bar{\mathbf{F}},\bar{\mathbf{B}}) = \min_{\bar{\mathbf{F}},\bar{\mathbf{B}}\in\mathbb{R}^{\infty\times\infty}} \bar{\Omega}(P_{U_m}\bar{\mathbf{F}}P_{V_n}, P_{V_n}\bar{\mathbf{B}}P_{U_m}) \,.$$

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

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

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,

$$(3.7) = \min_{(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\in\mathsf{C}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \geq \min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \,.$$

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

$$(3.7) = \min_{(\mathsf{F}_{m,n},\mathsf{B}_{n,m})\in\mathsf{C}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) = \min_{\substack{\mathsf{F}_{m,n}\in\mathbb{R}^{m\times n}\\ \mathsf{B}_{n,m}\in\mathbb{R}^{n\times m}}} \omega(\mathsf{F}_{m,n},\mathsf{B}_{n,m}) \,.$$

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. □

7 Numerical Examples

This section investigates numerical examples, one synthetic and one real-world, to showcase the effectiveness of our reversible Gromov-Monge sampler. The synthetic example is of a sanity check nature to see that RGM can effectively learn simple parametric distributions, whereas the real-world example is to generate high fidelity images that are drawn from the underlying probability distribution supported on the MNIST image manifold.

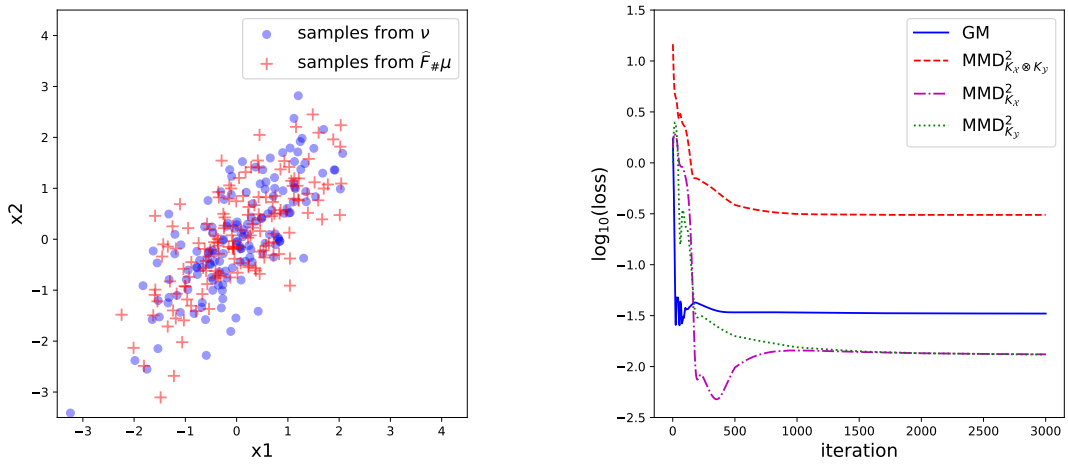
To implement our method, one needs to specify $c_{\mathcal{X}}$, $c_{\mathcal{Y}}$, $K_{\mathcal{X}}$, $K_{\mathcal{Y}}$ according to the nature of the data set. In practice, scaling $c_{\mathcal{X}}, c_{\mathcal{Y}}, K_{\mathcal{X}}, K_{\mathcal{Y}}$ similarly leads to lower empirical loss and more accurate samplers; hence proper tuning for cost functions and kernels is crucial. Here we offer some concrete suggestions on tuning cost functions and kernels: for cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$, if we know how to guarantee strong isomorphisms as in Gaussian example, we may simply choose the functions that ensure the existence of isomorphisms; otherwise we may choose first-stage cost functions as $c_{\mathcal{X},0} = d_{\mathcal{X}}^2/p$ and $c_{\mathcal{Y},0} = d_{\mathcal{Y}}^2/q$ (p, q are extrinsic dimensions of $\mathcal{X}$, $\mathcal{Y}$) or some scaled kernels as in MNIST example, and then standardize them by matching medians and standard errors. In other words, in the second case one can choose $c_{\mathcal{X}} = (c_{\mathcal{X},0} - m_{\mathcal{X}})/\mathrm{sd}_{\mathcal{X}}$, $c_{\mathcal{Y}} = (c_{\mathcal{Y},0} - m_{\mathcal{Y}})/\mathrm{sd}_{\mathcal{Y}}$, where $m_{\mathcal{X}}$, $\mathrm{sd}_{\mathcal{X}}$ are median and standard error of $\{c_{\mathcal{X},0}(x_i, x_j)\}_{i,j=1}^m$, and $m_{\mathcal{Y}}$, $\mathrm{sd}_{\mathcal{Y}}$ are defined analogously on $\mathcal{Y}$. For $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$, we suggest kernels that are characteristic, so to better enforce the equality between $(\mathrm{Id}, F)_{\#}\mu$ and $(B, \mathrm{Id})_{\#}\nu$.

Throughout the experiments, we highlight two main features of our formulation: first, an exemption from complicated tuning that is usually unavoidable for deep generative models; second, an approximate isomorphism $\widehat{F}$ that facilitates transform sampling. Let's clarify the term *approximate isomorphism*, which originates from Definition 3. We say $F : (\mathcal{X}, \mu, c_{\mathcal{X}}) \to (\mathcal{Y}, \nu, c_{\mathcal{Y}})$ is an approximate isomorphism if it satisfies $F_{\#}\mu \approx \nu$ and $c_{\mathcal{X}}(x, x') \approx c_{\mathcal{Y}}(F(x), F(x'))$ for x, $x' \in \mathcal{X}$. Note that how close the approximation in metrics is depends on the numerical value of the GM term in our examples. In both two examples below, we demonstrate that $\widehat{F}$, $\widehat{B}$ are approximate isomorphisms.

7.1 2D Gaussian

We first check our method on a synthetic Gaussian data set, for which a strong isomorphism can be guaranteed by our specification. Suppose the target distribution is $\nu = N(0, \Sigma)$ on $\mathcal{Y} = \mathbb{R}^2$, where Σ is a full-rank covariance matrix ($\Sigma = [1.0, 0.7; 0.7, 1.0]$). We select $\mathcal{X} = \mathbb{R}^2$, $\mu = N(0, I_2)$, $c_{\mathcal{X}}(x, x') = \langle x, x'\rangle$, $c_{\mathcal{Y}}(y, y') = \langle y, \Sigma^{-1}y'\rangle$ where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. Under this setting, two network spaces are strongly isomorphic due to Example 2. We further specify $K_{\mathcal{X}} = K_{\mathcal{Y}} = K_2$ where $K_2(x, y) = (\langle x, y\rangle + 1)^2$, a degree-2 polynomial kernel on $\mathbb{R}^2$, to ensure $(\mathrm{Id}, F)_{\#}\mu \approx (B, \mathrm{Id})_{\#}\nu$ in the empirical problem (3.1). Now we expect that our model learns strongly isomorphic maps, e.g., $F(x) = \Sigma^{1/2}Qx$ and $B(y) = Q^{\top}\Sigma^{-1/2}y$ up to the orthogonal group $Q \in O(2)$.

We set sample size $m = n = 1000$, tuning parameters $\lambda_1 = \lambda_2 = \lambda_3 = 1$, and use the gradient descent (GD) approach as mentioned in Section 3. To be specific, we restrict F and B to be linear transformations, a



(a) (b)

Figure 1: Generated samples and training loss on 2D Gaussian data. Subfigure (a) compares 150 samples generated by applying $\widehat{F}$ on i.i.d. new samples from μ (our reversible Gromov-Monge sampling approach) and 150 new samples from the target ν, and subfigure (b) shows the logarithmic loss for each component in the empirical loss (3.1) indexed by (gradient descent) iterations.

function class rich enough to contain strong isomorphisms, and run gradient descent algorithm for 3000 iterations. The learning rate at the initial iteration is 0.05, and will halve after every 500 iterations. Figure 1(a) compares 150 new samples from $\widehat{F}_{\#}\mu$ with 150 new samples from ν, which confirms that our model learns the distribution well. The value of each component in (3.1), which will be referred to as GM, $\mathrm{MMD}^2_{K_\mathcal{X}\otimes K_\mathcal{Y}}$, $\mathrm{MMD}^2_{K_\mathcal{X}}$, and $\mathrm{MMD}^2_{K_\mathcal{Y}}$ in the order, are shown in Figure 1(b) on a logarithmic scale against the number of GD iterations. For the estimated linear transformations $(\widehat{F}, \widehat{B})$, we have

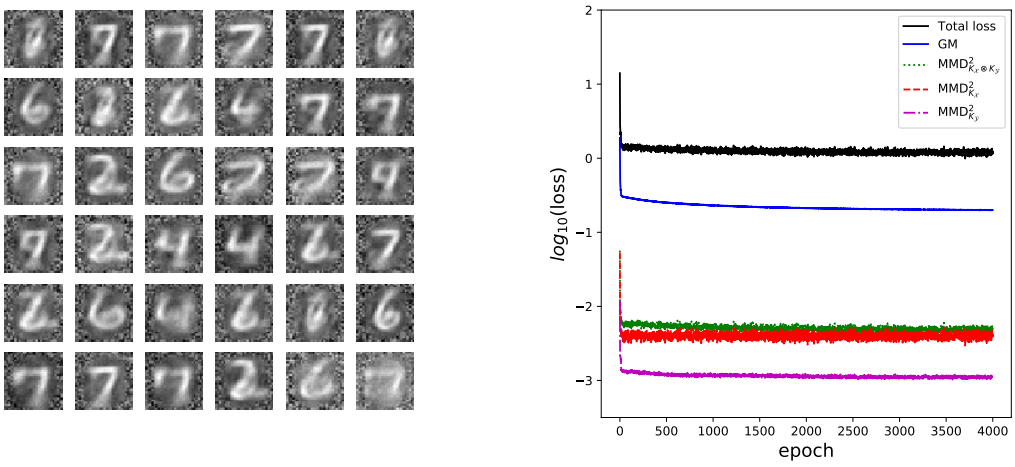
$$\widehat{F}\widehat{F}^\top \approx \begin{pmatrix} 1.0202 & 0.6968 \\ 0.6968 & 0.9669 \end{pmatrix}, \qquad \widehat{B}\Sigma\widehat{B}^\top \approx \begin{pmatrix} 0.9615 & -0.0044 \\ -0.0044 & 1.0776 \end{pmatrix}.$$

Hence $\widehat{F}\widehat{F}^\top \approx \Sigma$, $\widehat{B}\Sigma\widehat{B}^\top \approx I_2$, implying that our model indeed approximately captures strong isomorphisms.

7.2 MNIST

Now we apply our method to generate new MNIST images (images unseen in the data set), whose distribution might be a high dimensional distribution confined to a low dimensional image manifold. For simplicity, we focus on 4 digits $(2, 4, 6, 7)$ and choose ν to be the corresponding MNIST distribution supported on some manifold $\mathcal{Y} \subset \mathbb{R}^{784}$. Since we lack additional knowledge on the existence of strong isomorphisms in this example, we simply choose $\mathcal{X} = \mathbb{R}^2$, $\mu = N(0, I_2)$, $c_{\mathcal{X},0} = K_\mathcal{X} = K_2$, $c_{\mathcal{Y},0} = K_\mathcal{Y} = K_{784}$, where $K_d(x, y) = \exp\big(-\|x - y\|^2/d\big)$ for $d = 2$, 784 and $\|\cdot\|$ denotes the Euclidean distance. Finally, we compute $c_\mathcal{X}, c_\mathcal{Y}$ by rescaling $c_{\mathcal{X},0}, c_{\mathcal{Y},0}$. In words, we want to best match a two-dimensional space to the image manifold of digits $(2, 4, 6, 7)$ in the MNIST data set.

We choose the sample size $m = 40000$, $n = 8000$, tuning parameters $\lambda_1 = \lambda_2 = \lambda_3 = 100$, and again use the (stochastic) gradient descent approach. In addition, we parameterize F and B by multi-layer perceptrons (MLP), a class of feedforward neural networks such that every two nearby layers are fully connected. Under this parameterization, both $F\colon \mathbb{R}^2 \to \mathbb{R}^{784}$ and $B\colon \mathbb{R}^{784} \to \mathbb{R}^2$ have 3 hidden layers, and each hidden layer



(a) (b)

Figure 2: Generated images and training loss on MNIST data set for digits $2, 4, 6, 7$. Subfigure (a) visualizes our generated images by applying $\widehat{F}$ on i.i.d. new samples from $N(0, I_2)$: these are RGM simulated images that are distinct from the MNIST data. Subfigure (b) shows the logarithmic training loss for each component in the empirical loss (3.1) over epochs.

consists of 50 neurons. We apply the rectified linear unit (ReLU) activation function,

$$\sigma(x) = \max(x, 0) \ ,$$

to all hidden layers of F and B. To put it explicitly, F has the following form

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

with the parameter space $\{(W_l, b_l, 1 \le l \le 3) \mid W_1 \in \mathbb{R}^{50\times 2}, W_2 \in \mathbb{R}^{50\times 50}, W_3 \in \mathbb{R}^{784\times 50}, b_1, b_2 \in \mathbb{R}^{50\times 1}, b_3 \in \mathbb{R}^{784\times 1}\}$. Similarly, B has the following form

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

with $\widetilde{W}_1 \in \mathbb{R}^{50\times 784}, \widetilde{W}_2 \in \mathbb{R}^{50\times 50}, \widetilde{W}_3 \in \mathbb{R}^{2\times 50}, \widetilde{b}_1, \widetilde{b}_2 \in \mathbb{R}^{50\times 1}, \widetilde{b}_3 \in \mathbb{R}^{2\times 1}$. We use Adam [42], a variant of stochastic gradient descent, to train the neural networks. The training set is randomly divided into 20 batches, each batch containing 2000 samples from μ and 400 samples from ν. The learning rate is 0.01 in the first 500 iterations; it decreases to 0.001 from iteration 501 to iteration 1000, and further reduces to 0.0001 after 1000 iterations. Figure 2 includes: (a) newly generated samples from $\widehat{F}_{\#}\mu$, namely generating a fresh two-dimensional Gaussian $x \sim \mu$ and push-forwarding it via the learned transformation $\widehat{F}$; (b) total loss and component-wise loss in (3.1) on a logarithmic scale against the number of epochs. Here at each time stamp, each loss is computed as the average of the respective batch losses in a whole epoch. Subfigure

(a) demonstrates the generative power of our method: these are new, unseen images different from the $60K$ images in the MNIST data set. Our RGM balances among-class expressivity (the newly generated images can express different digits), as well as in-class variability (the newly generated images with the same digit differ from each other).

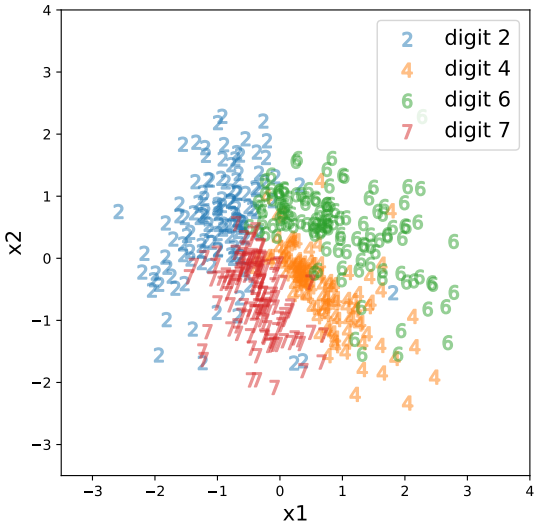


Figure 3: Embedding MNIST images into $\mathbb{R}^2$. We generate these points by applying $\widehat{B}$ on 500 MNIST test samples, where we pick up 125 samples for each digit.

Next, we examine the approximate isomorphism for $\widehat{B}$ in Figure 3, a scatter plot by applying $\widehat{B}$ on 500 new samples from MNIST test set. In plain language, we would like to see how to best embed a $\mathbb{R}^{784}$ MNIST image to a $\mathbb{R}^2$ space. We note two observations. First, the distribution of images of $\widehat{B}$, as a whole, is similar to $N(0, I_2)$, which can be easily seen by overlooking the color of the data cloud. Second, each digit forms a local cluster in $\mathbb{R}^2$ according to the angle. In other words, $\widehat{B}_{\#}\nu \approx N(0, I_2)$ and $K_2(\widehat{B}(y), \widehat{B}(y')) \approx K_{784}(y, y')$ hold for $y, y' \in \mathbb{R}^{784}$, hence indicating an approximate isomorphism.

8 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, we designed a transform sampler that can operate between distributions defined on heterogenous spaces. In addition, we introduced two concrete optimization methods for computing RGM given finite samples and proved their properties. Accordingly, our work not only provides a simple yet promising transform sampler, but also sheds light on tackling a notoriously difficult quadratic assignment problem.

Lastly, we mention a few directions that can lead to future research. First, it will be interesting to understand whether one can derive global convergence results for the gradient descent optimization method. Second, we should establish criteria that are theoretically justified for choosing the Lagrangian multipliers in (3.1). More generally, how to specify and tune appropriate functions $c_{\mathcal{X}}, c_{\mathcal{Y}}$ and kernels $K_{\mathcal{X}}, K_{\mathcal{Y}}$ for better empirical results is still not fully elucidated. We leave such questions as potential future work.

References

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

[2] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces, 2012.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[4] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[6] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. 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, 2015.

[7] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1718–1727, Lille, France, 07–09 Jul 2015. PMLR.

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

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

[10] Larry Wasserman. *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York, 2006.

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

[12] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[13] Carl Doersch. Tutorial on Variational Autoencoders. *arXiv:1606.05908 [cs, stat]*, August 2016.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[16] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

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

[18] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

[19] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017.

[20] Tengyuan Liang. How well generative adversarial networks learn distributions. *arXiv preprint arXiv:1811.03179, Journal of Machine Learning Research, to appear*, 2018.

[21] Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

[22] Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

[23] Jonathan Weed and Quentin Berthet. Estimation of smooth densities in wasserstein distance. *arXiv preprint arXiv:1902.01778*, 2019.

[24] Qi Lei, Jason D Lee, Alexandros G Dimakis, and Constantinos Daskalakis. Sgd learns one-layer networks in wgans. *arXiv preprint arXiv:1910.07030*, 2019.

[25] Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv preprint arXiv:2002.03938*, 2020.

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

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

[28] Christian Gouriéroux and Alain Monfort. *Simulation-Based Econometric Methods*. OUP/CORE Lecture Series. Oxford University Press, Oxford, 1997.

[29] Facundo Memoli. On the use of Gromov-Hausdorff Distances for Shape Comparison. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007.

[30] Justin Solomon, Gabriel Peyré, Vladimir G. Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4), 2016.

[31] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein learning for graph matching and node embedding. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6932–6941. PMLR, 09–15 Jun 2019.

[32] Samir Chowdhury and Facundo Mémoli. The Gromov–Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 11 2019.

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

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

[35] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning (ICML)*, 2016.

[36] Vayer Titouan, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced gromov-wasserstein. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[37] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[38] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 907–915. PMLR, April 2019.

[39] Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning, 2018.

[40] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[41] Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1064–1068. PMLR, June 2017.

[42] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[43] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

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

A Remaining Proofs

A.1 Proofs in Section 4

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

Proof of Proposition 2. Recall that

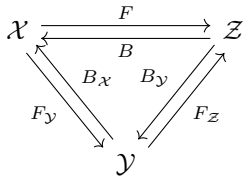
$$\mathrm{RGM}(\mu_{\mathcal{X}},\mu_{\mathcal{Z}}) = \inf_{(F,B)\in\mathcal{I}(\mu_{\mathcal{X}},\mu_{\mathcal{Z}})} C_{\mathcal{X}\mathcal{Z}}(F,B)\ ,$$

where

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

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 $(F_{\mathcal{Z}},B_{\mathcal{Y}}) \in \mathcal{I}(\mu_{\mathcal{Y}},\mu_{\mathcal{Z}})$[2] 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

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

[2]The subscript of a map denotes its range; for instance, $B_{\mathcal{X}}$ maps to $\mathcal{X}$.

since

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

Hence,

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

□

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 [32], 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

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

□

A.2 Proofs in Section 5

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

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

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

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

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

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

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

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

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:

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

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

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

Hence,

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

Due to the triangle inequality of MMD, we have

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

By Theorem 3.4 of [15],

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

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,

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

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 $n \ge m$. From the proof of Proposition 4,

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

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

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

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 $x_1,\ldots,x_m$:

$$d(F_k,F'_k) := \max_{i\in[m]}|F_k(x_i)-F_k(x'_i)|.$$

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 $d(F_k,F'_k)\leq\epsilon$. 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

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

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

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

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

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

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

$$\begin{aligned}&\mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i h_{F,B}(x_i, y_i) \right| \\ &\leq \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, \{x_i\}_{i=1}^m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, \{y_i\}_{i=1}^m)}{2m}} \\ &\leq \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, m)}{2m}} \,.\end{aligned}$$

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

$$\mathop{\mathbb{E}}_{x_i} \mathop{\mathbb{E}}_{y_i} \mathop{\mathbb{E}}_{\epsilon_i} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{m} \sum_{i=1}^m \epsilon_i h_{F,B}(x_i, y_i) \right| \leq \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, m)}{2m}} \,.$$

As a result,

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

holds with probability at leat $1 - \delta$. Here, $\log N_\infty(\epsilon, \mathcal{B}_\ell, m) \leq \log N_\infty(\epsilon, \mathcal{B}_\ell, n)$ holds since $n \geq m$, 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

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

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 a sub-Gaussian random variable with parameter $H^2/(4n)$. By the same argument as before,

$$\mathop{\mathbb{E}}_{y_j}\mathop{\mathbb{E}}_{\epsilon_j} \sup_{(F,B)\in\mathcal{F}\times\mathcal{B}} \left| \frac{1}{n}\sum_{j=1}^{n} \epsilon_j h_{F,B}(x_i, y_j) \right| \le \eta\epsilon + H\sqrt{\frac{\sum_{k=1}^{\dim(\mathcal{Y})} \log N_\infty(\epsilon, \mathcal{F}_k, m) + \sum_{\ell=1}^{\dim(\mathcal{X})} \log N_\infty(\epsilon, \mathcal{B}_\ell, n)}{2n}} .$$

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

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

holds with probability at least $1-\delta$. Combining (A.1) and (A.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

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

Hence,

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

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

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

$\mathrm{MMD}_{K_\mathcal{Y}}(\widehat{\nu}_n, \nu)$ 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 $\|f\|_{\mathcal{H}_\mathcal{Y}} \le 1$. Then, the reproducing property implies

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

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,

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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,

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

First, we obtain an upper bound on Δ:

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

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,

$$\Delta \ge |K_{\mathcal{Y}}(F(x_1), F(x_2)) - K_{\mathcal{Y}}(F_0(x_1), F_0(x_2))| = |K_{\mathcal{Y}}(y_0, y'_0) - K_{\mathcal{Y}}(y_0, y_0)| > 0 .$$

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

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

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. □

A.3 Auxiliary Lemmas

Lemma 3 (Theorem 4.10 of [43]). *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, $\sup_{g\in\mathcal{G}}\|g\|_\infty \le b$. Let $z_1,\dots,z_m$ are i.i.d. samples from ρ and let $\widehat{\rho}_m$ be the empirical measure constructed from them. Then, for any $\delta>0$,*

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

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

Lemma 4 (Theorem 12.2 of [40]). *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 $m\ge\mathrm{Pdim}(\mathcal{G})$,*

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

Lemma 5 (Example 2.12 of [44]). *For any $A\in\mathbb{S}_0^{m\times m}$ and $0\le\lambda<1/(2\|A\|_{\mathrm{op}})$,*

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

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

This lemma tells that $g^\top Ag$ is a sub-Gamma random variable with variance factor $2\|A\|^2$ and scale parameter $2\|A\|_{\mathrm{op}}$ (see Chapter 2.4 of [44] 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,*

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

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