SIAM J. MATH. DATA SCI.
Vol. 6, No. 2, pp. 283–310

Reversible Gromov–Monge Sampler for Simulation-Based Inference*

YoonHaeng Hur†, Wenxuan Guo‡, and Tengyuan Liang‡

Abstract. This paper introduces a new simulation-based inference procedure to model and sample from multidimensional probability distributions given access to independent and identically distributed samples, circumventing the usual approaches of explicitly modeling the density function or designing Markov chain Monte Carlo. Motivated by the seminal work on distance and isomorphism between metric measure spaces, we develop a new transform sampler to perform simulation-based inference, which estimates a notion of optimal alignments between two heterogeneous metric measure spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ from empirical data sets, with estimated maps that approximately push forward one measure μ to the other ν, and vice versa. We introduce a new notion called the reversible Gromov–Monge (RGM) distance, providing mathematical formalism behind the new sampler. We study the statistical rate of convergence of the new transform sampler, along with several analytic properties of the RGM distance and operator viewpoints of transform sampling. Synthetic and real-world examples showcasing the effectiveness of the new sampler are also demonstrated.

Key words. Gromov–Wasserstein metric, transform sampling, simulation-based inference, generative models, isomorphism, likelihood-free inference

MSC codes. 49Q22, 62R20, 90C20

DOI. 10.1137/23M1550384

1. Introduction.

One of the central tasks in statistics is to model and sample from a multidimensional 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 independent and identically distributed (i.i.d.) samples from MCMC methods is complex as one has to control correlations among successive samples or run parallel chains.

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

*Received by the editors January 31, 2023; accepted for publication (in revised form) November 20, 2023; published electronically April 5, 2024.
https://doi.org/10.1137/23M1550384
Funding: The third author received support from an NSF CAREER award (DMS-2042473) and from the William S. Fishman Faculty Research Fund at the University of Chicago Booth School of Business.
†Department of Statistics, University of Chicago, Chicago, IL 60637 USA (yoonhaenghur@uchicago.edu).
‡Booth School of Business, University of Chicago, Chicago, IL 60637 USA (wxguo@chicagobooth.edu, tengyuan.liang@chicagobooth.edu).

monotonic transformation $T : [0,1] \to \mathbb{R}$ called the inverse cumulative distribution function (CDF):

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

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

The transform sampling idea can be extended to the multidimensional setting: given a target probability measure ν supported on $\mathcal{Y}$, one can specify a probability measure μ on $\mathcal{X}$, which is easy to sample from such as a multivariate Gaussian, and then find a measurable map $T : \mathcal{X} \to \mathcal{Y}$ such that $T_{\#}\mu = \nu$, where the pushforward measure $T_{\#}\mu$ is defined analogously to the one-dimensional case above. Such a map T—called a *transport map* from μ to ν—transforms i.i.d. samples from μ into i.i.d. samples from ν. Over the past few years, the generative modeling literature in machine learning has been actively employing such transform sampling ideas by identifying $T_{\#}\mu = \nu$ through the following minimization:

$$\min_{T \in \mathcal{F}} \mathcal{L}(T_{\#}\mu, \nu)\,, \tag{1.2}$$

where $\mathcal{F}$ is a class of maps from $\mathcal{X}$ to $\mathcal{Y}$ parametrized by neural networks and $\mathcal{L}$ measures certain discrepancies between two distributions. Different choices of $\mathcal{L}$ have led to various models such as the Jensen–Shannon divergence for generative adversarial networks (GANs) [22], the Wasserstein-1 distance for Wasserstein-GAN [2], and the maximum mean discrepancy (MMD) for MMD-GAN [18, 29]. One caveat is that there can be infinitely many transport maps from μ to ν; for instance, when $\mu = \nu = \mathrm{Unif}([0,1])$, define $T : [0,1] \to [0,1]$ by $T(x) = |2x - 1|$, then the n-fold compositions of T are valid transport maps for all $n \in \mathbb{N}$. In other words, finding a map T satisfying $T_{\#}\mu = \nu$ is an overidentified problem, where (1.2) may have infinitely many minimizers. Though all minimizers are equivalent in terms of transform sampling, not all are equally preferred in light of the Occam's razor principle: one wishes to select simple, desirable transport maps among the overidentified set $\{T : T_{\#}\mu = \nu\}$.

Inductive biases tackle the aforementioned overidentified problem by restricting the search to transport maps with desirable properties. In this context, meaningful progress has been made based on optimal transport (OT) theory [52, 33]. The OT theory aims to identify an optimal transformation T, quantified by the transportation cost of moving mass from μ to ν; for instance, when μ and ν lie in the same space $\mathbb{R}^d$, each transport map T is associated with the transport cost $C(T) := \int_{\mathbb{R}^d} \|x - T(x)\|^2 \,\mathrm{d}\,\mu(x)$. Brenier [8] proved that, under mild regularity conditions, there exists a unique minimizer $T^\star$ of C among all transport maps, namely,

$$T^\star = \operatorname*{argmin}_{T_{\#}\mu = \nu} C(T)\,. \tag{1.3}$$

More importantly, $T^\star$ is the gradient of some convex function. On the one hand, Brenier's result extends the one-dimensional (inverse) transform sampling to the multidimensional case. When $d = 1$ and $\mu = \mathrm{Unif}([0,1])$, the inverse CDF map in (1.1) turns out to be exactly $T^\star$; for $d > 1$, the multidimensional map $T^\star : \mathbb{R}^d \to \mathbb{R}^d$ is the gradient of a convex function, generalizing

monotonic functions on the real line to multidimensions. On the other hand, Brenier's result naturally initiates an inductive bias in transform sampling: instead of searching any transport map, one may find $T^\star$, the optimal one with the smallest cost. To contrast this with the plain transform sampling (1.2), let us rewrite (1.3) using a suitable Lagrangian multiplier $\lambda > 0$ to enforce the equality constraint $T_\# \mu = \nu$:

$$\min_{T \in \mathcal{F}} C(T) + \lambda \cdot \mathcal{L}(T_\# \mu, \nu)\,. \tag{1.4}$$

Now, we can see that (1.4) incorporates an additional objective function of T—the transport cost—in (1.2), thereby introducing an inductive bias towards the OT map that achieves the minimum of C. Moreover, the Lagrangian provides an implementable formulation in practice to leverage OT ideas in generative modeling.

Such an OT-based approach, however, can be cumbersome in practice if the target ν is a high-dimensional embedding of some low-dimensional distribution. For instance, let ν be the distribution of handwritten digit images from the MNIST data set on $\mathbb{R}^{784}$.[1] To use the above OT-based approach, one must choose μ on $\mathbb{R}^{784}$ and find a map $T : \mathbb{R}^{784} \to \mathbb{R}^{784}$. However, the support of ν is intrinsically low-dimensional (roughly $\mathbb{R}^{15}$ as in [19]); hence, other transform samplers with $\mathcal{X} = \mathbb{R}^{15}$ yielding $T : \mathbb{R}^{15} \to \mathbb{R}^{784}$ are more efficient than the OT-based method in terms of estimating T and computing $T(X)$ for $X \sim \mu$.

In this paper, we propose and study a new transform sampler combining the best of both worlds: it introduces beneficial inductive biases like the OT approach, while operating when $\mathcal{X}$ and $\mathcal{Y}$ are heterogeneous spaces. The key to our approach is to utilize a notion of isomorphism and the Gromov–Wasserstein (GW) distance between μ and ν. Given two cost functions $c_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the GW distance [35, 13] is

$$\mathrm{GW}(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left(c_\mathcal{X}(x,x') - c_\mathcal{Y}(y,y')\right)^2 \mathrm{d}\,\gamma(x,y)\,\mathrm{d}\,\gamma(x',y') \right)^{1/2}, \tag{1.5}$$

where $\Pi(\mu,\nu)$ is the set of couplings between μ and ν. GW aims to match the cost functions defined on two heterogeneous spaces, intending to identify an isomorphism, namely, a transport map T such that $c_\mathcal{X}(x,x') = c_\mathcal{Y}(T(x), T(x'))$ for all $x, x' \in \mathcal{X}$. Inspired by these, one can define the following objective function of T to replace the transport cost C:

$$Q(T) := \int_\mathcal{X} \int_\mathcal{X} \left(c_\mathcal{X}(x,x') - c_\mathcal{Y}(T(x), T(x'))^2 \,\mathrm{d}\,\mu(x)\,\mathrm{d}\,\mu(x')\,. \right. \tag{1.6}$$

One way to design a transform sampler with inductive biases toward isomorphisms (the target when $\min_T Q(T) = 0$) is to utilize the Lagrangian form (1.4), but with Q instead of C. It turns out, however, that the objective function Q—which is quadratic in μ—results in several modeling subtleties when considering its plug-in estimation based on finite i.i.d. samples from μ and ν, which might not be favorable in practice (see Remark 3.3). To circumvent such issues, we develop a transform sampler with a different objective function, which we will detail in section 3.

[1] Images are normalized and fit into a 28×28 pixel bounding box, hence defined on $\mathbb{R}^{28\times 28} \equiv \mathbb{R}^{784}$ [27].

Organization. The rest of the paper is organized as follows. First, we briefly review other related studies omitted in the discussion above; section 2 briefly outlines some preliminary background on OT and the Gromov–Wasserstein (GW) distance. Then, section 3 introduces the primary methodology of this paper, where we develop a new transform sampler based on the new notion called the reversible Gromov–Monge (RGM) distance. Section 4 delineates the main theoretical result, providing the statistical rate of convergence for the plug-in estimation. Section 5 discusses further theoretical perspectives on the new sampler. Synthetic and real-world examples showcasing the effectiveness of the new sampler are demonstrated in section 6 as a proof of concept. The supplementary material (supplement.pdf [local/web 1.35MB]) collects details of the results in sections 4, 5, and 6 along with relevant discussions.

1.1. Related literature. 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 [45] and studies the minimax optimality of directly estimating such density functions [49]. It is also unclear how to proceed to sample from a possibly improper[2] density estimator, even with an optimal estimator at hand. One may employ MCMC techniques for sampling from specific models. However, on the computational front, it is highly nontrivial how to ensure the mixing properties of MCMC for a designed sampler [43].

A recent trend in unsupervised machine learning is to learn complex, high-dimensional distributions via (deep) generative models, either explicitly by parametrizing the sufficient statistics of the exponential families [16, 25], or implicitly by parametrizing the pushforward map transporting distributions [18, 22], with a focus on tractability in computation. Surprisingly, though lacking theoretical underpinning and optimality, generative models perform 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, [39, 29, 18] on MMDs, [48, 30] on integral probability metrics, and [38, 3, 31, 46, 4, 54, 28, 12] on GANs. Last but not least, we emphasize that an alternative implicit distribution estimation approach using the simulated method of moments has been formulated in the econometrics literature since [34, 40] and [23].

Originally introduced as a tool for comparing objects in computer graphics, analytic properties of the GW distance have been studied extensively [35, 50]; the most important one is that it defines a distance between metric measure spaces, namely, metric spaces endowed with probability measures. Since many real-world data sets can be modeled as metric measure spaces, the GW distance has been utilized in various problems such as shape correspondence [47], graph matching [56], and protein comparison [20]. Certain statistical aspects of comparing metric measure spaces have been studied in [7, 55].

Computation of the GW distance amounts to a relaxation of the quadratic assignment problem [26]; both are known to be NP-hard [11] in the worst case. Several approaches have been proposed for the approximate computation of the GW distance. [35] studies lower bounds on the GW distance that are easier to compute. [41] adds an entropic regularization term

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

to the GW distance, which leads to a fast iterative algorithm; [44] further modifies this by imposing a low-rank constraint on couplings. [53] proposes the sliced GW distance defined by integrating GW distances over one-dimensional projections. Last but not least, recent papers [56, 6, 14] study scalable partitioning schemes to approximately compute GW distances.

GW distances have been utilized as a discrepancy measure in the generative modeling [10]; roughly speaking, they use GW distances as $\mathcal{L}$ in (1.2), which is significantly different from the method developed in this paper.

2. Background.

This section provides background on the OT theory and the GW distance. We first introduce the notation that we use throughout the paper.

Notation. Let $\|A\|$ denote the Frobenius norm of a matrix A and $\|x\|$ denote the Euclidean norm of a vector x. Given a set $\mathcal{X}$ and a function $f : \mathcal{X} \to \mathbb{R}$, let $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$ denote the sup norm. For an integer $n \in \mathbb{N}$, we define $[n] = \{1, \dots, n\}$. For a metric space $\mathcal{X}$, we denote its metric as $d_\mathcal{X}$ and write $\mathcal{P}(\mathcal{X})$ to denote the collection of all Borel probability measures on $\mathcal{X}$. We call a pair $(\mathcal{X}, \mu)$ a Polish probability space if $\mathcal{X}$ is a metric space that is complete and separable 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 $\mathcal{T}(\mu, \nu) := \{T : \mathcal{X} \to \mathcal{Y} \mid T_\# \mu = \nu\}$; we call $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ a coupling between μ and ν if $\gamma(A \times \mathcal{Y}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$ for all Borel subsets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$, and we denote the collection of all such couplings as $\Pi(\mu, \nu)$. For a sequence of numbers $a(n), b(n) \in \mathbb{R}$, we use $a(n) \precsim b(n)$ to denote the relationship that $a(n)/b(n) \le C$ for all $n \in \mathbb{N}$ with some universal constant $C > 0$.

2.1. A brief overview of optimal transport theory.

A major goal of OT is minimizing the cost associated with the transport map between two Polish probability spaces, say, $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$. Consider a measurable function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$; 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, that is, $\int_\mathcal{X} c(x, T(x)) \, \mathrm{d}\mu(x)$. Minimizing the cost over $\mathcal{T}(\mu, \nu)$ is referred to as the Monge problem, named after Gaspard Monge. We call $T^\star$ an OT map if $T^\star$ is minimizer, that is,

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

Another important OT problem is minimizing the cost given by couplings. We define the average cost incurred by a coupling $\gamma \in \Pi(\mu, \nu)$ as the integration of the cost c with respect to γ, namely, $\int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y)$. Minimizing this cost over $\Pi(\mu, \nu)$ is called the Kantorovich problem, credited to Leonid Kantorovich. We call $\gamma^\star$ an optimal coupling if

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

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

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

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

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

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

Definition 2.1. *Given a metric space $\mathcal{X}$ that is complete and separable, we call*

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

the Wasserstein-2 distance[3] *between $\mu,\nu\in\mathcal{P}(\mathcal{X})$.*

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

Mémoli's pioneering work [35] 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 $(x,y)\in\mathcal{X}\times\mathcal{Y}$, we associate two pairs $(x,y),(x',y')\in\mathcal{X}\times\mathcal{Y}$ with the discrepancy of intraspace quantities $c_{\mathcal{X}}(x,x')$ and $c_{\mathcal{Y}}(y,y')$. In summary, by switching from the integration $\mathrm{d}\,\gamma$ to the double integration $\mathrm{d}\,\gamma\,\mathrm{d}\,\gamma$, we no longer need an otherwise interspace quantity $c:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$. Therefore, we can always define this objective function whenever we have proper $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ in each individual space, leading to the following definition.

[3]One can define the Wasserstein-p distance by replacing the exponent 2 above with $p\in[1,\infty]$.

Definition 2.2. *A triple $(\mathcal{X}, \mu, c_{\mathcal{X}})$ is called a network space if $(\mathcal{X}, \mu)$ is a Polish probability space such that $\operatorname{supp}(\mu) = \mathcal{X}$ and $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is measurable. The GW 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 2.3. We adopt the network space definition introduced in [13]. 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 [35] and [50]. In short, a network space is a generalization of a metric measure space.

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

Definition 2.4. *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 $T : \mathcal{X} \to \mathcal{Y}$ 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 check that $\cong$ is indeed an equivalence relation on the collection of network spaces. The following theorem states that the GW distance satisfies all metric axioms on the quotient space—under the equivalence relation $\cong$—of metric measure spaces.

Theorem 2.5 (Lemma 1.10 of [50]). *Let $\mathcal{M}$ be the collection of all metric measure spaces. Then, GW satisfies the three metric axioms on $\mathcal{M}/_{\cong}$, the collection of all equivalence classes of $\mathcal{M}$ induced by $\cong$.*

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 (GM) distance.

Definition 2.6. *The GM distance between spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ is defined as*

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

Loosely speaking, computing GM amounts to finding a transport map T such that $c_{\mathcal{X}}(x,x')$ best matches $c_{\mathcal{Y}}(T(x),T(x'))$ on average; we can view such a map T as a surrogate for an isomorphism. See section 2.4 of [36] for more details of GM. Observe that the objective function in the definition of GM is exactly (1.6) mentioned in section 1. As briefly hinted in section 1, one natural way to design a practical transform sampler with inductive biases toward isomorphisms is to replace the transport cost C in (1.4) with (1.6); this can also be viewed as the Lagrangian form of GM, which replaces the constraint $T \in \mathcal{T}(\mu,\nu)$ with a penalty term $\lambda \cdot \mathcal{L}(T_{\#}\mu, \nu)$. It turns out, however, such a formulation has some subtle issues which might be unfavorable in practice (see Remark 3.3). To circumvent such issues, we develop a transform sampler with slight modifications which we introduce in the next section.

3. Transform sampling via 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. Minimizing the above objective function under the binding constraint leads to the following definition.

Definition 3.1. *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 $F : \mathcal{X} \to \mathcal{Y}$ and $B : \mathcal{Y} \to \mathcal{X}$ satisfy the binding constraint $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$. We define the RGM distance between $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ as*

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

Remark 3.2. 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 $\{(|2y - 1|, y) : y \in [0, 1]\}$. Last, note that $\mathcal{I}(\mu, \nu)$ might be empty, for instance, if μ and ν are discrete and their supports have different cardinality, say, $\mu = \delta_x$ and $\nu = (\delta_{y_1} + \delta_{y_2})/2$, namely, Dirac measures supported on $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$; in such a case, $\mathrm{RGM}(\mu, \nu) = \infty$.

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 $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$; details on such analytic properties will be discussed in section 5.1.

RGM sampler. We design a transform sampling method based on finding a minimizing pair (F, B) of RGM to capture isomorphic relations between network spaces. To implement this method, we utilize the form similar to (1.4) mentioned in section 1, which leads to the Lagrangian form that allows efficient estimation of (F, B) using i.i.d. samples from μ and ν. The idea is to consider the Lagrangian of the minimization problem in the definition of RGM. First, we rewrite the minimization problem with the binding constraint as follows:

$$\begin{aligned} \min_{\substack{F:\mathcal{X}\to\mathcal{Y} \\ B:\mathcal{Y}\to\mathcal{X}}} \quad & \int_{\mathcal{X}\times\mathcal{Y}} (c_{\mathcal{X}}(x, B(y)) - c_{\mathcal{Y}}(F(x), y))^2 \,\mathrm{d}\,\mu \otimes \nu \\ \text{s.t.} \quad & \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}((\mathrm{Id}, F)_{\#}\mu, (B, \mathrm{Id})_{\#}\nu) = 0\,. \end{aligned} \tag{3.2}$$

Here, $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ is a suitable discrepancy measure on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ so that the constraint of (3.2) is a surrogate for the original constraint $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$. In practice, we do not require that $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}} = 0$ implies $(\mathrm{Id}, F)_{\#}\mu = (B, \mathrm{Id})_{\#}\nu$; in fact, the former constraint can be a relaxation of the latter. The choice of $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$ will be specified later. Now, we turn (3.2) into the Lagrangian:

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

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:

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

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 enhances empirical results:

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

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

Last, 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.[4] MMD between two measures is a distance between their embeddings in some reproducing kernel Hilbert space (RKHS), which is indeed a metric under mild conditions [39]. Also, MMD is representable via the reproducing kernel of the RKHS; hence one may simply choose a kernel function to define it. Concretely, for any kernel $K_{\mathcal{X}}$ on $\mathcal{X}$, the square of MMD between $\widehat{\mu}_m$ and $B_{\#}\widehat{\nu}_n$ is

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

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

$$\begin{aligned}\min_{\substack{F:\mathcal{X}\to\mathcal{Y}\\ B:\mathcal{Y}\to\mathcal{X}}} \quad & \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n (c_{\mathcal{X}}(x_i, B(y_j)) - c_{\mathcal{Y}}(F(x_i), y_j))^2\\ & + \lambda_1\cdot\mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id}, F)_{\#}\widehat{\mu}_m, (B, \mathrm{Id})_{\#}\widehat{\nu}_n)\\ & + \lambda_2\cdot\mathrm{MMD}^2_{K_{\mathcal{X}}}(\widehat{\mu}_m, B_{\#}\widehat{\nu}_n) + \lambda_3\cdot\mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n)\,.\end{aligned} \tag{3.3}$$

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.

[4] This is merely a proof of concept. One may use other quantities in practice as described in section 6.

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 GM.

Like other transform sampling approaches in generative modeling, a practical way to solve (3.3) is to restrict the maps F and B to vector-valued function classes $\mathcal{F}$ and $\mathcal{B}$ parametrized by neural networks, respectively, and then optimize using a gradient descent algorithm. We note the following difference between this minimization problem and 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 [15, 32]. In contrast, the proposed RGM sampler only solves a single minimization problem in network parameters. Although generally nonconvex in nature, the parameter minimization problem in neural networks can often be efficiently optimized by stochastic gradient descent, and can even provably achieve the global optima if the loss satisfies certain Polyak–Łojasiewicz conditions [5].

Remark 3.3. We conclude this section by explaining a subtle difference between the proposed RGM sampler and using the form (1.4) with C replaced by Q defined in (1.6). This subtlety also explains why we employ the RGM formulation rather than GM. The latter approach—which can be viewed as the Lagrangian of GM as mentioned at the end of section 2—aims to find an isomorphism $T:\mathcal{X}\to\mathcal{Y}$ such that $c_{\mathcal{X}}(x,x')$ best matches $c_{\mathcal{Y}}(T(x),T(x'))$ on average, whose plug-in version with empirical data is

$$\frac{1}{m^2}\sum_{i,i'=1}^{m}\left(c_{\mathcal{X}}(x_i,x_{i'})-c_{\mathcal{Y}}(T(x_i),T(x_{i'}))\right)^2. \tag{3.4}$$

It is desirable to use information from both samples $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ to capture any isomorphic relation between two spaces. However, the GM objective (3.4) only uses information—the samples $\{x_i\}_{i=1}^m$—from $\mathcal{X}$, not from the target space $\mathcal{Y}$. In contrast, our RGM objective uses all information—samples from both spaces—to capture an isomorphic relation, which is encoded by the first term in (3.3):

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(c_{\mathcal{X}}(x_i,B(y_j))-c_{\mathcal{Y}}(F(x_i),y_j)\right)^2.$$

As the most valuable information (given that our goal is to learn the target distribution ν) is the available samples $\{y_j\}_{j=1}^n$ from ν, it is favorable to utilize them to learn isomorphisms.

4. Statistical analysis of the RGM sampler.

This section provides the main theoretical analysis of this paper: we study the statistical rate of convergence for the empirical problem (3.3), assuming it can be solved accurately. First, define

$$\begin{aligned} C(\mu,\nu,F,B) := &\int \left(c_{\mathcal{X}}(x,B(y))-c_{\mathcal{Y}}(F(x),y)\right)^2 \mathrm{d}\,\mu\otimes\nu \\ &+\lambda_1\cdot \mathrm{MMD}^2_{K_{\mathcal{X}}\otimes K_{\mathcal{Y}}}((\mathrm{Id},F)_{\#}\mu,(B,\mathrm{Id})_{\#}\nu) \\ &+\lambda_2\cdot \mathrm{MMD}^2_{K_{\mathcal{X}}}(\mu,B_{\#}\nu)+\lambda_3\cdot \mathrm{MMD}^2_{K_{\mathcal{Y}}}(F_{\#}\mu,\nu)\,. \end{aligned} \tag{4.1}$$

Then, the objective function of (3.3) is a plug-in estimator $C(\widehat{\mu}_m, \widehat{\nu}_n, F, B)$. We consider solving (3.3) over the transformation class $\mathcal{F} \times \mathcal{B}$ given as follows, for which we will state our nonasymptotic results in full generality. From now on, let $\mathcal{X}$ and $\mathcal{Y}$ be subsets of Euclidean spaces of dimensions $\dim(\mathcal{X})$ and $\dim(\mathcal{Y})$, respectively. $\mathcal{F}$ (resp., $\mathcal{B}$) is a collection of vector-valued measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ (resp., from $\mathcal{Y}$ to $\mathcal{X}$). For each $F \in \mathcal{F}$ and $k \in [\dim(\mathcal{Y})]$, we write $F_k(x)$ to denote the kth 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 kth coordinate of $F \in \mathcal{F}$. For $\ell \in [\dim(\mathcal{X})]$, we define B_ℓ and $\mathcal{B}_\ell = \{B_\ell : \mathcal{Y} \to \mathbb{R} \mid B \in \mathcal{B}\}$ analogously. Then, solving (3.3) over $\mathcal{F} \times \mathcal{B}$ is written as $\min_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\widehat{\mu}_m, \widehat{\nu}_n, F, B)$. We prove that the empirical solution leads to an approximate infimum of $(F, B) \mapsto C(\mu, \nu, F, B)$ evaluated with the population measures μ, ν, with sufficiently large sample sizes m and n.

Overview of assumptions. Before stating the main theorem, we briefly outline technical assumptions; the complete statement of the assumptions and definitions will be provided shortly. First, we assume that the cost functions $c_{\mathcal{X}}, c_{\mathcal{Y}}$ are bounded and Lipschitz (Assumptions 1, 4). Similarly, we assume boundedness and Lipschitzness of the kernel functions $K_{\mathcal{X}}, K_{\mathcal{Y}}$ corresponding to the MMD term (Assumptions 2, 5). Last, we impose two assumptions on the classes of transformations $F : \mathcal{X} \to \mathcal{Y}$ and $B : \mathcal{Y} \to \mathcal{X}$: we assume that the transformation classes are uniformly bounded (Assumption 3) and should contain nontrivial maps (Assumption 6). We will also employ a notion of combinatorial dimension to measure the complexity of real-valued function classes, which is called the pseudodimension (Definition 4.7).

Theorem 4.1. *Let $(\widehat{F}, \widehat{B})$ be a solution to the empirical problem*

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

with $C : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \times \mathcal{F} \times \mathcal{B} \to \mathbb{R}$ defined in (4.1). Under Assumptions 1–6, the following inequality holds with probability at least $1 - \delta$ on $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$:

$$C(\mu, \nu, \widehat{F}, \widehat{B}) - \inf_{(F,B)\in\mathcal{F}\times\mathcal{B}} C(\mu, \nu, F, B) \precsim \mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta)\,. \tag{4.2}$$

Here, $\mathcal{M}(\mathcal{F}, \mathcal{B}, m, n, \delta)$ denotes a complexity measure of $(\mathcal{F}, \mathcal{B})$ given in terms of pseudodimensions (Pdim) of $\mathcal{F}_k$ and $\mathcal{B}_\ell$ defined in Definition 4.7:

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

Remark 4.2. When $\mathcal{F}$ and $\mathcal{B}$ are parametrized by neural network classes (the ones we will use for numerical demonstrations in section 6), tight pseudodimension bounds established in [1, 24] can be plugged into Theorem 4.1 for concrete nonasymptotic rates.

Overview of the proof of Theorem 4.1. The rest of this section presents the main ideas of the proof of Theorem 4.1. To derive an upper bound (4.2), we will decompose the left-hand side of (4.2) into several terms, derive upper bounds on them separately, and then combine

those upper bounds together to obtain the right-hand side of (4.2); details of the proofs are provided in the supplementary material (supplement.pdf [local/web 1.35MB]).

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 $\lambda_1, \lambda_2, \lambda_3 > 0$. For convenience, we denote

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

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

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

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 4.1. Therefore,

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

where the right-hand side 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 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. We derive 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.

4.1. Concentration inequalities and uniform deviations. We utilize the McDiarmid's inequality to derive bounds on $|\widehat{C}_0(F,B) - C_0(F,B)|$ and $|\widehat{M}(F,B) - M(F,B)|$ for fixed 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

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

Proposition 4.3. *Under Assumption 1, for any pair* $(F,B)\in\mathcal{F}\times\mathcal{B}$ *and* $\delta>0$,

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

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

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

Proposition 4.4. *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 $1-6\delta$.

We now derive uniform deviation bounds for

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

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

Definition 4.5 (uniform covering number). *Let* $\mathcal{G}$ *be a collection of real-valued functions defined on a set* $\mathcal{Z}$. *Given* m *points* $z_1,\dots,z_m\in\mathcal{Z}$ *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* $z_1,\dots,z_m$:

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

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

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

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 $\mathcal{B}_\ell$ 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

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

The above assumption ensures smoothness of the 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 4.6. *Under Assumptions 1, 3, and 4, for any $\epsilon > 0$ and $\delta > 0$,*

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

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

Now, the remaining task is to choose ϵ carefully in Proposition 4.6 for a concrete upper bound. To this end, we utilize the pseudodimension defined below.

Definition 4.7 (pseudodimension). *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 pseudoshattered 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 pseudodimension of $\mathcal{G}$, denoted as $\operatorname{Pdim}(\mathcal{G})$, as the maximum cardinality of a subset $S \subset \mathcal{Z}$ that is pseudoshattered by $\mathcal{G}$.*

Using a well-established relation of the uniform covering number and the pseudodimension (Lemma SM1.2 in the supplementary materials), we can simplify Proposition 4.6 as follows.

Corollary 4.8. *Under Assumptions 1, 3, and 4, for any $\delta > 0$,*

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

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 the Rademacher complexities defined below.

Definition 4.9 (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:*

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

Here, $z_1,\dots,z_m$ 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 4.10. *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.*

4.2. Bounding Rademacher complexities via chaining. Now, the only remaining task is to bound the four Rademacher complexities that appear in the upper bound of Proposition 4.10. We will derive upper bounds for the compositional function classes 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

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

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

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

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

For any $F:\mathcal{X}\to\mathcal{Y}$, let $A_F\in\mathbb{R}^{m\times m}$ be a matrix whose diagonal elements are zero and the (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 $g := [g_1,\dots,g_m]^\top \sim N(0,I_m)$. We rely on the following chaining bound for the quadratic processes.

Lemma 4.11 (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*

$$\mathop{\mathbb{E}}_{g}\sup_{A\in\mathcal{T}} g^\top A g \le \inf_{J\in\mathbb{N}}\left\{m\delta_J + 12\int_{\delta_J/2}^{\Delta/2}\sqrt{2\log N(\delta,\mathcal{T})}\,\mathrm{d}\,\delta + 24\int_{\delta_J/2}^{\Delta/2}\log N(\delta,\mathcal{T})\,\mathrm{d}\,\delta\right\}, \tag{4.3}$$

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 right-hand side of (4.3), 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 nonconstant 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 nonconstant 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 nonconstant maps. With these assumptions, we can derive the following result.

Proposition 4.12. *Under Assumptions 2, 3, 5, and 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 4.3, 4.4, 4.10, and 4.12 and Corollary 4.8 directly imply Theorem 4.1. Omitted proofs and details can be found in the supplementary material (supplement.pdf [local/web 1.35MB]).

5. Further studies: Metric properties and representer theorem.

This section discusses some further perspectives on the RGM sampler. First, we focus on the metric side of the new notion RGM; we formally state its metric properties and connections to the GW and GM distances. Next, by utilizing an operator viewpoint, we develop an infinite-dimensional convex relaxation of (3.3), where global optima can be found efficiently; we analyze the new formulation by establishing a representer theorem on a suitable RKHS.

5.1. Metric properties of RGM.

It turns out that RGM possesses metric properties similar to those of the GW distance; the following result is an analogue to Theorem 2.5.

Theorem 5.1. *Let $\mathcal{M}$ be the collection of all metric measure spaces. Then, RGM satisfies the three metric axioms on $\mathcal{M}/\cong$, the collection of all equivalence classes of $\mathcal{M}$ induced by $\cong$.*

Remark 5.2. In practice, $\mathcal{X}, \mathcal{Y}$ are usually Euclidean spaces and $d_{\mathcal{X}}, d_{\mathcal{Y}}$ are the standard Euclidean distances. In many applications, instead of comparing the metric measures spaces $(\mathcal{X}, \mu, d_{\mathcal{X}}), (\mathcal{Y}, \nu, d_{\mathcal{Y}})$, it is often more desirable to work with network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}}), (\mathcal{Y}, \nu, c_{\mathcal{Y}})$ with cost functions given by $c_{\mathcal{X}} = h(d_{\mathcal{X}})$ and $c_{\mathcal{Y}} = h(d_{\mathcal{Y}})$ for some transformation $h : \mathbb{R}_+ \to \mathbb{R}$; one of the most common choices is $h(x) = \exp(-\alpha x^2)$ with $\alpha > 0$, which makes $h(d_{\mathcal{X}}), h(d_{\mathcal{Y}})$ radial basis function (RBF) kernels on $\mathcal{X}, \mathcal{Y}$, respectively (we will use these in numerical experiments). As a result, it is desirable to consider a distance on the collection of network spaces equipped with cost functions given as a composition of a transformation h and the base metric. The next result provides a modification of Theorem 5.1 tailored to such a situation.

Theorem 5.3. *Let $h : \mathbb{R}_+ \to \mathbb{R}$ 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 $\mathcal{N}^h/\cong$, the collection of all equivalence classes of $\mathcal{N}^h$ induced by $\cong$.*

Now that we have three distances—GW, GM,[5] and RGM—between network spaces, one may wonder about the generic relations among three distances GW, GM, and RGM, which will be established in the next proposition.

Proposition 5.4. *For network spaces* $(\mathcal{X}, \mu, c_{\mathcal{X}})$ *and* $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ *as in Definition* 2.2,

$$\mathrm{GW}(\mu,\nu) \le \mathrm{GM}(\mu,\nu) \le \mathrm{RGM}(\mu,\nu)\,. \tag{5.1}$$

Interestingly, under mild conditions, the above inequalities in Proposition 5.4 hold as equality, thus showing that RGM provides the exact metric as GW.

Theorem 5.5. *Let* $(\mathcal{X}, \mu, c_{\mathcal{X}})$ *and* $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ *be two network spaces. Assume that* $c_{\mathcal{X}}$ *and* $c_{\mathcal{Y}}$ *are bounded and* $\mu(\{x\}) = \nu(\{y\}) = 0$ *for any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$. *Then,* $\mathrm{GW}(\mu,\nu) = \mathrm{GM}(\mu,\nu) = \mathrm{RGM}(\mu,\nu)$.

The proof details of Theorem 5.3, Proposition 5.4, and Theorem 5.5 can be found in the supplementary material (supplement.pdf [local/web 1.35MB]).

Remark 5.6. The proof of Theorem 5.5 is inspired by [37, 9], recent developments in the GW analysis literature that study conditions under which GW = GM holds. Another important topic in the GW analysis literature is to study the type of cost functions under which the optimal coupling of GW is induced by a transport map [17]. Finally, we clarify that this section aims to derive metric properties of RGM in connection with GW and isomorphisms, as opposed to establishing new analytical results for GW and GM.

Finally, we conclude this section by pointing out a connection between inductive biases in RGM motivated by [8]. Given two Polish probability spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$, there exist cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ (that depend on μ, ν) such that the resulting network spaces $(\mathcal{X}, \mu, c_{\mathcal{X}})$ and $(\mathcal{Y}, \nu, c_{\mathcal{Y}})$ are strongly isomorphic. Then, more importantly, among (possibly) infinitely many pairs (F, B)'s in $\mathcal{I}(\mu, \nu)$, which are all valid for transform sampling, any optimal pair $(F^\star, B^\star)$ minimizing the RGM term (3.1) achieves the strong isomorphism

$$\mathrm{RGM}(\mu,\nu) = \int_{\mathcal{X}\times\mathcal{Y}} \left(c_{\mathcal{X}}(x, B^\star(y)) - c_{\mathcal{Y}}(F^\star(x), y)\right)^2 \mathrm{d}\,\mu \otimes \nu = 0\,, \tag{5.2}$$

and thus $c_{\mathcal{X}}(x, B^\star(y)) = c_{\mathcal{Y}}(F^\star(x), y)$ almost surely. In plain language, the RGM introduces an inductive bias favoring strong isomorphisms, in the same spirit as the Wasserstein-2 metric favors the transport map with the optimal cost seen in the introduction. Please refer to the supplementary material (supplement.pdf [local/web 1.35MB]) for a detailed discussion in section SM5 in the supplementary material.

5.2. Convex relaxation and representer theorem. As the last bit of the study, we utilize an operator viewpoint and develop a convex relaxation of (3.3) by relaxing and lifting it to an infinite-dimensional space. There are two reasons behind our convex relaxation: first, as a computational alternative to the possibly nonconvex optimization, and 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)

[5] Technically speaking, GM is not a distance as it is not symmetric.

linear operator $\mathbf{F}: L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ that maps functions on $\mathcal{Y}$ to functions on $\mathcal{X}$, where $L^2_{\mathcal{X}}$ is the collection of real-valued measurable functions f defined on $\mathcal{X}$ such that $\int_{\mathcal{X}} f^2 \,\mathrm{d}\,\pi_{\mathcal{X}} < \infty$ given a Borel measure $\pi_{\mathcal{X}}$ on $\mathcal{X}$; similarly, define $L^2_{\mathcal{Y}}$ given a Borel measure $\pi_{\mathcal{Y}}$ on $\mathcal{Y}$. Then, for a measurable map $F: \mathcal{X} \to \mathcal{Y}$, we can define $\mathbf{F}: L^2_{\mathcal{Y}} \to L^2_{\mathcal{X}}$ by letting $\mathbf{F}(g) = g \circ F$ for all $g \in L^2_{\mathcal{Y}}$. Similarly, we define $\mathbf{B}: L^2_{\mathcal{X}} \to L^2_{\mathcal{Y}}$ for each measurable map $B: \mathcal{Y} \to \mathcal{X}$. Then, one can verify that $\mathbf{F}$ and $\mathbf{B}$ are well-defined bounded linear operators under a mild assumption (detailed proofs are provided in the supplementary material (supplement.pdf [local/web 1.35MB])).

We derive the representer theorem for the convex relaxation of (3.3) under $c_{\mathcal{X}} = K_{\mathcal{X}}$ and $c_{\mathcal{Y}} = K_{\mathcal{Y}}$, namely, the cost functions are the 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. Moreover, since finite-dimensional convex optimization can be optimized globally with provable guarantees, such a formulation can be solved numerically efficiently.

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:

$$K_{\mathcal{X}}(x,x') = \sum_k \lambda_k \phi_k(x)\phi_k(x')\,, \quad K_{\mathcal{Y}}(y,y') = \sum_\ell \gamma_\ell \psi_\ell(y)\psi_\ell(y') \tag{5.3}$$

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

$$\mathbf{F}[\psi_\ell] = \sum_{k=1}^{\infty} \mathbf{F}_{k\ell}\phi_k\,, \quad \mathbf{B}[\phi_k] = \sum_{\ell=1}^{\infty} \mathbf{B}_{\ell k}\psi_\ell\,. \tag{5.4}$$

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 $[\mathbf{B}_{\ell k}]$.[6] Then, we can prove that the objective function in (3.3) with $c_{\mathcal{X}} = K_{\mathcal{X}}$ and $c_{\mathcal{Y}} = K_{\mathcal{Y}}$ is

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

[6] In other words, $\mathbf{F}$ and $\mathbf{B}$ are infinite-dimensional matrices, in which the number of rows and the number of columns can be infinite.

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

Here, $\mathbf{F}$ and $\mathbf{B}$ are the matrices denoting the operators induced by F and B, respectively, $\Phi_x=[\cdots,\phi_k(x),\cdots]^\top\in\mathbb{R}^\infty$ and $\Psi_y=[\cdots,\psi_\ell(y),\cdots]^\top\in\mathbb{R}^\infty$ for any $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, and $\Lambda=\operatorname{diag}(\lambda_1,\lambda_2,\dots)$ and $\Gamma=\operatorname{diag}(\gamma_1,\gamma_2,\dots)$ are diagonal matrices; detailed proofs are provided in the supplementary material (supplement.pdf [local/web 1.35MB]). Hence, (3.3) can be lifted to an infinite-dimensional optimization problem

$$\min_{(\mathbf{F},\mathbf{B})\in\mathcal{C}}\ \Omega(\mathbf{F},\mathbf{B})\,, \tag{5.5}$$

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

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,

$$\min_{\mathbf{F},\mathbf{B}\in\mathbb{R}^{\infty\times\infty}}\ \Omega(\mathbf{F},\mathbf{B}). \tag{5.6}$$

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 (5.6) boils down to a finite-dimensional convex program; the proof and relevant details can be found in the supplementary material (supplement.pdf [local/web 1.35MB]).

Theorem 5.7. *Consider* (5.5) *under the assumptions in Proposition* SM3.2. *Then, for any minimizer* $(\mathbf{F}^\star,\mathbf{B}^\star)$ *to the relaxed problem* (5.6), *we can find finite-dimensional matrices* $\mathsf{F}^\star_{m,n}\in\mathbb{R}^{m\times n}$ *and* $\mathsf{B}^\star_{n,m}\in\mathbb{R}^{n\times m}$ *such that*

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

where $\Lambda=\operatorname{diag}(\lambda_1,\lambda_2,\dots)$, $\Gamma=\operatorname{diag}(\gamma_1,\gamma_2,\dots)$, *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* (5.3). *In this case,* $\Omega(\mathbf{F}^\star,\mathbf{B}^\star)$ *can be rewritten as* $\omega(\mathsf{F}^\star_{m,n},\mathsf{B}^\star_{n,m})$ *for some convex function* ω *defined over* $\mathbb{R}^{m\times n}\times\mathbb{R}^{n\times m}$. *Hence, by minimizing* ω *over* $\mathbb{R}^{m\times n}\times\mathbb{R}^{n\times m}$, *we obtain a relaxation of* (5.6), *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 right-hand side is a finite-dimensional convex optimization. Last, 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 5.8. Looking inside the proof of Theorem 5.7, we know the solution to the infinite-dimensional optimization is an operator taking the 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

$$\mathbf{F}^\star[g](x) = \underbrace{K_{\mathcal{X}}(x, X_m)}_{1\times m} \underbrace{\mathsf{F}^\star_{m,n}}_{m\times n} \underbrace{g(Y_n)}_{n\times 1}, \tag{5.7}$$

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

Now let's draw a connection between the classic Nadaraya–Watson estimator and (5.7). For now consider a special case: (x_i, y_i)'s are paired with $m = n$. In such a case, the Nadaraya–Watson estimator takes the form

$$\sum_{i,j} K_{\mathcal{X}}(x, x_i) \cdot \tfrac{1}{m}\delta_{i=j} \cdot g(y_j); \tag{5.8}$$

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 (5.7) extends the above nonparametric smoothing idea to the decoupled data case, where the coupling weights $\mathsf{F}^\star_{m,n}$ are based on a solution to a convex program, with

$$(5.7) = \sum_{i,j} K_{\mathcal{X}}(x, x_i) \cdot \mathsf{F}^\star_{m,n}[i, j] \cdot g(y_j). \tag{5.9}$$

Last, 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 evaluating $\mathbb{E}_{y\sim F^\star_\# \mu}[g(y)] = \mathbb{E}_{x\sim\mu}[g \circ F^\star(x)]$. The integration, cast in the induced operator form, has the expression

$$\mathop{\mathbb{E}}_{x\sim\mu}[\mathbf{F}^\star[g](x)] = \mathop{\mathbb{E}}_{x\sim\mu}\left[\underbrace{K_{\mathcal{X}}(x, X_m)\mathsf{F}^\star_{m,n}}_{=:W(x)\in\mathbb{R}^n} g(Y_n)\right] = \mathop{\mathbb{E}}_{x\sim\mu}\left[\sum_{j=1}^n W_j(x) g(y_j)\right], \tag{5.10}$$

where $W(x)$ can be interpreted as the importance weights in the Monte Carlo integration. We conclude with one more remark as a sanity check: if plug-in instead $x \sim \widehat{\mu}_m$ in (5.10), one can verify that under mild conditions,

$$\mathop{\mathbb{E}}_{x\sim\widehat{\mu}_m}[\mathbf{F}^\star[g](x)] = \frac{1}{n}\sum_{j=1}^n g(y_j)\,. \tag{5.11}$$

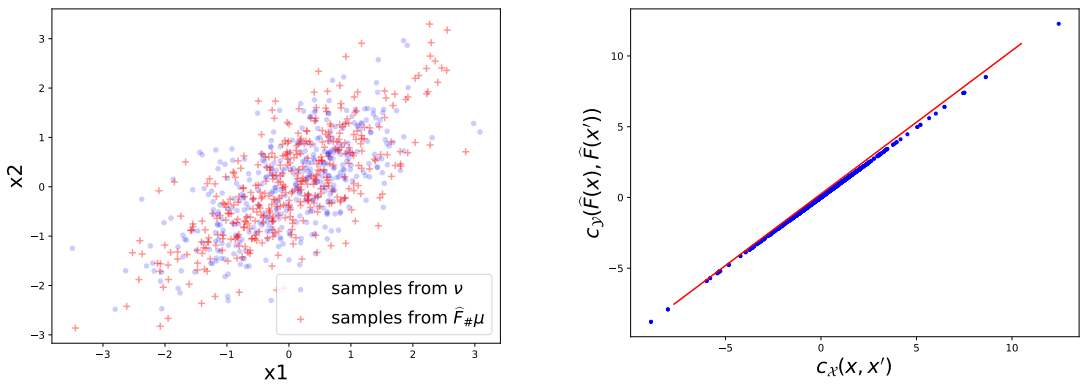
That is, with the empirical measure as input, (5.10) outputs the simple sample average.

6. Experiments. This section studies the empirical performance of the RGM sampler as a proof of concept. Following section 3, we find a minimum $(\widehat{F}, \widehat{B})$ of (3.3) over a suitable class $\mathcal{F} \times \mathcal{B}$ via gradient descent; then, we inspect the quality of transform sampling. Complete implementation details of the experiments are deferred to section SM4.

6.1. Gaussian distributions. Consider two strongly isomorphic Gaussian distributions on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$: the base measure $\mu = N(0, I_2)$ and the target distribution $\nu = N(0, \Sigma)$, where I_2 is the identity matrix and the entries of Σ are $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = 0.7$. We let $c_{\mathcal{X}}(x, x') = x^\top x'$ and $c_{\mathcal{Y}}(y, y') = y^\top \Sigma^{-1} y'$; then two network spaces are strongly isomorphic by design; indeed, any pair (F, B) given by $F(x) = \Sigma^{1/2} Q x$ and $B(y) = Q^\top \Sigma^{-1/2} y$ for $Q \in O(2)$, where $O(2)$ is the orthogonal group, yields $c_{\mathcal{X}}(x, B(y)) = c_{\mathcal{Y}}(F(x), y)$ for all $x, y \in \mathbb{R}^2$; hence F and B are strong isomorphisms. We aim at obtaining such a pair of (linear) isomorphisms by letting $\mathcal{F} = \mathcal{B} = \{x \mapsto Wx : W \in \mathbb{R}^{2\times 2}\}$, that is, the collection of all linear maps from $\mathbb{R}^2$ to $\mathbb{R}^2$. We set $K_{\mathcal{X}} = K_{\mathcal{Y}}$ as a degree-2 polynomial kernel that maps (x, y) to $(x^\top y + 1)^2$; the resulting MMD compares distributions by matching the first two moments, which is sufficient to distinguish Gaussian distributions. The linear maps found by solving the optimization problem (3.3) are given by $\widehat{F}(x) = \mathbf{F}x$ and $\widehat{B}(y) = \mathbf{B}y$ for some $\mathbf{F}, \mathbf{B} \in \mathbb{R}^{2\times 2}$ satisfying

$$\mathbf{F}\mathbf{F}^\top - \Sigma \approx \begin{pmatrix} -0.015 & -0.009 \\ -0.009 & -0.007 \end{pmatrix}, \quad \mathbf{B}\Sigma\mathbf{B}^\top - I_2 \approx \begin{pmatrix} 0.023 & -0.011 \\ -0.011 & 0.013 \end{pmatrix},$$
$$\mathbf{F}\mathbf{B} - I_2 \approx \begin{pmatrix} 0.006 & 0.002 \\ 0.004 & 0.002 \end{pmatrix}.$$

Since $\mathbf{F}\mathbf{F}^\top \approx \Sigma$, $\mathbf{B}\Sigma\mathbf{B}^\top \approx I_2$, and $\mathbf{F}\mathbf{B} \approx I_2$, the pair $(\widehat{F}, \widehat{B})$ can be seen as an instance of the pair of strong isomorphisms described above. Figure 1 illustrates that $\widehat{F}$ is a strong isomorphism (Definition 2.4): (a) shows that $\widehat{F}_{\#}\mu \approx \nu$, that is, $\widehat{F}$ is roughly a transport map, and (b) implies that $c_{\mathcal{X}}(x, x') \approx c_{\mathcal{Y}}(\widehat{F}(x), \widehat{F}(x'))$ holds.



(a) (b)

Figure 1. *Gaussian experiment:* $m = n = 50000$ *and* $\lambda_1 = \lambda_2 = \lambda_3 = 1$. (a) $\{\tilde{y}_j\}_{j=1}^{400}$ *versus* $\{\widehat{F}(\tilde{x}_i)\}_{i=1}^{400}$, *where* $\{\tilde{y}_j\}_{j=1}^{400}$ *and* $\{\tilde{x}_i\}_{i=1}^{400}$ *are i.i.d. from* $\nu = N(0, \Sigma)$ *and* $\mu = N(0, I_2)$, *respectively; they are new samples independent from* $\{y_j\}_{j=1}^{1000}$ *and* $\{x_i\}_{i=1}^{1000}$ *used in* (3.3). (b) *Points* $\{(c_{\mathcal{X}}(\tilde{x}_i, \tilde{x}_{i'}), c_{\mathcal{Y}}(\widehat{F}(\tilde{x}_i), \widehat{F}(\tilde{x}_{i'})))\}_{i,i'=1}^{40}$ *and a straight line* $y = x$.

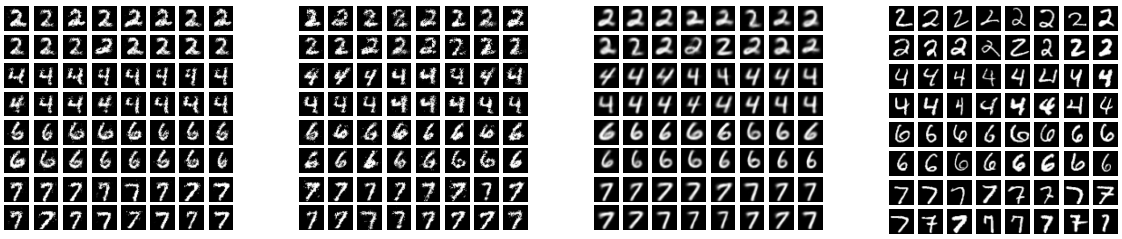
6.2. MNIST. Let ν be the distribution of images corresponding to four digits (2, 4, 6, 7) from the MNIST data set, which is supported on $\mathbb{R}^{784}$. Recall from section 1 that the support $\mathcal{Y}$ of ν is low-dimensional [19]; hence, choosing $\mathcal{X} = \mathbb{R}^d$ with $d \ll 784$ is reasonable. Here, for visualization, we first try an extreme embedding task with $d = 2$ and $\mu = N(0, I_2)$, that is, generate MNIST images by transforming two-dimensional Gaussian samples.

Model specifications. Unlike the Gaussian example, where we design the cost functions in advance to make the two spaces strongly isomorphic, specifying them can be more complicated in general cases, which might affect the quality of the RGM sampler. One common choice is the RBF kernel, also called the heat kernel, widely used in the object matching literature [47]. In this MNIST example, we have found that using RBF kernels as cost functions provides reasonable performance once they are scaled properly. Concretely, first define the RBF kernel $K_d(x, y) = \exp(-\|x-y\|^2/d)$ for $d \in \mathbb{N}$ and $x, y \in \mathbb{R}^d$; here, the constant $(1/d)$ serves as a scaling factor. Then, we define the cost functions as $c_{\mathcal{X}} = (K_2 - m_{\mathcal{X}})/\mathrm{sd}_{\mathcal{X}}$ and $c_{\mathcal{Y}} = (K_{784} - m_{\mathcal{Y}})/\mathrm{sd}_{\mathcal{Y}}$, where $m_{\mathcal{X}}$ and $\mathrm{sd}_{\mathcal{X}}$ are the median and the standard error of $\{K_{\mathcal{X}}(x_i, x_{i'})\}_{i,i'=1}^m$, respectively; $m_{\mathcal{Y}}$ and $\mathrm{sd}_{\mathcal{Y}}$ are defined analogously. This additional standardization process helps to align the cost functions. Similarly, $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ must be properly specified; comparing the first two moments using the degree-2 polynomial kernel is no longer sufficient as the target distribution is non-Gaussian. We also suggest using RBF kernels for the MMD terms; let $K_{\mathcal{X}} = K_2$ and $K_{\mathcal{Y}} = K_{784}$. The MMD induced by the RBF kernel indeed defines a metric between distributions under mild assumptions [39], which allows the resulting MMD terms to represent the binding constraint mentioned in section 3. We need richer classes than the linear maps used in the Gaussian case for the function classes $\mathcal{F}$ and $\mathcal{B}$. To this end, we will use the fully connected neural network (FCNN) with three hidden layers, each consisting of 50 neurons. Last, we let $m = n = 20000$.

Key features of the RGM sampler. Figure 2(a) shows the images generated by applying the resulting map $\widehat{F}$ to new i.i.d. samples from $\mu = N(0, I_2)$, which are independent of the samples used for training $\widehat{F}, \widehat{B}$. Though not perfect, we see that recognizable images can be generated by transforming two-dimensional Gaussian samples, efficient in computation.[7] Meanwhile, the map $\widehat{B}$ shows how the MNIST images can be embedded in $\mathbb{R}^2$. Figure 3(a) shows $\{\widehat{B}(\tilde{y}_j)\}_{j=1}^{500}$, where $\{\tilde{y}_j\}_{j=1}^{500}$ are drawn from ν (125 images for each digit), which are independent of the samples used for training $\widehat{F}, \widehat{B}$. We see that each digit forms a local cluster in $\mathbb{R}^2$, each of which is roughly representable according to the range of the angular coordinate. Last, though not perfect as in Figure 1(b) (strongly isomorphic case), Figure 3(b) shows that $\widehat{B}$ leads to a reasonable alignment of $c_{\mathcal{X}}(\widehat{B}(y), \widehat{B}(y'))$ versus $c_{\mathcal{Y}}(y, y')$.

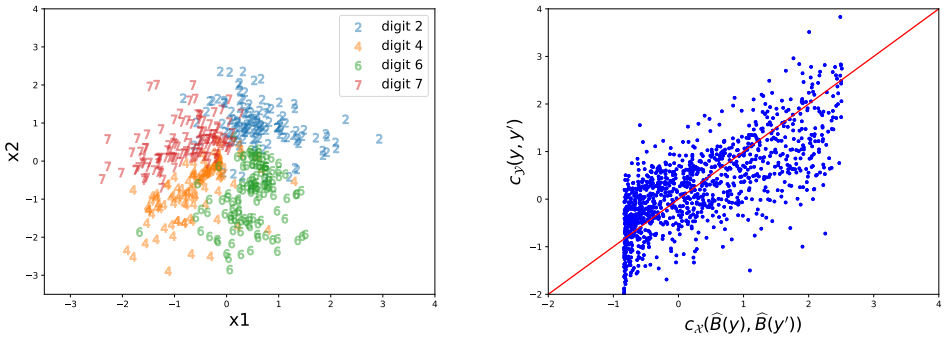
Quantitative evaluation and ablation analysis. We clarify that the above experiment with $\mu = N(0, I_2)$ is a proof of concept to demonstrate the key features of the RGM sampler. To obtain high-fidelity images comparable to those generated by dedicated MNIST generators, exhaustive tests should be done for tuning every component of the RGM sampler and optimization strategies, which is beyond the scope of this paper. Instead, we conclude this section by quantitatively evaluating some key components—discrepancy measures, the source dimension, and the neural network architecture—to inspect the performance of the RGM sampler more systematically.

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



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

Figure 2. *Panel* (a) *is generated by transforming new i.i.d. samples from* $\mu = N(0, I_2)$ *using* $\widehat{F}$*, trained under the aforementioned model specifications. Panel* (b) *is generated analogously, but with replacing the source dimension* $d = 2$ *with* $d = 4$*, namely,* $\mu = N(0, I_4)$*; for* (c)*, we further replace the three MMD terms in* (3.3) *with Sinkhorn divergences. Panel* (d) *shows real MNIST images.*



(a) (b)

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

The first component is choosing the discrepancies $\mathcal{L}_{\mathcal{X}}, \mathcal{L}_{\mathcal{Y}}, \mathcal{L}_{\mathcal{X}\times\mathcal{Y}}$. Though we have been using MMD as the leading example, one may use other discrepancies. Here, we consider a modification of (3.3) by replacing the MMD terms in (3.3) with Sinkhorn divergences, which has shown good performance for generative modeling tasks [21]. Next, we also try higher source dimensions, namely, $\mu = N(0, I_d)$ with $d = 4, 10, 20, 50$. Last, we test a different neural network architecture called the deep convolutional neural network (DCNN), which has shown good empirical performance in representing image-type data [42].[8] In summary, we test the combinations of two different discrepancies (MMD and Sinkhorn), five choices for the source dimension $d \in \{2, 4, 10, 20, 50\}$, and two choices for the neural network architecture (FCNN and DCNN). To quantitatively evaluate the performance of $\widehat{F}$ trained under each combination, we should quantify the closeness of $\widehat{F}_{\#}\mu$ and the target distribution ν; though there are many suggestions in the literature for such a task, we take a straightforward approach by comparing the moments and evaluating log likelihood scores.

[8]Roughly speaking, we adapt the structures of the generator and discriminator in [42] to construct $\mathcal{F}$ and $\mathcal{B}$, respectively.

Table 1

Evaluation scores using the squared differences between the first moments of the generated distribution $\widehat{F}_{\#}\mu$ and the target distribution ν; these are estimated by new samples $\{\tilde{x}_i\}_{i=1}^{2000}$ from $\mu = N(0, I_d)$ and $\{\tilde{y}_j\}_{j=1}^{2000}$ from the MNIST test set. We repeat the evaluation process for 100 times in each setting, and present the average evaluation score with the corresponding standard deviation in parentheses. The baseline is computed by randomly dividing the MNIST test set of 4000 samples into two subsets of 2000 samples and calculating the squared difference between the first moments of the two subsets analogously to the evaluation score. Here smaller evaluation scores imply better empirical results.

d	2	4	10	20	50
FCNN					
Sinkhorn	0.852(0.017)	0.743(0.011)	0.627(0.013)	0.587(0.012)	0.672(0.015)
MMD	0.784(0.015)	0.875(0.014)	0.941(0.020)	1.004(0.018)	1.131(0.021)
MMD-GAN	0.804(0.018)	1.016(0.024)	0.777(0.018)	1.121(0.024)	0.999(0.021)
DCNN					
Sinkhorn	0.841(0.014)	0.729(0.015)	0.671(0.014)	0.695(0.016)	0.647(0.015)
MMD	0.900(0.018)	0.913(0.017)	1.031(0.020)	1.249(0.025)	1.155(0.021)
MMD-GAN	0.927(0.022)	0.825(0.019)	1.044(0.020)	0.841(0.018)	1.069(0.023)
Baseline	0.432				

Table 1 summarizes the evaluation scores defined as $2\|\sum_i \widehat{F}(\tilde{x}_i)/\tilde{m} - \sum_j \tilde{y}_j/\tilde{n}\|^2$,[9] where $\{\tilde{x}_i\}_{i=1}^{\tilde{m}}$ and $\{\tilde{y}_j\}_{j=1}^{\tilde{n}}$ are i.i.d. samples from μ and ν, respectively; in words, we estimate the squared difference between the first moments of $\widehat{F}_{\#}\mu$ and ν by using new samples that are independent of the samples used to train $\widehat{F}, \widehat{B}$. By comparing the first two rows, we can observe that replacing the MMD terms with the Sinkhorn divergence leads to smaller (= better) evaluation scores (exception: FCNN with $d = 2$). By comparing Figures 2(b) and 2(c), we can see that images generated by Sinkhorn divergences tend to have smoother boundaries, which are visually more natural.[10] Next, we can see that the RGM sampler with the MMD terms often (but not always) has smaller evaluation scores than the MMD-GAN [18, 29], which aims to solve $\min_F \mathrm{MMD}^2_{K_\mathcal{Y}}(F_{\#}\widehat{\mu}_m, \widehat{\nu}_n)$, that is, only minimizing the last MMD term associated with $\mathcal{Y}$ in (3.3); note that the RGM sampler is better when d is very small, especially when $d = 2$.[11] Increasing the source dimension d or using a more sophisticated neural network architecture DCNN shows mixed results. For instance, for the Sinkhorn one, increasing d often (but not always) leads to smaller (= better) evaluation scores; however, for the MMD one, larger d results in the opposite. Similarly, DCNN often (but not always) leads to better scores for the Sinkhorn one, but the other way around for the MMD one. The main reason for such mixed results is likely related to the optimization; larger d and DCNN usually require more sophisticated and tailored optimization schemes. We can similarly analyze different evaluation scores defined by comparing the second moments or computing the log likelihood (estimated using a kernel density estimator [22]) to complement Table 1, which

[9] This formulation is motivated by the energy distance [51], which is obtained by replacing the distance in the energy distance formulation with the squared distance.

[10] Of course, given that many images in MNIST have uneven boundaries, one cannot conclude that Sinkhorn outperforms MMD.

[11] One should, however, be reminded of the following difference when comparing the RGM sampler and MMD-GAN: the former requires training of both maps F, B, whereas the latter is a minimization over F only.

we present in the supplementary material (supplement.pdf [local/web 1.35MB]). Omitted details of this section can also be found therein.

7. Discussions. In this work, we proposed the reversible Gromov–Monge (RGM) sampler, a new variant of transform sampling based on the RGM distance operating between distributions on heterogeneous spaces. The RGM sampler fuses the Gromov–Wasserstein (GW) idea with the transform sampling task, aiming at introducing inductive biases toward isomorphisms (see section SM5). We have established statistical results as the main theoretical contribution, along with several supporting results on analytic properties of the RGM distance and the convex relaxation based on the operator viewpoint and the representer theorem.

Last, we mention two directions for future research. First, as briefly mentioned in section 3, the RGM sampler is easily implementable by solving a minimization problem via the gradient descent algorithm; however, it is unclear whether one can derive global convergence results from this minimization, which we leave as future research. Next, another interesting direction is to study analytic properties of the RGM distance; investigating deeper connections with the GW distance and strong isomorphisms would be an interesting topic in the GW analysis literature.

REFERENCES

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, UK, 1999, https://doi.org/10.1017/CBO9780511624216.

[2] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

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

[4] Y. Bai, T. Ma, and A. Risteski, *Approximability of Discriminators Implies Diversity in GANs*, preprint, arXiv:1806.10586, 2018.

[5] R. Bassily, M. Belkin, and S. Ma, *On Exponential Convergence of SGD in Non-Convex Over-Parametrized Learning*, preprint, arXiv:1811.02564, 2018.

[6] A. J. Blumberg, M. Carriere, M. A. Mandell, R. Rabadan, and S. Villar, *MREC: A Fast and Versatile Framework for Aligning and Matching Point Clouds with Applications to Single Cell Molecular Data*, preprint, arXiv:2001.01666, 2020.

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

[8] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[9] Y. Brenier and W. Gangbo, *L^p approximation of maps by diffeomorphisms*, Calc. Var. Partial Differential Equations, 16 (2003), pp. 147–164, https://doi.org/10.1007/s005260100144.

[10] C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka, *Learning generative models across incomparable spaces*, in Proceedings of the International Conference on Machine Learning, PMLR, 2019.

[11] E. Çela, *The Quadratic Assignment Problem: Theory and Algorithms*, Springer, New York, 1998, https://doi.org/10.1007/978-1-4757-2787-6.

[12] M. Chen, W. Liao, H. Zha, and T. Zhao, *Statistical Guarantees of Generative Adversarial Networks for Distribution Estimation*, preprint, arXiv:2002.03938, 2020.

[13] S. Chowdhury and F. Mémoli, *The Gromov–Wasserstein distance between networks and stable network invariants*, Inf. Inference, 8 (2019), pp. 757–787.

[14] S. Chowdhury, D. W. Miller, and T. Needham, *Quantized Gromov-Wasserstein*, in Proceedings of ECML/PKDD, 2021.

[15] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, *Training GANs with Optimism*, preprint, arXiv:1711.00141, 2017.

[16] C. Doersch, *Tutorial on Variational Autoencoders*, preprint, arXiv:1606.05908, 2016.
[17] T. Dumont, T. Lacombe, and F.-X. Vialard, *On the Existence of Monge Maps for the Gromov-Wasserstein Distance*, preprint, arXiv:2210.11945, 2022.
[18] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, *Training generative neural networks via maximum mean discrepancy optimization*, in Proceedings of the 31st Annual Conference on Uncertainty in Artificial Intelligence (UAI), 2015, pp. 258–267.
[19] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, *Estimating the intrinsic dimension of datasets by a minimal neighborhood information*, Sci. Rep., 7 (2017), pp. 1–8.
[20] M. Gellert, M. F. Hossain, F. J. F. Berens, L. W. Bruhn, C. Urbainsky, V. Liebscher, and C. H. Lillig, *Substrate specificity of thioredoxins and glutaredoxins—towards a functional classification*, Heliyon, 5 (2019), e02943.
[21] A. Genevay, G. Peyré, and M. Cuturi, *Learning generative models with Sinkhorn divergences*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 1608–1617.
[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, Vol. 27, 2014, https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.
[23] C. Gouriéroux and A. Monfort, *Simulation-Based Econometric Methods*, Oxford University Press, New York, 1997, https://doi.org/10.1093/0198774753.001.0001.
[24] N. Harvey, C. Liaw, and A. Mehrabian, *Nearly-tight VC-dimension bounds for piecewise linear neural networks*, in Proceedings of the Conference on Learning Theory, PMLR, 2017, pp. 1064–1068.
[25] D. P. Kingma and M. Welling, *Auto-encoding Variational Bayes*, preprint, arXiv:1312.6114, 2013.
[26] T. C. Koopmans and M. Beckmann, *Assignment problems and the location of economic activities*, Econometrica, 25 (1957), pp. 53–76, https://doi.org/10.2307/1907742.
[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proc. IEEE, 86 (1998), pp. 2278–2324.
[28] Q. Lei, J. D. Lee, A. G. Dimakis, and C. Daskalakis, *SGD Learns One-Layer Networks in WGANs*, preprint, arXiv:1910.07030, 2019.
[29] Y. Li, K. Swersky, and R. Zemel, *Generative moment matching networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2015, pp. 1718–1727.
[30] T. Liang, *Estimating Certain Integral Probability Metric (IPM) Is as Hard as Estimating under the IPM*, preprint, arXiv:1911.00730, 2019.
[31] T. Liang, *How well generative adversarial networks learn distributions*, J. Mach. Learn. Res., 22 (2021), pp. 1–41.
[32] T. Liang and J. Stokes, *Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, Vol. 89, PMLR, 2019, pp. 907–915.
[33] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee, *Optimal transport mapping via input convex neural networks*, in Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 6672–6681, https://proceedings.mlr.press/v119/makkuva20a.html.
[34] D. McFadden, *A method of simulated moments for estimation of discrete response models without numerical integration*, Econometrica, 57 (1989), pp. 995–1026, https://doi.org/10.2307/1913621.
[35] F. Mémoli, *Gromov–Wasserstein distances and the metric approach to object matching*, Found. Comput. Math., 11 (2011), pp. 417–487, https://doi.org/10.1007/s10208-011-9093-5.
[36] F. Mémoli and T. Needham, *Distance Distributions and Inverse Problems for Metric Measure Spaces*, preprint, arXiv:1810.09646, 2018.
[37] F. Mémoli and T. Needham, *Comparison Results for Gromov-Wasserstein and Gromov-Monge Distances*, preprint, arXiv:2212.14123, 2022.
[38] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng, *Sobolev GAN*, preprint, arXiv:1711.04894, 2017.
[39] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, *Kernel mean embedding of distributions: A review and beyond*, Found. Trends Mach. Learn., 10 (2017), pp. 1–141, https://doi.org/10.1561/2200000060.

[40] A. Pakes and D. Pollard, *Simulation and the asymptotics of optimization estimators*, Econometrica, 57 (1989), pp. 1027–1057, https://doi.org/10.2307/1913622.
[41] G. Peyré, M. Cuturi, and J. Solomon, *Gromov-Wasserstein averaging of kernel and distance matrices*, in Proceedings of the International Conference on Machine Learning, PMLR, 2016, pp. 2664–2672.
[42] A. Radford, L. Metz, and S. Chintala, *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, https://arxiv.org/abs/1511.06434, 2016.
[43] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, 1999, https://doi.org/10.1007/978-1-4757-3071-5.
[44] M. Scetbon, G. Peyré, and M. Cuturi, *Linear-Time Gromov Wasserstein Distances Using Low Rank Couplings and Costs*, preprint, arXiv:2106.01128, 2021.
[45] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1986, https://www.routledge.com/Density-Estimation-for-Statistics-and-Data-Analysis/Silverman/p/book/9780412246203.
[46] S. Singh and B. Póczos, *Minimax Distribution Estimation in Wasserstein Distance*, preprint, arXiv:1802.08855, 2018.
[47] J. Solomon, G. Peyré, V. G. Kim, and S. Sra, *Entropic metric alignment for correspondence problems*, ACM Trans. Graph., 35 (2016), 72.
[48] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, *On the empirical estimation of integral probability metrics*, Electron. J. Stat., 6 (2012), pp. 1550–1599.
[49] C. J. Stone, *Optimal global rates of convergence for nonparametric regression*, Ann. Statist., 10 (1982), pp. 1040–1053.
[50] K.-T. Sturm, *The Space of Spaces: Curvature Bounds and Gradient Flows on the Space of Metric Measure Spaces*, preprint, arXiv:1208.0434, 2012.
[51] G. J. Székely and M. L. Rizzo, *Energy statistics: A class of statistics based on distances*, J. Statist. Plann. Inference, 143 (2013), pp. 1249–1272.
[52] A. Taghvaei and A. Jalali, 2-*Wasserstein Approximation via Restricted Convex Potentials with Application to Improved Training for GANs*, preprint, http://arxiv.org/abs/1902.07197, 2019.
[53] V. Titouan, R. Flamary, N. Courty, R. Tavenard, and L. Chapel, *Sliced Gromov-Wasserstein*, in Advances in Neural Information Processing Systems, Vol. 32, 2019.
[54] J. Weed and Q. Berthet, *Estimation of Smooth Densities in Wasserstein Distance*, preprint, arXiv:1902.01778, 2019.
[55] C. A. Weitkamp, K. Proksch, C. Tameling, and A. Munk, *Gromov-Wasserstein Distance Based Object Matching: Asymptotic Inference*, preprint, arXiv:2006.12287, 2020.
[56] H. Xu, D. Luo, and L. Carin, *Scalable Gromov-Wasserstein learning for graph partitioning and matching*, in Advances in Neural Information Processing Systems, Vol. 32, 2019.