METRICS FOR PARAMETRIC FAMILIES OF NETWORKS

MARIO GÓMEZ, GUANQUN MA, TOM NEEDHAM, AND BEI WANG

ABSTRACT. We introduce a general framework for analyzing data modeled as parameterized families of networks. Building on a Gromov-Wasserstein variant of optimal transport, we define a family of parameterized Gromov-Wasserstein distances for comparing such parametric data, including time-varying metric spaces induced by collective motion, temporally evolving weighted social networks, and random graph models. We establish foundational properties of these distances, showing that they subsume several existing metrics in the literature, and derive theoretical approximation guarantees. In particular, we develop computationally tractable lower bounds and relate them to graph statistics commonly used in random graph theory. Furthermore, we prove that our distances can be consistently approximated in random graph and random metric space settings via empirical estimates from generative models. Finally, we demonstrate the practical utility of our framework through a series of numerical experiments.

1. Introduction

Motivating examples. Consider the problem of mathematically modeling the collective spatial motion of multiple agents over time, such as a group of animals [5, 20, 35, 46], a population of cells [7, 33], or a fleet of vehicles [21]. It is desirable to adopt a representation that is invariant under ambient isometries when only the *intrinsic* features of the motion are of interest, rather than the *extrinsic* position of the group in ambient space. A natural choice in this setting is to record the pairwise distances between agents over time, which leads to the notion of a *time-varying metric space*: a one-parameter family of metrics defined on a common underlying set. A substantial body of work has developed methods for analyzing data in the form of time-varying metric spaces [23, 24, 32, 52].

While time-varying metric spaces offer a concrete and well-studied class of examples, they exemplify a broader and increasingly common scenario in data science and network analysis: the need to analyze parameterized families of complex structures. Additional examples include:

- Time-varying graphs. Applications in social network modeling [16,39] and neuroscience [6,53] often involve weighted graphs whose edge weights evolve over time, giving rise to data in the form of a family of graphs parameterized by real numbers, representing the time parameter.
- Heat kernels. The heat kernel on a Riemannian manifold M describes the diffusion of heat across the manifold and is widely used in geometry processing [30,34,37]. In this setting, the data naturally take the form of a family of maps $M \times M \to \mathbb{R}$ parameterized by the positive real numbers.
- Random graph models. Generative random graph models [18], such as the Erdős–Rényi model [12], can be viewed as drawing graphs from a (typically unknown) distribution over the space of all graphs on a fixed node set. Here, the data consist of a collection of graphs parameterized by this underlying state space according to the (unknown) distribution.

In practice, a dataset may consist of an *ensemble* of parameterized objects, such as a collection of timevarying fMRI brain connectivity graphs, in which case a metric is needed to compare elements within the ensemble.

To this end, we introduce the notion of a *parameterized measure network*, a flexible model for representing parameterized families of complex structures that encompasses the examples described above. We then define a novel and highly general family of distances on the space of parameterized measure networks, establish their metric, analytical, and statistical properties, and demonstrate their practical utility through a robust numerical implementation.

Gromov-Wasserstein distance. To describe our contributions in more detail, we now briefly review a key component of our approach, the Gromov-Wasserstein framework from optimal transport theory [27, 29].

1

Recall that a metric measure space (mm-space) [17] is a triple $\mathcal{X} = (X, \mu_X, d_X)^1$, where X is a Polish space (i.e., a separable completely metrizable topological space), $d_X : X \times X \to \mathbb{R}_{\geq 0}$ is a metric on X, and μ_X is a Borel probability measure on X. In other words, it is a metric space equipped with a probability measure.

Given another mm-space $\mathcal{Y}=(Y,\mu_Y,d_Y)$, the associated order-p Gromov-Wasserstein (GW) distance between mm-spaces \mathcal{X} and \mathcal{Y} [17,29], for $1 \leq p < \infty$, is

(1)
$$\mathsf{GW}_p(\mathcal{X}, \mathcal{Y}) \coloneqq \frac{1}{2} \inf_{\pi} \left(\int_{(X \times Y)^2} \int_{(X \times Y)^2} |d_X(x, x') - d_Y(y, y')|^p \pi(dx \otimes dy) \pi(dx' \otimes dy') \right)^{1/p},$$

where the infimum is over measure couplings (or joint measures) π on $X \times Y$ whose marginals agree with μ_X and μ_Y , respectively (see Sec. 2.2 for more details).

A general framework of comparing parameterized measure networks. Our proposed framework substantially generalizes the GW distance, extending it to a family of metrics designed to compare general parameterized measure networks. As a concrete illustration, we next show how the GW framework described in Eqn. (1) can be adapted to the setting of time-varying metric spaces.

Let $\mathcal{X} = (X, \mu_X, (d_X^t)_{t \in [0,1]})$, where (X, μ_X) is a Polish probability space as in the setting of metric measure spaces, but the metric structure is replaced by a one-parameter family $(d_X^t)_{t \in [0,1]}$ of metrics on X, assumed to vary continuously in t. Given another such structure $\mathcal{Y} = (Y, \mu_Y, (d_Y^t)_{t \in [0,1]})$, one can define a notion of GW distance between \mathcal{X} and \mathcal{Y} by

$$(2) \quad \mathsf{GW}_p(\mathcal{X},\mathcal{Y}) \coloneqq \frac{1}{2} \inf_{\pi} \int_0^1 \left(\int_{(X \times Y)^2} \int_{(X \times Y)^2} |d_X^t(x,x') - d_Y^t(y,y')|^p \pi(dx \otimes dy) \pi(dx' \otimes dy') \right)^{1/p} \nu(dt),$$

where ν is a Lebesgue measure on [0,1].

The structure of the metric in Eqn. (2) is quite natural, and related ideas have appeared in previous work [32,42,52]. In contrast, our general framework encompasses a broader class of novel metrics that have not yet been explored in the literature. In particular, it includes a variant that incorporates an additional optimal transport—based alignment step for comparing parameterized measure networks defined over different parameter spaces—an approach especially relevant for applications involving, for example, random graph models.

Contributions. We now provide a more detailed account of our contributions.

(1) **New model for parameterized data.** We introduce the notion of a *parameterized measure network* (*pm-net*) (Proposition 3.1), which unifies all of the data types described above. Analogous to a metric measure space, a pm-net is defined as a tuple

$$\mathcal{X} = (X, \mu_X, \Omega_X, \nu_X, \omega_X),$$

where (X, μ_X) is a Polish probability space, (Ω_X, ν_X) is a measured parameter space, and $\omega_X = (\omega_X^t)_{t \in \Omega_X}$ is a parameterized family of kernels $\omega_X^t : X \times X \to \mathbb{R}$. Examples of pm-nets, along with their connection to the motivating examples discussed earlier, are given in Sec. 3.1.2.

(2) A general family of Gromov-Wasserstein-type distances. We define a new family of GW-type distances, called parameterized Gromov-Wasserstein distances (Proposition 3.12). Members of this family are denoted GW_C, where the subscript C specifies a chosen cost structure—that is, a rule for quantifying the geometric distortion induced by a given probability coupling. The metric properties of GW_C, which depend on the choice of C, are established in Theorem 1 and Theorem 2. The first of these results is stated at a high level of generality and provides a category-theoretic interpretation of GW-type distances, which may be of independent interest in optimal transport theory (see Proposition 3.22). The second result is more specialized: its proof shares key ideas with standard arguments for existing GW variants (see, e.g., [4,8,29]), but requires substantial work to extend them to the parameterized setting. In particular, the proof relies on technical lemmas concerning the (lower semi-)continuity of a certain distortion functional (Proposition 3.15 and Proposition 3.24), as well as on a novel and somewhat subtle equivalence relation for parameterized networks (Proposition 3.26).

¹The original definition uses the tuple (X, d_X, μ_X) ; for convenience, we adopt the order (X, μ_X, d_X) in this paper.

- (3) Generalizations of existing metrics. Certain instances of the parameterized GW framework recover metrics that have previously appeared in the literature. In particular, we show that for specific choices of C under suitable technical assumptions, our metric coincides with:
 - the temporal alignment–based GW distance of [10] (Proposition 3.20);
 - certain instances of the Z-GW distances of [4] (Proposition 4.1);
 - metrics for time-varying metric spaces from [42] (Proposition 4.2) and [23,24] (Proposition 4.4);
 - GW-type distances for heat kernels from [30] (Proposition 4.5) and [9] (Proposition 4.6).
- (4) Lower bounds and stability of invariants. It is shown in Theorem 3 that (for a particularly useful choice of C), the distance GW_C can be lower bounded by a Wasserstein distance (see Sec. 2.2) defined over the (classical) GW space. It follows that the parameterized GW distance GW_C is lower bounded by a polynomial-time computable pseudometric, defined also in terms of Wasserstein distances (see Proposition 4.10 and Proposition 4.12). On one hand, these lower bounds give computationally tractable estimates of the parameterized GW distance. Alternatively, we interpret these lower bounds as proofs that certain invariants of random graph and random metric space models are stable. For example, Proposition 4.13 shows that the distribution of total edges in a random graph is a GW-stable invariant of the model.
- (5) Sampling convergence for random graphs and metric spaces. In the random graph model setting, one does not typically have access to the full parameter space (Ω_X, ν_X) , but rather samples from it. In Sec. 4.4, we study approximation of parameterized GW distances from random samples of the parameter space; in particular, Theorem 4 shows that estimates from random samples converge to the true distance as the number of samples goes to infinity.
- (6) Numerical experiments. We describe our implementation of parametrized GW distances in Sec. 5.1. We adapt components of the Python Optimal Transport (POT) library [13] (which implements algorithms from [36, 45]) to approximate GW_C using gradient descent for several choices of cost structure C. We provide explicit formulas for the gradient of GW_C and other auxiliary quantities.

We perform a number of numerical experiments, which illustrate the intuition that parametrized GW distances integrate information from all parameters $t \in \Omega$. We give qualitative examples which show that parameterized GW distances are able to pick up subtle structures of data at multiple scales in Sec. 5.2 and Sec. 5.4. In Sec. 5.3, we show that the parameterized GW distance serves as a meaningful invariant for comparing two samples of random graphs. Finally, we incorporate our metrics into a supervised learning framework: rather than fixing the measure ν on the parameter set Ω , we allow it to vary as a mechanism for feature selection. We evaluate this idea by clustering dynamic metric spaces (parameterized by time) that differ only at specific time intervals (Sec. 5.5).

Outline. We begin in Sec. 2 with a review of essential concepts from measure theory and optimal transport. In Sec. 3, we formally define parameterized measure networks and introduce a family of GW distances for comparing them, establishing their core metric properties. Sec. 4 presents our main theoretical results on the equivalence and estimation of these distances, including their connections to existing metrics, lower bounds, and sampling convergence guarantees. Finally, Sec. 5 details our computational framework for estimating the proposed distances and demonstrates their practical utility through several numerical examples.

2. Preliminary Concepts and Notations

2.1. **Basic Terminology from Measure Theory.** In this subsection, we review basic terminology from measure theory, which experts may safely skip. This also serves to standardize our notation for the rest of the paper.

A measurable space is a pair (X, \mathcal{A}) , where X is a set and \mathcal{A} is a σ -algebra on X (i.e., a collection of subsets of X, called measurable sets, that is closed under complements, countable unions and intersections). Given a measurable space (X, \mathcal{A}) , a measure μ_X is a function $\mu_X : \mathcal{A} \to [0, \infty]$ that assigns a non-negative value to each measurable set such that it is countably additive and $\mu_X(\emptyset) = 0$. A measure space is a measurable space together with a measure, denoted as a triple (X, \mathcal{A}, μ_X) ; sometimes the σ -algebra \mathcal{A} is omitted from the notation when it is clear from the context—we generally assume that X is endowed with a topology and that it is endowed with the Borel σ -algebra, generated by open sets. Let (X, \mathcal{A}, μ_X) be a measure space; a property is said to hold μ_X -almost everywhere (often written as μ_X -a.e.) on X if the set of points where

the property does not hold has measure zero. Given a pair of measure spaces (X, \mathcal{A}, μ_X) and (Y, \mathcal{B}, μ_Y) , a product measure is a measure $\mu_X \otimes \mu_Y : \mathcal{A} \otimes \mathcal{B} \to [0, \infty]$ satisfying $(\mu_X \otimes \mu_Y)(A \times B) = \mu_X(A) \mu_Y(B)$, $\forall A \in \mathcal{A}, B \in \mathcal{B}$. For a measure space (X, μ_X) and $p \in [1, \infty]$, we use $\|\cdot\|_{L^p(X; \mu_X)}$ to denote the standard L^p -norm. We use $L^p(X; \mu_X)$ to denote the set of all measurable functions $f : X \to \mathbb{R}$ with finite L^p -norm. In particular $L^\infty(X; \mu_X)$ is the set of essentially bounded functions.

A Polish space is a topological space (X, τ) that is separable (i.e., it contains a countable dense subset) and completely metrizable (i.e., there exists a metric d_X on X such that d_X induces the topology τ and (X, d_X) is complete). Here, the metric is not part of the structure; we only require that one exists. In contrast, a Polish metric space is a metric space (X, d_X) that is both complete (i.e., every Cauchy sequence converges) and separable. In this case, the metric d_X is part of the structure. Every Polish metric space gives rise to a Polish space by forgetting the metric, and every Polish space admits some compatible metric that turns it into a Polish metric space. A Polish probability space is a measure space (X, μ_X) where X is assumed to be a Polish space and μ_X is a Borel probability measure, i.e., $\mu_X(X) = 1$. We use $\mathcal{P}(X)$ to denote the set of probability measures on X. Given a $\mu_X \in \mathcal{P}(X)$, the support of μ_X , denoted supp (μ_X) , is the set of $x \in X$ such that every open neighborhood of x has positive measure.

Let (X, μ_X) be a Polish probability space. Given another Polish space Y and a (Borel) measurable map $f: X \to Y$, the pushforward of μ_X to Y, denoted $f_{\#}\mu_X$, is the measure on Y defined by $(f_{\#}\mu_X)(A) = \mu_X(f^{-1}(A))$ (for A being any Borel set). If μ_Y is a probability measure on Y, the map f is measure-preserving if $f_{\#}\mu_X = \mu_Y$.

More generally, given two Polish probability spaces (X, μ_X) and (Y, μ_Y) , a coupling of μ_X and μ_Y is a measure $\pi \in \mathcal{P}(X \times Y)$ whose left and right marginals are μ_X and μ_Y , respectively; that is, using $p_X : X \times Y \to X$ and $p_Y : X \times Y \to Y$ to denote the coordinate projections, $(p_X)_{\#}\pi = \mu_X$ and $(p_Y)_{\#}\pi = \mu_Y$. We use $\mathcal{C}(\mu_X, \mu_Y)$ to denote the collection of all couplings of μ_X and μ_Y .

2.2. Wasserstein Distances. The notion of measure coupling leads naturally to the Kantorovich formulation of transport distance between measures, the core object of study in optimal transport theory [47]. Let (X, d_X) be a Polish metric space and let $\mu_X, \mu_X' \in \mathcal{P}(X)$ be any two probability measures on X. For $p \in [1, \infty]$, the order-p Wasserstein distance [47, Definition 6.1] between μ_X and μ_X' is

(3)
$$\mathsf{W}_{p}^{d_{X}}(\mu_{X}, \mu'_{X}) \coloneqq \inf_{\pi \in \mathcal{C}(\mu_{X}, \mu'_{X})} \|d_{X}\|_{L^{p}(X \times X; \pi)} = \inf_{p < \infty} \inf_{\pi \in \mathcal{C}(\mu_{X}, \mu'_{X})} \left(\int_{X \times X} d_{X}(x, x')^{p} \pi(dx \otimes dx') \right)^{1/p}.$$

Here, and throughout the rest of the paper, we use the notation = to indicate that the integral formulation is valid for $p < \infty$, whereas the L^p -norm definition holds for any $p \in [1, \infty]$.

2.3. Measure Networks and Gromov-Wasserstein Distances. The Wasserstein distance described above is able to compare measures defined over the same (Polish) metric space, whereas this paper is primarily interested in comparing distributions defined over distinct spaces. This is handled with the Gromov-Wasserstein (GW) framework Eqn. (1), whose purview is extended beyond metric measure spaces, following the work of Chowdhury and Mémoli [8].

In order to handle kernel structures which are more general than metrics, Chowdhury and Mémoli introduced the following concept.

Definition 2.1 (Measure Network [8, Definition 2.1]). A measure network is a triple $\mathcal{X} = (X, \mu_X, \omega_X)$ such that X is a compact Polish space, μ_X is a fully supported Borel probability measure, and ω_X is a bounded measurable function on $X \times X$. In other words, (X, μ_X) is a Polish probability space and $\omega_X : X \times X \to \mathbb{R}$ is a kernel belonging to $L^{\infty}(X \times X, \mu_X \otimes \mu_X)$.

Example 2.2. If ω_X is a distance metric (inducing the given Polish space a topology τ on X), then $\mathcal{X} = (X, \mu_X, \omega_X)$ is a metric measure space (mm-space). However, the measure network formalism allows much more general structures than an mm-space. Of particular interest is the case where X is a finite set of nodes for a graph structure, and ω_X is a kernel encoding node interactions. For example, ω_X could be an adjacency function (possibly including edge weights), that is, $\omega_X : X \times X \to \mathbb{R}_{\geq 0}$, where $\omega_X(u, v) = w_e$ if nodes u and v are connected by an edge e with a weight w_e ; otherwise $\omega_X(u, v) = 0$. ω_X could also be a graph Laplacian or a graph heat kernel (see Proposition 3.7).

In [8], the GW distance (1) was extended to a pseudometric which is able to compare general measure networks. A similar idea was considered by Sturm [43], where some different regularity and symmetry assumptions on the kernels were imposed. To streamline notation for the rest of the paper, we define a preliminary concept (which goes back at least to Mémoli [27]).

Definition 2.3 (Distortion). Fix $1 \leq p < \infty$. Let (X, μ_X) and (Y, μ_Y) be probability spaces, and let $f_X : X \times X \to \mathbb{R}$ and $f_Y : Y \times Y \to \mathbb{R}$ be essentially bounded, measurable functions. Given a coupling $\pi \in \mathcal{C}(\mu_X, \mu_Y)$, define the *p*-distortion as

$$\operatorname{dis}_p(\pi, f_X, f_Y) := \left(\int_{(X \times Y)^2} |f_X(x, x') - f_Y(y, y')|^p \, \pi(dx \otimes dy) \, \pi(dx' \otimes dy') \right)^{1/p}.$$

For $p = \infty$, define

$$\operatorname{dis}_{\infty}(\pi, f_X, f_Y) \coloneqq \sup_{(x,y), (x',y') \in \operatorname{supp}(\pi)} |f_X(x, x') - f_Y(y, y')|,$$

where $\operatorname{supp}(\pi)$ denotes the support of π . In other words, we have the general definition

$$\operatorname{dis}_{p}(\pi, f_{X}, f_{Y}) := \|f_{X} \circ (p_{X}, p_{X}) - f_{Y} \circ (p_{Y}, p_{Y})\|_{L^{p}((X \times Y)^{2}; \pi \otimes \pi)},$$

where p_X and p_Y are the coordinate projections from $X \times Y$ to X and Y, respectively.

For $p \in [1, \infty]$, the associated **Gromov-Wasserstein (GW) distance** between measure networks \mathcal{X} and \mathcal{Y} is given by

$$\mathsf{GW}_{p}(\mathcal{X}, \mathcal{Y}) \coloneqq \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_{X}, \mu_{Y})} \operatorname{dis}_{p}(\pi, \omega_{X}, \omega_{Y}) \\
= \frac{1}{p < \infty} \inf_{\pi \in \mathcal{C}(\mu_{X}, \mu_{Y})} \left(\int_{(X \times Y)^{2}} |\omega_{X}(x, x') - \omega_{Y}(y, y')|^{p} \pi(dx \otimes dy) \pi(dx' \otimes dy') \right)^{1/p}.$$

It was shown in [8] that GW_p defines a pseudometric on the space of measure networks, with $\mathsf{GW}_p(\mathcal{X},\mathcal{Y}) = 0$ if and only if \mathcal{X} and \mathcal{Y} are weakly isomorphic, defined as follows [31, Definition 2.4].

Definition 2.4 (Weakly Isomorphic). Let \mathcal{X} and \mathcal{Y} be measure networks. A measure network \mathcal{Z} is called a **stabilization** of \mathcal{X} and \mathcal{Y} if there exist maps $\varphi_X : Z \to X$ and $\varphi_Y : Z \to Y$ such that (i) φ_X and φ_Y are measure-preserving, and (ii) $\omega_Z(z,z') := \omega_X(\varphi_X(z),\varphi_X(z')) = \omega_Y(\varphi_Y(z),\varphi_Y(z'))$ holds for $(\mu_Z \otimes \mu_Z)$ -almost every $(z,z') \in Z \times Z$. If a stabilization exists, we say that \mathcal{X} and \mathcal{Y} are **weakly isomorphic**.

3. Parameterized Gromov-Wasserstein Distances

- 3.1. Parameterized Measure Networks. Motivated by the examples described in Sec. 1, we now introduce a general framework for encoding objects consisting of a parameterized family of kernels over a fixed set.
- 3.1.1. Main Definition. The primary objects of interest in this paper are defined as follows.

Definition 3.1 (Parameterized Measure Network). A parameterized measure network (abbreviated as pm-net) is a 5-tuple of the form $\mathcal{X} = (X, \mu_X, \Omega_X, \nu_X, \omega_X)$, where:

- X is a Polish space endowed with a Borel probability measure $\mu_X \in \mathcal{P}(X)$, referred to as the underlying measure space.
- Ω_X is a compact Polish space endowed with a Borel probability measure ν_X , referred to as the *parameter space*, and
- ω_X is a function

$$\omega_X: \Omega_X \to L^{\infty}(X \times X; \mu_X \otimes \mu_X)$$
$$t \mapsto \omega_X^t$$

which is continuous with respect to the L^{∞} norm, referred to as a parameterized network kernel.

Remark 3.2. Various technical conditions in Proposition 3.1 could be relaxed—for example, one could assume each ω_X^t lies in L^p rather than L^{∞} , or weaken the compactness requirement on Ω_X —but we impose these conditions for convenience. They provide a sufficiently flexible framework while sparing us from having to verify an excess of intricate technical details in the proofs.

3.1.2. Examples of Parameterized Measure Networks. We now present several examples of parameterized measure networks, starting with the simple observation that this framework generalizes the standard measure network as a special case.

Example 3.3 (Measure Networks). Let (X, μ_X, ω_X) be a measure network, in the sense of [8] (see Sec. 2.3). This gives a trivial example of a parameterized measure network: let (Ω_X, ν_X) be a space consisting of a single point, $\Omega_X = \{t\}$, and define $\mathcal{X} = (X, \mu_X, \Omega_X, \nu_X, \omega_X)$, where (by a slight abuse of notation) $\omega_X^t = \omega_X$.

The next few examples take (a subset of) \mathbb{R} as the parameter space. In these cases, a parameter t intuitively represents a notion of "time" or "scale". These examples formalize concepts which were described informally in the introduction.

Example 3.4 (Time-Varying Metric Spaces). A time-varying metric space or dynamic metric space consists of a (finite) set X endowed with a collection $(d_X^t)_{t\in\Omega_X}$ of (pseudo-)metrics, parameterized by some compact subset Ω_X of \mathbb{R} . Such objects were studied as models for flocking behavior in [44] and metrics on the space of these objects were studied in [23–25,52], using constructions similar to Gromov-Hausdorff and Gromov-Wasserstein distances, as well as ideas from topological data analysis. Setting $\omega_X^t = d_X^t$ and imposing the mild assumption that $t \mapsto \omega_X^t$ is L^{∞} -continuous, one obtains a representation of a dynamic metric space as a parameterized measure network for any choices of $\nu_X \in \mathcal{P}(\Omega_X)$ and $\mu_X \in \mathcal{P}(X)$.

Example 3.5 (Time-Varying Networks). Similar to the above, one can consider data consisting of a weighted graph whose edge weights vary in time t, defined over some compact subset of real numbers (e.g., [6, 16, 25, 39, 53]). Taking ω_X^t to be a graph kernel for each t (e.g., the weighted adjacency matrix), and choosing necessary distributions, leads to a pm-network representation of this $time-varying\ network$ structure.

Example 3.6 (Riemannian Heat Kernels). Let X be a (say, compact) Riemannian manifold with associated Laplacian operator Δ . A solution $\omega:(0,\infty)\times X\times X\to\mathbb{R}$ of the heat equation

$$\frac{\partial}{\partial t}\omega(t,x,x') = \Delta_x\omega(t,x,x'),$$

which limits to the delta distribution at x as $t \to 0^+$, is called a *heat kernel* for X. Let Ω_X be a compact subset of $\mathbb{R}_{>0}$, endowed with some measure ν_X , let μ_X denote the normalized Riemannian volume measure on X, and let $\omega_X^t(x,x') = \omega(t,x,x')$ for a heat kernel ω . This data then defines a parameterized measure network

Example 3.7 (Graph Heat Kernels). The heat kernels described in Proposition 3.6 have a discrete counterpart in graph theory. Let G be a finite graph with node set X. The natural discrete version of the Riemannian Laplacian operator is the *graph Laplacian*. We formulate it in matrix notation as follows. Choosing an enumeration of $X = (x_1, \ldots, x_n)$, let A denote the associated $n \times n$ adjacency matrix, and D the $n \times n$ degree matrix (the diagonal matrix whose i-th diagonal entry is the degree of node x_i in G). Then the *graph Laplacian* is the matrix $\Delta = D - A$. The *graph heat equation* is typically written as

$$\frac{d}{dt}B(t) = -\Delta B(t),$$

and a solution B, understood to be a time-varying $n \times n$ matrix, is given by the graph heat kernel

$$B(t) = \exp(-t\Delta).$$

One thus obtains a parameterized measure network by taking μ_X to be some measure over X, ν_X to be some measure over a compact parameter space $\Omega_X \subset \mathbb{R}_{>0}$, and by setting

$$\omega_X^t(x_i, x_i) = B(t)_{i,i},$$

where $B(t)_{i,j}$ denotes the (i,j)-entry of the matrix B(t).

The following examples have a different flavor than those above, in that the parameter space is treated as a state space, so that the parameterized network kernel is naturally considered as a random variable (valued in a function space).

Example 3.8 (Random Graphs). Let X be a finite set and let Ω_X be the set of all (say, simple) graphs with node set X. This is a finite set, which we endow with the discrete topology. A distribution ν_X over Ω_X can then be understood as a random graph model; for example, the $Erd\Hos-R\acute{e}nyi$ model [12], the stochastic

block model [19], or the Watts-Strogatz model [50] (see [18] for a survey on the topic). We construct a pm-net \mathcal{X} by choosing a measure μ_X on X (e.g., the uniform measure), and taking ω_X^t to be the adjacency kernel associated to the graph $t \in \Omega_X$ (i.e., $\omega_X^t(x, x') = 1$ if the nodes $x, x' \in X$ are connected by an edge in the graph at time t, and $\omega_X^t(x, x') = 0$ otherwise).

We make the observation here that one typically does not have access to the full distribution ν_X in practice. Rather, the standard examples described above are generative models; one generally has access to iid samples t_1, \ldots, t_N from Ω_X , and hence to samples $\omega_X^{t_1}, \ldots, \omega_X^{t_N}$ of the the kernel defining the pm-net structure. This motivates the statistical questions that we consider in Sec. 4.4.

Example 3.9 (Random Metric Spaces). This example is similar to the random graph example above, extending it to consider random metric structures. Towards this end, let (X, μ_X) be a compact Polish probability space and let Ω_X be a compact collection of metrics inducing the given topology on X, endowed with the subspace topology coming from the inclusion $\Omega_X \subset L^{\infty}(X \times X; \mu_X \otimes \mu_X)$. A distribution ν_X on Ω_X defines a random metric space model over X. There is an associated pm-net \mathcal{X} given by taking ω_X^t to be equal to the metric $t \in \Omega_X$.

As in Proposition 3.8, one typically only has access to samples of metrics distributed according to ν_X , and statistical inferences come into play (see Sec. 4.4). In practice, these samples could arise from noisy measurements of some metric structure; to give a concrete example, the nuclear magnetic resonance problem in structural biology represents molecular conformation via many noisy measurements of pairwise distances between its atoms [11].

We conclude our list of examples by describing a situation where the parameter space consists of a set of modalities for representing a given dataset.

Example 3.10 (Graph Representations). Given a graph G with a node set X, there are many choices of kernels $X \times X \to \mathbb{R}$ for representing the structure of G, including: the adjacency kernel, the graph Laplacian, the graph heat kernel for various choices of t, the shortest path distance function, or other kernels induced by node features (if they exist). One can consider a multimodal representation of G by setting Ω_X to be a (say, finite) set of representation modalities t and defining ω_X^t to be the kernel for G under the given modality. For any choices of distributions ν_X and μ_X , this data determines a parameterized measure network.

- 3.2. Distances Between Parameterized Measure Networks. Our next objective is to define a suitable notion of distance between parameterized measure networks. Rather than tailoring a distinct distance for each specific class of objects, we adopt a unified and general framework that accommodates a wide range of structures, including those introduced in Proposition 3.3 through Proposition 3.10. This generality is essential, as the objects we seek to analyze are inherently diverse. By formulating a family of distances at this level of abstraction, we are able to derive theoretical guarantees applicable across multiple settings, specializing only when necessary to address particular cases.
- 3.2.1. Classes of Parameterized Measure Networks. Throughout the rest of the section, let \mathfrak{N} denote some fixed but arbitrary class of parameterized measure networks. The family of distances introduced below will be defined with respect to the class \mathfrak{N} .

Example 3.11 (Classes of Parameterized Measure Networks). We will return frequently to the following important classes \mathfrak{N} of pm-nets:

- (1) the class containing all pm-nets, which we denote as \mathfrak{N}_{all} ;
- (2) for a fixed parameter space (Ω, ν) , the class of pm-nets of the form $\mathcal{X} = (X, \mu_X, \Omega, \nu, \omega_X)$, which we denote as \mathfrak{N}_{ν} :
- (3) one of the more specific classes consisting of the objects described in Proposition 3.3–Proposition 3.9, for which we do not currently introduce any specialized notation.
- 3.2.2. General Family of Distances. Our general family of distances on a fixed class $\mathfrak N$ is defined as follows.

Definition 3.12 (Parameterized Gromov-Wasserstein Distance). A **cost structure** on \mathfrak{N} is an assignment C taking a pair of pm-nets $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}$ to a function

$$\mathsf{C}_{\mathcal{X},\mathcal{Y}}:\mathcal{C}(\mu_X,\mu_Y)\to\mathbb{R}_{>0}.$$

Given a cost structure C, the parameterized Gromov-Wasserstein distance induced by C is

(5)
$$\begin{aligned} \mathsf{GW}_\mathsf{C} : \mathfrak{N} \times \mathfrak{N} \to \mathbb{R}_{\geq 0} \\ (\mathcal{X}, \mathcal{Y}) \mapsto \mathsf{GW}_\mathsf{C}(\mathcal{X}, \mathcal{Y}) \coloneqq \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \mathsf{C}_{\mathcal{X}, \mathcal{Y}}(\pi). \end{aligned}$$

Remark 3.13. We refer to GW_C as a "distance" only in a colloquial sense, as we do not assert that it satisfies the axioms of a metric in general. However, we show below that under suitable assumptions on the cost function C, GW_C does indeed exhibit metric properties.

Remark 3.14. An optimization problem similar to Eqn. (5) was studied in a recent paper by Sebbouh, Cuturi, and Peyré [38], with a view toward generalizing duality properties in optimal transport-type problems. There, the additional assumption that $C_{\mathcal{X},\mathcal{Y}}$ is always a concave function was imposed for analytical purposes, but we make no such restriction.

3.2.3. Continuity of Distortion. Before providing examples of interesting cost structures C, we establish a useful property of the distortion function, introduced in Proposition 2.3.

Lemma 3.15. Let \mathcal{X} and \mathcal{Y} be pm-nets and fix $1 \leq p \leq \infty$. For all $s \in \Omega_X$ and $t \in \Omega_Y$, the quantity $\operatorname{dis}_p(\pi, \omega_X^s, \omega_Y^t)$ is well-defined; in particular, it is finite. Moreover, if $p < \infty$, the function

$$\operatorname{dis}_{p}: \mathcal{C}(\mu_{X}, \mu_{Y}) \times \Omega_{X} \times \Omega_{Y} \to \mathbb{R}$$
$$(\pi, s, t) \mapsto \operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t})$$

is continuous. If $p = \infty$, $\operatorname{dis}_{\infty}$ is lower semicontinuous.

Proof. The finiteness claim is straightforward, due to our regularity assumptions in Proposition 3.1:

$$\operatorname{dis}_p(\pi,\omega_X^s,\omega_Y^t) \leq \|\omega_X^s\|_{L^p(\mu_X\otimes\mu_X)} + \|\omega_Y^t\|_{L^p(\mu_Y\otimes\mu_Y)} \leq \|\omega_X^s\|_{L^\infty(\mu_X\otimes\mu_X)} + \|\omega_Y^t\|_{L^\infty(\mu_Y\otimes\mu_Y)} < \infty.$$

We proceed with the continuity claim. Suppose $p < \infty$. Let $\epsilon > 0$ and fix $\pi_0 \in \mathcal{C}(\mu_X, \mu_Y)$, $s_0 \in \Omega_X$, and $t_0 \in \Omega_Y$. By [8, Lemma 2.3], for any fixed s and t, the function

$$\operatorname{dis}_{p}(\bullet, \omega_{X}^{s}, \omega_{Y}^{t}) : \mathcal{C}(\mu_{X}, \mu_{Y}) \to \mathbb{R}$$
$$\pi \mapsto \operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t})$$

is continuous. Thus there exists $V \subset \mathcal{C}(\mu_X, \mu_Y)$ neighborhood of π_0 such that for all $\pi \in V$,

$$|\operatorname{dis}_p(\pi, \omega_X^{s_0}, \omega_Y^{t_0}) - \operatorname{dis}_p(\pi_0, \omega_X^{s_0}, \omega_Y^{t_0})| < \epsilon.$$

Similarly, since $\omega_X : \Omega_X \to L^{\infty}(X \times X, \mu_X \otimes \mu_X)$ is continuous, there exists $U_X \subset \Omega_X$ neighborhood of s_0 such that

$$\|\omega_X^s - \omega_X^{s_0}\|_{L^p(X^2, \mu_X \otimes \mu_X)} \le \|\omega_X^s - \omega_X^{s_0}\|_{L^\infty(X^2, \mu_X \otimes \mu_X)} < \epsilon$$

for all $s \in U_X$. We define U_Y analogously.

Let p_X and p_Y be the standard projections from $X \times Y$. For convenience, write $\overline{\omega}_X^s = \omega_X^s \circ (p_X, p_X)$ and $\overline{\omega}_Y^t = \omega_Y^t \circ (p_Y, p_Y)$. By the triangle inequality of the L_p norm,

$$\begin{aligned} \operatorname{dis}_{p}(\pi, \omega_{X}^{s_{0}}, \omega_{Y}^{t_{0}}) &= \|\overline{\omega}_{X}^{s_{0}} - \overline{\omega}_{Y}^{t_{0}}\|_{L^{p}((X \times Y)^{2}; \pi \otimes \pi)} \\ &\leq \|\overline{\omega}_{X}^{s_{0}} - \overline{\omega}_{X}^{s}\|_{L^{p}((X \times Y)^{2}; \pi \otimes \pi)} + \|\overline{\omega}_{X}^{s} - \overline{\omega}_{Y}^{t}\|_{L^{p}((X \times Y)^{2}; \pi \otimes \pi)} + \|\overline{\omega}_{Y}^{t} - \overline{\omega}_{Y}^{t_{0}}\|_{L^{p}((X \times Y)^{2}; \pi \otimes \pi)} \\ &= \|\omega_{X}^{s_{0}} - \omega_{X}^{s}\|_{L^{p}(X^{2}; \mu_{X} \otimes \mu_{X})} + \|\overline{\omega}_{X}^{s} - \overline{\omega}_{Y}^{t}\|_{L^{p}((X \times Y)^{2}; \pi \otimes \pi)} + \|\omega_{Y}^{t} - \omega_{Y}^{t_{0}}\|_{L^{p}(Y^{2}; \mu_{Y} \otimes \mu_{Y})} \\ &< \operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t}) + 2\epsilon. \end{aligned}$$

With a symmetric argument we obtain $|\operatorname{dis}_p(\pi,\omega_X^s,\omega_Y^t) - \operatorname{dis}_p(\pi,\omega_X^{s_0},\omega_Y^{t_0})| < 2\epsilon$. Then for all $\pi \in V$, $s \in U_X$, and $t \in U_Y$,

$$\begin{split} |\operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t}) - \operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s_{0}}, \omega_{Y}^{t_{0}})| \\ & \leq |\operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t}) - \operatorname{dis}_{p}(\pi, \omega_{X}^{s_{0}}, \omega_{Y}^{t_{0}})| + |\operatorname{dis}_{p}(\pi, \omega_{X}^{s_{0}}, \omega_{Y}^{t_{0}}) - \operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s_{0}}, \omega_{Y}^{t_{0}})| < 3\epsilon. \end{split}$$

This proves that dis_p is continuous for $p < \infty$. Since $\operatorname{dis}_\infty$ is the supremum over $p \geq 1$ of the family of continuous functions dis_p , it is lower semicontinuous.

3.2.4. Examples of Parameterized Gromov-Wasserstein Distances. The following examples of cost structures C, along with their associated parameterized Gromov-Wasserstein distances, will be referenced frequently throughout the paper.

Example 3.16 (Main Example: Fixed Parameter Space). Fix a parameter space (Ω, ν) and let \mathfrak{N}_{ν} be the class of pm-nets with this parameter space, that is, of the form $\mathcal{X} = (X, \mu_X, \Omega, \nu, \omega_X)$; see Proposition 3.11. A natural cost structure is given by

(6)
$$\mathsf{C}_{\mathcal{X},\mathcal{Y}}(\pi) = \frac{1}{2} \| \mathrm{dis}_p(\pi,\omega_X,\omega_Y) \|_{L^q(\Omega;\nu)} = \frac{1}{q < \infty} \frac{1}{2} \left(\int_{\Omega} \mathrm{dis}_p(\pi,\omega_X^t,\omega_Y^t)^q \nu(dt) \right)^{1/q}.$$

We record the full expression for GW_C in this case, for later reference:

$$\mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \| \mathrm{dis}_p(\pi, \omega_X, \omega_Y) \|_{L^q(\Omega; \nu)}$$

$$= \int_{p,q<\infty} \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \left(\int_{\Omega} \left(\int_{(X\times Y)^2} |\omega_X(x,x') - \omega_Y(y,y')|^p \pi(dx \times dy) \pi(dx' \times dy') \right)^{q/p} \nu(dt) \right)^{1/q}.$$

Observe that the cost structure is well-defined (i.e., finite). Indeed, Proposition 3.15 implies that, for fixed $\pi \in \mathcal{C}(\mu_X, \mu_Y)$, the map $\Omega \to \mathbb{R} : t \mapsto \operatorname{dis}_p(\pi, \omega_X^t, \omega_Y^t)$ is continuous. By compactness of Ω , it is therefore q-integrable.

Remark 3.17. Suppose that (Ω, ν) is a one-point space, $\Omega = \{t\}$. Let $\mathcal{X} = (X, \mu_X, \omega_X)$ be a measure network; as was observed in Proposition 3.3, \mathcal{X} is naturally represented as an element of the class \mathfrak{N}_{ν} , which we denote in this remark as $\overline{\mathcal{X}} = (X, \mu_X, \Omega, \nu, \omega_X)$. With the cost structure C from Proposition 3.16, the parameterized GW distance is equivalent to the standard GW distance for measure networks, as was described in Sec. 2.3. Indeed,

$$\mathsf{GW}_{\mathsf{C}}(\overline{\mathcal{X}}, \overline{\mathcal{Y}}) = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \| \mathrm{dis}_p(\pi, \omega_X, \omega_Y) \|_{L^q(\Omega; \nu)}$$
$$= \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \mathrm{dis}_p(\pi, \omega_X^t, \omega_Y^t) = \mathsf{GW}_p(\mathcal{X}, \mathcal{Y}).$$

Remark 3.18. For any parameter space, GW_C is related to a Z-Gromov-Wasserstein distance, in the sense of [4]. This is explained in detail below, in Sec. 4.1.

Example 3.19 (Main Example: General Parameter Spaces). Let \mathfrak{N}_{all} denote the class of all pm-nets, and let $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}_{all}$. For $p, q \in [1, \infty]$, we have the cost structure

(8)
$$C_{\mathcal{X},\mathcal{Y}}(\pi) = \frac{1}{2} \inf_{\xi \in \mathcal{C}(\nu_X,\nu_Y)} \|\operatorname{dis}_p(\pi,\omega_X,\omega_Y)\|_{L^q(\Omega_X \times \Omega_Y,\xi)} \\ = \frac{1}{2} \inf_{\xi \in \mathcal{C}(\nu_X,\nu_Y)} \left(\int_{\Omega_X \times \Omega_Y} \operatorname{dis}_p(\pi,\omega_X^t,\omega_Y^s)^q \xi(dt \otimes ds) \right)^{1/q},$$

in which case GW_C becomes

(9)
$$\mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \inf_{\xi \in \mathcal{C}(\nu_X, \nu_Y)} \| \mathrm{dis}_p(\pi, \omega_X, \omega_Y) \|_{L^q(\Omega_X \times \Omega_Y, \xi)}.$$

By arguments similar to the above, Proposition 3.15 implies that the cost structure is well-defined (i.e., finite).

Although our analysis and experiments in the remainder of the paper primarily focus on the cost structures presented in Proposition 3.16 and Proposition 3.19, a wide range of alternative cost structures can be constructed to suit specific applications. We largely leave detailed exploration of these possibilities for future work, but describe one such potential idea below. We also note that a discrete analogue of this construction was previously introduced in [10], motivated by the problem of aligning time series valued in different spaces.

Example 3.20 (Optimization Over Reparameterizations). Consider the class \mathfrak{N} consisting of pm-nets whose parameter space is $\Omega_X = [0, 1]$, endowed with Lebesgue measure ν_X ; for example, this class contains versions of the pm-nets described in Proposition 3.6, Proposition 3.7, and Proposition 3.4, when the parameter space is restricted to the interval (extending definitions to the parameter t = 0, as necessary). Let $\mathrm{Diff}_+([0,1])$ denote the group of orientation-preserving diffeomorphisms of [0,1]. The cost structure

$$\mathsf{C}_{\mathcal{X},\mathcal{Y}}(\pi) = \inf_{\alpha \in \mathsf{Diff}_+([0,1])} \left(\int_0^1 \mathsf{dis}_p(\pi,\omega_X^t,\omega_Y^{\alpha(t)})^q \nu(dt) \right)^{1/q}$$

leads to a distance GW_C involving a "temporal alignment" of the pm-nets. In particular, if \mathcal{X} and \mathcal{Y} only differ by an orientation-preserving reparameterization of their parameter spaces, then one has $\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y}) = 0$ under this cost structure. The associated optimization problem has connections to ideas from the field of statistical shape analysis [3,40], which we plan to explore in future work.

- 3.3. General metric properties. We now formally study metric-like structures arising from parameterized GW distances. Predictably, these depend on properties of the cost structure C. For the rest of this subsection, fix a class $\mathfrak N$ of pm-nets and a cost structure C on $\mathfrak N$.
- 3.3.1. Preliminary Concepts. To state the main theorem, we need some additional terminology and notation. Recall the Gluing Lemma [43, Lemma 1.4], a standard result in optimal transport theory which says that, given couplings $\pi_{XY} \in \mathcal{C}(\mu_X, \mu_Y)$ and $\pi_{YZ} \in \mathcal{C}(\mu_Y, \mu_Z)$, there exists a unique probability measure $\tilde{\pi}$ on $X \times Y \times Z$ whose (X,Y)- and (Y,Z)-marginals are π_{XY} and π_{YZ} , respectively. We use the notation $\pi_{XY} \bullet \pi_{YZ}$ for the $(X \times Z)$ -marginal of $\tilde{\pi}$; that is, letting $p_A : X \times Y \times Z \to A$ denote the coordinate projection for $A \in \{X, Y, Z\}$, we define

$$\pi_{XY} \bullet \pi_{YZ} := (p_X \times p_Z)_{\#} \tilde{\pi} \in \mathcal{C}(\mu_X, \mu_Z).$$

We now introduce several useful properties of a cost structure.

Definition 3.21 (Properties of Cost Structures). Let C be a cost structure on a class of pm-nets \mathfrak{N} .

(1) The cost structure C respects gluing if, for any $\pi_{XY} \in \mathcal{C}(\mu_X, \mu_Y)$ and $\pi_{YZ} \in \mathcal{C}(\mu_Y, \mu_Z)$, it holds that

$$C_{\mathcal{X},\mathcal{Z}}(\pi_{XY} \bullet \pi_{YZ}) \le C_{\mathcal{X},\mathcal{Y}}(\pi_{XY}) + C_{\mathcal{Y},\mathcal{Z}}(\pi_{YZ}).$$

(2) Recall that a map $\varphi: X \to Y$ that is measure-preserving with respect to μ_X and μ_Y induces a coupling via $(\mathrm{id}_X \times \varphi)_{\#} \mu_X \in \mathcal{C}(\mu_X, \mu_Y)$. Given a cost structure, we abuse notation and write

$$\mathsf{C}_{\mathcal{X},\mathcal{Y}}(\varphi) \coloneqq \mathsf{C}_{\mathcal{X},\mathcal{Y}}\left((\mathrm{id}_X \times \varphi)_{\#} \mu_X\right).$$

The cost structure respects identities if $C_{\mathcal{X},\mathcal{X}}(\mathrm{id}_X) = 0$ for all $\mathcal{X} \in \mathfrak{N}$.

- (3) If C respects gluing and respects identities, we call it a lax homomorphism.
- (4) Given a coupling $\pi \in \mathcal{C}(\mu_X, \mu_Y)$, there is a corresponding **adjoint coupling** $\pi^* \in \mathcal{C}(\mu_Y, \mu_X)$, given by $\pi^* = \operatorname{swap}_{\#} \pi$, where swap : $X \times Y \to Y \times X$ is the map $\operatorname{swap}(x,y) = (y,x)$. We say that the cost structure C is **symmetric** if

$$C_{\mathcal{X},\mathcal{Y}}(\pi) = C_{\mathcal{Y},\mathcal{X}}(\pi^*) \qquad \forall \ \pi \in \mathcal{C}(\mu_X,\mu_Y).$$

Remark 3.22 (Categorical Interpretation). The lax homomorphism terminology is inspired by its connection to category theory: C can be viewed as a lax 2-functor between a certain 2-category of pm-nets and a 2-category constructed via the monoidal structure of the non-negative real numbers (see, e.g., [22, 26] for general background on 2-categories). Intuitively, the idea is that the cost structure translates between the gluing composition of couplings and addition on \mathbb{R} , both considered as algebraic operations (hence it behaves like a homomorphism), but it only preserves the structure up to an inequality (this is the "lax-ness" of the homomorphism). As fully developing this perspective would require substantial effort and is not essential for the remainder of the paper, we defer its detailed treatment to future work.

3.3.2. Parameterized Gromov-Wasserstein Distances as Pseudometrics. We are now ready to state our result. Its proof is straightforward, owing to the careful design of our definitions. However, demonstrating that this general theorem applies to specific examples of interest requires additional work, which we present later.

Theorem 1. If C is a symmetric lax homomorphism, then the parameterized Gromov-Wasserstein distance GW_C defines a pseudometric on \mathfrak{N} .

Proof. Non-negativity and symmetry of GW_C are obvious and $\mathsf{C}_{\mathcal{X},\mathcal{X}}(\mathrm{id}_X) = 0$ implies that $\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{X}) = 0$. Finally, triangle inequality follows immediately from the assumption that C respects gluings.

- 3.4. **Specialized metric properties.** In this subsection, we examine special cases of the parameterized GW framework to which Theorem 1 can be applied.
- 3.4.1. Standard Examples of Parameterized Measure Networks and Cost Structures. Throughout this subsection, we consider the following pairs $(\mathfrak{N},\mathsf{C})$ of classes of pm-nets and cost structures, which we refer to as the standard examples:
 - (1) $\mathfrak{N} = \mathfrak{N}_{\nu}$ is the class of pm-nets over a fixed parameter space (Ω, ν) , and C is the cost structure from Proposition 3.16, for some choices of $p, q \in [1, \infty]$.
 - (2) $\mathfrak{N} = \mathfrak{N}_{all}$ is the class of pm-nets whose parameter spaces are endowed with probability measures (i.e., $\mathcal{X} \in \mathfrak{N}$ has $\nu_X(\Omega_X) = 1$), and C is the cost structure from Proposition 3.19, for some choice of $p, q \in [1, \infty]$.

Proposition 3.23. For the standard examples (\mathfrak{N}, C) , optimal couplings exist. That is, the infimum of the associated parameterized GW distance GW_C is realized.

The proof uses the following extension of Proposition 3.15.

Lemma 3.24. Let \mathcal{X} and \mathcal{Y} be pm-nets, and let $p, q \in [1, \infty]$. The function

$$G_{p,q}: \mathcal{C}(\nu_X, \nu_Y) \times \mathcal{C}(\mu_X, \mu_Y) \to \mathbb{R}$$

$$(\xi, \pi) \mapsto \left(\int_{\Omega_X \times \Omega_Y} \operatorname{dis}_p(\pi, \omega_X^s, \omega_Y^t)^q \, \xi(ds \otimes dt) \right)^{1/q}$$

is lower semicontinuous for $1 \le p, q \le \infty$ and continuous if $p, q < \infty$.

Proof. Suppose $p < \infty$ and fix $\epsilon > 0$, $\xi_0 \in \mathcal{C}(\nu_X, \nu_Y)$ and $\pi_0 \in \mathcal{C}(\mu_X, \mu_Y)$. The function

$$G_{p,q}(\bullet, \pi_0) : \mathcal{C}(\nu_X, \nu_Y) \to \mathbb{R}$$

$$\xi \mapsto \left(\int_{\Omega_X \times \Omega_Y} \operatorname{dis}_p(\pi_0, \omega_X^s, \omega_Y^t)^q \, \xi(ds \otimes dt) \right)^{1/q}$$

is continuous in the topology of weak convergence because the cost function $(s,t) \mapsto \operatorname{dis}_p(\pi_0, \omega_X^s, \omega_Y^t)^q$ is continuous, by Proposition 3.15, and bounded, by compactness of $\Omega_X \times \Omega_Y$. Hence, there exists $V_1 \subset \mathcal{C}(\Omega_X, \Omega_Y)$ neighborhood of ξ_0 such that for all $\xi \in V_1$, $|G_{p,q}(\xi, \pi_0) - G_{p,q}(\xi_0, \pi_0)| < \epsilon$.

For the second component of $G_{p,q}$, we claim that there exists $V_2 \subset \mathcal{C}(\mu_X, \mu_Y)$, a neighborhood of π_0 , such that

(10)
$$\left|\operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t}) - \operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s}, \omega_{Y}^{t})\right| < 2\epsilon$$

for all $\pi \in V_2$ and any $s \in \Omega_X$ and $t \in \Omega_Y$. Since $\pi \mapsto \operatorname{dis}_p(\pi, \omega_X^s, \omega_Y^t)$ is continuous by [8, Lemma 2.3], we can find $V_{s,t} \subset \mathcal{C}(\mu_X, \mu_Y)$ and $U_{s,t} \subset \Omega_X \times \Omega_Y$ neighborhoods of π_0 and $(s,t) \in \Omega_X \times \Omega_Y$, respectively, such that for all $\pi \in V_{s,t}$ and $(s',t') \in U_{s,t}$,

$$|\operatorname{dis}_{p}(\pi, \omega_{X}^{s'}, \omega_{Y}^{t'}) - \operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s}, \omega_{Y}^{t})| < \epsilon.$$

By compactness of Ω_X and Ω_Y , there exists a finite cover $\{U_{s_i,t_i}\}_{1\leq i\leq n}$ of $\Omega_X\times\Omega_Y$. We claim that the neighborhood of π_0 defined by $V_2=\bigcap_{1\leq i\leq n}V_{s_i,t_i}$ is the desired set. Fix $\pi\in V_2$. For any $(s,t)\in\Omega_X\times\Omega_Y$, there exists $1\leq k\leq n$ such that $(s,t)\in U_{s_k,t_k}$. Since $\pi\in V_2\subset V_{s_k,t_k}$ and $(s,t)\in U_{s_k,t_k}$, inequality Eqn. (11)

gives $|\operatorname{dis}_p(\pi, \omega_X^s, \omega_Y^t) - \operatorname{dis}_p(\pi_0, \omega_X^{s_k}, \omega_Y^{t_k})| < \epsilon$. By the same reasoning, $\pi_0 \in V_2 \subset V_{s_k, t_k}$ and $(s, t) \in U_{s_k, t_k}$ imply $|\operatorname{dis}_p(\pi_0, \omega_X^s, \omega_Y^t) - \operatorname{dis}_p(\pi_0, \omega_X^s, \omega_Y^{t_k})| < \epsilon$. Hence,

$$\begin{aligned} |\operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t}) - \operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s}, \omega_{Y}^{t})| \\ < |\operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t}) - \operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s_{k}}, \omega_{Y}^{t_{k}})| + |\operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s}, \omega_{Y}^{t}) - \operatorname{dis}_{p}(\pi_{0}, \omega_{X}^{s_{k}}, \omega_{Y}^{t_{k}})| < 2\epsilon. \end{aligned}$$

Now we finish the proof of the continuity of $G_{p,q}$. If $\xi \in V_1$ and $\pi \in V_2$, the reverse triangle inequality of the L^q norm gives

$$|G_{p,q}(\xi,\pi) - G_{p,q}(\xi_0,\pi_0)| \le |G_{p,q}(\xi,\pi) - G_{p,q}(\xi,\pi_0)| + |G_{p,q}(\xi,\pi_0) - G_{p,q}(\xi_0,\pi_0)|$$

$$< \|\operatorname{dis}_p(\pi,\omega_X^s,\omega_Y^t) - \operatorname{dis}_p(\pi_0,\omega_X^s,\omega_Y^t)\|_{L^q(\Omega_X \times \Omega_Y,\xi)} + \epsilon$$

$$< \|2\epsilon\|_{L^q(\Omega_X \times \Omega_Y,\xi)} + \epsilon = 3\epsilon.$$

Finally, we establish lower semicontinuity in the case that p or q is ∞ . For fixed $p < \infty$, standard properties of L^q -norms imply

$$G_{p,\infty}(\xi,\pi) = \sup_{1 \le q < \infty} G_{p,q}(\xi,\pi),$$

so that $G_{p,\infty}$ is a supremum of a family of continuous functions, hence lower semicontinuous. Now fix $q \in [1,\infty]$ and (ξ,π) . Observe that, by Hölder's inequality, $\operatorname{dis}_p(\pi,\omega_X^s,\omega_Y^t)$ is an increasing function of $p \in [1,\infty)$, with limit equal to $\operatorname{dis}_{\infty}(\pi,\omega_X^s,\omega_Y^t)$. By the Monotone Convergence Theorem,

$$G_{\infty,q}(\xi,\pi) = \|\operatorname{dis}_{\infty}(\pi,\omega_X^{\bullet},\omega_Y^{\bullet})\|_{L^q(\Omega_X\otimes\Omega_Y;\xi)} = \sup_{1\leq p<\infty} \|\operatorname{dis}_p(\pi,\omega_X^{\bullet},\omega_Y^{\bullet})\|_{L^q(\Omega_X\otimes\Omega_Y;\xi)} = \sup_{1\leq p<\infty} G_{p,q}(\xi,\pi),$$

so that lower semicontinuity follows.

We will also use the following general result.

Lemma 3.25. Suppose that, for all $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}$, $C_{\mathcal{X},\mathcal{Y}} : \mathcal{C}(\mu_X, \mu_Y) \to \mathbb{R}_{\geq 0}$ is lower semicontinuous in the topology of weak convergence. Then the infimum in the definition of GW_C is always realized.

Proof. By Prokhorov's theorem, the space $C(\mu_X, \mu_Y)$ is compact (see [43, Lemma 1.2] for details). The lower semicontinuous function $C_{\mathcal{X},\mathcal{Y}}$ must therefore achieve its minimum over this set.

Proof of Proposition 3.23. By Proposition 3.25, it suffices to show that C is lower semicontinuous. The conclusion follows immediately from Proposition 3.24 in the setting of $\mathfrak{N}_{\rm all}$, whereas \mathfrak{N}_{ν} requires a bit more work. Indeed, in the latter case, the associated cost structure C is obtained by restricting the first coordinate of $G_{p,q}$ (as defined in Proposition 3.24) to be the identity coupling in $\mathcal{C}(\nu,\nu)$. Lower semicontinuity of $G_{p,q}$ then implies lower semicontinuity of its restriction.

3.4.2. Metric Structure for the Standard Examples. Finally, we show that the standard examples fall under the purview of Theorem 1, so that the associated distances GW_C define pseudometrics. Moreover, in these settings, we can completely characterize the distance zero equivalence classes, using the following definition.

Definition 3.26 (Isomorphisms for Parameterized Measure Networks). Let \mathcal{X} and \mathcal{Z} be pm-nets. A pair of maps (Φ, φ) , with $\Phi : \Omega_Z \to \Omega_X$ and $\varphi : Z \to X$, is called **structure-preserving** if

- (1) Φ and φ are both measure-preserving maps, and
- (2) the pair (Φ, φ) preserves parameterized network kernels, in the sense that

$$\omega_Z^t(z,z') = \omega_X^{\Phi(t)}(\varphi(z),\varphi(z'))$$

holds almost everywhere, with respect to $\nu_Z \otimes \mu_Z \otimes \mu_Z$.

Let $\mathcal Y$ be another pm-net. We say that $\mathcal Z$ is a **stabilization** of $\mathcal X$ and $\mathcal Y$ if there exist structure-preserving maps $\Phi_A:\Omega_Z\to\Omega_A$ and $\varphi_A:Z\to A$ for $A\in\{X,Y\}$. If a stabilization exists, we say that $\mathcal X$ and $\mathcal Y$ are **isomorphic**. In the case that $\mathcal X$ and $\mathcal Y$ have the same parameter space (Ω,ν) , we say that $\mathcal X$ and $\mathcal Y$ are **fixed-parameter isomorphic** if there is a stabilization $\mathcal Z$ with parameter space (Ω,ν) such that the maps Φ_A are identity maps.

Remark 3.27 (Weak Isomorphism for Measure Networks). Let $\mathcal{X} = (X, \mu_X, \omega_X)$ and $\mathcal{Y} = (Y, \mu_Y, \omega_Y)$ be measure networks and let $\overline{\mathcal{X}}$ and $\overline{\mathcal{Y}}$ denote their representations as pm-nets parameterized over the one point space (Ω, ν) (cf. Proposition 3.17). Then \mathcal{X} and \mathcal{Y} are weakly isomorphic (Proposition 2.4) if and only if their pm-net representations $\overline{\mathcal{X}}$ and $\overline{\mathcal{Y}}$ are fixed-parameter isomorphic.

In light of Proposition 3.27, the following generalizes [8, Theorems 2.3 and 2.4], which characterize the pseudometric structure of GW_p on the space of measure networks.

Theorem 2. For the standard examples $(\mathfrak{N}, \mathsf{C})$, the associated parameterized GW distance GW_C defines a pseudometric. Moreover, $\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y}) = 0$ if and only if:

- (1) \mathcal{X} and \mathcal{Y} are isomorphic, in the case $\mathfrak{N} = \mathfrak{N}_{all}$;
- (2) \mathcal{X} and \mathcal{Y} are fixed-parameter isomorphic, in the case $\mathfrak{N} = \mathfrak{N}_{\nu}$.

Proof. By Theorem 1, to show that GW_C is a pseudometric, we need to show that C is symmetric, respects identities, and respects gluings. The first two properties are obvious, so we focus on the last.

Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z} \in \mathfrak{N}$. For ease of notation, let $\overline{\omega}_A^t := \omega_A^t \circ (p_A, p_A)$ for $t \in \Omega_A$ and $A \in \{X, Y, Z\}$, with p_A denoting projection onto the A factor of any product of spaces including A. Let $\pi_{XY} \in \mathcal{C}(\mu_X, \mu_Y)$, $\pi_{YZ} \in \mathcal{C}(\mu_Y, \mu_Z)$, $\pi_{XY} \bullet \pi_{YZ} \in \mathcal{C}(\mu_X, \mu_Y)$ and $\tilde{\pi}$ as in Proposition 3.21. Then for all $r \in \Omega_X$, $s \in \Omega_Y$ and $t \in \Omega_Z$, the triangle inequality of the L^p norm gives

(12)
$$\operatorname{dis}_{p}(\pi_{XY} \bullet \pi_{YZ}, \omega_{X}^{r}, \omega_{Z}^{t}) = \|\overline{\omega}_{X}^{r} - \overline{\omega}_{Z}^{t}\|_{L^{p}(X \times Z; \pi_{XY} \bullet \pi_{YZ})}$$

$$= \|\overline{\omega}_{X}^{r} - \overline{\omega}_{Z}^{t}\|_{L^{p}(X \times Y \times Z; \tilde{\pi})}$$

$$\leq \|\overline{\omega}_{X}^{r} - \overline{\omega}_{Y}^{s}\|_{L^{p}(X \times Y \times Z; \tilde{\pi})} + \|\overline{\omega}_{Y}^{s} - \overline{\omega}_{Z}^{t}\|_{L^{p}(X \times Y \times Z; \tilde{\pi})}$$

$$= \|\overline{\omega}_{X}^{r} - \overline{\omega}_{Y}^{s}\|_{L^{p}(X \times Y; \pi_{XY})} + \|\overline{\omega}_{Y}^{s} - \overline{\omega}_{Z}^{t}\|_{L^{p}(Y \times Z; \pi_{YZ})}$$

$$= \operatorname{dis}_{p}(\pi_{XY}, \omega_{X}^{r}, \omega_{Y}^{s}) + \operatorname{dis}_{p}(\pi_{YZ}, \omega_{Y}^{s}, \omega_{Z}^{t}).$$

Analogously, given $\xi_{XY} \in \mathcal{C}(\nu_X, \nu_Y)$ and $\xi_{YZ} \in \mathcal{C}(\nu_Y, \nu_Z)$, we get

(13)
$$\|\operatorname{dis}_{p}(\pi_{XY} \bullet \pi_{YZ}, \omega_{X}, \omega_{Z})\|_{L^{q}(\Omega_{X} \times \Omega_{Z}; \xi_{XY} \bullet \xi_{YZ})}$$

$$\leq \|\operatorname{dis}_{p}(\pi_{XY}, \omega_{X}, \omega_{Y})\|_{L^{q}(\Omega_{X} \times \Omega_{Y}; \xi_{XY})} + \|\operatorname{dis}_{p}(\pi_{YZ}, \omega_{Y}, \omega_{Z})\|_{L^{q}(\Omega_{Y} \times \Omega_{Z}; \xi_{YZ})}.$$

Recall that in the case of \mathfrak{N}_{ν} , we have a common parameter space (Ω, ν) . If we set r = s = t and $\xi_{XY} = \xi_{YZ} = (\mathrm{id}_{\Omega} \times \mathrm{id}_{\Omega})_{\#}\nu$, Eqn. (13) becomes

$$\begin{aligned} \mathsf{C}_{\mathcal{X}\mathcal{Z}}(\pi_{XY} \bullet \pi_{YZ}) &= \| \operatorname{dis}_{p}(\pi_{XY} \bullet \pi_{YZ}, \omega_{X}^{t}, \omega_{Z}^{t}) \|_{L^{q}(\Omega; \nu)} \\ &\leq \| \operatorname{dis}_{p}(\pi_{XY}, \omega_{X}^{t}, \omega_{Y}^{t}) \|_{L^{q}(\Omega; \nu)} + \| \operatorname{dis}_{p}(\pi_{YZ}, \omega_{Y}^{t}, \omega_{Z}^{t}) \|_{L^{q}(\Omega; \nu)} \\ &= \mathsf{C}_{\mathcal{X}\mathcal{Y}}(\pi_{XY}) + \mathsf{C}_{\mathcal{Y}\mathcal{Z}}(\pi_{YZ}). \end{aligned}$$

In the case of \mathfrak{N}_{all} , we infimize the right side of Eqn. (13) over $\xi_{XY} \in \mathcal{C}(\nu_X, \nu_Y)$ and $\xi_{YZ} \in \mathcal{C}(\nu_Y, \nu_Z)$ to obtain

$$C_{\mathcal{X}\mathcal{Z}}(\pi_{XY} \bullet \pi_{YZ}) = \inf_{\xi \in \mathcal{C}(\nu_X, \nu_Z)} \|\operatorname{dis}_p(\pi_{XY} \bullet \pi_{YZ}, \omega_X, \omega_Z)\|_{L^q(\Omega_X \times \Omega_Z; \xi)}$$

$$\leq \|\operatorname{dis}_p(\pi_{XY} \bullet \pi_{YZ}, \omega_X, \omega_Z)\|_{L^q(\Omega_X \times \Omega_Z; \xi_{XY} \bullet \xi_{YZ})}$$

$$\leq C_{\mathcal{X}\mathcal{V}}(\pi_{XY}) + C_{\mathcal{V}\mathcal{Z}}(\pi_{YZ}).$$

Hence, both cost structures in the standard examples respect gluings.

It remains to characterize the distance zero conditions. First consider the case of $\mathfrak{N}_{\rm all}$. For $\mathcal{X}, \mathcal{Z} \in \mathfrak{N}_{\rm all}$, suppose that there exist a structure-preserving pair of maps $\Phi: \Omega_Z \to \Omega_X$ and $\varphi: Z \to X$. We use these to construct couplings as pushforwards:

$$\xi = (\mathrm{id}_{\Omega_Z} \times \Phi)_{\#} \nu_Z \quad \text{and} \quad \pi = (\mathrm{id}_Z \times \varphi)_{\#} \mu_Z.$$

It is then straightforward to verify, via the change-of-variables formula, that

$$\mathsf{C}_{\mathcal{X},\mathcal{Z}}(\pi) \le \|\mathrm{dis}_p(\pi,\omega_Z,\omega_X)\|_{L^q(\Omega_Z \times \Omega_X; \nu_Z \otimes \nu_X)} = 0,$$

so that $\mathsf{GW}_\mathsf{C}(\mathcal{Z},\mathcal{X}) = 0$. Thus, if \mathcal{X} and \mathcal{Y} are isomorphic pm-nets, it follows by the triangle inequality of GW_C that $\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y}) = 0$. Conversely, suppose that $\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y}) = 0$. By Proposition 3.23, there exist couplings $\xi \in \mathcal{C}(\nu_X, \nu_Y)$ and $\pi \in \mathcal{C}(\mu_X, \mu_Y)$ such that

(14)
$$\|\operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t})\|_{L^{q}(\Omega_{X} \times \Omega_{Y}; \mathcal{E})} = 0.$$

Define $Z := X \times Y$, $\mu_Z := \pi$, $\Omega_Z := \Omega_X \times \Omega_Y$, and $\nu_Z := \xi$. Let $\varphi_A : Z \to A$ and $\Phi_A : \Omega_Z \to \Omega_A$ be the standard projections for $A \in \{X,Y\}$, and define ω_Z^t as the pullback $\varphi_X^* \omega_X^{\Phi(t)}$ for every $t \in \Omega_Z$. We claim that $\mathcal{Z} := (Z, \mu_Z, \Omega_Z, \nu_Z, \omega_Z)$ is a stabilization of \mathcal{X} and \mathcal{Y} . The maps φ_A and Φ_A are measure-preserving

for both $A \in \{X,Y\}$ because of the marginal constraints of π and ξ , respectively. In addition, (Φ_X, φ_X) satisfies condition (Item 2) of Proposition 3.26 for every $(t, z, z') \in \Omega_Z \times Z \times Z$ by definition of ω_Z . Hence, (Φ_X, φ_X) is structure-preserving. The fact that (Φ_Y, φ_Y) is structure-preserving follows from Eqn. (14), as

$$\begin{split} &\int_{\Omega_Z} \left(\int_{Z^2} |\varphi_X^* \omega_X^{\Phi_X(t)}(z,z') - \varphi_Y^* \omega_Y^{\Phi_Y(t)}(z,z')|^p \mu_Z(dz) \mu_Z(dz') \right)^{q/p} \nu_Z(dt) \\ &= \int_{\Omega_X \times \Omega_Y} \left(\int_{(X \times Y)^2} |\omega_X^r(x,x') - \omega_Y^s(y,y')|^p \pi(dx \otimes dy) \pi(dx' \otimes dy') \right)^{q/p} \xi(dr \otimes ds) \\ &= 0. \end{split}$$

The above implies $\omega_Z^t(z,z') = \varphi_X^* \omega_X^{\Phi_X(t)}(z,z') = \varphi_Y^* \omega_Y^{\Phi_Y(t)}(z,z')$ for $(\nu_Z \otimes \mu_Z \otimes \mu_Z)$ -almost every (t,z,z'). Hence, $\mathcal Z$ is a stabilization of $\mathcal X$ and $\mathcal Y$, so $\mathcal X$ and $\mathcal Y$ are isomorphic.

Now consider the case of \mathfrak{N}_{ν} . If \mathcal{X} and \mathcal{Y} are fixed-parameter isomorphic, then constructions similar to those that were used above can be used to show that $\mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})=0$. On the other hand, suppose that $\mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})=0$. The proof that \mathcal{X} and \mathcal{Y} are fixed-parameter isomorphic is similar to the proof above, except that we have $\Omega_X=\Omega_Y=\Omega$ and $\nu_X=\nu_Y=\nu$, so we define $\Omega_Z=\Omega$, $\nu_Z=\nu$, and $\Phi_X=\Phi_Y=\mathrm{id}_{\Omega}$. Proposition 3.23 yields the existence of a coupling π such that $\|\operatorname{dis}_p(\pi,\omega_X^t,\omega_Y^t)\|_{L^q(\Omega;\nu)}=0$, and the above equation becomes

$$\int_{\Omega} \left(\int_{Z^2} |\varphi_X^* \omega_X^{\Phi_X(t)}(z, z') - \varphi_Y^* \omega_Y^{\Phi_Y(t)}(z, z')|^p \mu_Z(dz) \mu_Z(dz') \right)^{q/p} \nu(dt)
= \int_{\Omega} \left(\int_{(X \times Y)^2} |\omega_X^t(x, x') - \omega_Y^t(y, y')|^p \pi(dx \otimes dy) \pi(dx' \otimes dy') \right)^{q/p} \nu(dt) = 0,$$

so that \mathcal{Z} gives the desired stabilization.

4. Comparisons and Approximations for Parametric Gromov-Wasserstein Distances

This section compares our parameterized GW distances for general pm-nets with existing distances proposed in the literature for specific classes of pm-nets. We also examine several approximation strategies for parameterized GW distances.

- 4.1. Realization as a Z-Gromov-Wasserstein Distance. In this subsection, we fix a parameter space (Ω, ν) and consider the class of pm-nets \mathfrak{N}_{ν} from Proposition 3.11. We also fix the cost structure C from Proposition 3.16, for some choices of $p, q \in [1, \infty]$. As noted in Proposition 3.18, the associated parameterized GW distance is related to the Z-Gromov-Wasserstein distance [4], and we now make this connection precise.
- 4.1.1. Z-Gromov-Wasserstein Distances. We take a brief detour to recall the notion of Z-GW distance. Fix a complete metric space (Z, d_Z) and consider a structure of the form $\mathcal{X} = (X, \mu_X, \omega_X)$, where (X, μ_X) is a (say, compact) Polish measure space and $\omega_X : X \times X \to Z$ is a Z-valued kernel, which we assume to be bounded (more generally, [4] allows p-integrable kernels, for a given p). Such a structure is referred to in [4] as a Z-network. For $p \in [1, \infty]$, the p-Gromov Wasserstein distance between two Z-networks \mathcal{X} and \mathcal{Y} is

$$\mathsf{GW}^Z_p(\mathcal{X},\mathcal{Y}) \coloneqq \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \|d_Z \circ (\omega_X,\omega_Y)\|_{L^p((X \times Y)^2;\pi \otimes \pi)},$$

where the function in the norm is given by

$$d_Z \circ (\omega_X, \omega_Y) : X \times Y \times X \times Y \to \mathbb{R}$$
$$(x, y, x', y') \mapsto d_Z (\omega_X(x, x), \omega_Y(y, y')).$$

For context, note that when (Z, d_Z) is the real line equipped with its standard metric, the Z-GW framework reduces to the classical GW distance on measure networks.

4.1.2. Parameterized GW Distances as Z-GW Distances. We now return to the setting of interest, to understand the parameterized GW distance on \mathfrak{N}_{ν} as a Z-GW distance. To this end, set (Z, d_Z) to be $(L^q(\Omega; \nu), \|\cdot\|_{L^q(\Omega; \nu)})$, where we are abusing notation and using $\|\cdot\|_{L^q(\Omega; \nu)}$ as a placeholder for its induced metric. Let $\mathcal{X} = (X, \mu_X, \Omega, \nu, \omega_X) \in \mathfrak{N}_{\nu}$. This can be understood as a Z-network $\overline{\mathcal{X}} = (\overline{X}, \overline{\mu}_X, \overline{\omega}_X)$, where $\overline{X} = X, \overline{\mu}_X = \mu_X$, and $\overline{\omega}_X$ is the Z-valued kernel defined, for all $x, x' \in X$, $t \in \Omega_X$, $\omega_X^t(x, x') \in \mathbb{R}$ by

$$\overline{\omega}_X(x,x') := (t \mapsto \omega_X^t(x,x')) \in L^q(\Omega;\nu).$$

This leads to the following comparison result for parameterized GW distance and Z-GW distance, where notation is the same as above:

Proposition 4.1. For any $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}_{\nu}$, if $p \square q$ then $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}) \square \mathsf{GW}_{p}^{Z}(\overline{\mathcal{X}}, \overline{\mathcal{Y}})$, where $\square \in \{\leq, \geq, =\}$.

Proof. Assuming $p \leq q$, we have

$$\begin{split} 2 \cdot \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}) &= \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \| \mathrm{dis}_p(\pi, \omega_X, \omega_Y) \|_{L^q(\Omega, \nu)} \\ &= \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \| \| \omega_X \circ (p_X, p_X) - \omega_Y \circ (p_Y, p_Y) \|_{L^p((X \times Y)^2; \pi \otimes \pi)} \|_{L^q(\Omega, \nu)} \\ &\leq \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \| \| \omega_X \circ (p_X, p_X) - \omega_Y \circ (p_Y, p_Y) \|_{L^q(\Omega, \nu)} \|_{L^p((X \times Y)^2; \pi \otimes \pi)} \\ &= \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \| d_Z \circ (\overline{\omega}_X, \overline{\omega}_Y) \|_{L^p((X \times Y)^2; \pi \otimes \pi)} \\ &= 2 \cdot \mathsf{GW}_p^2(\overline{\mathcal{X}}, \overline{\mathcal{Y}}), \end{split}$$

where we have applied a generalized version of Minkowski's inequality [2, Proposition 1.3]. The case $p \ge q$ follows by a similar argument and these together imply the p = q case.

- 4.1.3. Followup Remarks. We conclude this subsection with some remarks on the connection described above.
- (1) Theoretical properties of GW_p^Z are derived for general choices of (Z, d_Z) in [4]. These properties therefore apply to GW_C on \mathfrak{N}_ν in the case where p=q: the induced metric space (after modding out by fixed-parameter isomorphism; see Theorem 2) is complete, contractible and geodesic, for example.
- (2) One could instead define a cost structure C on \mathfrak{N}_{ν} by integrating in a different order; that is, for $p, q < \infty$,

$$\mathsf{C}_{\mathcal{X},\mathcal{Y}}(\pi) \coloneqq \frac{1}{2} \left(\int_{(X \times Y)^2} \left(\int_{\Omega} |\omega_X(x,x') - \omega_Y(y,y')|^q \pi(dx \times dy) \pi(dx' \times dy') \right)^{p/q} \nu(dt) \right)^{1/p}$$

In this case, one has $\mathsf{GW}_\mathsf{C} = \mathsf{GW}_p^Z$ (with (Z, d_Z) defined as above) for all choices of p, q. The rationale for our particular choice of cost structure on \mathfrak{N}_ν is that it more naturally generalizes to the cost structure on the collection of all pm-nets $\mathfrak{N}_{\mathrm{all}}$ defined in Proposition 3.19.

- (3) The parameterized GW distance defined on \mathfrak{N}_{all} via Proposition 3.19 is not realized as an instance of a Z-GW distance (in any obvious way).
- 4.2. Metrics for Spaces Parameterized by Real Numbers. The pm-nets described in Proposition 3.4 through Proposition 3.7 are all defined over a parameter space (Ω, ν) consisting of a compact set of real numbers, endowed with some measure. In this subsection, we compare our approach to existing metrics in the literature, which were designed to compare specialized pm-net structures.
- 4.2.1. Time-Varying Metric Spaces. Fix a parameter space (Ω, ν) consisting of a compact interval of real numbers endowed with some probability measure (e.g., $\Omega = [0, 1]$, ν is Lebesgue measure) and consider the class of pm-nets $\mathfrak{N}_{\text{tvm}}$ consisting of time-varying metric measure spaces (Proposition 3.4); that is, an element of $\mathfrak{N}_{\text{tvm}}$ is a pm-net of the form $\mathcal{X} = (X, \mu_X, \Omega, \nu, d_X)$, where, for each $t \in \Omega$, d_X^t is a metric inducing the given topology on X. Observe that this is a proper subclass of \mathfrak{N}_{ν} , due to the constraint that the kernel is a metric for each parameter value. We now provide examples of distances between objects of $\mathfrak{N}_{\text{tvm}}$ which have been previously introduced in the literature, and compare them to our approach.

Example 4.2 (Sturm's Distance). In [42], Sturm introduced a distance on a more general class of objects of the form $\mathcal{X} = (X, (\mu_X^t)_{t \in \Omega}, (d_X^t)_{t \in \Omega})$, where

• X is a Polish space;

- Each d_X^t is a metric generating the topology of X; Sturm also assumed that each d_X^t is geodesic, but this property is not important for our discussion;
- μ_X^t is a time-varying family of Borel measures which are absolutely continuous with respect to some reference probability measure μ_X .

We refer to such a structure as a generalized time-varying metric measure space. Clearly, this concepts restricts to our notion of time-varying metric space by imposing the constraint that $\mu_X^t = \mu_X$ for all t.

Given two generalized time-varying metric measure spaces \mathcal{X} and \mathcal{Y} , Sturm defines a distance between them as (15)

$$\mathsf{D}_{\mathrm{St}}(\mathcal{X},\mathcal{Y}) \coloneqq \inf_{d_{X\sqcup Y},\pi} \left(\int_{\Omega} \int_{X\times Y} d_{X\sqcup Y}^t(x,y)^2 \pi(dx\otimes dy) \nu(dt) \right)^{1/2} + \int_{\Omega} \int_{X\times Y} |f_X^t(x) - f_Y^t(y)| \pi(dx\otimes dy) \nu(dt),$$

where the infimum is over

- $\pi \in \mathcal{C}(\mu_X, \mu_Y)$, i.e., couplings of the reference measures,
- $(d^t_{X\sqcup Y})_{t\in\Omega}$ is a parameterized family of *metric couplings*, or metrics $d^t_{X\sqcup Y}$ on the disjoint union which restrict to d^t_X and d^t_Y , respectively, on the appropriate subsets, for almost every $t\in\Omega$, and
- $(f_X^t)_{t\in\Omega}$ is chosen so that $\mu_X^t = e^{f_X^t} \mu_X$ for all t, and similarly for $(f_Y^t)_{t\in\Omega}$.

Restricting to the subspace \mathfrak{N}_{tvm} , the second term of Eqn. (15) vanishes and Eqn. (15) further simplifies to

$$\mathsf{D}_{\mathrm{St}}(\mathcal{X},\mathcal{Y}) = \inf_{d_{X \sqcup Y},\pi} \left(\int_{\Omega} \int_{X \times Y} d_{X \sqcup Y}^t(x,y)^2 \pi(dx \otimes dy) \nu(dt) \right)^{1/2} = \left\| \inf_{d_{X \sqcup Y}} \mathsf{W}_2^{d_{X \sqcup Y}^t}(\mu_X,\mu_Y) \right\|_{L^2(\Omega;\nu)},$$

that is, an integrated version of Sturm's well-known L^2 -transportation distance on the space of metric measure spaces [41] (we abuse notation here, and consider μ_X and μ_Y as measures on the disjoint union, in the obvious way). The transportation distance is known to upper bound (classical) GW distance, and the proof in [29] can be extended to show

$$\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y}) \leq \mathsf{D}_{\mathrm{St}}(\mathcal{X},\mathcal{Y}),$$

where the cost structure C comes from Proposition 3.16, with p = q = 2.

Example 4.3 (Integrated Gromov-Hausdorff Distance). A simple metric on \mathfrak{N}_{tvm} , which has primarily been used in the topological data analysis (TDA) literature [32,52], is the **integrated Gromov-Hausdorff distance**, defined as

$$\mathsf{IGH}(\mathcal{X},\mathcal{Y}) \coloneqq \int_{\Omega} \mathsf{GH}\big((X,d_X^t),(Y,d_Y^t)\big) \nu(dt),$$

where GH is the standard Gromov-Hausdorff distance between compact metric spaces. Arguments presented in [29] can be adapted to show that

$$GW_{C}(\mathcal{X}, \mathcal{Y}) < IGH(\mathcal{X}, \mathcal{Y}),$$

where C is as in Proposition 3.16 with $p = \infty$ and q = 1.

Example 4.4 (Slack Interleaving Distance). An alternative metric on $\mathfrak{N}_{\text{tvm}}$, based on constructions used in TDA, was in introduced in [23]. We review the details here. Given continuous functions $f, g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and $\lambda \geq 0$, the λ -slack interleaving distance between the functions is defined by

$$d_{\lambda}(f_1, f_2) := \inf \left\{ \varepsilon \in [0, \infty] \mid \forall t \in \mathbb{R}, \min_{s \in [t-\varepsilon, t+\varepsilon]} f_i(s) \le f_j(t) + \lambda \varepsilon, i, j = 1, 2 \right\}.$$

For $p \in [1, \infty]$, the (p, λ) -Gromov-Wasserstein distance between $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}_{\text{tym}}$ is defined as

$$\inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \left\| d_{\lambda} \circ (d_X \times d_Y) \right\|_{L^p(\pi \otimes \pi)}.$$

Taking the cost structure

$$\mathsf{C}_{\mathcal{X},\mathcal{Y}}(\pi) = \left\| d_{\lambda} \circ (d_X \times d_Y) \right\|_{L^p(\pi \otimes \pi)},$$

we see that the (p, λ) -GW distance is an instance of a parameterized GW distance. We note that it was shown in [4, Proposition 3.13] that the (p, λ) -GW distance can also be realized as a Z-GW distance.

4.2.2. Heat Kernels. We consider classes of pm-nets induced by heat kernels. For the rest of this subsection, we temporarily abuse terminology and drop the assumptions that the parameter space in a pm-net is compact, and that the measure on this space is a probability measure. This is only for the sake of avoiding technical details; in this setting, the basic definitions introduced in the paper related to parameterized GW distances are still valid, but certain theoretical properties are no longer guaranteed. As we are only concerned here with estimates of the parameterized GW distances, this issue is not a concern.

First, consider the class \mathfrak{N}_{HK} of pm-nets of the form $\mathcal{X} = (X, \mu_X, \Omega, \nu, \omega_X)$, where

- $\Omega = \mathbb{R}_{>0} := (0, +\infty)$ (with its usual topology, hence a non-compact space) and ν is Lebesgue measure (not a probability measure),
- X is a compact Riemannian manifold endowed with normalized Riemannian volume μ_X , and
- $\omega_X^t: X \times X \to \mathbb{R}$ is the normalized heat kernel of X (with respect to the given Riemannian metric), where normalized means that $\lim_{t\to\infty} \omega^t(x,x') = 1$ for all $x,x'\in X$.

Example 4.5 (Spectral Gromov-Wasserstein distance). Mémoli defined a distance $\mathsf{GW}_p^{\mathrm{spec}}$ to compare two Riemannian manifolds using their heat kernels [28, 30]; in our terminology, this is a distance on the class $\mathfrak{N}_{\mathrm{HK}}$ defined above. Here, we recall the definition of Mémoli's distance and rewrite it as a parametrized GW distance. To recall the original definition, we first define an auxiliary function $c(t) \coloneqq e^{-(t+t^{-1})}$. Given $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}_{\mathrm{HK}}$, we define a cost

$$\Gamma_{X,Y,t}^{\text{spec}}(x,y,x',y') := |\omega_X^t(x,x') - \omega_Y^t(y,y')|$$

and

$$\mathsf{GW}_p^{\mathrm{spec}}(X,Y) \coloneqq \inf_{\pi \in \mathcal{C}(\mathrm{Vol}_X,\mathrm{Vol}_Y)} \sup_{t>0} c^2(t) \cdot \|\Gamma_{X,Y,t}^{\mathrm{spec}}\|_{L^p((X \times Y)^2; \pi \otimes \pi)},$$

where $p \in [1, \infty]$.

Given $p, q \in [1, \infty]$, we define a new cost structure C by

$$\mathsf{C}_{\mathcal{X},\mathcal{Y}}(\pi) \coloneqq \left\| c^2(t) \cdot \| \Gamma_{X,Y,t}^{\mathrm{spec}} \|_{L^p((X \times Y)^2, \pi \otimes \pi)} \right\|_{L^q(\Omega,\nu)}$$

for some $\pi \in \mathcal{C}(\operatorname{Vol}_X, \operatorname{Vol}_Y)$. Then $\operatorname{\mathsf{GW}}_p^{\operatorname{spec}}(X,Y) = \operatorname{\mathsf{GW}}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})$ when $q = \infty$. We note that a similar interpretation of $\operatorname{\mathsf{GW}}_p^{\operatorname{spec}}$ as a Z-GW distance was provided in [4, Proposition 3.12], but that this result only gives an upper bound due to issues similar to those that arose in the proof of Proposition 4.1.

Next we consider the class \mathfrak{N}_{GHK} of graph heat kernels $\mathcal{X} = (X, \mu_X, \Omega, \nu, \omega_X)$, where (Ω, ν) is as above, X is the set of nodes of a graph, endowed with some probability measure μ_X , and ω_X^t is the graph heat kernel at scale t.

Example 4.6 (Graph Heat Kernels). Chowdhury and Needham [9] studied GW distances between graph heat kernels at a fixed scale parameter t. Their approach contrasts with that of the present paper, which aims to incorporate information across all scales when comparing pm-nets. In [9], the scale parameter was treated as a tunable hyperparameter. From a high-level perspective, this can be viewed as computing GW_C while allowing additional flexibility in the choice of the measure ν ; for example, permitting it to collapse to a Dirac measure at a single scale t. This viewpoint aligns with the feature selection approach described in Sec. 5.5.

- 4.3. **Approximation by Wasserstein Distance.** In this subsection, we derive a lower bound on the parameterized GW distance by a certain Wasserstein distance.
- 4.3.1. Wasserstein Distance Over the GW Space. We begin by setting up necessary preliminary concepts. Let \mathfrak{M} denote the class of measure networks (by Proposition 3.3, \mathfrak{M} is equivalent to the class \mathfrak{N}_{ν} , where (Ω, ν) is a one-point space, but we introduce this additional notation for bookkeeping purposes). For the rest of this subsection, we let \sim denote the weak isomorphism equivalence relation on \mathfrak{M} (Proposition 2.4); for a measure network \mathcal{X} , we let $[\mathcal{X}]$ denote its equivalence class, and we let $[\mathcal{M}]$ denote the set of equivalence classes, that is, measure networks, considered up to weak isomorphism.

Given $p \in [1, \infty]$, let GW_p be the p-GW distance on \mathfrak{M} . Considering elements of \mathfrak{M} up to weak isomorphism, GW_p induces a complete and separable metric on \mathfrak{M}_{\sim} : completeness follows essentially from [43, Theorem 5.8], which enforces an additional symmetry condition on the kernels that is not intrinsically necessary for the proof; separability is a standard argument, but a precise reference is [4, Proposition 4.8]. We abuse

notation and continue to denote the induced metric by GW_p . Therefore $(\mathfrak{M}_{\sim}, \mathsf{GW}_p)$ is a Polish metric space, and measure theory on it is sufficiently well-behaved to consider the Wasserstein distance between distributions defined over it. Throughout the rest of this subsection, we use $\mathsf{W}_q^{\mathsf{GW}_p}$ to denote the q-Wasserstein distance on $(\mathfrak{M}_{\sim}, \mathsf{GW}_p)$ (see Sec. 2.2).

4.3.2. Lower Bound on Parameterized GW Distance. Now consider an arbitrary pm-net $\mathcal{X} \in \mathfrak{N}_{all}$. For any $t \in \Omega_X$, the triplet $\mathcal{X}_t := (X, \mu_X, \omega_X^t)$ is a measure network. We define the map

$$m_{\mathcal{X}}: \Omega_X \to \mathfrak{M}_{\sim}$$

$$t \mapsto [\mathcal{X}_t],$$

and the associated pushforward measure $\overline{\nu}_X := (m_{\mathcal{X}})_{\#} \nu_X$ on \mathfrak{M}_{\sim} . For this to be well-defined, we require $m_{\mathcal{X}}$ to be measurable, which is verified by the following Proposition 4.7.

Lemma 4.7. The map $m_{\mathcal{X}}$ is continuous.

Proof. We factor $m_{\mathcal{X}}$ as $m_{\mathcal{X}} = q \circ \hat{m}_{\mathcal{X}}$, where $\hat{m}_{\mathcal{X}} : \Omega_X \to \mathfrak{M}$ is the map $t \mapsto \mathcal{X}_t$ and $q : \mathfrak{M} \to \mathfrak{M}/\sim$ is the quotient map. It suffices to prove that $\hat{m}_{\mathcal{X}}$ is continuous, with respect to the topology induced by the pseudometric GW_p . Let us metrize Ω_X via some choice $d_{\Omega_X,p}$; the particular metric is irrelevant, and we only assume that it induces the given Polish topology on Ω_X . Let $\epsilon > 0$. As parameterized network kernels are assumed to be L^{∞} continuous, there exists $\delta > 0$ such that $d_{\Omega_X,p}(s,t) < \delta$ implies $\|\omega_X^s - \omega_X^t\|_{L^{\infty}(X \times X; \mu_X \otimes \mu_X)} < \epsilon$. We have

$$\mathsf{GW}_p(\mathcal{X}_s, \mathcal{X}_t) \leq \mathsf{GW}_\infty(\mathcal{X}_s, \mathcal{X}_t) \leq \|\omega_X^s - \omega_X^t\|_{L^\infty(X \times X; \mu_X \otimes \mu_X)}.$$

Here, the first inequality follows from [29, Theorem 5.1 (h)]; that result proves the inequality in the case of metric measure spaces, but the proof is based on properties of L^p -norms and applies to more general measure networks. The second inequality follows by considering the ∞ -distortion of the identity coupling of μ_X with itself.

We are now able to state the main result of this subsection.

Theorem 3. Let $p, q \in [1, \infty]$ and let C be the cost structure from Proposition 3.19. For $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}_{\text{all}}$, we have

$$\mathsf{W}_q^{\mathsf{GW}_p}(\overline{\nu}_X, \overline{\nu}_Y) \leq \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}).$$

The proof will use the following general measure theory result.

Lemma 4.8. Let (X, μ_X) and (Y, μ_Y) be Polish probability spaces, X' and Y' Polish spaces, and $f: X \to X'$ and $g: Y \to Y'$ measurable maps. Then

$$C(f_{\#}\mu_X, g_{\#}\mu_Y) = \{ (f \times g)_{\#}\pi \mid \pi \in C(\mu_X, \mu_Y) \},$$

where $f \times g : X \times Y \to X' \times Y'$ is the product map $(x, y) \mapsto (f(x), g(y))$.

Proof. One inclusion is straightforward: it is easy to show that a measure of the form $(f \times g)_{\#}\pi$ is a coupling of $f_{\#}\mu_X$ and $g_{\#}\mu_Y$, for $\pi \in \mathcal{C}(\mu_X, \mu_Y)$.

To prove the remaining inclusion, let $\{\mu_X(\cdot \mid x')\}_{x' \in X'}$ denote the disintegration of μ_X with respect to f. That is, for each $x' \in X'$, $\mu_X(\cdot \mid x')$ is a Borel probability measure on X satisfying

$$\mu_X(A) = \int_{X'} \mu_X(A \mid x') f_{\#} \mu_X(dx')$$
 and $f_{\#} \mu_X(\cdot \mid x') = \delta_{x'}$,

where $\delta_{x'}$ denotes the Dirac measure on X' (see, e.g., [1, Section 5.3]). Likewise, let $\{\mu_Y(\cdot \mid y')\}_{y' \in Y'}$ denote the disintegration of μ_Y with respect to g.

Given $\xi \in \mathcal{C}(f_{\#}\mu_X, g_{\#}\mu_Y)$, define a measure π on $X \times Y$ for a product Borel set $A \times B$ as

$$\pi(A \times B) = \int_{X' \times Y'} \mu_X(A \mid x') \mu_Y(B \mid y') \xi(dx' \otimes dy').$$

We claim that $\pi \in \mathcal{C}(\mu_X, \mu_Y)$ and that $(f \times g)_{\#}\pi = \xi$. The first point follows by checking marginals:

$$\pi(A \times Y) = \int_{X' \times Y'} \mu_X(A \mid x') \mu_Y(Y \mid y') \xi(dx' \otimes dy')$$

$$= \int_{X' \times Y'} \mu_X(A \mid x') \xi(dx' \otimes dy')$$

$$= \int_{X'} \mu_X(A \mid x') f_\# \mu_X(dx') = \mu_X(A),$$

where we have used that $\mu_Y(\cdot \mid y')$ is a probability measure and the marginal condition on ξ . The fact that $\pi(X \times B) = \mu_Y(B)$ follows similarly.

Finally, we prove that $(f \times g)_{\#}\pi = \xi$. We will use the following identity, which holds for any $x' \in X'$, and which follows by the property that $f_{\#}\mu_X(\cdot \mid x') = \delta_{x'}$:

(16)
$$\int_{f^{-1}(A')} \mu_X(dx \mid x') = 1_{A'}(x'),$$

where $1_{A'}$ denotes the indicator function for a Borel set A'. Proceeding with the proof, let $A' \times B'$ be a product Borel set in $X' \times Y'$. Then

$$(f \times g)_{\#}\pi(A' \times B') = \int_{A' \times B'} (f \times g)_{\#}\pi(dx' \otimes dy')$$

$$= \int_{f^{-1}(A') \times g^{-1}(B')} \pi(dx \otimes dy)$$

$$= \int_{X' \times Y'} \int_{g^{-1}(B')} \int_{f^{-1}(A')} \mu_{X}(dx \mid x') \mu_{Y}(dy \mid y') \xi(dx' \otimes dy')$$

$$= \int_{X' \times Y'} \int_{g^{-1}(B')} 1_{A'}(x') \mu_{Y}(dy \mid y') \xi(dx' \otimes dy')$$

$$= \int_{X' \times Y'} 1_{A'}(x') 1_{B'}(y') \xi(dx' \otimes dy')$$

$$= \int_{A' \times B'} \xi(dx' \otimes dy') = \xi(A' \times B'),$$

where Eqn. (17) follows from the definition of π and Fubini's Theorem, and Eqn. (18) and Eqn. (19) both follow by applying Eqn. (16) to the iterated integrals.

Proof of Theorem 3. First consider the $q < \infty$ case $(p \in [1, \infty])$ is arbitrary). For any $\pi \in \mathcal{C}(\mu_X, \mu_Y)$, we have

$$W_{q}^{\mathsf{GW}_{p}}(\overline{\nu}_{X}, \overline{\nu}_{Y}) = \inf_{\overline{\xi} \in \mathcal{C}(\overline{\nu}_{X}, \overline{\nu}_{Y})} \left(\int_{\mathfrak{M} \times \mathfrak{M}} \mathsf{GW}_{p}([\mathcal{X}], [\mathcal{Y}])^{q} \, \overline{\xi}(d[\mathcal{X}] \otimes d[\mathcal{Y}]) \right)^{1/q}$$

$$= \inf_{\xi \in \mathcal{C}(\nu_{X}, \nu_{Y})} \left(\int_{\mathfrak{M} \times \mathfrak{M}} \mathsf{GW}_{p}([\mathcal{X}], [\mathcal{Y}])^{q} \, (m_{\mathcal{X}} \times m_{\mathcal{Y}})_{\#} \xi(d[\mathcal{X}] \otimes d[\mathcal{Y}]) \right)^{1/q}$$

$$= \inf_{\xi \in \mathcal{C}(\nu_{X}, \nu_{Y})} \left(\int_{\Omega_{X} \times \Omega_{Y}} \mathsf{GW}_{p}(\mathcal{X}_{s}, \mathcal{Y}_{t})^{q} \, \xi(ds \otimes dt) \right)^{1/q}$$

$$\leq \inf_{\xi \in \mathcal{C}(\nu_{X}, \nu_{Y})} \left(\int_{\Omega_{X} \times \Omega_{Y}} \frac{1}{2} \operatorname{dis}_{p}(\pi, \omega_{X}^{s}, \omega_{Y}^{t})^{q} \, \xi(ds \otimes dt) \right)^{1/q},$$

where Eqn. (20) follows from Proposition 4.8 and Eqn. (21) follows from a change of variables, and the fact that GW_p is invariant over equivalence classes. The result then follows by infimizing over $\pi \in \mathcal{C}(\mu_X, \mu_Y)$ on the right-hand side.

Finally, consider the $q=\infty$ case. The work above shows that for any $\pi\in\mathcal{C}(\mu_X,\mu_Y)$ and any $\xi\in\mathcal{C}(\nu_X,\nu_Y)$,

$$\mathsf{W}^{\mathsf{GW}_p}_{\infty}(\overline{\nu}_X,\overline{\nu}_Y) = \lim_{q \to \infty} \mathsf{W}^{\mathsf{GW}_p}_q(\overline{\nu}_X,\overline{\nu}_Y) \leq \lim_{q \to \infty} \frac{1}{2} \|\mathrm{dis}_p(\pi,\omega_X,\omega_Y)\|_{L^q(\xi)} = \frac{1}{2} \|\mathrm{dis}_p(\pi,\omega_X,\omega_Y)\|_{L^\infty(\xi)},$$

where the first equality follows from [15, Proposition 3]. The result once again follows by infimizing the right-hand side. \Box

4.3.3. Lower Bound by Weight Distribution. We now utilize Theorem 3 to give further lower bounds on parameterized GW distances. These are given in terms of a new invariant of pm-nets, defined below.

Definition 4.9 (Weight Distribution). Consider the function (see [8, 29])

(22)
$$\mathfrak{M}_{\sim} \to \mathcal{P}(\mathbb{R})$$

$$[\mathcal{Y}] \mapsto (\omega_Y)_{\#}(\mu_Y \otimes \mu_Y).$$

Let \mathcal{X} be a pm-net and let $\overline{\nu}_X \in \mathcal{P}(\mathbb{R})$ be as in Theorem 3 (Sec. 4.3.2). The weight distribution of \mathcal{X} is the measure $\Delta_{\mathcal{X}} \in \mathcal{P}(\mathcal{P}(\mathbb{R}))$ given by the pushforward of $\overline{\nu}_X$ by the map Eqn. (22).

In Proposition 4.9, the weight distribution defines an invariant of pm-nets which takes the form of a probability distribution over the space of probability distributions, denoted as $\mathcal{P}(\mathcal{P}(\mathbb{R}))$. Such an invariant may appear to be rather unwieldy, but we show below that it is stable and, moreover, that it is easy to compute in certain circumstances.

Corollary 4.10. The weight distribution is a stable invariant of pm-nets. That is, let \mathcal{X} and \mathcal{Y} be pm-nets, let $\overline{\nu}_X$ and $\overline{\nu}_Y$ be as in Theorem 3, and let C be the cost structure from Proposition 3.19, for some $p, q \in [1, \infty]$. Then

$$\mathsf{W}_q^{\mathsf{W}_p^{\mathcal{P}(\mathbb{R})}}(\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}) \leq 2 \cdot \mathsf{GW}_\mathsf{C}(\mathcal{X}, \mathcal{Y}).$$

In Eqn. (23), the superscript in $W_q^{W_p^{\mathcal{P}(\mathbb{R})}}$ indicates that the Wasserstein distance W_q is taken over the metric space $(\mathcal{P}(\mathbb{R}), W_p^{\mathcal{P}(\mathbb{R})})$; that is, it is a Wasserstein distance on the Wasserstein space. The proof will use a lemma, whose proof follows directly from (the easy direction of) Proposition 4.8.

Lemma 4.11. Let $f: X \to Y$ be a k-Lipschitz map between Polish metric spaces (X, d_X) and (Y, d_Y) . Then the pushforward $f_\#: \mathcal{P}(X) \to \mathcal{P}(Y)$ is a k-Lipschitz map with respect to $\mathsf{W}_q^{d_X}$ and $\mathsf{W}_q^{d_Y}$.

Proof of Proposition 4.10. It is shown in [8, Theorem 3.1] that the map Eqn. (22) is 2-Lipschitz, with respect to GW_p and $\mathsf{W}_p^{\mathcal{P}(\mathbb{R})}$. By Proposition 4.11, the associated pushforward map $\mathcal{P}(\mathfrak{M}_{\sim}) \to \mathcal{P}(\mathcal{P}(\mathbb{R}))$ is 2-Lipschitz with respect to $\mathsf{W}_q^{\mathsf{GW}_p}$ and $\mathsf{W}_q^{\mathsf{W}_p^{\mathcal{P}(\mathbb{R})}}$. From Theorem 3, we have

$$\mathsf{W}_q^{\mathsf{W}_p^{\mathcal{P}(\mathbb{R})}}(\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}}) \leq 2 \cdot \mathsf{W}_q^{\mathsf{GW}_p}(\overline{\nu}_X, \overline{\nu}_Y) \leq 2 \cdot \mathsf{GW}_\mathsf{C}(\mathcal{X}, \mathcal{Y}).$$

Remark 4.12 (Computation). For pm-nets \mathcal{X} and \mathcal{Y} defined over finite sets X and Y and finite parameter spaces Ω_X and Ω_Y , respectively, the right-hand-side of Eqn. (23) involves computing Wasserstein distance between finitely-supported distributions on the Wasserstein space $\left(\mathcal{P}\left(\mathcal{P}\left(\mathbb{R}\right)\right), \mathsf{W}_p^{\mathcal{P}(\mathbb{R})}\right)$. It is therefore polynomial-time computable (in the magnitudes of X, Y, Ω_X, Ω_Y). This gives a tractable lower estimate of the parameterized GW distance.

4.3.4. Weight Distributions for Random Graph Models. We now specialize the results of Sec. 4.3.3 to the setting of random graphs. Consider a random graph model \mathcal{X} , in the sense of Proposition 3.8, and suppose that |X| = n and that μ_X is uniform. For the sake of simplifying the discussion, choose an ordering of X and consider the kernels ω_X^t arising in this model as symmetric, binary, $n \times n$ matrices with zeros on their diagonals (recall that we use adjacency kernels in this example). In this case, the distribution $\overline{\nu}_X \in \mathcal{P}(\mathfrak{M}/\sim)$ can be considered as a discrete probability measure on the finite set of such matrices (the set has cardinality n(n-1)/2). Given such a matrix ω_X^t , the distribution $(\omega_X^t)_{\#}(\mu_X \otimes \mu_X)$ arising from Eqn. (22) is supported on $\{0,1\}$. Indeed, the weight on the point 1 is exactly k/n^2 , where k is the number of non-zero entries appearing in the matrix; that is, k is twice the number of edges appearing in the graph t. In light of this interpretation, the information contained in $(\omega_X^t)_{\#}(\mu_X \otimes \mu_X)$ is exactly the total number of edges in the graph t, and we refer to the weight distribution Δ_X in this case as the distribution of total edges.

The discussion above immediately leads to the following specialization of Proposition 4.10, stated here somewhat informally.

Corollary 4.13. The distribution of total edges is a stable invariant of a random graph model.

Example 4.14 (Erdős-Rényi Model). For a concrete example, consider an Erdős-Rényi random graph model \mathcal{X} on n nodes with probability $\rho \in [0,1]$ that any pair of nodes is connected. A straightforward calculation shows that the distribution of total edges is given by

$$\Delta_{\mathcal{X}} = \sum_{k=0}^{N} {N \choose k} \rho^k (1-\rho)^{N-k} \delta_{\frac{k}{n^2}},$$

where N = n(n-1)/2, and $\delta_{\frac{k}{n^2}}$ is a shorthand for the Dirac mass on $\mathcal{P}(\mathbb{R})$ located at the distribution supported on $\{0,1\}$ with weight $k/(n^2)$ at 1 and $1-k/(n^2)$ at 0. This recovers the well-known fact that the total number of edges in an Erdős-Rényi graph is binomially distributed.

Remark 4.15 (Computation for Random Graph Models). When \mathcal{X} is a random graph model, in particular, with each ω_X^t taking values only in $\{0,1\}$, the distribution of total edges $\Delta_{\mathcal{X}}$ is especially easy to work with from a computational perspective. In this setting, the cost matrix used in the computation of $W_q^{W_1^{\mathcal{P}(\mathbb{R})}}(\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}})$ involves 1-Wasserstein distances between the measures $(\omega_X^t)_\#(\mu_X \otimes \mu_X)$ and $(\omega_Y^t)_\#(\mu_Y \otimes \mu_Y)$. Each of these measures is supported on $\{0,1\}$, and it is not hard to show (via the semi-explicit formula for Wasserstein distances on the real line [48, Remark 2.19]) that this is given by

$$|(\omega_X^t)_{\#}(\mu_X \otimes \mu_X)(\{1\}) - (\omega_Y^t)_{\#}(\mu_Y \otimes \mu_Y))(\{1\})|.$$

This makes the lower bound of Proposition 4.10 (or Proposition 4.13) very efficient to work with in the random graph setting; see Sec. 5.3.2 for a numerical example.

4.4. Sample Approximation. In this subsection, we consider random graphs and random metric spaces, as introduced in Proposition 3.8 and Proposition 3.9. As previously noted, one typically does not have access to the full distribution over the parameter space, but only to i.i.d. samples of measure networks drawn from ν_X . Accordingly, this subsection focuses on convergence results for this sampling process. Throughout, we employ the cost structure C from Proposition 3.19.

4.4.1. Random Measure Networks. We begin by focusing on a broadened version of the random metric space model of Proposition 3.9.

Definition 4.16 (Random Measure Network Model). A random measure network model is a pm-net \mathcal{X} with the property that $\Omega_X \subset L^{\infty}(X \times X; \mu_X \otimes \mu_X)$ and $\omega_X^t = t$. We view Ω_X as a metric space equipped with the metric $d_{\Omega_X,p}$ induced by the L^p -norm $\|\cdot\|_{L^p(X \times X; \mu_X \otimes \mu_X)}$.

Let \mathcal{X} be a random measure network model. We define the **empirical pm-net**

$$\mathcal{X}_{T_N} := (X, \mu_X, T_N, \nu_N, \omega_N)$$

by sampling N kernels $T_N = \{t_1, \dots, t_N\}$ i.i.d. according to ν_X and setting $\nu_N := \sum_{i=1}^N \frac{1}{N} \delta_{t_i}$ and $\omega_N^{t_i} = t_i$. For any $t \in \Omega_X$, let \mathcal{X}_t be the trivial pm-net from Proposition 3.3 (alternatively, this is equivalent to the associated measure network, utilized in Sec. 4.3.2).

We begin with some basic lemmas.

Lemma 4.17. Let (A, α) and (B, β) be probability spaces such that |B| = 1 (i.e., B is a one-point set). Then $\mathcal{C}(\alpha, \beta)$ consists of a single element $\xi = (p_A^{-1})_{\#}\alpha$ that satisfies $\int_{A \times B} F(a, b) \, \xi(da \otimes db) = \mathbb{E}_{\alpha}[F(\bullet, b)] := \int_A F(a, b) \, \alpha(da)$ for any measurable map $F: A \times B \to \mathbb{R}$.

Proof. Let $\xi \in \mathcal{C}(\alpha, \beta)$. Since B is a one-point set, the projection $p_A : A \times B \to A$ is a bijection, so $(p_A)_\# \xi = \alpha$ forces $\xi = (p_A^{-1})_\# \alpha$. Then $\int_{A \times B} F(a,b) \; \xi(da \otimes db) = \int_{A \times B} F(a,b) \; (p_A^{-1})_\# \alpha(da \otimes db) = \int_A F(p_A^{-1}(a)) \; \alpha(da) = \int_A F(a,b) \; \alpha(da)$.

Lemma 4.18. For $p \in [1, \infty]$ and $q \in [1, \infty)$, $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{X}_t) = \inf_{\pi \in \mathcal{C}(\mu_X, \mu_X)} \mathbb{E}_{\nu_X}[\mathrm{dis}_p(\pi, \bullet, t)^q]^{1/q}$.

Proof. This follows by setting $A = \Omega_X$, $\alpha = \nu_X$, $B = \{t\}$ and $F_{\pi}(s,t) = \operatorname{dis}_p(\pi,s,t)^q$ in Proposition 4.17. \square This leads to our first sampling convergence result.

Proposition 4.19. For $p \in [1, \infty]$ and $q \in [1, \infty)$, $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N}, \mathcal{X}_t) \to \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{X}_t)$ almost surely as $N \to \infty$.

Proof. Given $\pi \in \mathcal{C}(\mu_X, \mu_X)$, define $F_{\pi}: T_N \to \mathbb{R}_{\geq 0}$ by $F_{\pi}(t_i) := \operatorname{dis}_p(\pi, t_i, t)$. By the Strong Law of Large Numbers, $\mathbb{E}_{\nu_N}[F_{\pi}^q] \to \mathbb{E}_{\nu_X}[F_{\pi}^q]$ almost surely. Taking infimum over $\mathcal{C}(\mu_X, \mu_X)$ and using Proposition 4.17 yields

$$\mathsf{GW}_\mathsf{C}(\mathcal{X}_{T_N},\mathcal{X}_t) = \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \mathbb{E}_{\nu_N}[F^q_\pi(s)]^{1/q} \to \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \mathbb{E}_{\nu_X}[F^q_\pi(s)]^{1/q} = \mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{X}_t)$$

almost surely.

Remark 4.20 ($W_q^{d_{\Omega_X,p}}$ metrizes weak convergence on $\mathcal{P}(\Omega_X)$). Let $p \in [1,\infty]$ and $q \in [1,\infty)$. [47, Theorem 6.9] states that $W_q^{d_{\Omega_X,p}}$ parametrizes weak convergence on the set of measures $\nu_X \in \mathcal{P}(\Omega_X)$ that have finite q moment. However, any $\nu_X \in \mathcal{P}(\Omega_X)$ has finite moments of all orders by compactness of Ω_X because if D_p is the diameter of $(\Omega_X, d_{\Omega_X,p})$ and $q < \infty$,

$$\int_{\Omega_X} \|t - t_0\|_{\Omega_X, p}^q \nu_X(dt) \le \int_{\Omega_X} D_p^q \nu_X(dt) = D_p^q < \infty.$$

Hence, $W_q^{d_{\Omega_X,p}}$ parametrizes weak convergence on $\mathcal{P}(\Omega_X)$ for any $1 \leq q < \infty$.

Hence, we can state our next sampling convergence result for all measures $\nu_X \in \mathcal{P}(\Omega_X)$.

Proposition 4.21. Let $p \in [1, \infty]$ and $q \in [1, \infty)$. For any $\nu_X \in \mathcal{P}(\Omega_X)$, $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{T_N}, \mathcal{X}) \leq \mathsf{W}_q^{d_{\Omega_X, p}}(\nu_N, \nu_X)$ and, thus, $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{T_N}, \mathcal{X}) \to 0$ almost surely as $N \to \infty$.

Proof. Let $\Delta := (\mathrm{id}_X \times \mathrm{id}_X)_{\#} \mu_X$ be the diagonal coupling in $\mathcal{C}(\mu_X, \mu_X)$. For $p < \infty$, we have

$$\operatorname{dis}_{p}(\Delta, t_{i}, t) = \left(\int_{(X \times X)^{2}} |t_{i}(x, x') - t(y, y')|^{p} \Delta(dx \otimes dy) \Delta(dx' \otimes dy')\right)^{1/p}$$
$$= \left(\int_{X \times X} |t_{i}(x, x') - t(x, x')|^{p} \mu_{X}(dx) \mu_{X}(dx')\right)^{1/p} = d_{\Omega_{X}, p}(t_{i}, t),$$

while

$$\operatorname{dis}_{\infty}(\Delta, t_i, t) = \sup_{(x, x'), (y, y') \in \operatorname{supp}(\Delta)} |t_i(x, y) - t(x', y')| = \sup_{x, x' \in X} |t_i(x, x') - t(x, x')| = d_{\Omega_X, \infty}(t_i, t).$$

Then

$$\begin{aligned} \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{X})^{q} &= \inf_{\xi \in \mathcal{C}(\nu_{N}, \nu_{X})} \inf_{\pi \in \mathcal{C}(\mu_{X}, \mu_{X})} \int_{T_{N} \times \Omega_{X}} \mathrm{dis}_{p}(\pi, t_{i}, t)^{q} \ \xi(dt_{i} \otimes dt) \\ &\leq \inf_{\xi \in \mathcal{C}(\nu_{N}, \nu_{X})} \int_{T_{N} \times \Omega_{X}} \mathrm{dis}_{p}(\Delta, t_{i}, t)^{q} \ \xi(dt_{i} \otimes dt) \\ &= \inf_{\xi \in \mathcal{C}(\nu_{N}, \nu_{X})} \int_{T_{N} \times \Omega_{X}} \|t_{i} - t\|_{\Omega_{X}, p}^{q} \ \xi(dt_{i} \otimes dt) \\ &= \mathsf{W}_{q}^{d_{\Omega_{X}, p}}(\nu_{N}, \nu_{X})^{q}. \end{aligned}$$

This yields the first claim. Since $\mathsf{W}_q^{d_{\Omega_X,p}}$ metrizes weak convergence on $\mathcal{P}(\Omega_X)$ by Proposition 4.20, and the empirical measures ν_N converge (set-wise, hence weakly) to ν_X almost surely, $\mathsf{W}_q^{d_{\Omega_X,p}}(\nu_N,\nu_X) \to 0$ almost surely as $N \to \infty$.

This quickly leads to our main sampling convergence result.

Theorem 4. Let $p \in [1, \infty]$ and $q \in [1, \infty)$. Let \mathcal{X} and \mathcal{Y} be random measure network models, as in Proposition 4.16. Let \mathcal{X}_{T_N} and \mathcal{Y}_{S_N} be their respective empirical pm-nets. Then

$$|\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N}, \mathcal{Y}_{S_N}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y})| \leq \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N}, \mathcal{X}) + \mathsf{GW}_{\mathsf{C}}(\mathcal{Y}_{S_N}, \mathcal{Y}),$$

and thus, $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N}, \mathcal{Y}_{S_N}) \to \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y})$ almost surely as $N \to \infty$.

Proof. By the triangle inequality.

$$\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{Y}_{S_{N}}) \leq \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{X}) + \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}) + \mathsf{GW}_{\mathsf{C}}(\mathcal{Y}, \mathcal{Y}_{S_{N}}), \text{ and }$$
$$\mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}) \leq \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{X}_{T_{N}}) + \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{Y}_{S_{N}}) + \mathsf{GW}_{\mathsf{C}}(\mathcal{Y}_{S_{N}}, \mathcal{Y}).$$

Hence,

$$|\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N}, \mathcal{Y}_{S_N}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y})| \leq \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N}, \mathcal{X}) + \mathsf{GW}_{\mathsf{C}}(\mathcal{Y}_{S_N}, \mathcal{Y}),$$

and the right-hand-side converges almost surely to zero by Proposition 4.21.

From Proposition 4.21, the triangle inequality for Wasserstein distances, the argument in the proof of Theorem 4, and Proposition 4.10, we deduce the following corollary.

Corollary 4.22. With the same notation and setup as Theorem 4, the weight distributions satisfy

$$\mathsf{W}_q^{\mathsf{W}_p^{\mathcal{P}(\mathbb{R})}}(\Delta_{\mathcal{X}_{T_N}}, \Delta_{\mathcal{X}}) \to 0 \quad \text{and} \quad \mathsf{W}_q^{\mathsf{W}_p^{\mathcal{P}(\mathbb{R})}}(\Delta_{\mathcal{X}_{T_N}}, \Delta_{\mathcal{Y}_{S_N}}) \to \mathsf{W}_q^{\mathsf{W}_p^{\mathcal{P}(\mathbb{R})}}(\Delta_{\mathcal{X}}, \Delta_{\mathcal{Y}})$$

almost surely as $N \to \infty$.

Proposition 4.21 implies that $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{T_N},\mathcal{X})$ converges towards 0 no faster than $\mathsf{W}_q^{d_{\Omega_X,p}}(\nu_N,\nu_X)$ does. Likewise, the convergence of $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{T_N},\mathcal{Y}_{S_N})$ towards $\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y})$ in Theorem 4 is dominated by terms of the form $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{T_N},\mathcal{X})$ and thus, by Wasserstein distances. Thanks to existing results on rates of convergence of Wasserstein distances, we can quantify the rate of convergence in Proposition 4.21 and Theorem 4. Although the results we cite hold for measures on \mathbb{R}^d and ν_X is defined on Ω_X , the elements of Ω_X are network functions which can be embedded into \mathbb{R}^d for some fixed d. We just need to restrict to finite pm-nets in order for $d < \infty$. Following [14], we use the notation $M_{\mathbb{R}^d,r}(\mu) \coloneqq \int_{\mathbb{R}^d} \|v\|^r \ \mu(dv)$ where μ is a measure and $\|\cdot\|$ is a norm, both defined on \mathbb{R}^d . Below, we use $\|\cdot\|_{\Omega_X,p}$ as short hand for the L^p -norm, restricted to Ω_X .

Proposition 4.23. Let $p \in [1, \infty]$ and $q \in [1, \infty)$. Let \mathcal{X} and \mathcal{Y} be random measure network models with n := |X|, m := |Y| and $m \le n < \infty$. Let \mathcal{X}_{T_N} and \mathcal{Y}_{S_N} be their respective empirical pm-nets for some $T_N \subset \Omega_X$ and $S_N \subset \Omega_Y$ with $|T_N| = |S_N| = N$. Define $D_X := \sup_{t \in \Omega_X} ||t||_{\Omega_X, p}$ and $D_Y := \sup_{t \in \Omega_Y} ||t||_{\Omega_Y, p}$. Let $d := n^2$. Then there exist constants C = C(p, q, n) and C' = C'(p, q, n, m) such that

$$\mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{X})^{q}\right] \leq CD_{X}^{q} \cdot \begin{cases} N^{-1/2} & \text{if } q > d/2, \\ N^{-1/2} \log(1+N) & \text{if } q = d/2, \\ N^{-q/d} & \text{if } 0 < q < d/2. \end{cases}$$

and

$$\mathbb{E}\left[|\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{Y}_{S_{N}}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y})|\right] \leq C'(D_{X} + D_{Y}) \cdot \begin{cases} N^{-1/2q} & \text{if } q > d/2, \\ N^{-1/2q} \log(1+N)^{1/q} & \text{if } q = d/2, \\ N^{-1/d} & \text{if } 0 < q < d/2. \end{cases}$$

If \mathcal{X} and \mathcal{Y} are random metric space models instead (see Proposition 3.9), then the same bounds hold with d = n(n-1)/2 instead of $d = n^2$.

Proof. Given a labelling $X = \{x_1, x_2, \dots, x_n\}$, any function $f \in L^p(X \times X, \mu_X \otimes \mu_X)$ is represented by a matrix $M_f \in \mathbb{R}^{n \times n}$ defined by $(M_f)_{ij} = f(x_i, x_j)$. We define the linear function $\Phi_{n \times n} : L^p(X \times X, \mu_X \otimes \mu_X) \to \mathbb{R}^{n \times n}$ by sending f to M_f . Since X is finite, any $f : X \times X \to \mathbb{R}$ belongs to $L^p(X \times X, \mu_X \otimes \mu_X)$, so $\Phi_{n \times n}$ has an inverse given by $\Phi_{n \times n}^{-1}(M)(x_i, x_j) = M_{ij}$ for all $M \in \mathbb{R}^{n \times n}$. This makes $\Phi_{n \times n}$ into an isomorphism of vector spaces.

If \mathcal{X} and \mathcal{Y} are random metric space models, let L^p_{Sym} be the linear subspace of $L^p(X \times X, \mu_X \otimes \mu_X)$ of symmetric functions with 0 diagonal. As the dimension of L^p_{Sym} is n(n-1)/2, define $\Phi_{\mathrm{Sym}}: L^p_{\mathrm{Sym}} \to \mathbb{R}^{n(n-1)/2}$ to be the coordinate map of L^p_{Sym} with respect to the standard basis of $\mathbb{R}^{n(n-1)/2}$. As before, Φ_{Sym} is an isomorphism of vector spaces.

The rest of the proof proceeds analogously for both random measure network and metric space models, so we will fix the notation L_X^p and Φ to mean $L^p(X \times X, \mu_X \otimes \mu_X)$ and $\Phi_{n \times n}$ if \mathcal{X} and \mathcal{Y} are random measure network models and L_{Sym}^p and Φ_{Sym} if \mathcal{X} and \mathcal{Y} are random metric space models instead. Let $\|\cdot\|_{\Omega_X,p}$ be the L^p norm on L_X^p . Since Φ is an isomorphism, the function $\|v\|_{\mathbb{R}^d,p} := \|\Phi^{-1}(v)\|_{\Omega_X,p}$ defines a norm on \mathbb{R}^d . Recall that $\Omega_X \subset L_X^p$ by definition, so any measure $\nu_X \in \mathcal{P}(\Omega_X)$ extends to a measure on L_X^p . Hence, $\Phi_{\#}\nu_X$ is a measure on \mathbb{R}^d with support $\Phi(\Omega_X)$ such that

$$M_{\mathbb{R}^d,r}(\Phi_{\#}\nu_X) = \int_{\mathbb{R}^d} \|v\|_{\mathbb{R}^d,p}^r \ \Phi_{\#}\nu_X(dv) = \int_{\Omega_X} \|\Phi(t)\|_{\mathbb{R}^d,p}^r \ \nu_X(dt) = \int_{\Omega_X} \|t\|_{\Omega_X,p}^r \ \nu_X(dt).$$

Moreover, $\sup_{r>0} M_{\mathbb{R}^d,r} (\Phi_{\#} \nu_X)^{1/r} = \sup_{t \in \Omega_X} \|t\|_{\Omega_X,p} = D_X$. For any other measure $\nu_X' \in \mathcal{P}(\Omega_X) \subset \mathcal{P}(L_X^p)$, we have, for W_q denoting the Wasserstein distance over the appropriate Euclidean space,

$$W_{q}(\Phi_{\#}\nu_{X}, \Phi_{\#}\nu'_{X})^{q} = \inf_{\xi \in \mathcal{C}(\Phi_{\#}\nu_{X}, \Phi_{\#}\nu'_{X})} \int_{\mathbb{R}^{d} \times \mathbb{R}^{d}} \|v - v'\|_{\mathbb{R}^{d}, p}^{q} \, \xi(dv \otimes dv')
= \inf_{\xi' \in \mathcal{C}(\nu_{X}, \nu'_{X})} \int_{\mathbb{R}^{d} \times \mathbb{R}^{d}} \|v - v'\|_{\mathbb{R}^{d}, p}^{q} \, (\Phi \times \Phi)_{\#} \xi'(dv \otimes dv')
= \inf_{\xi' \in \mathcal{C}(\nu_{X}, \nu'_{X})} \int_{L_{X}^{p} \times L_{X}^{p}} \|\Phi(t) - \Phi(t')\|_{\mathbb{R}^{d}, p}^{q} \, \xi'(dt \otimes dt')
= \inf_{\xi' \in \mathcal{C}(\nu_{X}, \nu'_{X})} \int_{L_{X}^{p} \times L_{X}^{p}} \|t - t'\|_{\Omega_{X}, p}^{q} \, \xi'(dt \otimes dt')
= W_{q}^{d_{\Omega_{X}, p}}(\nu_{X}, \nu'_{X})^{q}.$$

We used Proposition 4.8 in the second line and the fact that the support of ξ' is $\Omega_X \times \Omega_X$ in the last. Now that we have pushed our measures into \mathbb{R}^d , we can use the bounds of Fournier [14]. Let ν_N be the empirical measure defined by T_N . By Proposition 4.21,

$$\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N},\mathcal{X})^q \leq \mathsf{W}_q^{d_{\Omega_X,p}}(\nu_N,\nu_X)^q = \mathsf{W}_q(\Phi_\#\nu_N,\Phi_\#\nu_X)^q,$$

and thus, $\mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N},\mathcal{X})^q\right] \leq \mathbb{E}\left[\mathsf{W}_q(\Phi_{\#}\nu_N,\Phi_{\#}\nu_X)^q\right]$. If $q \neq d/2$, we apply [14, Theorem 2.1] (replacing their p,q,m with q,r,p respectively) to get

$$\mathbb{E}\left[\mathsf{W}_{q}(\Phi_{\#}\nu_{N},\Phi_{\#}\nu_{X})^{q}\right] \leq 2^{q}\kappa_{d,q}^{(p)}[M_{r}^{(p)}(\Phi_{\#}\nu_{X})]^{q/r}\theta_{d,q,r}^{(p)}\cdot N^{-e}$$

for some functions $\kappa_{d,q}^{(p)}$ and $\theta_{d,q,r}^{(p)}$; see [14, Theorem 2.1] for their closed-form expressions. The value of e depends on q and d: e=1/2 if q>d/2 and e=q/d otherwise. Fournier noted in [14, Section 2.4] that $\theta_{d,q,r}^{(p)}$ is a decreasing function with $\lim_{r\to\infty}\theta_{d,q,r}^{(p)}=1$, so together with $M_{\mathbb{R}^d,r}(\Phi_{\#}\nu_X)^{1/r}\leq D_X$, we remove the dependency on r from the inequality above:

$$\begin{split} \mathbb{E}\left[\mathsf{W}_{q}(\Phi_{\#}\nu_{N}, \Phi_{\#}\nu_{X})^{q} \right] &\leq \inf_{r>0} 2^{q} \kappa_{d,q}^{(p)} [M_{r}^{(p)}(\Phi_{\#}\nu_{X})]^{q/r} \theta_{d,q,r}^{(p)} \cdot N^{-e} \\ &\leq \inf_{r>0} 2^{q} \kappa_{d,q}^{(p)} D_{X}^{q} \theta_{d,q,r}^{(p)} \cdot N^{-e} \\ &= 2^{q} \kappa_{d,q}^{(p)} D_{X}^{q} \cdot N^{-e}. \end{split}$$

Hence, we obtain the first claim with $C(p,q,n) \coloneqq 2^q \kappa_{d,q}^{(p)} D_X^q$ if $q \neq d/2$. If q = d/2, we get an analogous bound with e = 1/2, but the functions $\kappa_{d,p,N}^{(p)}$ and $\theta_{d,q,r,N}^{(p)}$ depend on N. However, $\theta_{d,q,r,N}^{(p)}$ is still decreasing in r and has $\lim_{r \to \infty} \theta_{d,q,r,N}^{(p)} = 1$, while $\kappa_{d,p,N}^{(p)} = O(\ln(1+N))$. The result follows as above. For the second claim, we embed Ω_Y in \mathbb{R}^d instead of the smaller space $\mathbb{R}^{m \times m}$ (or $\mathbb{R}^{m(m-1)/2}$ for random

For the second claim, we embed Ω_Y in \mathbb{R}^d instead of the smaller space $\mathbb{R}^{m \times m}$ (or $\mathbb{R}^{m(m-1)/2}$ for random metric space models) as the rates are dominated by the convergence in \mathbb{R}^d anyways. Then by Theorem 4 and Jensen's inequality $\mathbb{E}[X]^q \leq \mathbb{E}[X^q]$ for $q \geq 1$, we get

$$\begin{split} \mathbb{E}\left|\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}},\mathcal{Y}_{S_{N}}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})\right| &\leq \mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}},\mathcal{X})\right] + \mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{Y}_{S_{N}},\mathcal{Y})\right] \\ &\leq \mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}},\mathcal{X})^{q}\right]^{1/q} + \mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{Y}_{S_{N}},\mathcal{Y})^{q}\right]^{1/q}. \end{split}$$

The bound on $\mathbb{E} |\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N}, \mathcal{Y}_{S_N}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y})|$ follows by applying the first claim to each term.

Since p = q = 2 is a common choice in the upcoming experiments, we specialize the result above to these values.

Corollary 4.24. With the same notation as Proposition 4.21, but with the specialization p = q = 2 and n > 2, we have

$$\mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N},\mathcal{X})^2\right] \leq CD_X^2 \cdot N^{-2/n^2}, \text{ and}$$

$$\mathbb{E}\left|\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N},\mathcal{Y}_{S_N}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})\right| \leq C'(D_X + D_Y)N^{-1/n^2}$$

Proof. These bounds are obtained by simplifying the conclusion of Proposition 4.23 with $d = n^2$ and q = 2. For example, the conditions q > d/2, q = d/2 and 0 < q < d/2 become $n^2 < 4$, $n^2 = 4$ and $4 < n^2$, which in turn simplify to n < 2, n = 2 and n > 2, respectively. We only retain the n > 2 case.

Once again, random metric space models have slightly better convergence rates.

Corollary 4.25. Under the same assumptions as Proposition 4.21, except that \mathcal{X} and \mathcal{Y} are random metric space models, d = n(n-1)/2, p = q = 2 and $n \ge 4$, we have

$$\mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{X})^{2}\right] \leq CD_{X}^{2} \cdot N^{-2/d}, \text{ and}$$

$$\mathbb{E}\left|\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_{N}}, \mathcal{Y}_{S_{N}}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y})\right| \leq C'(D_{X} + D_{Y}) \cdot N^{-1/d}.$$

Proof. Note that with q=2, the conditions q>d/2 and 0< q< d/2 become n(n-1)/2<4 and 4< n(n-1)/2, which are equivalent to $n\leq 3$ and $n\geq 4$. Once again, the result follows by simplifying Proposition 4.21.

Remark 4.26. The convergence rates for the quantity $\mathbb{E}\left[\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_{T_N},\mathcal{Y}_{S_N}) - \mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})\right]$ in the previous propositions originate from the rates for $\mathsf{W}_q^{d_{\Omega_X,p}}(\nu_X,\nu_N)$. One may wonder if there exists another estimator of $\mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})$ that has better convergence rates. However, [51, Chapter 3] proves that there exists no estimator of $\mathsf{W}_q(\mu_X,\mu_Y)$ that improves the convergence by more than a logarithmic factor. We have no reason to believe that we can improve the situation for $\mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})$.

Remark 4.27. The convergence results of this section would not hold if we replaced parametrized GW distances with the average GW distance. Let $\hat{\omega}_N := \frac{1}{N} \sum_{i=1}^N t_i$ and let $\hat{\mathcal{X}}_N$ be the measure network $(X, \mu_X, \hat{\omega}_N)$. To condense notation, we use $\|\cdot\|_p$ to denote the L^p norm in $L^p\left((X\times X)^2; \mu_X\otimes\mu_X\right)$ and denote the coordinate projections as $p_1, p_2: X\times X\to X$. By the triangle inequality,

$$\operatorname{dis}_{p}(\pi, \hat{\omega}_{N}, t) = \|\hat{\omega}_{N} \circ (p_{1}, p_{1}) - t \circ (p_{2}, p_{2})\|_{p} = \left\| \sum_{i=1}^{N} \frac{1}{N} [t_{i} \circ (p_{1}, p_{1}) - t \circ (p_{2}, p_{2})] \right\|_{p}$$

$$\leq \sum_{i=1}^{N} \frac{1}{N} \|t_{i} \circ (p_{1}, p_{1}) - t \circ (p_{2}, p_{2})\|_{p} = \frac{1}{N} \sum_{i=1}^{N} \operatorname{dis}_{p}(\pi, t_{i}, t).$$

Hence, infimizing over $\mathcal{C}(\mu_X, \mu_Y)$ yields $\mathsf{GW}_p(\hat{\mathcal{X}}_N, \mathcal{X}_t) \leq \inf_{\pi \in \mathcal{C}(\mu_X, \mu_X)} \frac{1}{N} \sum_{i=1}^N \mathrm{dis}_p(\pi, t_i, t)$.

However, $\inf f + \inf g \leq \inf(f + g)$, so

$$\frac{1}{N} \sum_{i=1}^{N} \mathsf{GW}_{p}(X_{t_{i}}, X_{t}) \le \inf_{\pi \in \mathcal{C}(\mu_{X}, \mu_{X})} \frac{1}{N} \sum_{i=1}^{N} \mathrm{dis}_{p}(\pi, t_{i}, t).$$

Note that $\mathsf{GW}_p(\hat{\mathcal{X}}_N, \mathcal{X}_t)$ and $\frac{1}{N} \sum_{i=1}^N \mathsf{GW}_p(X_{t_i}, X_t)$ are not comparable in these inequalities, so even if the average GW distance converges, we can say nothing about $\mathsf{GW}_p(\hat{\mathcal{X}}_N, \mathcal{X}_t)$.

4.4.2. Extending Beyond Random Measure Network Models. The results of the previous subsection apply specifically to random measure network models. In particular, the convergence result does not apply directly to random graph models, as formulated in Proposition 3.8. Clearly, the set of graphs over a node set X is in bijective equivalence with the set of adjacency kernels $X \times X \to \{0,1\}$, so that one can trivially reformulate any random graph model as a random measure network model. Our convergence theorem therefore easily translates to this setting. We record this, somewhat informally, as a corollary.

Corollary 4.28. Let \mathcal{X} and \mathcal{Y} be random graph models and let \mathcal{X}_{T_N} and \mathcal{Y}_{S_N} be their respective empirical pm-nets. Then $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{T_N},\mathcal{Y}_{S_N}) \to \mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y})$ almost surely as $N \to \infty$.

Moreover, we make the observation that, when working with the cost structure C, the convergence results described in this subsection apply broadly when considering pm-nets up to isomorphism. This is formalized as follows.

Proposition 4.29. Any pm-net is isomorphic to a random measure network model.

Proof. Let \mathcal{X} be an arbitrary pm-net. We define an associated pm-net $\widetilde{\mathcal{X}} = (X, \mu_X, \widetilde{\Omega}_X, \widetilde{\nu}_X, \widetilde{\omega}_X)$ with:

• $\widetilde{\Omega}_X = L^{\infty}(X \times X; \mu_X \otimes \mu_X)$ and $\widetilde{\nu}_X = (\widetilde{m}_{\mathcal{X}})_{\#}\mu_X$, where $\widetilde{m}_{\mathcal{X}} : \Omega_X \to \widetilde{\Omega}_X$ is a map which is closely related to the maps used in the proof of Proposition 4.7, namely,

$$\widetilde{m}_{\mathcal{X}}(t) = \omega_X^t$$
.

Here, the map $\widetilde{m}_{\mathcal{X}}$ is continuous, by the same arguments used in Proposition 4.7, so that the measure $\widetilde{\nu}_X$ is well-defined;

• $\widetilde{\omega}_X$ is defined in the obvious way: given a point ω_X^t in the support of $\widetilde{\nu}_X$, we define $\widetilde{\omega}_X^{\omega_X^t} = \omega_X^t$.

Then $\widetilde{\mathcal{X}}$ is a random measure network model.

We claim that \mathcal{X} is a stabilization of $\widetilde{\mathcal{X}}$, hence that \mathcal{X} is isomorphic to a random measure network model. Indeed, the structure-preserving maps $\Phi: \Omega_X \to \widetilde{\Omega}_X$ and $\varphi: X \to X$ in the definition of stabilization are given by $\Phi = \widetilde{m}_{\mathcal{X}}$ and $\varphi = \mathrm{id}_X$. Clearly, these are both measure-preserving maps. The second condition in the definition of structure-preserving maps reads in this case as

$$\omega_X^t(x, x') = \widetilde{\omega}_X^{\omega_t}(x, x'),$$

which is also obvious from the definition. This verifies that \mathcal{X} is a stabilization of $\widetilde{\mathcal{X}}$ and completes the proof.

Proposition 4.29 says that, when working with the cost structure C, we can replace an arbitrary measure network with a random measure network model at GW_C -distance zero. Employing these replacements, the sampling result Theorem 4 then applies to general measure networks.

5. Numerical Experiments

5.1. **Implementation.** We provide Python implementations of the distances described in ?? 3.16?? 3.19 with p = q = 2. Our implementation builds on the ot.gromov_gromov_wasserstein function from the Python Optimal Transport (POT) library [13], which in turn implements the algorithms of [36, 45]. We briefly review the key results from these works before presenting our own algorithms in detail.

Let $\mathcal{X} = (X, \mu_X, \omega_X)$ and $\mathcal{Y} = (Y, \mu_Y, \omega_Y)$ be measure networks (recall Sec. 2.3) and let N := |X| and M := |Y|. Following the notation of [36], define $C \in \mathbb{R}^{N \times N}$ and $\overline{C} \in \mathbb{R}^{M \times M}$ by $C_{ik} = \omega_X(x_i, x_k)$ and $\overline{C}_{jl} = \omega_Y(y_j, y_l)$. Recall that a coupling $\pi \in \mathcal{C}(\mu_X, \mu_Y)$ is represented by a matrix $\pi \in \mathbb{R}^{N \times M}$ that satisfies $\pi \cdot \mathbb{1}_M = \mu_X$ and $\pi^{\mathsf{T}} \cdot \mathbb{1}_N = \mu_Y$ where $\mathbb{1}_N \in \mathbb{R}^N$ and $\mathbb{1}_M$ are all-one vectors. Given a function $L : \mathbb{R}^2 \to \mathbb{R}$, define the 4-way tensor

$$\mathcal{L}(C, \overline{C}) := \left(L(C_{ik}, \overline{C}_{jl}) \right)_{ijkl} \in \mathbb{R}^{N \times M \times N \times M}$$

and the tensor-matrix multiplication

(24)
$$\mathcal{L}(C,\overline{C}) \otimes \pi := \left(\sum_{kl} L(C_{ik},\overline{C}_{jl}) \pi_{ik}\right)_{ij} \in \mathbb{R}^{N \times M}.$$

Let \mathcal{L}_p be the 4-way tensor induced by $L_p(x,y) := |x-y|^p$ and let $\langle \bullet, \bullet \rangle$ be the Frobenius inner product. The distortion functional satisfies

$$(25) \operatorname{dis}_{p}(\pi, \omega_{X}, \omega_{Y})^{p} = \sum_{ijkl} |\omega_{X}(x_{i}, x_{k}) - \omega_{Y}(y_{j}, y_{l})|^{p} \pi_{ij}\pi_{kl} = \sum_{ijkl} L(C_{ik}, \overline{C}_{jl})\pi_{ij}\pi_{kl} = \langle \mathcal{L}_{p}(C, \overline{C}) \otimes \pi, \pi \rangle.$$

 $\mathcal{L}_2(C,\overline{C})\otimes\pi$ has a simplified form that is an order of magnitude faster to compute than Eqn. (24); see [36, Remark 1].

Lemma 5.1 ([36, Proposition 1]). Let $f_1(a) = a^2$, $f_2(b) = b^2$, $h_1(a) = a$ and $h_2(b) = 2b$. Then:

$$\mathcal{L}_2(C,\overline{C})\otimes\pi=c_{C,\overline{C}}-h_1(C)\cdot\pi\cdot h_2(\overline{C})^{\intercal},$$

where $c_{C,\overline{C}} = f_1(C) \cdot \mu_X \cdot \mathbb{1}_N^{\mathsf{T}} + \mathbb{1}_M \cdot \mu_Y \cdot f_2(\overline{C})^{\mathsf{T}}$.

To find $\mathsf{GW}_2(\mathcal{X},\mathcal{Y})$, Peyré et al. [36] used projected gradient descent to minimize the function $\mathcal{E}_{C,\overline{C}}(\pi) := \langle \mathcal{L}_2(C,\overline{C}) \otimes \pi,\pi \rangle$; note that Eqn. (25) implies $\mathsf{GW}_2(\mathcal{X},\mathcal{Y}) = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \mathcal{E}_{C,\overline{C}}(\pi)^{1/2}$. A useful observation is that the line-search step, i.e. minimizing the objective function in the direction of the projected gradient, has an explicit solution [45, Algorithm 2] that we specify in Proposition 5.2 Item 3. This Lemma collects

other results from [36,45] that we use to implement parameterized GW distances. Since these previous works solve more general problems, we also specify the parameter values that yield Proposition 5.2.

Lemma 5.2.

- (1) $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X, \mu_Y)} \mathcal{E}_{C\overline{C}}(\pi)^{1/2}$.
- $(2) \ \nabla \mathcal{E}_{C,\overline{C}}(\pi) = 2\mathcal{L}(C,\overline{C}) \otimes \pi.$
- (3) Given $\tau = \underset{\tau \in \mathcal{C}(\mu_X, \mu_Y)}{\operatorname{argmin}} \langle \tau, \nabla \mathcal{E}_{C, \overline{C}}(\pi) \rangle$ and $\tau_{\gamma} = (1 \gamma)\pi + \gamma\tau = \pi + \gamma(\tau \pi)$ for $0 \le \gamma \le 1$, the function

 $f(\gamma) := \mathcal{E}_{C,\overline{C}}(\tau_{\gamma})$ expands as a second degree polynomial $f(\gamma) = a\gamma^2 + b\gamma + c$ with coefficients

$$a = -\langle h_1(C) \cdot (\tau - \pi) \cdot h_2(\overline{C})^{\mathsf{T}}, \tau - \pi \rangle$$

$$b = -\langle h_1(C) \cdot \pi \cdot h_2(\overline{C})^{\mathsf{T}}, \tau - \pi \rangle - \langle h_1(C) \cdot \pi \cdot h_2(\overline{C})^{\mathsf{T}}, \pi \rangle.$$

If a > 0, f is minimized when γ is either 0, 1 or -b/2a. Otherwise, f is minimized at $\gamma = 0$ or $\gamma = 1$.

Proof. Item 1 follows from Eqn. (25) and the definition of $\mathcal{E}_{C,\overline{C}}(\pi)$. Contrary to the above, Peyré et al. [36] defined the GW distance in terms of $\mathcal{E}_{C,\overline{C}}$. Item 2 originally appears in a formula in [45, Proposition 2] after setting $\varepsilon = 0$. However, [45] does not contain the derivation and their final formula is missing a factor of 2. The detailed and corrected calculations are found in [49, Section 1.2]. Finally, considering item Item 3: the original solution of the line-search step appears in [45, Algorithm 2] when setting $\alpha = 1$. Instead, we use the formulas from [49, Section 1.3], which also come with a detailed derivation.

5.1.1. Gradient descent for Proposition 3.16. Let (Ω, ν) be a fixed parameter space. When Ω is finite, the cost structure in Proposition 3.16 becomes a sum of terms of the form $\operatorname{dis}_p(\pi, \omega_X^t, \omega_Y^t)$, and the formulas above generalize accordingly. Consequently, we compute $\operatorname{\mathsf{GW}}_\mathsf{C}$ with projected gradient descent, using the formulas in Proposition 5.3 to find the gradient and the explicit solution of the line-search step.

Fix p=q=2 and suppose, for simplicity, that $\Omega=\{1,\ldots,T\}$. Let $\mathcal{X}=(X,\mu_X,\Omega,\nu,\omega_X)$ and $\mathcal{Y}=(Y,\mu_Y,\Omega,\nu,\omega_Y)$ be pm-nets in \mathfrak{N}_{ν} with N:=|X| and M:=|Y|. Define $C\in\mathbb{R}^{T\times N\times N}$ and $\overline{C}\in\mathbb{R}^{T\times M\times M}$ by $C_{t,i,k}=\omega_X^t(x_i,x_k)$ and $\overline{C}_{t,j,l}=\omega_Y^t(y_j,y_l)$, and let

$$\mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\nu) := \sum_{t} \langle \mathcal{L}_2(C_{t,*,*},\overline{C}_{t,*,*}) \otimes \pi, \pi \rangle \cdot \nu_t.$$

Lemma 5.3.

- $(1) \ \mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y}) = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\nu)^{1/2}.$
- (2) $\nabla_{\pi} \mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\nu) = \sum_{t} 2\mathcal{L}_{2}(C_{t,*,*},\overline{C}_{t,*,*}) \otimes \pi \cdot \nu_{t}.$
- (3) Given $\tau = \operatorname*{argmin}_{\tau \in \mathcal{C}(\mu_X, \mu_Y)} \langle \tau, \nabla_{\pi} \mathcal{E}_{\mathsf{C}, C, \overline{C}}(\pi, \nu) \rangle$ and $\tau_{\gamma} = (1 \gamma)\pi + \gamma\tau = \pi + \gamma(\tau \pi)$ for $0 \le \gamma \le 1$,

the function $f(\gamma) := \mathcal{E}_{\mathsf{C},C,\overline{C}}(\tau_{\gamma},\nu)$ expands as a second degree polynomial $f(\gamma) = a\gamma^2 + b\gamma + c$ with coefficients

$$\begin{split} a &= -\sum_t \langle h_1(C_{t,*,*}) \cdot (\tau - \pi) \cdot h_2(\overline{C}_{t,*,*})^\intercal, \tau - \pi \rangle \cdot \nu_t \\ b &= -\sum_t \left[\langle h_1(C_{t,*,*}) \cdot \pi \cdot h_2(\overline{C}_{t,*,*})^\intercal, \tau - \pi \rangle + \langle h_1(C_{t,*,*}) \cdot \pi \cdot h_2(\overline{C}_{t,*,*})^\intercal, \pi \rangle \right] \cdot \nu_t. \end{split}$$

If a > 0, f is minimized when γ is either 0, 1 or -b/2a. Otherwise, f is minimized at $\gamma = 0$ or $\gamma = 1$.

Proof. Recall from Proposition 3.16 (setting p=2) that

$$\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y}) = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \| \mathrm{dis}_2(\pi,\omega_X,\omega_Y) \|_{L^2(\Omega;\nu)} = \frac{1}{2} \inf_{\pi \in \mathcal{C}(\mu_X,\mu_Y)} \left(\sum_t \mathrm{dis}_2(\pi,\omega_X^t,\omega_Y^t)^2 \cdot \nu_t \right)^{1/2}.$$

Item 1 follows by applying Eqn. (25) to each term above. Since the gradient is linear, using Proposition 5.2 Item 2 on each summand of $\nabla_{\pi} \mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\nu)$ yields Item 2. Likewise, $f(\gamma) = \mathcal{E}_{\mathsf{C},C,\overline{C}}(\tau_{\gamma},\nu)$ is a sum of second degree polynomials in γ with coefficients given by Proposition 5.2 Item 3, so Item 3 follows.

5.1.2. Alternating optimization for Proposition 3.19. Similar to Proposition 5.3, it is straightforward to generalize the objective function and the formulas in Proposition 5.2 to compute the parametrized GW_C from Proposition 3.19. However, this time we need to find two couplings $\xi \in \mathcal{C}(\nu_X, \nu_Y)$ and $\pi \in \mathcal{C}(\mu_X, \mu_Y)$ that jointly minimize the objective function, so we have to update the optimization procedure. We set up notation and generalize Proposition 5.2 before explaining the algorithm.

Once again, fix p=q=2. Let (Ω_X, ν_X) and (Ω_Y, ν_Y) be parameter spaces with $\Omega_X=\{1,\ldots,T\}$ and $\Omega_Y=\{1,\ldots,S\}$, and let $\mathcal{X}=(X,\mu_X,\Omega_X,\nu_X,\omega_X), \mathcal{Y}=(Y,\mu_Y,\Omega_Y,\nu_Y,\omega_Y)\in\mathfrak{N}_{all}$. Define $C\in\mathbb{R}^{T\times N\times N}$ and $\overline{C}\in\mathbb{R}^{S\times M\times M}$ by $C_{t,i,k}=\omega_X^t(x_i,x_k)$ and $\overline{C}_{s,j,l}=\omega_Y^s(y_j,y_l)$. Given $\pi\in\mathcal{C}(\mu_X,\mu_Y)$ and $\xi\in\mathcal{C}(\nu_X,\nu_Y)$, define

$$\mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\xi) := \sum_{t,s} \langle \mathcal{L}_2(C_{t,*,*},\overline{C}_{s,*,*}) \otimes \pi, \pi \rangle \cdot \xi_{ts}.$$

Lemma 5.4.

(1) $\mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y}) = \frac{1}{2}\inf_{\pi,\xi} \mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\xi)^{1/2}$ where the inf runs over $\pi \in \mathcal{C}(\mu_X,\mu_Y)$ and $\xi \in \mathcal{C}(\nu_X,\nu_Y)$.

(2)
$$\nabla_{\pi} \mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\xi) = \sum_{t,s} 2\mathcal{L}_2(C_{t,*,*},\overline{C}_{s,*,*}) \otimes \pi \cdot \xi_{ts}.$$

(3) Given $\tau = \underset{\tau \in \mathcal{C}(\mu_X, \mu_Y)}{\operatorname{argmin}} \langle \tau, \nabla_{\pi} \mathcal{E}_{\mathsf{C}, C, \overline{C}}(\pi, \xi) \rangle$ and $\tau_{\gamma} = (1 - \gamma)\pi + \gamma\tau = \pi + \gamma(\tau - \pi)$ for $0 \le \gamma \le 1$, the function $f(\gamma) := \mathcal{E}_{\mathsf{C}, C, \overline{C}}(\tau_{\gamma}, \xi)$ expands as a second degree polynomial $f(\gamma) = a\gamma^2 + b\gamma + c$ with

$$\begin{split} a &= -\sum_{t,s} \langle h_1(C_{t,*,*}) \cdot (\tau - \pi) \cdot h_2(\overline{C}_{s,*,*})^\intercal, \tau - \pi \rangle \cdot \xi_{ts} \\ b &= -\sum_{t,s} \left[\langle h_1(C_{t,*,*}) \cdot \pi \cdot h_2(\overline{C}_{s,*,*})^\intercal, \tau - \pi \rangle + \langle h_1(C_{t,*,*}) \cdot \pi \cdot h_2(\overline{C}_{s,*,*})^\intercal, \pi \rangle \right] \cdot \xi_{ts}. \end{split}$$

If a > 0, f is minimized when γ is either 0, 1 or -b/2a. Otherwise, f is minimized at $\gamma = 0$ or $\gamma = 1$.

We minimize the two-variable objective $\mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\xi)$ with an alternating optimization procedure. The minimization with respect to $\pi \in \mathcal{C}(\mu_X,\mu_Y)$ is solved using projected gradient descent with the updated formulas in Proposition 5.4. The minimization $\inf_{\xi \in \mathcal{C}(\nu_X,\nu_Y)} \mathcal{E}_{\mathsf{C},C,\overline{C}}(\pi,\xi)$ is a standard optimal transport problem with cost matrix $M_{\pi,C,\overline{C}} \in \mathbb{R}^{T \times S}$ given by

$$(M_{\pi,C,\overline{C}})_{ts} = \langle \mathcal{L}_2(C_{t,*,*},\overline{C}_{s,*,*}) \otimes \pi, \pi \rangle.$$

In other words, we solve

$$\begin{array}{ll} \text{Minimize: } \sum_{ts} (M_{\pi,C,\overline{C}})_{ts} \cdot \xi_{ts} \\ \text{Subject to: } \xi \in \mathcal{C}(\nu_X,\nu_Y). \end{array}$$

5.2. **Pandas.** We begin with a proof of concept that the parametrized GW distance incorporates information that is spread across multiple scales. By "spreading information" we mean that given a metric space (X, d_X) and an expression $X = X_1 \cup \cdots \cup X_\ell$, we define the pseudo-metrics $\omega_X^i : X \times X \to \mathbb{R}_{\geq 0}$ by $\omega_X^i(x, x') = d_X(x, x')$ if $x, x' \in X_i$ and 0 otherwise. Each ω_X^i only remembers the distances between points in X_i , so if we have another metric space (Y, d_Y) with an analogous expression $Y = Y_1 \cup \cdots \cup Y_\ell$ and pseudo-metrics ω_Y^i , the GW coupling between ω_X^i and ω_Y^i only has information on X_i and Y_i . We use the parametrized GW distance to incorporate the information of all ω_X^i and ω_Y^i in one coupling.

For this experiment, we use a graph that we call a panda. Let P_1 be a cycle graph of size N and select two vertices of C at distance $e \leq \lfloor N/2 \rfloor$. Given integers N, n, e with n < N and $e \leq \lfloor N/2 \rfloor$, an (N, n, e)-panda graph P is formed by gluing two cycle graphs P_2 and P_3 of size n to P_1 , one at each distinguished vertex. We say that the N-cycle P_1 is the head of the panda, and that each of the smaller n-cycles P_2 and P_3 is an ear. Consequently, $P = P_1 \cup P_2 \cup P_3$ and $|P_1 \cap P_i| = 1$ for i = 2, 3. The pseudo-metrics ω_P^t are defined as above. We define a number of pm-nets from this setup. Let $\Omega := \{1, 2, 3\}$. Let μ_P and ν be the uniform measures on P and Ω , respectively, and let d_P be the shortest path distance on P. We define the metric measure spaces (mm-spaces) $\mathcal{P}_0 := (P, \mu_P, d_P)$ and $\mathcal{P}_t := (P, \mu_P, \omega_P^t)$ for t = 1, 2, 3. We also define the pm-net $\mathcal{P}_{MS} := (P, \mu_P, \Omega, \nu, (\omega_P^t)_{t \in \Omega})$.

We perform our experiments on a (25, 10, 6)-panda graph X and a (30, 12, 6)-panda Y. The pm-nets \mathcal{X}_{MS} , \mathcal{Y}_{MS} and the mm-spaces \mathcal{X}_t and \mathcal{Y}_t are defined as above. Fig. 1 shows the graph, distance matrix, and the pseudo-metrics ω_t of X in the top row and those of Y, in the bottom.

We compute $\mathsf{GW}_2(\mathcal{X}_t,\mathcal{Y}_t)$ for t=1,2,3 and $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{MS},\mathcal{Y}_{MS})$ where C is the cost structure of Proposition 3.16 with p=q=2. Fig. 2 has the optimal couplings π_C for $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{MS},\mathcal{Y}_{MS})$ and π_t for $\mathsf{GW}_2(\mathcal{X}_t,\mathcal{Y}_t)$, $0 \le t \le 3$. We observe that a single π_t with $1 \le t \le 3$ only sees the points from X_t and Y_t , so every π_t is a random coupling outside of a single block. The coupling π_C combines the information from these couplings into one. We remark that π_C is still not an optimal coupling for $\mathsf{GW}_2(\mathcal{X}_t,\mathcal{Y}_t)$ because the computation of GW_C still has no access to interactions between X_i and Y_j for $i \ne j$.

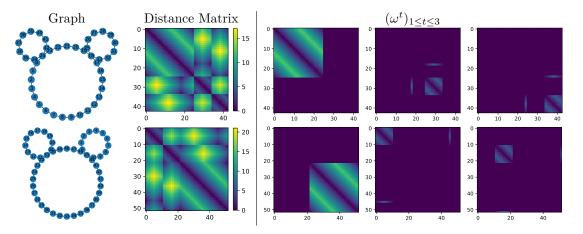


FIGURE 1. From left to right, each row shows a panda graph, its distance matrix, and the network functions ω^1 , ω^2 , and ω^3 . The function ω^1 is the restriction of the distance matrix to the vertices of the head, while ω^2 and ω^3 are the restrictions to the ears. In the top row, the head consists of 25 vertices and each ear of 10 vertices; in the bottom row, the head has 30 vertices and each ear 12. The ears are formed by the vertex sets $\{18\} \cup \{25, 26, \ldots, 33\}$ and $\{24\} \cup \{34, \ldots, 42\}$ in the top panda, and by $\{0, \ldots, 10\} \cup \{45\}$ and $\{11, \ldots, 21\} \cup \{51\}$ in the bottom panda. The corresponding heads are given by $\{0, \ldots, 24\}$ in the top panda and $\{22, \ldots, 51\}$ in the bottom panda.

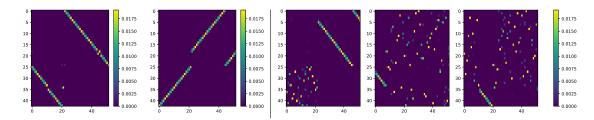


FIGURE 2. From left to right, the optimal couplings for $\mathsf{GW}_\mathsf{C}(\mathcal{X}_{MS},\mathcal{Y}_{MS})$ and $\mathsf{GW}_\mathsf{C}(\mathcal{X}_t,\mathcal{Y}_t)$ for t=0,1,2,3.

5.3. Random Graph Models. The next experiments illustrate the behavior of parameterized GW distances on random graph models.

5.3.1. Perturbation Model. For the first experiment, we study the behavior of the parameterized GW distance on a perturbative random graph model. Given a graph G = (V, E), we construct a generative random graph model as follows. For a positive integer k < |E|, a random sample is generated by deleting k existing edges and adding k new edges, both uniformly at random. Each graph is represented by a binary adjacency kernel, yielding a pm-net $\mathcal{V}_k = (V, \mu_V, \Omega_V, \nu_V, \omega_V)$, where:

- μ_V is the uniform distribution on V;
- Ω_V is the set of all adjacency kernels on V;
- ν_V is the (unknown) distribution from which the graph kernels are being sampled under the perturbation model with parameter k;

• $\omega_V^t = t$ for all $t \in \Omega_V$, i.e., an adjacency kernel.

The goal of this experiment is to understand the behavior of the parameterized GW distance on empirical estimates of this pm-net, as in Sec. 4.4. We begin with the well-known Karate Club graph G=(V,E), from [54]. For a fixed $k\in\{1,2,3,4,5\}$ and $n\in\{10,20,50,100,150\}$, we construct an empirical estimate $\mathcal{V}_{k,n}$ of the pm-net \mathcal{V}_k by drawing n samples of the perturbation model with addition/deletion parameter k. We then construct an additional empirical estimate $\mathcal{V}'_{k,n}$ via the same procedure (i.e., $\mathcal{V}_{k,n}$ and $\mathcal{V}'_{k,n}$ are both estimates of the same pm-net \mathcal{V}_k) and compute $\mathsf{GW}_\mathsf{C}(\mathcal{V}_{k,n},\mathcal{V}'_{k,n})$, where C is the cost structure defined in Proposition 3.19, with p=q=2. This calculation is repeated 10 times for each choice of parameters (k,n), and results are reported in Fig. 3. As a baseline, we compare the empirical estimates using the standard p=2 GW distance: for each pair of sampled graphs in $\mathcal{V}_{k,n}$ and $\mathcal{V}'_{k,n}$ (in the arbitrary order they were sampled), we compute the GW distance between their adjacency kernels and then average the results. These results are also recorded in Fig. 3.

The results of this experiment are rather intuitive. The standard GW distance (denoted as GW) is essentially constant as the number of samples increases, with the only difference being a tightening of the standard deviations over trials for larger numbers of samples. On the other hand, the parameterized GW distance (denoted as PGW) decreases as the number of samples increases—indeed, in theory, this should converge to zero as the number of samples goes to infinity. The difference between the standard GW and parameterized GW distances is more pronounced as k increases, i.e., as the underlying distribution becomes more complicated. This experiment illustrates the benefit of incorporating global information in the distance computation via the parameterized GW framework.

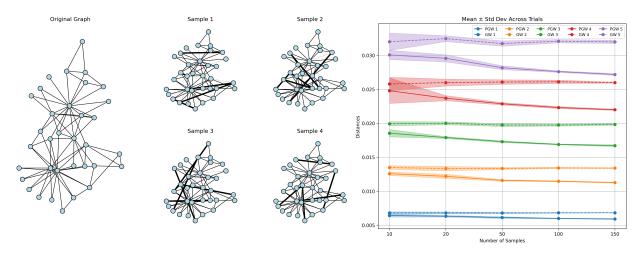


FIGURE 3. Illustration of the perturbation model from Sec. 5.3.1. The Karate Club graph G appears on the left. The center panel shows samples from the perturbation model, where k=5 edges are deleted from and added to G in each instance; added edges are drawn in bold, and deleted edges in dotted style. The right panel plots the standard GW distance (dashed, denoted as \mathbf{GW}) and the parameterized GW distance (solid, denoted as \mathbf{PGW}) between empirical estimates of the random graph model, as a function of the number of samples (x-axis) and the number of edge additions/deletions (encoded by color).

5.3.2. Clustering Random Graphs via Distributions of Total Edges. We use the distribution of total edges invariants described in Sec. 4.3.4 to cluster random graph models by their parameters. By Proposition 4.10 and Proposition 4.13, this serves a proxy for the parameterized GW distance, and by Proposition 4.15, it is efficiently computable. Here, we use parameters q = 2 and p = 1 when computing Wasserstein distances between distributions of total edges.

In the first version of the experiment, we use the Erdős-Rényi random graph model. We consider four instances of this model: in each instance, the underlying graphs have 50 nodes, with the probability of connecting any two nodes given by $\rho \in \{0.44, 0.46, 0.48, 0.5\}$. A single trial of the experiment is described as follows. For each $k \in \{1, 5, 10, 15, 20\}$, we draw k graphs from each model (i.e., each choice of ρ), and then repeat this a total of 10 times. This gives a total of 40 empirical random graph models, but these really come from only 4 classes—we expect that empirical models with the same ρ should cluster tightly together, and

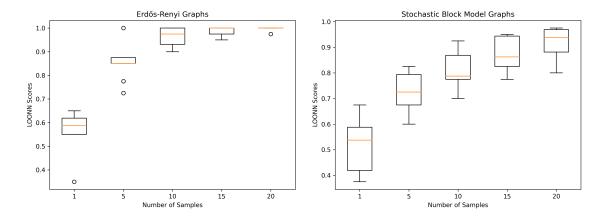


FIGURE 4. LOONN clustering scores (higher is better) versus number of samples in the empirical approximations for the Erdős-Rényi random graph model (Left) and the stochastic block model (Right).

that this clustering should become more pronounced for larger values of k. We compute pairwise Wasserstein distances between the distributions of total edges across the dataset of 40 random graph models. Clustering is measured by Leave One Out Nearest Neighbor (LOONN) score (higher is better): for a fixed random model, we determine which model among the remaining 39 is closest to the fixed one (using Wasserstein distance between distributions of total edges); if the closest model has the same ρ -value as the fixed one, this is treated as a success, and total success percentage across the dataset is the reported score. We repeat the full experiment 10 times, and the results are provided in Fig. 4. Observe that the results agree with intuition. Indeed, increasing k yields higher clustering scores, and the distribution of total edges appears to capture the dependence of the models on ρ quite well (this is unsurprising, given Proposition 4.14).

We next run the same experiment on a different random graph model. In the second version of the experiment, we use *stochastic block models*. In each instance, we have a graph on 50 nodes which have been partitioned into even groups of 25. An instance of the model depends on parameters $\rho_1, \rho_2 \in [0, 1]$, where ρ_1 is the probability of connecting any two nodes within a partition block, and ρ_2 is the probability of connecting two nodes lying in distinct blocks. Here, we also use four classes, with

$$(\rho_1, \rho_2) \in \{(0.5, 0.28), (0.5, 0.3), (0.6, 0.28), (0.6, 0.3)\}.$$

The experimental setup is then identical to the above. The results (also reported in Fig. 4) are qualitatively similar to the Erdős-Rényi case, but quantiatively indicate that this classification task is slightly more difficult.

5.4. **Nested Cycles.** The graph heat kernel represents the diffusion of heat in a graph across time, and it captures graph features at increasing scales as time advances. In many applications, people perform tests on graphs using the heat kernel at a single time, which raises the question of how to compare graphs that have features at multiple scales that the heat kernel cannot capture simultaneously.

We study this question with the following family of graphs. Given a sequence of graphs G_1, \ldots, G_n and basepoints $v_1 \in G_1, \ldots, v_n \in G_n$, we define an n-cycle of graphs as the result of attaching each G_i to an n-cycle C_n by gluing $v_i \in G_i$ to the i-th vertex of C_n . We set v_1 as the basepoint of the resulting graph. For a fixed set of positive integers n_1, \ldots, n_ℓ and m, we define a 1-nested cycle of cliques of type (n_1, m) as an n_1 -cycle of m-cliques and an ℓ -nested cycle of cliques of type (n_1, \ldots, n_ℓ, m) as an n_1 -cycle of cliques of type (n_2, \ldots, n_ℓ, m) . Note that this construction is independent of the choice of basepoint in the m-cliques. See Fig. 5. We refer to the n_i -cycles and m-cliques as features at scale i and $\ell + 1$, respectively. The heat kernel of an ℓ -nested cycle of cliques has features at ℓ different times because, in order for heat to diffuse through each n_i -cycle, it first has to diffuse through $(\ell - i)$ -nested cycles of cliques, and this process takes longer for smaller i.

To set a benchmark for the upcoming experiments, suppose we compute the GW distance between the distance matrices of two ℓ -nested cycles of cliques of types (n_1, \ldots, n_ℓ, m) and $(n_1, \ldots, n_\ell, m')$ with $m \neq m'$ and equipped with the uniform measure. The optimal coupling π captures the multiscale structure of these graphs through a hierarchy of nested blocks. At the coarsest level, these graphs are n_1 -cycles of subgraphs,

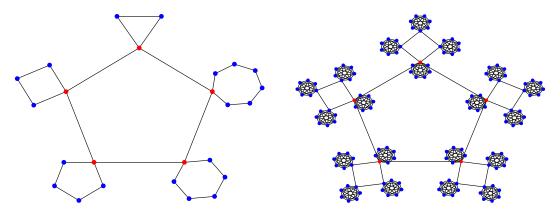


FIGURE 5. **Left:** A 5-cycle of graphs formed by the cycle graphs C_3 , C_4 , C_5 , C_6 , C_7 . The basepoint of each cycle graph is marked in red. **Right:** A 2-nested cycle of cliques of type (5,4,7). This graph is a 5-cycle of 1-nested cycles of type (4,7), each of which is a 4-cycle of 7-cliques. The basepoint of each 1-nested cycle is marked in red.

so π should be an n_1 -by- n_1 grid of blocks of size (N_1m) -by- (N_1m') , where $N_i = n_{i+1} \cdots n_\ell$. These outer blocks form a coupling between two n_1 -cycles. After normalizing, each non-zero sub-block of size (N_2m) -by- (N_2m') is a coupling between $(\ell-1)$ -nested cycles of cliques, and we can iterate this description on each block. Thus, π has a nested block structure where the outer blocks form a coupling of n_1 -cycles, each non-zero block thereof is a coupling of n_2 -cycles, and so on. At the smallest scale, the blocks that form a coupling of n_ℓ -cycles are themselves couplings between cliques of sizes m and m'. Note that an optimal coupling between n-cycles with uniform measure is a cyclic permutation of $\{1,\ldots,n\}$, while an optimal coupling between cliques is random. We say that the outer blocks of size (N_1m) -by- (N_1m') occur at scale 1, their sub-blocks of size (N_2m) -by- (N_2m') occur at scale 2, and so on.

To test how well the GW distance captures multiscale features, we compare heat kernels at multiple values of t. Concretely, let G_1 and G_2 be 2-nested cycles of cliques of types (10,5,5) and (10,5,20), and let $H_{i,t}$ denote the heat kernel of G_i at time t. Based on the discussion above, we expect the optimal coupling π to have 10 blocks, each with 5 sub-blocks of random noise of size 5-by-20, and the blocks at scales 1 and 2 should form cyclic permutations of 10- and 5-cycles, respectively. We summarize this structure using the binary vector cyclic above each coupling in Fig. 6 and Fig. 7. The i-th entry of cyclic indicates whether all non-zero blocks of π at scale i form cyclic permutations of n_i -cycles. In line with our intuition, the GW couplings in Fig. 6 capture small and large scale features at different times. Specifically, only the blocks at scale 2 are cyclic permutations when $30 \le t \le 50$ (as indicated by cyclic = [0, 1]), while the opposite is true when $t \ge 170$ (cyclic = [1, 0]).

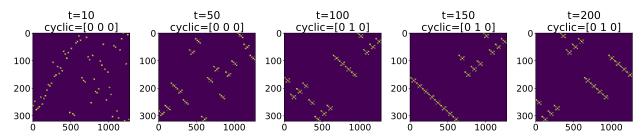


FIGURE 6. Optimal couplings for $\mathsf{GW}_2(H_{1,t},H_{2,t})$, where $H_{i,t}$ is the heat kernel of the 2-nested cycle of cliques G_i , and G_1 and G_2 have types (10,5,5) and (10,5,20). Each panel has the time parameter of $H_{i,t}$, and the vector cyclic indicates whether the coupling is a cyclic permutation at each scale.

We now attempt to capture features at both scales simultaneously with the parametrized GW distance on a fixed parameter space. Let C be the cost structure from Proposition 3.16 with p=q=2. We set $t_1=50,\ t_2=200$ and $\Omega=\{t_1,t_2\}$, but select ν later. We manually chose t_1 and t_2 because the GW

²More precisely, the n_1 -by- n_1 matrix of block sums is a coupling between two n_1 -cycles.

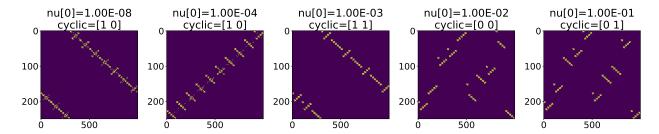


FIGURE 7. Optimal couplings for $\mathsf{GW}_\mathsf{C}(\mathcal{H}_1,\mathcal{H}_2)$. \mathcal{H}_i is a sequence of two heat kernels $H_{i,t}$ with $t_1=50$ and $t_2=200$. The title of each panel contains the value of ν_1 (nu[0] in 0-indexing) and the vector cyclic that indicates whether the coupling is a cyclic permutation at each scale. The coupling induced by $\nu_1=10^{-3}$ (and $\nu_2=0.999$) has the desired block structure at all scales (i.e. cyclic=[1,1]).

couplings satisfy cyclic = [0,1] when t = 50 and cyclic = [1,0] when t = 200. We construct pm-nets $\mathcal{H}_i = (G_i, \mu_i, (H_{i,t_i})_{i=1,2}, \Omega, \nu)$ with the uniform measure μ_i for each i = 1, 2.

We have to carefully choose ν because of a numerical issue in the computation of $\mathsf{GW}_\mathsf{C}(\mathcal{H}_1,\mathcal{H}_2)$. The extreme values of H_{1,t_1} and H_{2,t_1} are several orders of magnitude larger than those of H_{1,t_2} and H_{2,t_2} , even after normalization (e.g. with the Frobenius norm). When we set ν as the uniform measure, the values at t_1 dominate the optimization and the resulting coupling resemble the GW coupling at t_1 . We resolve this issue by doing a grid search on ν . Fig. 7 has the optimal couplings for $\mathsf{GW}_\mathsf{C}(\mathcal{H}_1,\mathcal{H}_2)$ for several choices of ν . In particular, the coupling with $\nu_1 = 10^{-3}$ and $\nu_2 = 1 - \nu_1$ has the expected block structure (cyclic = [1, 1]).

Therefore, after some parameter tuning, the parametrized GW distance with a fixed parameter space (Proposition 3.16) captures information that is spread across multiple heat kernels with a single coupling.

5.4.1. 3-nested cycles of cliques. We repeat the experiments above with 3-nested cycles of cliques of types (4, 4, 4, 5) and (4, 4, 4, 20) to see if the parametrized GW distance needs 3 levels to capture features at three scales.

For the standard GW framework, we construct the heat kernels $H_{1,t}$ and $H_{2,t}$ as above with $10 \le t \le 500$ and compute $GW_2(H_{1,t}, H_{2,t})$; see Fig. 8. The blocks of the GW couplings form cyclic permutations of 4-cycles at scale 3 when t = 20,30 (cyclic=[0,0,1]), at scale 2 when $100 \le t \le 280$ (cyclic=[0,1,0]) and at scale 1 when t = 410,430 (cyclic=[1,0,0]).

For the parametrized GW framework, we manually select $t_1=30,\ t_2=100,\ \text{and}\ t_3=410,\ \text{and}\ \Omega=\{t_1,t_2,t_3\}$ and $\mathcal{H}_i=(G_i,\mu_i,(H_{i,t_j})_{1\leq j\leq 3},\Omega,\nu)$. We perform a grid search over ν and found that $\nu_1=1.29\times 10^{-4},\ \nu_2=3.39\times 10^{-5}$ and $\nu_3=1-\nu_1-\nu_2\approx 9.998\times 10^{-1}$ produces a coupling with the correct block structure at all scales; see Fig. 9.

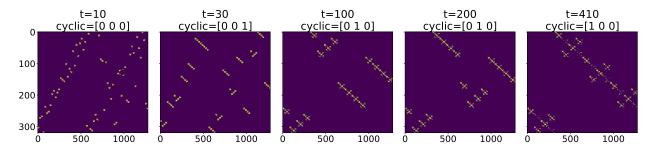


FIGURE 8. Optimal couplings for $\mathsf{GW}_2(H_{1,t},H_{2,t})$, where $H_{i,t}$ is the heat kernel of the 3-nested cycle of cliques G_i . G_1 and G_2 have types (4,4,4,5) and (4,4,4,20). Each panel has the time parameter of $H_{i,t}$, and the vector cyclic indicates whether the coupling is a cyclic permutation at each scale.

5.5. **Feature Selection.** In this subsection, we propose the parameterized GW distance as a cost function for feature selection. We interpret Ω as the set of feature labels (invariants) and ν as their relative importance. For instance, in the case of graph data one may take $\Omega = \{ \mathrm{adj}, \mathrm{Lap}, \mathrm{dist} \}$, where ω_G^{adj} denotes the adjacency matrix, ω_G^{Lap} the Laplacian, and ω_G^{dist} the shortest-path distance matrix of a given graph G.

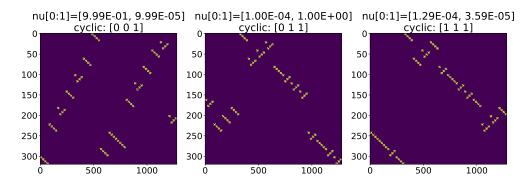


FIGURE 9. Optimal couplings for $\mathsf{GW}_\mathsf{C}(\mathcal{H}_1,\mathcal{H}_2)$. \mathcal{H}_i is a sequence of three heat kernels $H_{i,t}$ with $t_1=30$, $t_2=100$, and $t_3=410$. The title of each panel contains the value of ν_1 and ν_2 (nu[0:1] in 0-indexing, and rounded to 2 decimal places) and the vector cyclic that indicates whether the coupling is a cyclic permutation at each scale. The coupling induced by $\nu_1=1.29\times 10^{-4},\ \nu_2=3.39\times 10^{-5}$ and $\nu_3=1-\nu_1-\nu_2\approx 9.998\times 10^{-1}$ has the desired block structure at all scales (cyclic=[1,1,1]).

Our objective is to find the weights ν that optimize a downstream task. Invariants that hinder performance should get small or zero weights, while those that contribute positively should get larger weights. We work on the class \mathfrak{N}_{ν} of pm-nets parameterized by (Ω, ν) (recall Proposition 3.11) and use the cost structure from Proposition 3.16 with p=q=2. For any $\mathcal{X}=(X,\mu_X,\Omega,\nu,\omega_X)\in\mathfrak{N}_{\nu}$, the set $\{\omega_X^t:t\in\Omega\}$ contains all invariants of (X,μ_X) , and the parametrized GW distance $\mathsf{GW}_{\mathsf{C}}(\mathcal{X},\mathcal{Y})$ measures the difference between the corresponding invariants of $\mathcal{X},\mathcal{Y}\in\mathfrak{N}_{\nu}$.

To choose a concrete task, suppose we have pm-nets $\mathcal{X}_1, \ldots, \mathcal{X}_n$ with class labels $y_1, \ldots, y_n \in \{1, \ldots, m\}$, and we want to determine which invariants from Ω correctly classify them. Let M_{ν} be the *n*-by-*n* matrix given by $(M_{\nu})_{ij} = \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_i, \mathcal{X}_j)$. Suppose that $y_1 \leq \cdots \leq y_n$ so that M_{ν} has the block structure

(26)
$$M_{\nu} = \begin{pmatrix} B_{11} & \cdots & B_{1m} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mm} \end{pmatrix}$$

where B_{ij} is the matrix of distances between elements of classes i and j. The best clustering is achieved when the intra-cluster distances are small relative to the inter-cluster distances, so we want to find the ν that minimizes

$$\cot_p(\nu) := \frac{\|B_{11}\|_p^p + \dots + \|B_{mm}\|_p^p}{\|M_{\nu}\|_p^p}$$

where $\| \bullet \|_2$ is the Frobenius norm. Depending on the context, we may add a regularization term and minimize

(27)
$$\operatorname{cost}_{p,\lambda}(\nu) := \frac{\|B_{11}\|_p^p + \dots + \|B_{mm}\|_p^p}{\|M_{\nu}\|_p^p} + \lambda \cdot KL(\nu|q)$$

instead, where $\lambda \geq 0$, q is the uniform measure on Ω , and $KL(\nu|q)$ is the KL divergence.

5.5.1. Dynamic Metric Spaces. Recall that \mathcal{X} is a dynamic metric space if Ω is a compact subset of $\mathbb{R}_{\geq 0}$ and every ω_X^t is a (pseudo)-metric. Suppose that a set of drones flies through one of two corridors that have the same shape, except that one corridor has an obstacle. If the drones maintain roughly the same speed and direction, their behavior only changes significantly when they dodge the obstruction. We use the pipeline above to identify the times when the drones find the obstacle.

We define two types of pm-nets that represent the flight of the drones through one of the two corridors; see Fig. 10. Fix the indexing sets $X = \{0, \dots, 4\} \times \{0, \dots, 4\}$ and $\Omega = \{0, \dots, 4\}$, and let μ_X and ν_0 be their uniform probability measures. The corridor is the rectangle $[-1, 4] \times [-2, 2]$ in \mathbb{R}^2 and the obstacle is the ellipse $\left(\frac{x-1.5}{0.5}\right)^2 + \left(\frac{y}{0.7}\right)^2 = 1$. The initial drone configuration is the 5×5 grid $P(X) \subset [0, 1] \times [-1, 1]$ where $P(x, y) = \left(\frac{x}{4}, -1 + \frac{y}{2}\right)$. Each point represents a drone. At each time step, all drones move $\Delta x = 0.6$ units to the right resulting in 5 grids X_0, \dots, X_4 of 25 points each. In the corridor without obstacles, we apply Gaussian noise $\mathcal{N}(0, 0.05)$ independently to every point of X_i , and define ω_X^i as the distance matrix

λ	ν_0	ν_1	ν_2	ν_3	$ u_4$	$\cot_{\lambda}(\nu)$
0.01	0.007	0.001	0.968	0.022	0.001	0.095149
0.1	0.011	0.423	0.525	0.027	0.014	0.187433
1	0.146	0.258	0.298	0.151	0.146	0.283082
10	0.195	0.206	0.210	0.195	0.195	0.095149 0.187433 0.283082 0.296712

Table 1. Minimizer of $\cot_{\lambda}(\nu)$ for several values of λ . If λ is small, the measure ν assigns the most weight to the single time that distinguishes classes 1 and 2 the best. Conversely, if λ is too big, there is no time that has significantly larger weight than the others. In the intermediate range ($\lambda = 0.1, 1$), the times when the drones avoid the obstacle (t = 1, 2) have the largest weights.

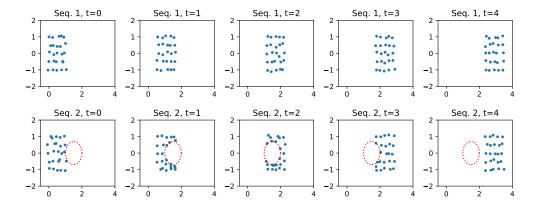


FIGURE 10. Examples of obstructed and unobstructed drone flights. Each row shows 5 snapshots of a 5×5 grid of drones flying through the square $[0,4] \times [-2,2]$. The first row shows unobstructed flight, while the second row shows the drones avoiding an obstacle at the ellipse $\left(\frac{x-1.5}{0.5}\right)^2 + \left(\frac{y}{0.7}\right)^2 = 1$.

of X_i . This process defines a random variable $\mathcal{X} = (X, \mu_X, \Omega, \nu_0, \omega_X)$ valued in pm-nets that we call *clear flight*.

In the presence of an obstacle, a drone avoids collision by moving above the ellipse if its y-coordinate is positive and below the ellipse otherwise. Within each vertical column of drones (i.e. drones with the same x coordinate), those that go above remain evenly spaced between the top of the obstacle and the line y = 1, while those that go below remain evenly spaced between the line y = -1 and the bottom of the obstacle. As before, this process produces 5 grids Y_0, \ldots, Y_4 of drones to which we apply independent Gaussian noise $\mathcal{N}(0,0.05)$, except when it would cause a collision. Let $\overline{\omega}_X^i$ be the distance matrix of Y_i ; the tuple $\overline{\mathcal{X}} = (X, \mu_X, \Omega, \nu_0, \overline{\omega}_X^i)$ defines another random variable valued in pm-nets called obstructed flight.

In our experiment, we sample 10 instances of pm-nets: 5 clear flights $\mathcal{X}_1, \ldots, \mathcal{X}_5$ and 5 obstructed flights $\mathcal{X}_6, \ldots, \mathcal{X}_{10}$. We then apply alternating optimization to minimize $\cot_{\lambda}(\nu)$ for several values of λ , with results summarized in Table 1. As λ increases from 0.01 to 10, the optimal measure ν assigns different weights to the features. For $\lambda = 0.01$, the regularization is too weak, and ν concentrates on the time steps where the drone configurations differ the most (see Fig. 10). In contrast, when $\lambda = 10$, the strong regularization drives ν close to the uniform measure, failing to distinguish between time steps. The most informative range is between 0.1 and 1, where the weights in ν capture the differences between the two flight classes. Specifically, in obstructed flights the drones begin crossing the obstacle at time 1 and nearly clear it by time 3, making times 1 and 2 the most distinctive, time 3 moderately distinctive, and times 0 and 4 indistinguishable. This pattern is reflected in the weights, with ν_1 and ν_2 largest, ν_3 intermediate, and ν_0 and ν_0 smallest.

5.5.2. Supervised Classification. We use the above dataset in a supervised classification experiment. We sample 15 instances of each flight pattern and reserve 5 of each as test set. Using the rest as training set, we obtain the $\nu_{\rm opt}$ that minimizes $\cot_{\lambda}(\nu)$ with $\lambda = 10^{-1}$.

We then classify each entry of the test set by its nearest neighbor under the parametrized GW distance with parameter space $(\Omega, \nu_{\text{opt}})$. For comparison, we build the following ensemble classifier. For every $\mathcal{X} = (X, \mu_X, \Omega, \nu_{\text{opt}}, \omega_X)$ in the training set and every $\mathcal{Y} = (X, \mu_X, \Omega, \nu_{\text{opt}}, \omega_Y)$ in the test set, we can

compute the standard GW distances $\mathsf{GW}_2(\omega_X^t, \omega_Y^t)$ for every $t \in \Omega$. This results in one set of distances between the training and test sets for every $t \in \Omega$, and thus, one nearest neighbor classifier for each $t \in \Omega$ that we call the *t-th GW classifier*. The ensemble classifier then labels each entry of the test set with the most frequent label among the labels given by the GW classifiers.

After repeating the above experiment 10 times, the parametrized GW distance classifier reaches an average of 95% classification accuracy with a 7% standard deviation, while the ensemble classifier (based on standard GW) manages only 87% accuracy with a 8% standard deviation.

5.5.3. Implementation. Minimizing $\cot_{\lambda}(\nu)$ requires solving several optimizations that we describe now. Let $\nu \in \mathcal{P}(\Omega)$. For every $i=1,\ldots,n$, let $\mathcal{X}_i=(X_i,\mu_i,\Omega,\nu,\omega_i)$ be a pm-net in \mathfrak{N}_{ν} with class label y_i such that $1 \leq y_1 \leq \cdots \leq y_n \leq m$. We begin by finding couplings $\pi_{ij} \in \mathcal{C}(\mu_i,\mu_j)$ that realize $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_i,\mathcal{X}_j)$ for $1 \leq i,j \leq n$ using gradient descent as detailed in Sec. 5.1.1. Then we assemble the matrix M_{ν} as in Eqn. (26) and minimize $\cot_{\lambda}(\nu)$ subject to $\nu \in \mathcal{P}(\Omega)$ using gradient descent. We alternate these optimizations until a convergence criterion is satisfied. Each gradient descent starts from the previous optimal coupling (resp. measure).

To complement Sec. 5.1, we record the gradient of $\cot_{\lambda}(\nu)$ with respect to ν . To simplify the presentation below, we assume $\Omega = \{1, \ldots, T\}$. The results below hold for arbitrary $1 \leq p, q < \infty$, but our code only implements the version for p = q = 2.

Lemma 5.5. Let $M \in \mathbb{R}^{n \times n}$ be a matrix with block structure

$$M = \begin{pmatrix} B_{11} & \cdots & B_{1m} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mm} \end{pmatrix}$$

where $B_{ij} \in \mathbb{R}^{n_i \times n_j}$ and $n_1 + \dots + n_m = n$. Let $S(M) := \frac{\|B_{11}\|_1 + \dots + \|B_{mm}\|_1}{\|M\|_1}$. Then:

- If M_{ij} belongs to a block B_{kk} , $\frac{\partial S}{\partial M_{ij}} = \frac{1 S(M)}{\|M\|_1}$.
- Otherwise, $\frac{\partial S}{\partial M_{ij}} = -\frac{S(M)}{\|M\|_1}$.

Proof. Suppose that the entry M_{ij} belongs to the block B_{kk} for some $1 \le k \le m$. Note that $\partial \|B_{kk}\|_1/\partial M_{ij} = 1$ and $\partial \|B_{hh}\|_1/\partial M_{ij} = 0$ for any $h \ne k$. Likewise, $\partial \|M\|_1/\partial M_{ij} = 1$. Then

$$\frac{\partial S}{\partial M_{ij}} = \left[\frac{\partial \|B_{kk}\|_1}{\partial M_{ij}} \cdot \|M\|_1 - \left(\sum_{h=1}^m \|B_{hh}\|_1 \right) \cdot \frac{\partial \|M\|_1}{\partial M_{ij}} \right] / \|M\|_1^2
= \left[1 - \left(\sum_{h=1}^m \|B_{hh}\|_1 \right) / \|M\|_1 \right] \frac{1}{\|M\|_1}
= \frac{1 - S(M)}{\|M\|_1}.$$

If M_{ij} does not belong to any block B_{kk} , then $\partial \|B_{hh}\|_1/\partial M_{ij}=0$ for all $1 \leq h \leq m$ and $\partial \|M\|_1/\partial M_{ij}=1$. Hence

$$\frac{\partial S}{\partial M_{ij}} = \left[0 \cdot \|M\|_1 - \left(\sum_{h=1}^m \|B_{hh}\|_1 \right) \cdot \frac{\partial \|M\|_1}{\partial M_{ij}} \right] / \|M\|_1^2
= -\frac{\sum_{h=1}^m \|B_{hh}\|_1}{\|M\|_1} \cdot \frac{1}{\|M\|_1}
= -\frac{S(M)}{\|M\|_1}.$$

Lemma 5.6. Let $\nu \in \mathcal{P}(\Omega)$, and let $\mathcal{X}, \mathcal{Y} \in \mathfrak{N}_{\nu}$. Let $\pi \in \mathcal{C}(\mu_X, \mu_Y)$ be the coupling that realizes $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_i, \mathcal{X}_i)$. Then for any $1 \leq t \leq T$,

$$\frac{\partial}{\partial \nu_t} \mathsf{GW}_\mathsf{C}(\mathcal{X}_i, \mathcal{X}_j)^p = \mathrm{dis}_p(\pi, \omega_X^t, \omega_Y^t)^p.$$

Proof. When Ω is finite, the equation in Proposition 3.16 becomes $\mathsf{GW}_\mathsf{C}(\mathcal{X},\mathcal{Y})^p = \sum_{s=1}^T \mathrm{dis}_p(\pi,\omega_X^s,\omega_Y^s)^p \cdot \nu_s.$

The result is immediate from here.

Lemma 5.7. Let $\nu \in \mathcal{P}(\Omega)$ and fix $1 \leq p < \infty$. For every $i = 1, \ldots, n$, let $\mathcal{X}_i = (X_i, \mu_i, \Omega, \nu, \omega_i)$ be a pm-net in \mathfrak{N}_{ν} with class label y_i such that $1 \leq y_1 \leq \cdots \leq y_n \leq m$. Let $\pi_{ij} \in \mathcal{C}(\mu_i, \mu_j)$ be the coupling that realizes $\mathsf{GW}_{\mathsf{C}}(\mathcal{X}_i, \mathcal{X}_j)$ and define $M \in \mathbb{R}^{n \times n}$ by $M_{ij} = \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_i, \mathcal{X}_j)^p$. Then for any $1 \leq t \leq T$,

$$\frac{\partial}{\partial \nu_t} \operatorname{cost}_p(\nu) = \sum_{\substack{M_{ij} \in B_{kk} \\ \text{for some } k}} \frac{1 - S(M)}{\|M\|_1} \cdot \operatorname{dis}_p(\pi_{ij}, \omega_i^t, \omega_j^t)^p + \sum_{\substack{M_{ij} \notin B_{kk} \\ \text{for any } k}} -\frac{S(M)}{\|M\|_1} \cdot \operatorname{dis}_p(\pi_{ij}, \omega_i^t, \omega_j^t)^p$$

Proof. Since we define $M_{ij} = \mathsf{GW}_{\mathsf{C}}(\mathcal{X}_i, \mathcal{X}_j)^p$, $\mathsf{cost}_p(\nu) = S(M)$, where S is the function defined in Proposition 5.5. Hence, using the chain rule and Proposition 5.5 and Proposition 5.6 yields

$$\frac{\partial}{\partial \nu_{t}} \operatorname{cost}_{p}(\nu) = \sum_{i,j=1}^{n} \frac{\partial S}{\partial M_{ij}} \cdot \frac{\partial M_{ij}}{\partial \nu_{t}} = \sum_{i,j=1}^{n} \frac{\partial S}{\partial M_{ij}} \cdot \frac{\partial}{\partial \nu_{t}} \operatorname{GW}_{\mathsf{C}}(\mathcal{X}_{i}, \mathcal{X}_{j})^{p}$$

$$= \sum_{\substack{M_{ij} \in B_{kk} \\ \text{for some } k}} \frac{1 - S(M)}{\|M\|_{1}} \cdot \operatorname{dis}_{p}(\pi_{ij}, \omega_{i}^{t}, \omega_{j}^{t})^{p} + \sum_{\substack{M_{ij} \notin B_{kk} \\ \text{for any } k}} -\frac{S(M)}{\|M\|_{1}} \cdot \operatorname{dis}_{p}(\pi_{ij}, \omega_{i}^{t}, \omega_{j}^{t})^{p}$$

- 5.6. **Discussions, Limitations, and Future Work.** In this section, we demonstrated four applications of the parameterized GW distance. The utility of the parameterized GW framework in each case is summarized as follows.
- (1) **Block matrices**: It defines distances by decomposing each block matrix into submatrices and combining information from all parts (Sec. 5.2).
- (2) Random graphs: It serves as a meaningful invariant for comparing and clustering random graph samples, tested on Erdős–Rényi and stochastic block models (Sec. 5.3).
- (3) Heat kernels: By aligning sets of heat kernels, it captures features across multiple timescales (e.g., cycles of different lengths or nesting structures), with ν tuned for effective couplings (Sec. 5.4).
- (4) Feature selection: Allowing ν to vary enables identification of discriminative time intervals in dynamic data, illustrated by classifying obstructed vs. unobstructed drone flights (Sec. 5.5).

In the block matrix experiments, the computation of GW_C ignores interactions between blocks X_i and Y_j for $i \neq j$, and consequently does not yield the expected optimal alignment. Further work is needed to determine the *minimal* interaction information required to approach the global optimum. In the heat kernel and feature selection experiments, manual tuning of the parameter ν is necessary both to capture features across scales and to identify discriminative time intervals.

We remark that the focus of this paper is on fundamental theory and establishing core methods for its application. In particular, the experiments above are qualitative in nature and only deal with synthentic data. An important direction for future research is the development of more efficient learning frameworks that support automatic parameter tuning, so that these methods are more applicable to real-world data. Applications to time-varying metric space and network data, e.g., in the form of longitudinal fMRI data, will be the goal of a followup project.

ACKNOWLEDGMENTS

This work was partially supported by grants from National Science Foundation (NSF) projects IIS-2145499, DMS-2324962, and CIF-2526630, and Department of Energy (DOE) projects DE-SC0023157 and DE-SC0021015. We thank Cédric Vincent-Cuaz for clarifying the Python Optimal Transport library and for identifying errors in our initial implementation.

References

- [1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows: in metric spaces and in the space of probability measures. Lectures in Mathematics ETH Zürich. Birkhäuser, 2005.
- [2] Hajer Bahouri, Jean-Yves Chemin, and Raphaël Danchin. Fourier analysis and nonlinear partial differential equations, volume 343 of Grundlehren der mathematischen Wissenschaften. Springer, Jan 2011.
- [3] Martin Bauer, Nicolas Charon, Eric Klassen, Sebastian Kurtek, Tom Needham, and Thomas Pierron. Elastic metrics on spaces of euclidean curves: Theory and algorithms. *Journal of Nonlinear Science*, 34(3):56, 2024.
- [4] Martin Bauer, Facundo Mémoli, Tom Needham, and Mao Nishino. The Z-Gromov-Wasserstein distance. arXiv preprint arXiv:2408.08233, 2024.
- [5] Marc Benkert, Joachim Gudmundsson, Florian Hübner, and Thomas Wolle. Reporting flock patterns. Computational Geometry, 41(3):111–125, 2008.
- [6] Jacob Billings, Manish Saggar, Jaroslav Hlinka, Shella Keilholz, and Giovanni Petri. Simplicial and topological descriptions of human brain dynamics. Network Neuroscience, 5(2):549–568, 2021.
- [7] Luis L Bonilla, Ana Carpio, and Carolina Trenado. Tracking collective cell motion by topological data analysis. PLOS Computational Biology, 16(12):e1008407, 2020.
- [8] Samir Chowdhury and Facundo Mémoli. The Gromov-Wasserstein distance between networks and stable network invariants. Information and Inference: A Journal of the IMA, 8(4):757–787, 11 2019.
- [9] Samir Chowdhury and Tom Needham. Generalized spectral clustering via Gromov-Wasserstein learning. In Arindam Banerjee and Kenji Fukumizu, editors, Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pages 712–720. PMLR, 13–15 Apr 2021.
- [10] Samuel Cohen, Giulia Luise, Alexander Terenin, Brandon Amos, and Marc Deisenroth. Aligning time series on incomparable spaces. In *International conference on artificial intelligence and statistics*, pages 1036–1044. PMLR, 2021.
- [11] Gordon M Crippen, Timothy F Havel, et al. Distance geometry and molecular conformation, volume 74. Research Studies Press Taunton, 1988.
- [12] Paul Erdös and Alfréd Rényi. On random graphs I. Publicationes Mathematicae Debrecen, 6:290-297, 1959.
- [13] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. Journal of Machine Learning Research, 22(78):1–8, 2021.
- [14] Nicolas Fournier. Convergence of the empirical measure in expected Wasserstein distance: non-asymptotic explicit bounds in \mathbb{R}^d . ESAIM: Probability and Statistics (ESAIM: P&S), 27:749–775, 2023.
- [15] Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. Michigan Mathematical Journal, 31(2):231–240, 1984.
- [16] Derek Greene, Donal Doyle, and Padraig Cunningham. Tracking the evolution of communities in dynamic social networks. In 2010 international conference on advances in social networks analysis and mining, pages 176–183. IEEE, 2010.
- [17] Mikhail Gromov. Metric structures for Riemannian and non-Riemannian spaces. Springer Science & Business Media, 2007.
- [18] Remco van der Hofstad. Random Graphs and Complex Networks, volume 1 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016.
- [19] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social networks, 5(2):109–137, 1983.
- [20] Yan Huang, Cai Chen, and Pinliang Dong. Modeling herds and their evolvements from trajectory data. In *International Conference on Geographic Information Science*, pages 90–105. Springer, 2008.
- [21] Hoyoung Jeung, Man Lung Yiu, Xiaofang Zhou, Christian S Jensen, and Heng Tao Shen. Discovery of convoys in trajectory databases. Proceedings of the VLDB Endowment, 1(1):1068–1080, 2008.
- [22] Niles Johnson and Donald Yau. 2-Dimensional Categories. arXiv preprint arXiv:2002.06055, 01 2021.
- [23] Woojin Kim. The Persistent Topology of Dynamic Data. PhD thesis, The Ohio State University, 2020.
- [24] Woojin Kim and Facundo Mémoli. Spatiotemporal persistent homology for dynamic metric spaces. Discrete & Computational Geometry, 66(3):831–875, 2021.
- [25] Woojin Kim, Facundo Mémoli, and Zane Smith. Analysis of dynamic graphs and dynamic metric spaces via zigzag persistence. In Topological Data Analysis: The Abel Symposium 2018, pages 371–389. Springer, 2020.
- [26] Patrick K McFaddin and Tom Needham. Interleaving distances, monoidal actions and 2-categories. To Appear: Algebraic and Geometric Topology. arXiv preprint arXiv:2311.11936, 2023.
- [27] Facundo Mémoli. On the use of Gromov-Hausdorff distances for shape comparison. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, Eurographics Symposium on Point-Based Graphics. The Eurographics Association, 2007.
- [28] Facundo Mémoli. Spectral Gromov-Wasserstein distances for shape matching. In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pages 256–263. IEEE, 2009.
- [29] Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. Foundations of Computational Mathematics, 11(4):417-487, 8 2011.
- [30] Facundo Mémoli. A spectral notion of Gromov-Wasserstein distance and related methods. Applied and Computational Harmonic Analysis, 30(3):363-401, 2011.

- [31] Facundo Mémoli and Tom Needham. Comparison results for Gromov-Wasserstein and Gromov-Monge distances. European Series in Applied and Industrial Mathematics: Control, Optimisation and Calculus of Variations (ESAIM: COCV), 30:78, 2024.
- [32] Elizabeth Munch. Applications of persistent homology to time varying systems. PhD thesis, Duke University, 2013.
- [33] Kyle C Nguyen, Carter D Jameson, Scott A Baldwin, John T Nardini, Ralph C Smith, Jason M Haugh, and Kevin B Flores. Quantifying collective motion patterns in mesenchymal cell populations using topological data analysis and agent-based modeling. Mathematical biosciences, 370:109158, 2024.
- [34] Maks Ovsjanikov, Quentin Mérigot, Facundo Mémoli, and Leonidas Guibas. One point isometric matching with the heat kernel. Computer Graphics Forum, 29(5):1555–1564, 2010.
- [35] Julia K Parrish, William Hamner, and William M Hamner. Animal groups in three dimensions: how species aggregate. Cambridge University Press, 1997.
- [36] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, pages 2664-2672. JMLR.org, 2016.
- [37] Dan Raviv, Michael M Bronstein, Alexander M Bronstein, and Ron Kimmel. Volumetric heat kernel signatures. In Proceedings of the ACM workshop on 3D object retrieval, pages 39–44, 2010.
- [38] Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Structured transforms across spaces with cost-regularized optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 586–594. PMLR, 2024.
- [39] Joakim Skarding, Bogdan Gabrys, and Katarzyna Musial. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. IEEE Access, 9:79143-79168, 2021.
- [40] Anuj Srivastava and Eric P Klassen. Functional and shape data analysis, volume 1. Springer, 2016.
- [41] Karl-Theodor Sturm. On the geometry of metric measure spaces. I. Acta Mathematica, 196:65–131, 2006.
- [42] Karl-Theodor Sturm. Super-Ricci flows for metric measure spaces. Journal of Functional Analysis, 275(12):3504–3569, 2018
- [43] Karl-Theodor Sturm. The Space of Spaces: Curvature Bounds and Gradient Flows on the Space of Metric Measure Spaces, volume 290. Memoirs of the American Mathematical Society, 2023.
- [44] David JT Sumpter. Collective animal behavior. In Collective animal behavior. Princeton University Press, 2010.
- [45] Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal transport for structured data with application on graphs. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 6275–6284. PMLR, 09–15 Jun 2019.
- [46] M Ulmer, Lori Ziegelmeier, and Chad M Topaz. A topological approach to selecting models of biological experiments. PloS one, 14(3):e0213679, 2019.
- [47] Cédric Villani. Optimal Transport: Old and New. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 2009.
- [48] Cédric Villani. Topics in optimal transportation, volume 58. American Mathematical Soc., 2021.
- [49] Cédric Vincent-Cuaz. Python Optimal Transport: Fused Gromov-Wasserstein conditional gradient solver. https://github.com/cedricvincentcuaz/cedricvincentcuaz.github.io/blob/master/POT/FGW___POT.pdf. Accessed: 2025-08-15.
- [50] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. Nature, 393(6684):440-442, 1998.
- [51] Jonathan Daniel Weed. Statistical problems in transport and alignment. PhD thesis, MIT, 2019.
- [52] Lu Xian, Henry Adams, Chad M Topaz, and Lori Ziegelmeier. Capturing dynamics of time-varying data via topology. Foundations of Data Science, 4(1):1–36, 2022.
- [53] Jaejun Yoo, Eun Young Kim, Yong Min Ahn, and Jong Chul Ye. Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. *Journal of Neuroscience Methods*, 267:1–13, 2016.
- [54] Wayne W Zachary. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33(4):452–473, 1977.