Journal of Machine Learning Research 27 (2026) 1-79 Submitted 4/25; Published 8/26

Weisfeiler-Lehman Meets Optimal Transport: Theory, Algorithms, and Implications

Samantha Chen SAC003@UCSD.EDU
Department of Computer Science and Engineering
University of California San Diego
La Jolla, CA 92093, USA

Sunhyuk Lim LSH3109@SKKU.EDU
Department of Mathematics
Sungkyunkwan University (SKKU)
Suwon-si, Gyeonggi-do, 16419, Republic of Korea

Facundo Mémoli FACUNDO.MEMOLI@RUTGERS.EDU
Department of Mathematics
Rutgers University
Piscataway, NJ 08854, USA

Zhengchao Wan* ZWAN@MISSOURI.EDU
Department of Mathematics
University of Missouri
Columbia, MO 65211, USA

Yusu Wang YUSUWANG@UCSD.EDU
Department of Computer Science and Engineering & Halıcıoğlu Data Science Institute
University of California San Diego
La Jolla, CA 92093, USA

Editor: Jian Tang

Abstract

We introduce the Weisfeiler–Lehman (WL) distance, a pseudometric defined on labeled measure Markov chains (LMMCs)—a general class of structures encompassing labeled graphs. This distance is computable in polynomial time and is closely aligned with the classical WL graph isomorphism test: it assigns a positive value if and only if the WL test distinguishes the two input graphs.

The WL distance captures fine structural differences in the underlying LMMCs and is, consequently, more discriminative than the state-of-the-art Wasserstein WL graph kernel. The distance motivates a neural network architecture on LMMCs, which we call *Markov chain neural networks (MCNNs)*, that is universal for continuous functions over LMMCs (hence all graphs) under the topology induced by the WL distance. This finding further extends to the analysis of graph neural networks (GNNs).

Additionally, the WL distance admits an alternative characterization through Markovian couplings and Causal Optimal Transport (COT), offering a stochastic-process perspective that uncovers a surprising connection between the seemingly unrelated WL test and COT. Finally, we show that the WL distance is stable under a natural variant of the

*. Corresponding author.

©2026 Samantha Chen, Sunhyuk Lim, Facundo Mémoli, Zhengchao Wan, and Yusu Wang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-0947.html.

Gromov-Wasserstein (GW) distance that we propose. The GW distance is generally NP-hard to compute and hence the WL distance provides a polynomial-time computable lower bound for it.

Keywords: Weisfeiler-Lehman test, optimal transport, causal optimal transport, Gromov-Wasserstein, graph neural networks

Contents

1 Introduction **2**
1.1 Contributions and Connections to Related Work 3

2 Preliminaries **7**
2.1 Notation 7
2.2 The Weisfeiler-Lehman Test 7
2.3 Optimal Transport 8
2.4 Markov Chains 9

3 The Weisfeiler-Lehman Distance **11**
3.1 Properties of the WL Distance 13
3.2 An Algorithm and a Computationally Efficient Lower Bound 16
3.3 Wasserstein Weisfeiler-Lehman Graph Kernel and Relationship with d_{WL} . 18
3.4 Graph Classification Experiments 21

4 Markov Chain Neural Networks and Graph Neural Networks **22**
4.1 Markov Chain Neural Networks 23
4.2 Implications for Message Passing GNNs 27

5 Reinterpretation of the WL Distance via Markovian Couplings **29**
5.1 Markovian Couplings and a Stochastic Process Interpretation of the WL Distance 29
5.2 The Connection with Causal Optimal Transport 31

6 Relationship with the Gromov Wasserstein Distance **33**
6.1 The GW Distance between MCMSs 33
6.2 The WL Distance vs. the GW Distance 36

7 Conclusion and Future Directions **38**

1. Introduction

The Weisfeiler-Lehman (WL) test (Lehman and Weisfeiler, 1968) is a classical procedure which provides a polynomial time proxy for testing graph isomorphism. It is efficient, capable of distinguishing most pairs of graphs in linear time (Babai and Kucera, 1979; Babai and Luks, 1983). The WL test is closely related to graph neural networks (GNNs), both in the design of GNN architectures and in terms of characterizing their expressive power. For example, Xu et al. (2019) showed that graph isomorphism networks (GINs) have the same

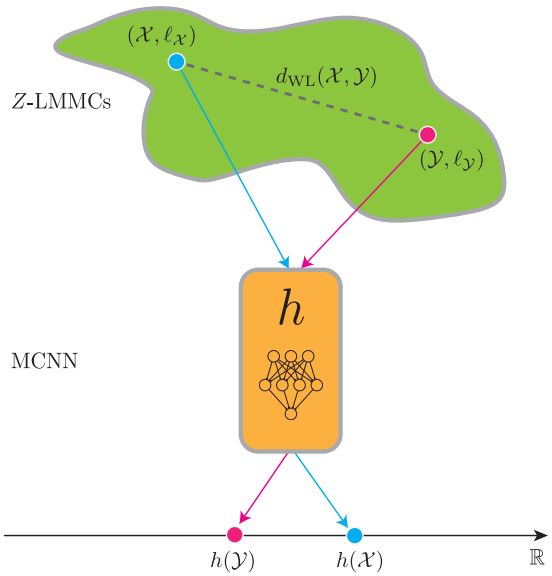


Figure 1: The WL distance metrizes the collection of all labeled measure Markov chains (LMMCs). Markov chain neural networks (MCNNs) are real valued functions h, with domain equal to the collection of all LMMCs, which are universal in the collection of functions that are continuous relative to the topology generated by the WL distance.

discriminative power as the WL test in ascertaining whether two graphs are isomorphic or not. Recently, Azizian and Lelarge (2021) demonstrated that message passing graph neural networks (MPNNs) are universal approximators for continuous functions defined on the set of graphs with bounded number of vertices, equipped with a topology induced by a specific variant of the graph edit distance, provided that the functions' discriminative power does not exceed that of the Weisfeiler-Lehman test.

However, while the WL test is well-suited for testing graph isomorphism, it does not directly support a *quantitative comparison* of graphs. This state of affairs naturally suggests identifying a distance function between graphs so that two graphs have positive distance if and only if they can be distinguished by the WL test. Although several WL-inspired graph kernels have been proposed to quantitatively compare graphs (Shervashidze et al., 2011; Togninalli et al., 2019), they either fail to handle continuous node features naturally, or do not have the same discriminative power as the WL test.

1.1 Contributions and Connections to Related Work

Building on our earlier work (Chen et al., 2022, 2023), this journal version provides a unified treatment of the Weisfeiler-Lehman distance, combining its iterative construction with its interpretation through Markovian couplings and causal optimal transport. This framework connects WL distinguishability with Lipschitz stability and universal approximation for Markov chain and message-passing graph neural networks, and with Gromov-Wasserstein comparisons of Markov chains. The additions in this version include extending the WL measure hierarchy to compact state spaces and establishing its well-definedness (Proposi-

tion 11), together with an analysis of the sharpness and limitations of the WL-based GW lower bounds (Remark 66). In particular, we exhibit equality cases and examples where these lower bounds vanish despite a positive GW distance.

We organize this unified treatment around four main themes:

(1) A new distance with the same discriminative power as the WL test. We introduce the *Weisfeiler-Lehman distance* by merging the core iterative idea of the WL test (which provides a 0 or 1 answer) with optimal transport (OT) (Villani, 2009). This new distance quantifies the WL test in the following sense:

- two graphs are at zero WL distance if and only if they cannot be distinguished by the WL test; and
- when the distance is positive, its value can be interpreted as the *degree to which the two graphs fail to pass the WL test*.

Moreover, the WL distance can be computed in polynomial time. Furthermore, our WL distance is actually defined on a more general and flexible type of objects called the *labeled measure Markov chains* (LMMCs), of which labeled graphs (i.e, graph with node features) are special cases; see Figure 1. LMMCs can naturally model the interaction between graphs and their node labels (node features), and thus provides a new OT based perspective for comparing node labeled graphs. Besides graphs, the LMMC framework also encompasses continuous objects such as Riemannian manifolds and graphons. It is worth noting that the idea of combining OT and Markov chains/heat kernels has long been used to study notions of curvature of geometric objects, e.g., graphs or Riemannian manifolds (von Renesse and Sturm, 2005; Ollivier, 2009) and has been utilized to quantify the dissimilarity between shapes in Mémoli (2009, 2011b).

We further establish an efficiently computable lower bound for the WL distance and carry out experiments to demonstrate the effectiveness of our WL distance (and its lower bounds) in graph comparison tasks.[1] These experiment validate that our WL distance provides a meaningful *quantitative* measure of graph dissimilarity, thereby extending beyond the binary classification given by the standard WL isomorphism test.

(2) Connection with graph kernels and graph neural networks. We identify a neural network architecture on the collection of all LMMCs, which we call MCNNs (for Markov chain neural networks). We show that MCNNs inherits the discrimination strength of the WL test when restricted to graphs. Simultaneously, this architecture universally approximates continuous functions on the LMMC space equipped with the topology generated by the WL distance; see Figure 1. Because MCNNs are closely related to graph neural networks (Scarselli et al., 2008), these results transfer to standard GNNs as well. Moreover, from MCNNs one can recover Weisfeiler-Lehman graph kernels (Togninalli et al., 2019), and we show that a key pseudometric in (Togninalli et al., 2019) is a strict lower bound for the WL distance (see Section 3.3), implying that the WL distance is strictly more discriminative than WWL graph kernels.

1. Our code is available at https://github.com/chens5/WL-distance

(3) Reinterpretation via Markovian couplings and causal optimal transport. Interestingly, the WL distance admits an alternative characterization that does not directly rely on the mechanics of the WL test: it can be understood through *Markovian couplings* and *causal optimal transport* (COT) (Lassalle, 2018; Backhoff et al., 2017).

Specifically, we demonstrate that the WL distance can be characterized by the minimal expected label discrepancy under pairs of Markov processes evolving jointly. This stochastic-process perspective not only provides an independent way to characterize the WL distance but also paves the way for leveraging existing COT methods in related algorithmic settings.

(4) Connections to Gromov-Wasserstein distance. Finally, we observe that our formulation of the WL distance resembles the Gromov-Wasserstein (GW) distance (Mémoli, 2007, 2011a; Peyré et al., 2016; Sturm, 2023; Vayer et al., 2020a; Chowdhury and Mémoli, 2019; Arya et al., 2024; Bauer et al., 2025) which is an OT-based distance between *metric measure spaces* which has been recently widely used in shape matching and machine learning. Exploring this, we devise a specific variant of the GW distance which can be used to compare *Markov chain metric spaces* (MCMSs, which includes all graphs). Our version of the GW distance implements a certain *multiscale comparison* of MCMSs and vanishes only when two MCMSs are isomorphic. Similarly to the standard GW distance, our version leads to NP-hard problems and, interestingly, it turns out that the poly-time computable WL distance is not only stable w.r.t. (i.e., upper bounded by) this variant of the GW distance, but can also be construed as a variant of the *third lower bound* of the GW distance described in (Mémoli, 2007, 2011a).

After the publication of a conference version of this work in 2022 (Chen et al., 2022), several concurrent and subsequent studies have emerged addressing the comparison of graphs (or graphons) via the WL test, employing approaches closely related to our construction of the WL distance. We summarize some of these in Table 1.

Organization of the paper. In Section 2, we begin with preliminaries, covering the Weisfeiler-Lehman (WL) test and optimal transport. In Section 3, we define the WL distance, explore its key properties, and present an algorithm for its computation. Additionally, we compare the WL distance with the Weisfeiler-Lehman Optimal Transport (WWL) graph kernel both theoretically and empirically. In Section 4, we introduce the Markov Chain Neural Network (MCNN) and leverage the WL distance to analyze its distinguishing and expressive capabilities. In Section 5, we reinterpret the WL distance using Markovian couplings, highlighting its deep connection to causal optimal transport. This reinterpretation serves as a unifying framework, underpinning most of the proofs in the paper. In Section 6, we extend our discussion to the Gromov-Wasserstein distance for Markov chains and elucidate its relationship with the WL distance. Finally, in Section 7 we discuss about future research plans. All proofs are deferred to the appendix.

2. This work is a collaboration between the 4th and 5th authors and Tristan Brugère.

Paper and follow-up work	**Key Contributions and Relation to the WL Distance**
Tree Mover's Distance Chuang and Jegelka (2022)	Proposes the Tree Mover's Distance (TMD) as a pseudometric on attributed graphs. Bounds the Lipschitz constant of GNNs with respect to TMD and provides a generalization bound for GNNs under distribution shifts. Combines the 1-WL test with optimal transport techniques in a way reminiscent of our WL distance. Follow up work by Vasileiou et al. (2025) uses graph pseudometrics, including a forest formulation of TMD, to derive generalization bounds for message-passing GNNs.
Graphon Distance Böker et al. (2023)	Integrates MPNNs into iterated degree measures and defines metrics based on Prokhorov and Wasserstein metrics. Shows that two graphons are close if MPNN outputs with bounded Lipschitz constants are close. Proves a two-way equivalence between graph similarity in the proposed metrics and MPNN outputs. The construction is inspired on our iterative definition of the WL distance.
Optimal Transport Markov Distance Brugère et al. (2024)	Proposes the Optimal Transport Markov (OTM) framework for comparing Markov chains, extending the WL distance by comparing chains over multiple steps rather than at a single step. A notable instance, the discounted WL distance, is identified within this framework and can be differentiated easily, which is advantageous for learning algorithms.[2]

Table 1: Summary of subsequent or follow-up works related to the WL Distance.

2. Preliminaries

2.1 Notation

We use $\{\cdot\}$ to denote a set and $\{\!\{\cdot\}\!\}$ to denote a multiset.

For any measurable space Z, we will denote by $\mathcal{P}(Z)$ the collection of all probability measures on Z. When Z is a metric space (Z, d_Z), we further require that every $\alpha \in \mathcal{P}(Z)$ has *finite 1-moment*, i.e., $\int_Z d_Z(z, z_0)\alpha(dz) < \infty$ for any fixed $z_0 \in Z$. When X is a finite set and $\alpha \in \mathcal{P}(X)$, we let $\alpha(x) \coloneqq \alpha(\{x\})$ for any $x \in X$.

Given two measurable spaces X and Y and a measurable map $\psi : X \to Y$, the *pushforward* map induced by ψ is the map $\psi_\# : \mathcal{P}(X) \to \mathcal{P}(Y)$ sending α to $\psi_\#\alpha$ where for any measurable $B \subseteq Y$, $\psi_\#\alpha(B) \coloneqq \alpha\left(\psi^{-1}(B)\right)$. If X is compact, Y is a metric space, and ψ is continuous, then $\psi_\#\alpha$ obviously has finite 1-moment and is thus an element of $\mathcal{P}(Y)$.

2.2 The Weisfeiler-Lehman Test

A labeled graph is a finite graph $G = (V_G, E_G)$ endowed with a *label function* $\ell_G : V_G \to Z$, where the labels (i.e., node features) are taken from some set Z. Common label functions include the degree label (i.e., $\ell_G : V_G \to \mathbb{N}$ sends each $v \in V_G$ to its degree, denoted by $\deg_G(v)$) and the constant label (assigning a constant to all vertices). For a node $v \in V_G$, let $N_G(v)$ denote the set of neighbors of v in G. Below, we recall the Weisfeiler-Lehman test (Lehman and Weisfeiler, 1968) in our terminology. In particular, we define the *Weisfeiler-Lehman hierarchy* for a labeled graph (G, ℓ_G) as follows.

Definition 1 (Weisfeiler-Lehman (WL) hierarchy) *Given any labeled graph (G, ℓ_G), we define the following hierarchy of maps, which we call the Weisfeiler-Lehman hierarchy:*

Step 0. *For each $v \in V_G$, we define*

$$\ell^{(0)}_{(G,\ell_G)}(v) \coloneqq \ell_G(v).$$

Step 1. *For each $v \in V_G$, we compute the pair*

$$\ell^{(1)}_{(G,\ell_G)}(v) \coloneqq \left(\ell_G(v), \left\{\!\!\left\{\ell_G(v') : v' \in N_G(v)\right\}\!\!\right\}\right).$$

...

Step k. *For each $v \in V$, we compute the pair*

$$\ell^{(k)}_{(G,\ell_G)}(v) \coloneqq \left(\ell^{(k-1)}_{(G,\ell_G)}(v), \left\{\!\!\left\{\ell^{(k-1)}_{(G,\ell_G)}(v') : v' \in N_G(v)\right\}\!\!\right\}\right).$$

...

In the literature (e.g. Shervashidze et al. (2011)), $\ell^{(k)}_{(G,\ell_G)}(v)$ is usually mapped into a common space of labels such as $\mathbb{N}$ through a hash function, a step which we omit in this paper.

We now induce, at each step $k \geq 0$, a multiset consisting of the images of $\ell^{(k)}_{(G,\ell_G)}$:

$$L_k((G, \ell_G)) \coloneqq \left\{\!\!\left\{\ell^{(k)}_{(G,\ell_G)}(v) : v \in V_G\right\}\!\!\right\}.$$

Definition 2 (Weisfeiler-Lehman test) *Let (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) be two labeled graphs. For each integer $k \geq 0$, check whether $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$. If there exists any $k \geq 0$ such that $L_k((G_1, \ell_{G_1})) \neq L_k((G_2, \ell_{G_2}))$, we conclude that the two labeled graphs are non-isomorphic; otherwise we say that the two labeled graphs* pass the WL test *and that the two graphs are "possibly isomorphic".*

2.3 Optimal Transport

For measurable spaces X and Y, given $\alpha \in \mathcal{P}(X)$ and $\beta \in \mathcal{P}(Y)$, $\gamma \in \mathcal{P}(X \times Y)$ is called a *coupling* between α and β if $(p_X)_{\#}\gamma = \alpha$ and $(p_Y)_{\#}\gamma = \beta$, where $p_X : X \times Y \to X$ and $p_Y : X \times Y \to Y$ are the canonical projections. For example, the product measure $\alpha \otimes \beta$ is one such coupling. Let $\mathcal{C}(\alpha, \beta)$ denote the set of all couplings between α and β.

Given a metric space (Z, d_Z), for $\alpha, \beta \in \mathcal{P}(Z)$, we define the ($\ell^1$-)Wasserstein distance between them as follows:

$$d_{\mathrm{W}}(\alpha, \beta) \coloneqq \inf_{\gamma \in \mathcal{C}(\alpha,\beta)} \int_{Z \times Z} d_Z(z, z')\gamma(dz \times dz').$$

By (Villani, 2009, Theorem 4.1), the infimum is always attained by some $\gamma \in \mathcal{C}(\alpha, \beta)$ which we call an *optimal coupling* between α and β.

> We will adopt the probabilistic perspective for optimal transport: rather than focusing on probability measures directly, we consider random variables distributed according to these probability measures. This approach affords more intuitive descriptions when describing some of our constructions.

A probabilistic description of the Wasserstein distance. A random variable $\boldsymbol{X}$ valued in a metric space Z is any measurable map $\boldsymbol{X} : (\Omega, \mathbb{P}) \to Z$ where $(\Omega, \mathbb{P})$ is a probability space. Recall that the law of $\boldsymbol{X}$ is defined as $\mathrm{law}(\boldsymbol{X}) \coloneqq \boldsymbol{X}_{\#}\mathbb{P}$. It is well known that for any probability measure α on a complete and separable metric space Z, there exists a random variable $\boldsymbol{X}$ with law α. We now reinterpret the notion of coupling in terms of random variables (see also Den Hollander (2012) for details).

Definition 3 (Coupling in terms of random variables) *Given two probability measures α, β on a complete and separable metric space Z, we call any pair of random variables $(\boldsymbol{X}, \boldsymbol{Y}) : (\Omega, \mathbb{P}) \to Z \times Z$ a coupling between α and β if* $\mathrm{law}(\boldsymbol{X}) = \alpha$ *and* $\mathrm{law}(\boldsymbol{Y}) = \beta$.

Of course, given a coupling $(\boldsymbol{X}, \boldsymbol{Y})$, one has that $\mathrm{law}((\boldsymbol{X}, \boldsymbol{Y})) \in \mathcal{C}(\alpha, \beta)$. Conversely, note that given any (measure-theoretic) coupling $\pi \in \mathcal{C}(\alpha, \beta)$, one can always find a coupling $(\boldsymbol{X}, \boldsymbol{Y})$ between α and β such that $\mathrm{law}\,((\boldsymbol{X}, \boldsymbol{Y})) = \pi$.

Now, given two probability measures α and β on a metric space Z, the Wasserstein distance between α and β can be rewritten using the language of random variables as follows:

$$d_{\mathrm{W}}(\alpha, \beta) = \inf_{(\boldsymbol{X}, \boldsymbol{Y})} \mathbb{E}\, d_Z(\boldsymbol{X}, \boldsymbol{Y}), \tag{1}$$

where the infimum is taken over all couplings $(\boldsymbol{X}, \boldsymbol{Y})$ between α and β.

Hierarchy of Probability Measures. An important ingredient in this paper is the following construction: Given a compact space X and a *metric space* Z, a map of the form $\psi : X \to \mathcal{P}(Z)$ induces $\psi_{\#} : \mathcal{P}(X) \to \mathcal{P}(\mathcal{P}(Z))$ which involves the space of probability measures over probability measures, i.e., $\mathcal{P}(\mathcal{P}(Z))$. Inductively, we define the family of spaces $\mathcal{P}^{\circ k}(Z)$, called the *hierarchy of probability measures*:

1. $\mathcal{P}^{\circ 0}(Z) \coloneqq Z$;
2. $\mathcal{P}^{\circ(k+1)}(Z) \coloneqq \mathcal{P}\left(\mathcal{P}^{\circ k}(Z)\right)$ for $k \geq 0$.

If Z is complete and separable then, when endowed with d_{W}, $\mathcal{P}(Z)$ is also complete and separable (Villani (2009, Theorem 6.18)). By induction, for each $k \in \mathbb{N}$, $\mathcal{P}^{\circ k}(Z)$ is also a complete and separable metric space. These hierarchies will be critical in our development of the WL distance (cf. Definition 10).

The Gromov-Wasserstein Distance. A *metric measure space* (MMS) is a triple $\mathbf{X} = (X, d_X, \mu_X)$, where (X, d_X) is a metric space and μ_X is a (Borel) probability measure on X with full support. Given any $\mathbf{X} = (X, d_X, \mu_X)$ and $\mathbf{Y} = (Y, d_Y, \mu_Y)$, for any coupling $(\boldsymbol{X}, \boldsymbol{Y})$ between μ_X and μ_Y, we define its (ℓ^1-)distortion as follows:

$$\mathrm{dis}((\boldsymbol{X}, \boldsymbol{Y})) \coloneqq \mathbb{E}|d_X(\boldsymbol{X}, \bar{\boldsymbol{X}}) - d_Y(\boldsymbol{Y}, \bar{\boldsymbol{Y}})|,$$

where $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}})$ denotes an independent copy of $(\boldsymbol{X}, \boldsymbol{Y})$, i.e., an independent random variable with the same law as $(\boldsymbol{X}, \boldsymbol{Y})$. It is straightforward to check that the above distortion is independent of the choice of independent copies and hence well-defined.

Then, the (ℓ^1-)Gromov-Wasserstein (GW) distance between $\mathbf{X} = (X, d_X, \mu_X)$ and $\mathbf{Y} = (Y, d_Y, \mu_Y)$ can be expressed as follows

$$d_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) \coloneqq \inf_{(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{C}(\mu_X, \mu_Y)} \mathrm{dis}((\boldsymbol{X}, \boldsymbol{Y})), \tag{2}$$

where we omit the usual $\frac{1}{2}$ factor from its original definition in (Mémoli, 2011a) for simplicity.

2.4 Markov Chains

Given a *compact* space X, we call any continuous map $m_\bullet^X : X \to \mathcal{P}(X)$ (where $\mathcal{P}(X)$ is endowed with the weak topology) a *Markov kernel* on X. Note that when X is finite, Markov kernels can be represented as transition matrices. A probability measure $\mu_X \in \mathcal{P}(X)$ is called a *stationary distribution w.r.t.* $m_\bullet^X$ if for every measurable subset $A \subseteq X$ we have:

$$\mu_X(A) = \int_X m_x^X(A) \mu_X(dx).$$

A *measure Markov chain (MMC)* is any tuple $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ where X is a compact space, $m_\bullet^X$ is a Markov kernel on X and μ_X is a fully supported stationary distribution w.r.t. $m_\bullet^X$.

Remark 4 (Existence of stationary distributions) *When X is finite, the existence of stationary distributions is guaranteed by the Perron-Frobenius Theorem (Saloff-Coste, 1997, Theorem 1.2.1). For a general compact X, one can prove that there exists a stationary*

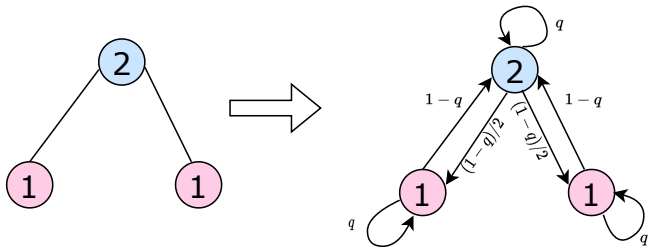


Figure 2: **Illustration of a graph induced LMMC.**

distribution μ_X by the fact that $\mathcal{P}(X)$ is compact (cf. Prokhorov's theorem (Billingsley, 2013, Theorem 5.1)) and using a Krylov-Bogoliubov type argument (Da Prato and Zabczyk, 1996, Theorem 3.1.1).

Definition 5 (Labeled measure Markov chain) *Given any metric space Z, which we refer to as the* metric space of labels, *a Z-*labeled measure Markov chain *((Z-)LMMC for short) is a tuple $(\mathcal{X}, \ell_X)$ where $\mathcal{X}$ is a MMC and $\ell_X : X \to Z$ is a* continuous *map. For technical reasons, throughout this paper, we assume that the metric space of labels Z is* complete *and* separable[3]. *We let $\mathcal{M}^L(Z)$ denote the collection of all Z-LMMCs.*

We define the notion of isomorphism between LMMCs as follows:

Definition 6 (Isomorphism of LMMCs) *Two Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ are said to be* isomorphic *if there exists a homeomorphism $\psi : X \to Y$ such that for all $x \in X$*

- $\ell_X(x) = \ell_Y(\psi(x))$;
- $\psi_\# m_x^X = m_{\psi(x)}^Y$ *and* $\psi_\# \mu_X = \mu_Y$.

Labeled Graphs as LMMCs. Any finite labeled graph induces a family of LMMCs as follows; see also Figure 2 for an illustration.

Definition 7 (q-Markov chains on graphs) *For any graph G and parameter $q \in [0, 1)$, we define the q-Markov chain $m_\bullet^{G,q}$ associated to G as follows: for any $v \in V_G$,*

$$m_v^{G,q} := \begin{cases} q\,\delta_v + \frac{1-q}{\deg_G(v)} \sum_{v' \in N_G(v)} \delta_{v'}, & N_G(v) \neq \emptyset \\ \delta_v, & N_G(v) = \emptyset \end{cases}.$$

We further let $\overline{\deg}_G(v) := \deg_G(v)$ if $N_G(v) \neq \emptyset$ and $\overline{\deg}_G(v) := 1$ otherwise. Then, it is easy to see that

$$\mu_G := \sum_{v \in V_G} \frac{\overline{\deg}_G(v)}{\sum_{v' \in V_G} \overline{\deg}_G(v')} \delta_v \tag{3}$$

is a stationary distribution for $m_\bullet^{G,q}$ for all $q \in [0, 1)$.

3. These assumptions are mild: they encompass finite metric spaces, compact metric spaces, closed subsets of Euclidean spaces and also the set of all integers with the usual metric.

For any $q \in [0,1)$, we define a LMMC $\mathcal{X}_q(G) := \left(V_G, m_\bullet^{G,q}, \mu_G\right)$ and call $(\mathcal{X}_q(G), \ell_G)$ a *graph induced LMMC*. When $q = 0$, we also let $m_\bullet^G := m_\bullet^{G,q}$ and let $\mathcal{X}(G) := \mathcal{X}_0(G)$.

It turns out this process of constructing LMMCs from graphs preserves distinguishability of the underlying labeled graphs. To see this precisely, we first recall the notion of isomorphism for labeled graphs.

Definition 8 (Isomorphism of labeled graphs) *Two finite labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) are said to be* isomorphic *if there exists a bijective map $\psi : V_{G_1} \to V_{G_2}$ such that for all $v \in V_{G_1}$, $\ell_{G_1}(v) = \ell_{G_2}(\psi(v))$ and $(v, v') \in E_{G_1}$ iff $(\psi(v), \psi(v')) \in E_{G_2}$.*

Then, we see that the labeled graph isomorphism is equivalent to the isomorphism of the induced LMMCs.

Proposition 9 *Consider two labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}). Then, for any fixed $q \in [0,1)$, (G_1, ℓ_{G_1}) is isomorphic to (G_2, ℓ_{G_2}) as labeled graphs if and only if $(\mathcal{X}_q(G_1), \ell_{G_1})$ is isomorphic to $(\mathcal{X}_q(G_2), \ell_{G_2})$ as LMMCs.*

3. The Weisfeiler-Lehman Distance

A non-empty finite multiset M of elements taken from a given set S encodes information about the *multiplicity* of each $s \in S$ in M. This suggests that one might consider the *probability measure* μ_M on S induced by M:

$$\mu_M(s) := \frac{m(s)}{\sum_{t \in S} m(t)}, \ \forall s \in S$$

where $m(s)$ denotes the multiplicity of s in S.

By interpreting the multisets in the WL hierarchy (see Definition 1) as probability measures, we naturally shift from a combinatorial framework to a measure-theoretic one. This perspective enables us to leverage optimal transport techniques and ultimately define a meaningful notion of graph distance.

Definition 10 (Weisfeiler-Lehman measure hierarchy) *Given any Z-LMMC $(\mathcal{X}, \ell_X)$, we let $\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)} := \ell_X$ and produce the following label functions whose codomains span a certain hierarchy of probability measures:*

Step 0 *We define*

$$\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)} : X \to Z.$$

Step 1 *We define*

$$\mathfrak{l}^{(1)}_{(\mathcal{X},\ell_X)} := \left(\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)}\right)_{\#} m_\bullet^X : X \to \mathcal{P}(Z).$$

...

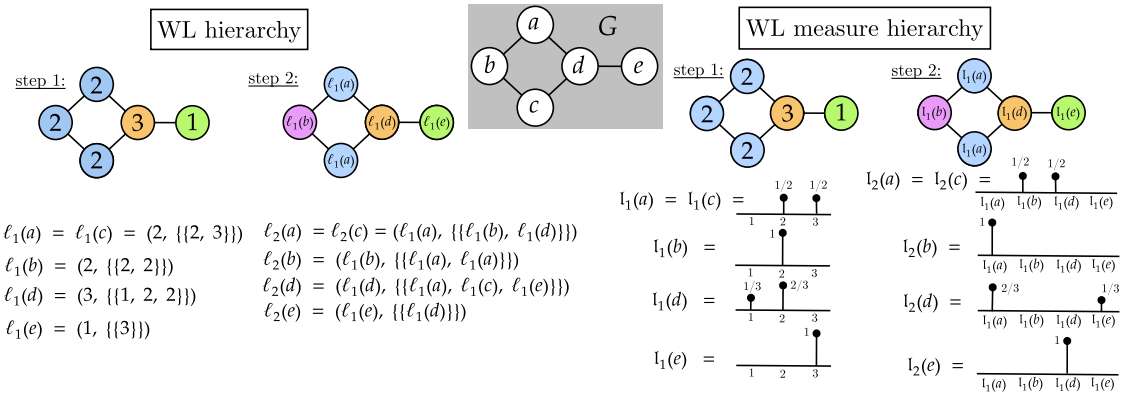


Figure 3: **Illustration of the WL (measure) hierarchy.** The graph G shown in the middle of the figure is assigned the degree label ℓ_G. We explicitly present two steps of the WL hierarchy of (G, ℓ_G) (on the left) and of the WL measure hierarchy of $(\mathcal{X}(G), \ell_G)$ (on the right). Every probability measure is represented as a histogram. For $i = 1, 2$, ℓ_i and $\mathfrak{l}_i$ are abbreviations for $\ell^{(i)}_{(G,\ell_G)}$ and $\mathfrak{l}^{(i)}_{(\mathcal{X}(G),\ell_G)}$, respectively. Notice how the WL measure hierarchy interprets the multisets from the WL hierarchy as probability measures.

Step k *For each integer* $k \geq 2$*, we inductively define*

$$\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)} \coloneqq \left(\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)}\right)_{\#} m^X_\bullet : X \to \mathcal{P}^{\circ k}(Z).$$

...

We verify that the WL measure hierarchy is well-defined by showing that all the push-forward measures are probability measures with finite 1-moment.

Proposition 11 (Well-definedness of the WL measure hierarchy) *For any Z-LMMC $(\mathcal{X}, \ell_X)$, consider any $x \in X$, then for any $k \geq 1$, $\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x) \in \mathcal{P}^{\circ k}(Z)$, i.e., $\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x)$ is a probability measure on the space $\mathcal{P}^{\circ(k-1)}(Z)$ with finite 1-moment.*

We then induce at each step $k \geq 0$ a probability measure

$$\mathfrak{L}_k((\mathcal{X}, \ell_X)) \coloneqq \left(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\right)_{\#} \mu_X \in \mathcal{P}^{\circ(k+1)}(Z).$$

$\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}$ is analogous to $\ell^{(k)}_{(G,\ell_G)}$ and $\mathfrak{L}_k((\mathcal{X}, \ell_X))$ is analogous to $L_k((G, \ell_G))$ from the WL hierarchy (cf. Definition 1). See Figure 3 for an illustration of the WL measure hierarchy of a graph induced LMMC and its comparison with the corresponding WL hierarchy. We will show later that, up to certain change of labels, the WL measure hierarchy for a graph induced LMMC captures all the information contained in the WL hierarchy of the original graph (see Proposition 17).

We now define the Weisfeiler-Lehman distance based on the WL measure hierarchy.

Definition 12 (Weisfeiler-Lehman distance) *For each integer $k \geq 0$ and any metric space of labels Z we define the* Weisfeiler-Lehman (WL) distance of depth k *between the Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ as*

$$d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \coloneqq d_{\mathrm{W}}\left(\mathfrak{L}_k((\mathcal{X}, \ell_X)), \mathfrak{L}_k((\mathcal{Y}, \ell_Y))\right) \tag{4}$$

where d_{W} above takes place in $\mathcal{P}^{\circ(k+1)}(Z)$. We also define the (absolute) ***Weisfeiler-Lehman distance*** *by*

$$d_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) \coloneqq \sup_{k\geq 0} d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y))\,.$$

Example 1 *We show explicit formulas for $d_{\mathrm{WL}}^{(k)}$ when $k=0$ and 1. When $k=0$, we have*

$$d_{\mathrm{WL}}^{(0)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = d_{\mathrm{W}}((\ell_X)_{\#}\mu_X,(\ell_Y)_{\#}\mu_Y)\,,$$

which is nothing but the Wasserstein distance between the label distributions $(\ell_X)_{\#}\mu_X$ and $(\ell_Y)_{\#}\mu_Y$. When $k=1$, we have that

$$d_{\mathrm{WL}}^{(1)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big((\ell_X)_{\#}m_x^X,(\ell_Y)_{\#}m_y^Y\big)\ \gamma(dx\times dy),$$

implementing the comparison of local *label distributions. See Section B.2.2 for a derivation.*

Example 2 (Potential triviality of the WL distance) *The WL distance can become trivial when the label functions are not chosen in a meaningful manner. For example, consider two LMMCs $(\mathcal{X},\ell_X)$ and $(\mathcal{Y},\ell_Y)$ where ℓ_X and ℓ_Y are constant functions taking the values $z_1\in Z$ and $z_2\in Z$, respectively. Then, we have that for any $k\geq 1$,*

$$d_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = d_Z(z_1,z_2).$$

This can be seen easily from Lemma 67 and the fact that $d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(k)}(x),\mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(k)}(y)\big) = d_Z(z_1,z_2)$ which can be verified by induction on k and the following property: $d_{\mathrm{W}}(\delta_{z_1},\delta_{z_2}) = d_Z(z_1,z_2)$.

Example 3 (Distance to a one point space) *Consider a Z-LMMC $(\mathcal{X},\ell_X)$ and a one point Z-LMMC $(\mathcal{Y},\ell_Y)\coloneqq(*,\ell_*)$ where $z\coloneqq\ell_*(*)$. Then, we have that for any $k\geq 1$,*

$$d_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = \int_X d_Z(\ell_X(x),z)\mu_X(dx).$$

This can be seen from Lemma 67 and the fact that $d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(k)}(x),\mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(k)}(*)\big) = d_Z(\ell_X(x),z)$ for all $x\in X$ and $k\geq 1$ which can be verified by induction on k and the assumption that μ_X is stationary.

3.1 Properties of the WL Distance

In this subsection we first study some basic properties of the WL distance and then establish its connection with the WL test.

3.1.1 Basic Properties of the WL Distance

Since LMMCs are assumed to be compact, the WL distance turns out always bounded.

Proposition 13 *For any Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we let*

$$C_{Z,\ell}(X,Y) := \sup_{x\in X, y\in Y} d_Z(\ell_X(x), \ell_Y(y)) \geq 0.$$

Since X, Y are compact and ℓ_X, ℓ_Y are continuous, we have $C_{Z,\ell}(X,Y) < \infty$. Then, we have that

$$d_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \leq C_{Z,\ell}(X,Y) < \infty.$$

Next we establish that d_{WL} is a pseudometric.

Theorem 14 *d_{WL} (resp. $d_{\mathrm{WL}}^{(k)}$ for $k \geq 0$) defines a pseudometric on the collection $\mathcal{M}^L(Z)$, i.e., d_{WL} (resp. $d_{\mathrm{WL}}^{(k)}$) (1) is bounded and symmetric, (2) satisfies the triangle inequality, (3) and is zero for isomorphic LMMCs (but non-isomorphic LMMCs can have zero d_{WL} (resp. $d_{\mathrm{WL}}^{(k)}$) distance).*

The following proposition states that the WL distance becomes more discriminating as the depth increases.

Proposition 15 *Let $k \geq 0$ be any integer. Given any two Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \leq d_{\mathrm{WL}}^{(k+1)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$.*

As a direct consequence of this monotonicity, we have that

Corollary 16 *Given any two Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that*

$$d_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = \lim_{k\to\infty} d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)).$$

3.1.2 Comparison with the WL Test

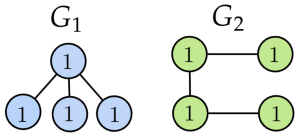
Given the apparent similarity between the WL hierarchy and the WL measure hierarchy, it should not be surprising that (as we will show later in Proposition 17), on graphs, d_{WL} essentially has the same discriminative power as the WL test, i.e., those pairs of graphs which can be distinguished by the WL test are the same as those at positive d_{WL}. The WL distance should therefore be interpreted as a quantification of the *degree* to which the graphs fail to pass the WL test.

However, there is a certain possible loss of information in the process of turning labeled graphs into a LMMC (cf. Section 2.4): a normalization step via the degree function was utilized to construct the Markov kernel and the stationary distribution. In certain cases, this normalization step implies that $d_{\mathrm{WL}}((\mathcal{X}_q(G_1), \ell_{G_1}), (\mathcal{X}_q(G_2), \ell_{G_2})) = 0$ even though the WL test distinguishes (G_1, ℓ_{G_1}) from (G_2, ℓ_{G_2}). We first exemplify this point under two scenarios and will then show a simple way to eliminate this issue altogether.

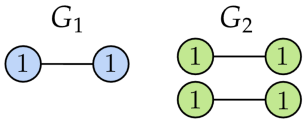
Example 4 (Constant labels) *Consider two graphs G_1 and G_2 with four nodes as shown in Figure 4a. Let the label functions ℓ_{G_i} for $i = 1, 2$ be constant and equal to 1 for both graphs.*

In the first step of the WL hierarchy, we find

$$L_1((G_1, \ell_{G_1})) = \{\!\{(1, \{\!\{1,1,1\}\!\}), (1, \{\!\{1\}\!\}), (1, \{\!\{1\}\!\}), (1, \{\!\{1\}\!\})\}\!\}$$



(a) Labeled graphs in Example 4.



(b) Labeled graphs in Example 5.

Figure 4: Labeled graphs in different examples.

and

$$L_1((G_2, \ell_{G_2})) = \{\!\{(1, \{\!\{1\}\!\}), (1, \{\!\{1\}\!\}), (1, \{\!\{1, 1\}\!\}), (1, \{\!\{1, 1\}\!\})\}\!\}\,.$$

Since $L_1((G_1, \ell_{G_1})) \neq L_1((G_2, \ell_{G_2}))$, (G_1, ℓ_{G_1}) *and* (G_2, ℓ_{G_2}) *are recognized as non-isomorphic by the WL test. Notice that in the first step, the WL test collects degree information and comparing* $L_1((G_1, \ell_{G_1}))$ *and* $L_1((G_2, \ell_{G_2}))$ *is equivalent to comparing the multisets of degrees w.r.t.* G_1 *and* G_2*, i.e.,* $\{\!\{3, 1, 1, 1\}\!\}$ *vs* $\{\!\{1, 1, 2, 2, \}\!\}$.

Example 5 (Degree label) *Consider two graphs* G_1 *and* G_2 *as shown in Figure 4b. For each* $i = 1, 2$*, let* ℓ_{G_i} *be the degree label function for* G_i*. In the first step of the WL test,*

$$L_1((G_1, \ell_{G_1})) = \{\!\{(1, \{\!\{1\}\!\}), (1, \{\!\{1\}\!\})\}\!\}$$

and

$$L_1((G_2, \ell_{G_2})) = \{\!\{(1, \{\!\{1\}\!\}), (1, \{\!\{1\}\!\}), (1, \{\!\{1\}\!\}), (1, \{\!\{1\}\!\})\}\!\}\,.$$

Then in the first step of the WL test, the two labeled graphs are already distinguished as non-isomorphic.

In either of the example above, however, when transforming (G_i, ℓ_{G_i}) $(i = 1, 2)$ into the LMMC $(\mathcal{X}_q(G_i), \ell_{G_i})$ for a fixed $q \in [0, 1)$, the degree information or the the size information are lost and we obtain two LMMCs with constant label functions and hence 0 WL distance (cf. Example 2). In order to resolve this issue, we hence introduce a simple augmentation method for the label functions so that these information are preserved in the label functions of the LMMCs.

Label augmentation. For any metric space of labels Z, consider any *injective* map $g : Z \times \mathbb{N} \times \mathbb{N} \to Z_1$ where Z_1 is any another metric space. A trivial example of such injective g is given by letting $Z_1 \coloneqq Z \times \mathbb{N} \times \mathbb{N}$ and letting g be the identity map.

Now, given any labeled graph $(G, \ell_G : V_G \to Z)$, we generate Z_1 and a new label function $\ell_G^g \coloneqq g(\ell_G, \deg_G(\bullet), |V_G|) : V_G \to Z_1$, which we call an *augmentation map*. Intuitively, this is augmenting the label ℓ_G with the size and degree information of G. After this augmentation, we establish that d_{WL} has the *same* discriminative power as the WL test.

Proposition 17 *For any* $q \in (\frac{1}{2}, 1)$ *and any augmentation map* g*, the WL test distinguishes two labeled graphs* (G_1, ℓ_{G_1}) *and* (G_2, ℓ_{G_2}) *iff* $d_{\mathrm{WL}}\Big(\Big(\mathcal{X}_q(G_1), \ell_{G_1}^g\Big), \Big(\mathcal{X}_q(G_2), \ell_{G_2}^g\Big)\Big) > 0$.

Note here that we require $q \in (\frac{1}{2}, 1)$ instead of $q \in [0, 1)$ as in Definition 7. This is due to the fact that we need to emphasize the label information of the given vertex v when

constructing the Markov kernel $m_v^{G,q}$ so that the WL distance has the same discriminative power as the WL test, which stores such information in the first argument of the pair when constructing a new label in the WL hierarchy.

Although it may seem that we are injecting more information into labels, this extra augmentation, in fact, does not affect the outcome of the WL test:

Lemma 18 *Let g be any augmentation map. The WL test distinguishes (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) iff it distinguishes $(G_1, \ell^g_{G_1})$ and $(G_2, \ell^g_{G_2})$.*

By Proposition 17 and a convergence result pertaining to the WL test (Krebs and Verbitsky, 2015), one has the following convergence result for $d_{\mathrm{WL}}^{(k)}$ (which implies that to determine whether two graph induced LMMCs satisfy $d_{\mathrm{WL}} = 0$ one only needs to inspect $d_{\mathrm{WL}}^{(k)}$ for finitely many k).

Corollary 19 *For any $q \in (\frac{1}{2}, 1)$ and any two labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), if $d_{\mathrm{WL}}^{(k)}\left(\left(\mathcal{X}_q(G_1), \ell^g_{G_1}\right), \left(\mathcal{X}_q(G_2), \ell^g_{G_2}\right)\right) = 0$ holds for each $k = 0, \ldots, |V_{G_1}| + |V_{G_2}|$, we then have that $d_{\mathrm{WL}}\left(\left(\mathcal{X}_q(G_1), \ell^g_{G_1}\right), \left(\mathcal{X}_q(G_2), \ell^g_{G_2}\right)\right) = 0$.*

Results from (Babai and Kucera, 1979) imply that the WL test can certify isomorphism of random graphs (with degree labels) with high probability. Then, immediately by Proposition 17, we have that with high probability d_{WL} generates positive distance for non-isomorphic random-graph-induced LMMCs.

3.2 An Algorithm and a Computationally Efficient Lower Bound

The hierarchical nature of the WL distance gives rise to the following iterative Algorithm 1 for computing the WL distance between finite LMMCs (with $Z = \mathbb{R}$ for simplicity).

Algorithm 1 Computation of $d_{\mathrm{WL}}^{(k)}$

Input: The depth $k \in \mathbb{N}$, and two finite LMMCs $\left(X = \{x_1, x_2, ..., x_n\}, m_\bullet^X, \mu_X, \ell_X : X \to \mathbb{R}\right)$ and $\left(Y = \{y_1, ..., y_m\}, m_\bullet^Y, \mu_Y, \ell_Y : Y \to \mathbb{R}\right)$
Initialization: $P = C = \mathrm{zeros}(n, m)$
for $i \in [n], j \in [m]$ **do**
  $P(i,j) = |\ell_X(x_i) - \ell_Y(y_j)|$
end for
for $l \in [k]$ **do**
  for $i \in [n], j \in [m]$ **do**
    $C(i,j) = \min_{\gamma \in \mathcal{C}\left(m^X_{x_i}, m^Y_{y_j}\right)} \sum_{a \in [n], b \in [m]} P(a,b)\gamma(a,b)$
  end for
  $P = C$
end for
$D = \min_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \sum_{i \in [n], j \in [m]} C(i,j)\gamma(i,j)$
Output: D

We establish the following complexity analysis for the algorithm below.

Proposition 20 *For any fixed $k \in \mathbb{N}$, computing $d_{\mathrm{WL}}^{(k)}$ between any finite LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ through Algorithm 1 can be achieved in time at most $O(k\, n^5 \log(n))$ where $n := \max(|X|, |Y|)$.*

Notice that directly computing the WL distance between two labeled graphs is computationally expensive. However, we provide below a computationally efficient lower bound for the WL distance.

A Lower Bound for the WL Distance. The WL measure hierarchy, defined through consecutive steps of pushforward maps, can be related to a certain sequence of Markov kernels which we explain next. Given a MMC $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ and any $k \in \mathbb{N}$, the k-step Markov kernel, denoted by $m_\bullet^{X,\otimes k}$, is defined inductively as follows: for any $x \in X$, when $k = 1$, $m_x^{X\otimes 1} \coloneqq m_x^X$ and when $k \geq 2$, for $A \subseteq X$,

$$m_x^{X,\otimes k}(A) \coloneqq \int_X m_{x'}^{X,\otimes(k-1)}(A)\, m_x^X(dx').$$

If we represent $m_\bullet^X$ by a transition matrix $M_{\mathcal{X}}$, then the matrix corresponding to $m_\bullet^{X,\otimes k}$ is the k-power of $M_{\mathcal{X}}$.

Recall the formula for $d_{\mathrm{WL}}^{(1)}$ from Example 1. We define a certain quantity (for each $k \in \mathbb{N}$) which arises by replacing the Markov kernels in that formula with k-step Markov kernels:

$$d_{\mathrm{WLLB}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \coloneqq \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\Big((\ell_X)_\# m_x^{X,\otimes k}; (\ell_Y)_\# m_y^{Y,\otimes k}\Big)\gamma(dx \times dy).$$

Notice that d_{W} above takes place in $\mathcal{P}(Z)$ whereas d_{W} in Equation (4) for defining $d_{\mathrm{WL}}^{(k)}$ takes place in $\mathcal{P}^{\circ k}(Z)$. Of course, we have that $d_{\mathrm{WLLB}}^{(1)} = d_{\mathrm{WL}}^{(1)}$. It turns out that for each $k \geq 1$, $d_{\mathrm{WLLB}}^{(k)}$ is a lower bound for $d_{\mathrm{WL}}^{(k)}$.

Proposition 21 *For any $(\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y) \in \mathcal{M}^L(Z)$ and any integer $k \geq 1$ we have that $d_{\mathrm{WLLB}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \leq d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$.*

For any fixed $k \in \mathbb{N}$ and any two finite $\mathbb{R}$-LMMCs, $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, if we let $n := \max(|X|, |Y|)$, then computing $d_{\mathrm{WL}}^{(k)}$ between the pair can be done in $O(kn^5 \log(n))$ (cf. Proposition 20) time whereas the total time complexity for computing $d_{\mathrm{WLLB}}^{(k)}$ is far less:

Proposition 22 *For any fixed $k \in \mathbb{N}$, computing $d_{\mathrm{WLLB}}^{(k)}$ between any finite LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ can be achieved in time at most $O(n^3 \log(nk))$ where $n = \max(|X|, |Y|)$.*

We empirically compare the runtimes of $d_{\mathrm{WL}}^{(k)}$ and $d_{\mathrm{WLLB}}^{(k)}$ for $k = 1, 2$ (see Figure 5). For our runtime comparisons, we use LMMCs induced by Erdös-Renyi graphs of sizes varying from 5 nodes to 100 nodes (with the degree label function and $q = 0.6$). Note that while the runtime for $d_{\mathrm{WLLB}}^{(k)}$ does not change much between $k = 1$ and $k = 2$, the $d_{\mathrm{WL}}^{(k)}$ distance shows a significant increase in the time needed to compute distance between two graphs from $k = 1$ to $k = 2$.

We will further empirically compare the WL distance with this lower bound in terms of performance in the graph classification task in Section 3.4.

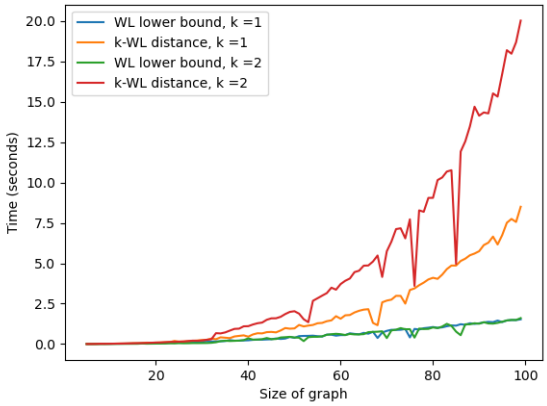


Figure 5: Comparison of runtime of WWL distance against $d_{\mathrm{WL}}^{(k)}$ and its lower bound $d_{\mathrm{WLLB}}^{(k)}$.

3.3 Wasserstein Weisfeiler-Lehman Graph Kernel and Relationship with d_{WL}

In this subsection, we show that a slight variant of the Wasserstein Weisfeiler-Lehman (WWL) graph kernel (Togninalli et al., 2019) provides a strict lower bound for the WL distance. In this way, we argue that the WL distance is more discriminative than the WWL graph kernel.

The WWL graph kernel deals with graphs with either categorical or "continuous" (i.e., Euclidean) labels (Togninalli et al., 2019). We only consider continuous labels since categorical labels usually do not come equipped with a natural distance and the WWL graph kernel will invoke a hash map to simplify the categorical labels at each step and hence a metric for such labels will not make sense.

Now, we describe the WWL framework w.r.t. continuous labels as follows. For technical reasons, we assume that all graphs involved in this section are such that all their connected components have cardinality at least 2 (i.e. no graph contains an isolated vertex).

Definition 23 (WWL Graph Kernels Togninalli et al. (2019)) *Given a labeled graph $(G, \ell_G : V_G \to \mathbb{R}^d)$ and any fixed integer $k > 0$, define iteratively label maps for $i = 0, \dots, k-1$ as follows: $\ell_G^0 := \ell_G$ and*

$$\forall v \in V_G \text{ and } i \geq 0, \quad \ell_G^{i+1}(v) := \frac{1}{2}\left(\ell_G^i(v) + \frac{1}{\deg_G(v)} \sum_{v' \in N_G(v)} \ell_G^i(v') \right).$$

Then, define the stacked label function L_G^k as follows:

$$L_G^k := \left(\ell_G^0, \dots, \ell_G^k \right) : V_G \to \mathbb{R}^{d \times (k+1)}.$$

Now, given any two labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), let λ_{G_i} denote the uniform measure on V_{G_i} for $i = 1, 2$ and define what we call the WWL distance between (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) as

$$D^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) := d_{\mathrm{W}}\left(\left(L_{G_1}^k\right)_{\#} \lambda_{G_1}, \left(L_{G_2}^k\right)_{\#} \lambda_{G_2} \right). \tag{5}$$

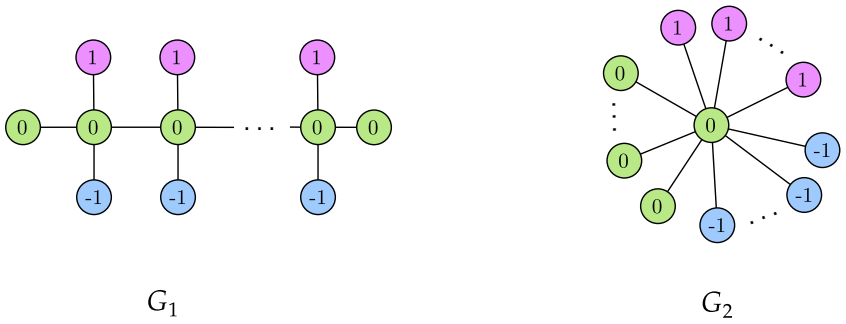


Figure 6: Each labeled graph has $3n+2$ vertices. Each of the graphs has n vertices with label 1, n vertices with label -1, and $n+2$ vertices with label 0.

Finally, for some hyperparameter $\gamma > 0$, the WWL kernel is defined as

$$\mathrm{WWL}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) \coloneqq \exp\Big(-\gamma D^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2}))^2\Big). \tag{6}$$

If we instead of uniform measures consider the stationary distributions μ_{G_1} and μ_{G_2} w.r.t. $m_\bullet^{G_1,\frac{1}{2}}$ and $m_\bullet^{G_2,\frac{1}{2}}$, respectively, we define the following variant of the WWL distance:

Definition 24 (A Variant of the WWL Distance) *Given any two labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), let μ_{G_1} and μ_{G_2} denote the stationary distributions w.r.t. $m_\bullet^{G_1,\frac{1}{2}}$ and $m_\bullet^{G_2,\frac{1}{2}}$, respectively, as defined in Equation (3). Then, we define the following variant of $D^{(k)}$:*

$$\hat{D}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) \coloneqq d_{\mathrm{W}}\left(\left(L^k_{G_1}\right)_{\#}\mu_{G_1}, \left(L^k_{G_2}\right)_{\#}\mu_{G_2}\right). \tag{7}$$

We then prove that the above slight variant of the WWL kernel $\hat{D}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2}))$ actually provides a strict lower bound for $d_{\mathrm{WL}}^{(k)}((\mathcal{X}_q(G_1), \ell_{G_1}), (\mathcal{X}_q(G_2), \ell_{G_2}))$ when $q = \frac{1}{2}$.

Proposition 25 *For any two labeled graphs $(G_1, \ell_{G_1} : V_{G_1} \to \mathbb{R}^d)$ and $(G_2, \ell_{G_2} : V_{G_2} \to \mathbb{R}^d)$, one has that for $q = \frac{1}{2}$ and any $k \in \mathbb{N}$,*

$$\hat{D}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) \leq k \cdot d_{\mathrm{WL}}^{(k)}((\mathcal{X}_q(G_1), \ell_{G_1}), (\mathcal{X}_q(G_2), \ell_{G_2})).$$

Example 6 ($d_{\mathrm{WL}}^{(k)}$ is strictly more discriminating than $\hat{D}^{(k)}$) *For any $n \geq 2$, consider the two $(3n+2)$-point labeled graphs shown in Figure 6. It is easy to see that for each $i = 1, 2$, and any $v \in V_{G_i}$,*

1. *if $\ell_{G_i}(v) = 0$, then $\ell^k_{G_i}(v) = 0$ for all $k = 0, 1, \ldots$;*
2. *if $\ell_{G_i}(v) = \pm 1$, then $\ell^k_{G_i}(v) = \pm\frac{1}{2^k}$ for all $k = 0, 1, \ldots$.*

Hence, for any $k = 0, \ldots,$ we have that

$$\left(L_{G_1}^k\right)_{\#}\mu_{G_1} = \left(L_{G_2}^k\right)_{\#}\mu_{G_2} = \frac{4n+2}{6n+2}\delta_{(0,\cdots,0)} + \frac{n}{6n+2}\delta_{(1,\cdots,2^{-k})} + \frac{n}{6n+2}\delta_{(-1,\cdots,-2^{-k})}.$$

Therefore, $\hat{D}^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2})) = 0$ for all $k = 0, \ldots$.

Now, we analyze $\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1),\ell_{G_1})}(v_1)$ for any $v_1 \in V_{G_1}$.

1. *If $\ell_{G_1}(v_1) = \pm 1$, then*
$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1),\ell_{G_1})}(v_1) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\pm 1}.$$
2. *If $\ell_{G_1}(v_1) = 0$ and v_1 is neither the leftmost nor the rightmost vertex, then*
$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1),\ell_{G_1})}(v_1) = \frac{6}{8}\delta_0 + \frac{1}{8}\delta_1 + \frac{1}{8}\delta_{-1}.$$
3. *If v_1 is either the leftmost or the rightmost vertex, then*
$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1),\ell_{G_1})}(v_1) = \delta_0.$$

We then analyze $\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}(v_2)$ for any $v_2 \in V_{G_2}$.

1. *If $\ell_{G_2}(v_2) = \pm 1$, then*
$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}(v_2) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\pm 1}.$$
2. *If v_2 is the center vertex, then*
$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}(v_2) = \frac{2n+1}{3n+1}\delta_0 + \frac{n}{2(3n+1)}\delta_1 + \frac{n}{2(3n+1)}\delta_{-1}.$$
3. *If $\ell_{G_2}(v_2) = 0$ and v_2 is not the center vertex, then*
$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}(v_2) = \delta_0.$$

Hence, it is clear that when $n > 1$

$$d_{\mathrm{WL}}^{(1)}\big((\mathcal{X}_q(G_1),\ell_{G_1}),(\mathcal{X}_q(G_2),\ell_{G_2})\big) = d_{\mathrm{W}}\left(\left(\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1),\ell_{G_1})}\right)_{\#}\mu_{G_1}, \left(\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}\right)_{\#}\mu_{G_2}\right) > 0.$$

Remark 26 *Recall that earlier we modified the definition from Togninalli et al. (2019) by imposing a stationary distribution instead of the uniform distribution originally used. This modification enables a direct comparison between the variant $\hat{D}^{(k)}$ and the WL distance. Our formulation of the WL distance is flexible, however, and can be further relaxed by comparing measure Markov chains with any reference probability measures—not just stationary ones. This more general approach has been further developed by the last two authors and their collaborator in Brugère et al. (2024).*

In this relaxed setting, we can directly compare the distance $D^{(k)}$ defined in Equation (5) with $d_{\mathrm{WL}}^{(k)}$ (without the stationary assumption). Specifically, for any graph G, we can replace the stationary distribution in $\mathcal{X}_q(G)$ with the uniform measure to obtain a new measure Markov chain $\mathcal{X}_q^{\mathrm{u}}(G)$. Then, using the same proof technique as in Proposition 25, one can show that

$$D^{(k)}\left((G_1, \ell_{G_1}), (G_2, \ell_{G_2})\right) \leq k \cdot d_{\mathrm{WL}}^{(k)}\left((\mathcal{X}_q^{\mathrm{u}}(G_1), \ell_{G_1}), (\mathcal{X}_q^{\mathrm{u}}(G_2), \ell_{G_2})\right).$$

Moreover, it still holds that in this setting, $d_{\mathrm{WL}}^{(k)}$ is strictly more discriminating than $D^{(k)}$ using the same pairs of graphs as in Example 6.

3.4 Graph Classification Experiments

In this section, we validate the usefulness of the WL distance in the task of graph classification on several publicly available graph benchmark datasets from TUDatasets (Morris et al., 2020); see Table 2 for dataset statistics. We will compare the performance of the WL distance with its lower bound defined in Section 3.2 as well as other graph kernel methods.

The accompanying code is available at `https://github.com/chens5/WL-distance`.

Name	# of graphs	Average # of Nodes	Average # of Edges
MUTAG	188	17.93	19.79
PROTEINS	1113	39.06	72.82
IMDB-B	1000	19.77	96.53
IMDB-M	1500	13.00	65.94
PTC-FM	349	14.11	14.48
PTC-MR	344	14.29	14.69
COX2	467	41.22	43.45

Table 2: Graph classification data sets. The table reports the number of graphs and the mean numbers of nodes and edges per graph for each data set.

3.4.1 Experimental Setup

Both $d_{\mathrm{WL}}^{(k)}$ and $d_{\mathrm{WLLB}}^{(k)}$ are pseudometrics whereas the WWL kernel is a kernel. We design two experiments to compare them in a fair manner.

1-Nearest Neighbor (1NN) Classifier. Recall from Equation (5) that the WWL kernel (Togninalli et al., 2019) is constructed from a pseudometric $D^{(k)}$. Therefore, in our first experiment, we first compute pairwise dissimilarity between graphs in each given dataset using three different pseudometrics, namely, $d_{\mathrm{WL}}^{(k)}$, $d_{\mathrm{WLLB}}^{(k)}$ and $D^{(k)}$, then we apply the 1NN Classifier to obtain the final classification scores.

Support Vector Machines (SVM) In the second experiment, we turn both $d_{\mathrm{WL}}^{(k)}$ and $d_{\mathrm{WLLB}}^{(k)}$ into kernels and then compare them with the WWL kernel. For this purpose, we consider the lower bound kernel $\exp\left(-\gamma d_{\mathrm{WLLB}}^{(k)}\right)$ and the WL distance kernel $\exp\left(-\gamma d_{\mathrm{WL}}^{(k)}\right)$ where γ is a hyperparameter. Notice that while the WWL kernel is positive definite, the

WL distance kernel and the lower bound kernel may be indefinite. Hence, for the WWL kernel, we apply directly the SVM classification method. For the WL distance kernel and the lower bound kernel, we apply two classification methods: (1) the vanilla SVM and (2) the KSVM (Loosli et al., 2015) which is designed to handle indefinite kernels.

For each data set, we tune the SVM regularization parameter C and kernel coefficient γ by grid search, with both parameters ranging over $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. Besides the WWL kernel, we also compare with the WL kernel (Shervashidze et al., 2011), and the Weisfeiler-Lehman optimal assignment kernel (WL-OA) (Kriege et al., 2016).

Note we do not include WL-OA and WL kernels into the 1NN experiments since they are not defined in terms of a pesudodistance like the WWL kernel.

Preprocessing. For all of our experiments, we use $q = 0.6$ to transform every graph G into the Markov chain $\mathcal{X}_q(G)$ (cf. Section 2.4).

In regards of the label augmentation mentioned in Section 3.1.2, we use two types of label functions:

- $f_1(G, v) := \deg_G(v)$ is the degree label function.
- $f_2(G, v) := \frac{1}{|V_G|} + \deg_G(v)$ is further augmented by the size of the graph.

For WWL, WL, and WL-OA kernels, we use the degree label function f_1. For $d_{\mathrm{WL}}^{(k)}$ and $d_{\mathrm{WLLB}}^{(k)}$, we use both f_1 and f_2. Note that f_2 is a legitimate label augmentation as described in Section 3.1.2 via the injective map $g : \{c\} \times \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ sending (c, n_1, n_2) to $n_1 + \frac{1}{n_2}$.

We report the best accuracies for WWL, WL, and WL-OA for iterations $1, \ldots, 4$.

3.4.2 Experimental Results

In Table 3 and Table 4, we have included the 1-NN and SVM classification accuracies for $k \in \{1, 2, 3, 4\}$, respectively.

1-NN Classification. Overall, the classification accuracies of $d_{\mathrm{WLLB}}^{(k)}$ and $d_{\mathrm{WL}}^{(k)}$ are close to those of the WWL distance, consistent with the close relationship between $d_{\mathrm{WL}}^{(k)}$ and the WWL distance outlined in Section 3.3. Comparing the best classification accuracies in Table 3, at least one of $d_{\mathrm{WLLB}}^{(k)}$ and $d_{\mathrm{WL}}^{(k)}$ outperforms the WWL distance on every dataset except MUTAG, where WWL performs better than both.

SVM Classification. As shown in Table 4, the kernels $\exp\left(-\gamma d_{\mathrm{WLLB}}^{(k)}\right)$ and $\exp\left(-\gamma d_{\mathrm{WL}}^{(k)}\right)$ generally yield comparable classification accuracies when used with either SVM or KSVM. Our lower bound kernel outperforms the WWL kernel on all datasets in terms of the best classification accuracy. Compared with the lower bound kernel, the $d_{\mathrm{WL}}^{(k)}$ kernel achieves slightly higher best accuracies on MUTAG and PTC-FM and slightly lower ones on PTC-MR and IMDB-B. Their best reported accuracies are the same on PROTEINS, IMDB-M, and COX2.

4. Markov Chain Neural Networks and Graph Neural Networks

As mentioned in the introduction, one of the major reasons for the popularity of the WL test stems from the theoretical analysis of graph neural networks (GNNs). In this section,

Method	MUTAG	PROTEINS	PTC-FM	PTC-MR	IMDB-B	IMDB-M	COX2
$d_{\mathrm{WL}}^{(1)}$, f_1	91.6 ± 7.9	62.2 ± 4.0	60.0 ± 10.2	54.0 ± 5.8	70.5 ± 5.9	38.7 ± 4.7	72.6 ± 2.9
$d_{\mathrm{WL}}^{(2)}$, f_1	92.1 ± 7.5	60.3 ± 3.9	**61.1 ± 8.9**	56.3 ± 4.1	**71.9 ± 4.7**	40.3 ± 6.1	74.5 ± 4.8
$d_{\mathrm{WL}}^{(3)}$, f_1	91.6 ± 6.3	60.7 ± 3.3	59.4 ± 8.7	56.6 ± 7.1	70.4 ± 4.9	38.9 ± 3.7	75.1 ± 3.6
$d_{\mathrm{WL}}^{(4)}$, f_1	92.6 ± 5.4	62.9 ± 3.6	58.3 ± 8.6	56.0 ± 5.6	71.2 ± 4.1	38.7 ± 4.7	75.1 ± 5.5
$d_{\mathrm{WL}}^{(1)}$, f_2	91.1 ± 7.8	62.9 ± 3.8	56.9 ± 7.5	53.4 ± 8.3	71.2 ± 5.5	42.1 ± 5.3	71.5 ± 3.8
$d_{\mathrm{WL}}^{(2)}$, f_2	91.6 ± 5.9	61.8 ± 3.1	57.4 ± 8.5	**57.1 ± 6.1**	71.5 ± 3.5	41.1 ± 2.3	73.2 ± 4.2
$d_{\mathrm{WL}}^{(3)}$, f_2	92.1 ± 6.3	63.9 ± 3.1	58.0 ± 7.9	**57.1 ± 8.1**	70.9 ± 4.3	41.4 ± 3.7	75.3 ± 4.7
$d_{\mathrm{WL}}^{(4)}$, f_2	92.1 ± 4.2	**64.0 ± 4.1**	58.0 ± 8.4	56.9 ± 7.9	70.4 ± 3.1	39.7 ± 2.1	75.1 ± 5.0
$d_{\mathrm{WLLB}}^{(1)}$, f_1	91.6 ± 7.9	61.6 ± 3.7	60.0 ± 10.2	54.0 ± 5.8	69.4 ± 5.4	38.7 ± 4.7	72.6 ± 2.9
$d_{\mathrm{WLLB}}^{(2)}$, f_1	92.1 ± 7.5	60.3 ± 4.4	60.0 ± 8.8	54.9 ± 6.9	71.8 ± 4.7	**43.5 ± 6.5**	76.0 ± 4.9
$d_{\mathrm{WLLB}}^{(3)}$, f_1	91.1 ± 6.2	60.2 ± 3.5	60.0 ± 7.0	55.4 ± 6.8	70.1 ± 3.9	40.2 ± 3.7	77.2 ± 3.8
$d_{\mathrm{WLLB}}^{(4)}$, f_1	92.1 ± 5.9	60.1 ± 2.4	60.0 ± 5.6	54.6 ± 7.8	71.4 ± 3.9	40.5 ± 3.2	**78.1 ± 3.8**
$d_{\mathrm{WLLB}}^{(1)}$, f_2	91.1 ± 7.8	62.9 ± 3.8	56.9 ± 7.5	53.4 ± 8.3	70.3 ± 6.4	39.6 ± 4.6	71.5 ± 3.8
$d_{\mathrm{WLLB}}^{(2)}$, f_2	92.1 ± 6.3	61.8 ± 3.8	58.3 ± 8.5	55.4 ± 8.3	71.7 ± 4.6	42.5 ± 4.6	74.5 ± 4.9
$d_{\mathrm{WLLB}}^{(3)}$, f_2	92.6 ± 5.4	63.7 ± 4.2	59.1 ± 8.0	56.0 ± 7.5	70.5 ± 3.4	42.1 ± 6.2	77.0 ± 3.4
$d_{\mathrm{WLLB}}^{(4)}$, f_2	92.1 ± 5.9	61.6 ± 3.5	59.1 ± 7.9	53.1 ± 7.6	71.1 ± 2.8	41.2 ± 4.9	76.6 ± 4.4
WWL	**93.7 ± 5.2**	63.3 ± 3.5	60.6 ± 5.8	56.6 ± 6.0	69.0 ± 4.3	43.1 ± 4.5	72.8 ± 3.5

Table 3: **1-Nearest Neighbor classification accuracy.** WWL represents using the pseudometric $D^{(k)}$ for constructing the WWL kernel (cf. Equation (5)). Values are mean ± standard deviation in percent; the highest mean in each column is bolded, including ties.

we first utilize the WL distance to analyze properties of a special type of neural networks for Markov chain inputs, which we refer to as *Markov Chain Neural Networks* (MCNNs), and then present certain implications for graph neural networks (GNNs).

In particular, in this section we describe three key contributions of our work: (1) we prove that certain GNN architectures exhibit Lipschitz continuity with respect to the WL distance and that (2) their distinguishing power is precisely characterized by the WL distance. Furthermore we prove that (3) they achieve universal approximation relative to the topology induced by the WL distance.

4.1 Markov Chain Neural Networks

We focus on the case when the metric space of labels Z is Euclidean, i.e., $Z = \mathbb{R}^d$, and define a family of real functions on $\mathcal{M}^L(\mathbb{R}^d)$ called *Markov Chain Neural Networks (*MCNN*s)*. We study the discriminative power, Lipschitz property and the universality of this family of functions.

For any function $\varphi : \mathbb{R}^i \to \mathbb{R}^j$, we define the map $q_\varphi : \mathcal{P}(\mathbb{R}^i) \to \mathbb{R}^j$ sending $\alpha \in \mathcal{P}(\mathbb{R}^i)$ to the average $\int_{\mathbb{R}^i} \varphi(x)\alpha(dx)$. Based on q_φ, we define two types of maps:

- $F_\varphi : \mathcal{M}^L(\mathbb{R}^i) \to \mathcal{M}^L(\mathbb{R}^j)$ sending $(\mathcal{X}, \ell_X)$ to $(\mathcal{X}, \ell_X^\varphi)$, where $\ell_X^\varphi : X \to \mathbb{R}^j$ is defined by $x \mapsto q_\varphi((\ell_X)_\# m_x^X)$.
- $S_\varphi : \mathcal{M}^L(\mathbb{R}^i) \to \mathbb{R}^j$ sending $(\mathcal{X}, \ell_X)$ to $q_\varphi((\ell_X)_\# \mu_X)$.

Method	MUTAG	PROTEINS	PTC-FM	PTC-MR	IMDB-B	IMDB-M	COX2
$d_{\mathrm{WL}}^{(1)}$, f_1	88.2 ± 6.8	72.1 ± 4.2	58.4 ± 5.7	55.8 ± 6.4	75.7 ± 2.7	51.3 ± 4.6	77.5 ± 2.5
$d_{\mathrm{WL}}^{(2)}$, f_1	**88.7 ± 6.7**	72.8 ± 4.3	58.1 ± 5.0	55.5 ± 6.7	75.7 ± 3.6	51.2 ± 4.4	78.2 ± 0.8
$d_{\mathrm{WL}}^{(3)}$, f_1	87.7 ± 6.5	72.7 ± 4.8	58.5 ± 5.1	54.3 ± 6.9	75.9 ± 3.8	51.3 ± 4.6	78.2 ± 0.8
$d_{\mathrm{WL}}^{(4)}$, f_1	**88.7 ± 6.7**	72.2 ± 5.1	58.7 ± 5.6	54.9 ± 6.7	76.1 ± 4.0	51.1 ± 5.1	78.2 ± 0.8
$d_{\mathrm{WL}}^{(1)}$, f_1, K	87.2 ± 5.6	72.4 ± 4.1	58.4 ± 6.5	53.8 ± 6.9	75.5 ± 3.0	**51.5 ± 4.4**	77.5 ± 2.5
$d_{\mathrm{WL}}^{(2)}$, f_1, K	88.2 ± 5.5	73.2 ± 3.9	58.1 ± 5.9	54.0 ± 8.6	75.3 ± 4.2	51.2 ± 4.4	78.2 ± 0.8
$d_{\mathrm{WL}}^{(3)}$, f_1, K	87.7 ± 5.6	73.3 ± 4.2	59.0 ± 5.7	53.5 ± 8.1	75.6 ± 4.1	51.3 ± 4.6	78.2 ± 0.8
$d_{\mathrm{WL}}^{(4)}$, f_1, K	88.2 ± 6.4	73.8 ± 4.4	58.5 ± 5.1	54.1 ± 7.6	75.9 ± 4.4	51.1 ± 5.0	78.2 ± 0.8
$d_{\mathrm{WL}}^{(1)}$, f_2	85.5 ± 9.7	74.3 ± 4.0	60.8 ± 5.4	57.2 ± 5.5	76.0 ± 2.8	51.3 ± 4.8	77.5 ± 2.5
$d_{\mathrm{WL}}^{(2)}$, f_2	85.0 ± 9.6	74.3 ± 4.5	59.0 ± 4.6	56.1 ± 7.9	75.8 ± 3.6	50.5 ± 4.3	78.2 ± 0.8
$d_{\mathrm{WL}}^{(3)}$, f_2	84.4 ± 9.5	**74.5 ± 4.4**	**61.6 ± 6.1**	56.1 ± 4.6	76.2 ± 3.8	51.1 ± 4.8	78.2 ± 0.8
$d_{\mathrm{WL}}^{(4)}$, f_2	85.5 ± 8.2	74.4 ± 4.6	**61.6 ± 6.1**	56.1 ± 5.2	76.1 ± 4.1	51.1 ± 4.7	78.2 ± 0.8
$d_{\mathrm{WL}}^{(1)}$, f_2, K	85.5 ± 8.2	73.9 ± 4.1	61.1 ± 6.0	57.8 ± 8.9	75.9 ± 2.7	51.3 ± 4.9	77.5 ± 2.5
$d_{\mathrm{WL}}^{(2)}$, f_2, K	86.1 ± 8.3	73.9 ± 3.7	59.6 ± 6.2	56.4 ± 7.9	75.2 ± 3.9	50.8 ± 4.6	78.2 ± 0.8
$d_{\mathrm{WL}}^{(3)}$, f_2, K	86.1 ± 8.3	74.1 ± 4.1	59.0 ± 6.5	55.8 ± 6.4	75.7 ± 3.9	50.9 ± 4.5	78.2 ± 0.8
$d_{\mathrm{WL}}^{(4)}$, f_2, K	84.5 ± 7.6	74.4 ± 4.3	59.3 ± 5.5	55.0 ± 7.0	75.9 ± 4.3	51.3 ± 4.7	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(1)}$, f_1	88.2 ± 6.8	72.1 ± 4.2	58.4 ± 5.7	55.8 ± 6.4	75.7 ± 2.7	51.3 ± 4.6	77.5 ± 2.5
$d_{\mathrm{WLLB}}^{(2)}$, f_1	87.2 ± 5.6	72.2 ± 4.2	59.6 ± 7.0	54.0 ± 7.4	75.6 ± 3.4	50.7 ± 4.9	77.9 ± 1.3
$d_{\mathrm{WLLB}}^{(3)}$, f_1	86.6 ± 6.9	72.5 ± 4.6	58.7 ± 6.6	53.8 ± 6.6	76.0 ± 3.3	50.7 ± 3.3	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(4)}$, f_1	86.6 ± 7.1	71.6 ± 5.4	57.0 ± 5.7	53.8 ± 5.8	76.1 ± 2.7	50.7 ± 4.8	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(1)}$, f_1, K	87.2 ± 5.6	72.4 ± 4.1	58.4 ± 6.5	53.8 ± 6.9	75.5 ± 3.0	**51.5 ± 4.4**	77.5 ± 2.5
$d_{\mathrm{WLLB}}^{(2)}$, f_1, K	87.2 ± 5.1	73.2 ± 4.6	59.9 ± 7.0	55.5 ± 7.2	75.2 ± 3.8	50.7 ± 4.9	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(3)}$, f_1, K	87.7 ± 6.9	73.0 ± 5.0	58.7 ± 6.6	54.9 ± 6.8	76.1 ± 3.2	51.3 ± 4.5	77.9 ± 1.3
$d_{\mathrm{WLLB}}^{(4)}$, f_1, K	86.6 ± 5.6	72.3 ± 5.6	56.7 ± 6.5	54.1 ± 5.7	76.0 ± 2.8	50.9 ± 4.8	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(1)}$, f_2	85.5 ± 9.7	74.3 ± 4.0	60.8 ± 5.4	57.2 ± 5.5	76.0 ± 2.8	51.3 ± 4.8	77.5 ± 2.5
$d_{\mathrm{WLLB}}^{(2)}$, f_2	85.0 ± 9.6	74.0 ± 4.6	59.6 ± 4.3	57.5 ± 7.1	75.9 ± 3.8	50.9 ± 4.3	77.9 ± 1.3
$d_{\mathrm{WLLB}}^{(3)}$, f_2	86.1 ± 8.3	73.7 ± 4.3	60.2 ± 5.5	54.9 ± 5.1	75.9 ± 3.6	50.3 ± 4.6	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(4)}$, f_2	87.1 ± 7.1	72.9 ± 5.0	59.6 ± 6.5	57.0 ± 6.4	**76.3 ± 2.8**	**51.5 ± 4.4**	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(1)}$, f_2, K	85.5 ± 8.2	73.9 ± 4.1	61.1 ± 6.0	57.8 ± 8.9	75.9 ± 2.7	51.3 ± 4.9	77.5 ± 2.5
$d_{\mathrm{WLLB}}^{(2)}$, f_2, K	85.5 ± 8.2	**74.5 ± 3.8**	60.5 ± 5.7	**58.4 ± 5.6**	75.3 ± 3.3	50.5 ± 4.5	77.9 ± 1.3
$d_{\mathrm{WLLB}}^{(3)}$, f_2, K	86.6 ± 8.3	74.2 ± 3.7	59.0 ± 7.6	56.1 ± 5.0	75.9 ± 3.3	50.6 ± 4.6	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(4)}$, f_2, K	86.6 ± 7.1	73.8 ± 4.0	59.3 ± 6.2	57.0 ± 5.3	76.2 ± 3.1	51.4 ± 4.3	78.2 ± 0.8
WWL	83.5 ± 3.0	72.5 ± 3.9	60.2 ± 5.6	57.5 ± 8.1	74.5 ± 4.1	48.7 ± 3.5	77.3 ± 3.1
WL	86.7 ± 4.4	70.8 ± 2.9	54.4 ± 6.1	54.7 ± 7.8	71.6 ± 4.7	50.7 ± 3.6	**78.8 ± 1.7**
WL-OA	85.0 ± 9.7	72.7 ± 4.1	57.9 ± 7.0	52.6 ± 8.4	73.4 ± 3.5	50.7 ± 3.7	78.2 ± 0.8

Table 4: **SVM classification accuracy.** For $d_{\mathrm{WL}}^{(k)}$ and d_{WLLB}, we show results using both the standard SVM and the KSVM (indicated by K in the table) (Loosli et al., 2015). Values are mean ± standard deviation in percent; the highest mean in each column is bolded, including ties.

We first define the general class of Markov Chain Neural Networks using Lipschitz maps.

Definition 27 (General Markov Chain Neural Networks) *For any family of Lipschitz maps $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ for $i = 1, \ldots, k+1$, and any Lipschitz map $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$, we define a map of the following form, which we call a k-layer* general Markov Chain Neural Network (gMCNN$_k$)*:*

$$h := \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} : \mathcal{M}^L(\mathbb{R}^d) \to \mathbb{R}. \tag{8}$$

We denote the collection of all such networks by $\mathcal{NN}_k^{gen}(\mathbb{R}^d)$.

In practice, functions φ_i and ψ are usually learnable functions instead of general Lipschitz maps. Hence, we would focus on cases when $\varphi = \varphi_\sigma$ is a multilayer perceptron (MLP) with a single hidden layer, i.e., any map of the form $\varphi_\sigma(x) := C\sigma * (Wx + b)$, where C, W are matrices, b is any vector, and $\sigma*$ represents elementwise application of a **fixed** activation function σ. For technical reasons, we also assume that σ is *Lipschitz*[4] and *non-polynomial*. For example, one can choose σ to be ReLU or sigmoid.

Definition 28 (Markov Chain Neural Networks) *If φ_i and ψ are restricted to be multilayer perceptrons (MLPs) with Lipschitz activation functions such as ReLU or sigmoid, the resulting k-layer network defined in Equation* (8) *is referred to as a* Markov Chain Neural Network (MCNN$_k$)*. The collection of all such networks is denoted by $\mathcal{NN}_k(\mathbb{R}^d)$, where:*

$$\mathcal{NN}_k(\mathbb{R}^d) \subseteq \mathcal{NN}_k^{gen}(\mathbb{R}^d). \tag{9}$$

Note the resemblance between our MCNNs and message passing neural networks (MPNNs) for graphs (Gilmer et al., 2017): Specifically, $(\ell_X)_\# m_x^X$ is analogous to the Aggregation operation, q_ϕ is analogous to the Update operation, and $\psi \circ S_\phi$ corresponds to the readout function that appears in the context of MPNN. We will specify how MCNNs reduce to MPNNs in more detail in Section 4.2.

Discriminative Power of MCNNs. It turns out that $\mathcal{NN}_k(\mathbb{R}^d)$ has the same discriminative power as the WL distance.

Theorem 29 *Given any $(\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y) \in \mathcal{M}^L(\mathbb{R}^d)$,*

1. *if $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = 0$, then for every $h \in \mathcal{NN}_k(\mathbb{R}^d)$ one has that $h((\mathcal{X}, \ell_X)) = h((\mathcal{Y}, \ell_Y))$;*
2. *if $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) > 0$, then there exists $h \in \mathcal{NN}_k(\mathbb{R}^d)$ such that $h((\mathcal{X}, \ell_X)) \neq h((\mathcal{Y}, \ell_Y))$.*

Recall from Corollary 19 that given any $q \in (\frac{1}{2}, 1)$, any injective map g and any labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), we need at most $2n$ steps to determine whether $d_{\mathrm{WL}}(\left(\mathcal{X}_q(G_1), \ell_{G_1}^g\right), \left(\mathcal{X}_q(G_2), \ell_{G_2}^g\right)) = 0$, where $n = \max(|V_{G_1}|, |V_{G_2}|)$. Consequently, MCNNs have the same discriminative power as the WL test:

Corollary 30 *Given any finite labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), for any $\frac{1}{2} < q < 1$, the WL test distinguishes (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) iff there exists $h \in \mathcal{NN}_{2n}(\mathbb{R}^d)$ for which $h\Big(\Big(\mathcal{X}_q(G_1), \ell_{G_1}^g\Big)\Big) \neq h\Big(\Big(\mathcal{X}_q(G_2), \ell_{G_2}^g\Big)\Big)$.*

4. Note that any MLP with a Lipschitz activation function is Lipschitz. This fact will be used in later proofs.

A stronger result on discriminative power. Utilizing the general MCNNs, we establish a result stronger than Theorem 29 under a setting similar to (but more flexible than) the one used in (Xu et al., 2019). Choose any *countable* subset $Z \subseteq \mathbb{R}^d$ and any *countable* subset $P \subseteq [0,1]$. Let $\mathcal{M}_P^L(Z)$ denote the collection of all LMMCs $(\mathcal{X}, \ell_X)$ so that the codomain of ℓ_X is Z and for any $x, x' \in X$, $m_x^X(x') \in P$ and $\mu_X(x) \in P$.

Example 7 *Fix any parameter $q \in [0,1)$ and any countable label space $Z \subseteq \mathbb{R}^d$. Let $P := \{0,1,q\} \cup \left\{ \frac{(1-q)m}{n} : m, n \in \mathbb{N} \right\}$. Then, it is easy to see that $\mathcal{M}_P^L(Z)$ contains every graph induced LMMCs (see Definition 7) with parameter q.*

Let $\mathcal{NN}_k^{1,\mathrm{gen}}(\mathbb{R}^d)$ denote the collection of maps in $\mathcal{NN}_k^{\mathrm{gen}}(\mathbb{R}^d)$ where φ_is and ψ have output dimension 1. Then, we establish the following main result:

Theorem 31 *For each $k \geq 0$, there exists $h \in \mathcal{NN}_k^{1,gen}(\mathbb{R}^d)$ such that for any $(\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y) \in \mathcal{M}_P^L(Z)$,*

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) > 0 \iff h((\mathcal{X}, \ell_X)) \neq h((\mathcal{Y}, \ell_Y)).$$

Lipschitz property of MCNNs. In addition to the study of the discriminative power, in other words, the study of the zero sets of functions in $\mathcal{NN}_k(\mathbb{R}^d)$, one can generalize the results above in a quantitative manner. More specifically, we establish that MCNNs are Lipschitz w.r.t. the WL distance. This in particular indicates that MCNNs are stable w.r.t. small perturbations of LMMCs in the sense of the WL distance.

Theorem 32 (Lipschitz property) *Let $h := \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} : \mathcal{M}^L(\mathbb{R}^d) \to \mathbb{R}$ be a k-layer MCNN. Assume that for $i = 1, \ldots, k+1$, φ_i is C_i-Lipschitz for some $C_i > 0$ and ψ is C-Lipschitz for some $C > 0$. Then,*

$$|h(\mathcal{X}, \ell_X) - h(\mathcal{Y}, \ell_Y)| \leq C \cdot \Pi_{i=1}^{k+1} C_i \cdot d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)).$$

We also remark that the Lipschitz property for MCNNs above is essential in proving the universality of MCNNs which we establish below.

Universality of MCNNs. Now we establish that MCNNs are universal. We start by introducing some notation. We let $\mathcal{M}_k^L(\mathbb{R}^d) := (\mathcal{M}^L(\mathbb{R}^d), d_{\mathrm{WL}}^{(k)})$. Then, it turns out that MCNNs can approximate continuous functions defined on $\mathcal{M}_k^L(\mathbb{R}^d)$.

Theorem 33 (Universal approximation) *For any $k \in \mathbb{N}$, let $\mathcal{K} \subseteq \mathcal{M}_k^L(\mathbb{R}^d)$ be any compact subspace.*[5] *Denote $C(\mathcal{K}, \mathbb{R})$ the set of all continuous functions on $\mathcal{K}$ and denote*

$$\mathcal{NN}_k(\mathbb{R}^d)|_{\mathcal{K}} := \left\{ h|_{\mathcal{K}} : h \in \mathcal{NN}_k(\mathbb{R}^d) \right\}.$$

Then,

$$\overline{\mathcal{NN}_k(\mathbb{R}^d)|_{\mathcal{K}}} = C(\mathcal{K}, \mathbb{R}) \text{ where } \overline{S} \text{ denote the closure of a set } S.$$

5. We stress that here compactness is relative to the the WL distance.

4.2 Implications for Message Passing GNNs

MCNNs reduce to a special type of message passing GNNs (MP-GNNs) when restricting to graph induced LMMCs. In this way, the properties of MCNNs studied in the previous subsection can be applied to analyze GNNs. In particular, in this subsection, we establish discriminative power, Lipschitz property and the universal property of GNNs.

More precisely, we let $\mathcal{G}(\mathbb{R}^d)$ denote the collection of all $\mathbb{R}^d$-labeled graphs and given $q > 0$, let $I_q : \mathcal{G}(\mathbb{R}^d) \to \mathcal{M}^L(\mathbb{R}^d)$ denote the map which sends a $\mathbb{R}^d$-labeled graph into a $\mathbb{R}^d$-LMMC via the method described in Definition 7. Let $\mathcal{G}_q(\mathbb{R}^d) := I_q(\mathcal{G}(\mathbb{R}^d)) \subseteq \mathcal{M}^L(\mathbb{R}^d)$. Then, for any $k \geq 0$, $d_{\mathrm{WL}}^{(k)}$ restricted to $\mathcal{G}_q(\mathbb{R}^d)$ induces a pseudometric, which we denote by $d_{\mathcal{G},q}^{(k)}$. We then call $d_{\mathcal{G},q}^{(k)}$ the (q-damped) WL distance of depth k between labeled graphs.

Now, we consider the following special type of MP-GNNs.

Definition 34 (MP-GNNs) *Given $q > 0$, we define a k-layer general Message Passing Graph Neural Network (MP-GNN) as follows. For each layer i, the message passing step is defined as:*

$$\textit{Message Passing} \quad \ell_G^{i+1}(v) := \begin{cases} q\varphi_{i+1}(\ell_G^i(v)) + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} w_{vv'} \varphi_{i+1}(\ell_G^i(v')) \\ \qquad \textit{if } \deg(v) > 0, \\ \varphi_{i+1}(\ell_G^i(v)) \textit{ if } \deg(v) = 0 \end{cases}$$

The final readout step is defined as:

$$\textit{Readout} \quad h((G, \ell_G)) := \psi \left(\sum_{v \in V} \frac{\overline{\deg}(v)}{\sum_{v' \in V} \overline{\deg}(v')} \varphi_{k+1}(\ell_G^k(v)) \right),$$

where $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ and $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$ are Lipschitz functions. The collection of all such networks with k layers is denoted by $\mathcal{NN}_k^{q,gen}(\mathbb{R}^d)$. When φ_i and ψ are MLPs with Lipschitz activation functions, we refer to the resulting network as a k-layer (MLP-based) MP-GNN *whose collection is denoted by $\mathcal{NN}_k^q(\mathbb{R}^d)$.*

Now, we show how MCNNs reduce to MP-GNNs when restricted to graph induced LMMCs. Let $h : \mathcal{G}(\mathbb{R}^d) \to \mathbb{R}$ be a k-layer MP-GNN as defined above with maps $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ for $i = 1, \ldots, k+1$ and $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$. Then, we define the following k-layer MCNN:

$$H := \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} : \mathcal{M}^L(\mathbb{R}^d) \to \mathbb{R}.$$

Proposition 35 *Recall that $I_q : \mathcal{G}(\mathbb{R}^d) \to \mathcal{M}^L(\mathbb{R}^d)$ sends a labeled graph to a LMMC and the image $\mathrm{Im}(I_q) = \mathcal{G}_q(\mathbb{R}^d)$. Then,*

$$h = H \circ I_q. \tag{10}$$

Based on this transition from MCNNs to MP-GNNs, we establish properties of MP-GNNs as follows.

Discriminative Power of MP-GNNs. The result below follows directly from Theorem 29.

Corollary 36 *Given any* $(G_1, \ell_{G_1}), (G_2, \ell_{G_2}) \in \mathcal{G}(\mathbb{R}^d)$,

1. *if* $d_{\mathcal{G},q}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) = 0$, *then for every* $h \in \mathcal{NN}_k^q(\mathbb{R}^d)$ *one has that* $h((G_1, \ell_{G_1})) = h((G_2, \ell_{G_2}))$;
2. *if* $d_{\mathcal{G},q}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) > 0$, *then there exists* $h \in \mathcal{NN}_k(\mathbb{R}^d)$ *such that* $h((G_1, \ell_{G_1})) \neq h((G_2, \ell_{G_2}))$.

Similarly to Theorem 31, we can establish a stronger result on discriminative power. Let $Z \subseteq \mathbb{R}^d$ and $P \subseteq \mathbb{R}$ be countable subsets. Define $\mathcal{G}_P(Z)$ as the set of all Z-labeled weighted graphs with edge weights in P. Let $\mathcal{NN}_k^{q,1,\mathrm{gen}}(\mathbb{R}^d)$ denote the collection of maps in $\mathcal{NN}_k^{q,\mathrm{gen}}(\mathbb{R}^d)$ where each φ_i has output dimension 1. As a direct consequence of Theorem 31, we can establish a stronger result for general MP-GNNs.

Corollary 37 *For any* $k \geq 0$, *there exists* $h \in \mathcal{NN}_k^{q,1,\mathrm{gen}}(\mathbb{R}^d)$ *such that for any labeled graphs* $(G_1, \ell_{G_1}), (G_2, \ell_{G_2}) \in \mathcal{G}_P(Z)$:

$$d_{\mathcal{G},q}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) > 0 \iff h((G_1, \ell_{G_1})) \neq h((G_2, \ell_{G_2})).$$

Notice that whereas in Corollary 36 the existence of h may depend on the choice of labeled graphs, h in Corollary 37 is universal for all pairs of labeled graphs. Furthermore, while maps φ_i involved in intermediate layers of h in Corollary 36 could potentially have large dimensions (d_i could be very large), φ_i can be chosen to have very low dimension ($d_i = 1$) in Corollary 37. In this way, the latter result is stronger than the previous one.

Remark 38 *Xu et al. (2019) established a result similar to Corollary 37: for any* $k \geq 0$, *there exists an MP-GNN* h *with the same discriminative power as the* k*-step WL test when restricted to graphs with labels from a common* countable *set and with a* uniformly bounded *number of vertices. However, the MP-GNN* h *they constructed relies on aggregation functions defined on a* countable *set with an inherent discrete nature: when the countable set is a subset of* $\mathbb{R}^d$, *the functions may not extend continuously to a domain such as* $\mathbb{R}^d$. *In contrast, each function in* $h \in \mathcal{NN}_k^{q,1,\mathrm{gen}}(\mathbb{R}^d)$ *is a continuous map between Euclidean spaces.*

Lipschitz property of MP-GNNs. The result below follows directly from Theorem 32:

Corollary 39 *Given a* k*-layer MP-GNN* $h : \mathcal{G}(\mathbb{R}^d) \to \mathbb{R}$ *as described above, assume that* φ_i *is* C_i*-Lipschitz for* $i = 1, \ldots, k+1$ *and that* ψ *is* C*-Lipschitz. Then, for any two labeled graphs* (G_1, ℓ_{G_1}) *and* (G_2, ℓ_{G_2}), *one has that*

$$|h((G_1, \ell_{G_1})) - h((G_2, \ell_{G_2}))| \leq C \cdot \Pi_{i=1}^{k+1} C_i \cdot d_{\mathcal{G},q}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) \,.$$

Remark 40 *We note that the WL distance could potentially be used to study the Lipschitz property of other types of MP-GNNs. See Section A.3 for such a study of a normalized version of Graph Isomorphism Network (Xu et al., 2019).*

Universal approximation Based on Theorem 33, we finally establish the universal approximation property of MP-GNNs.

Corollary 41 (Universal approximation of k-layer MP-GNNs) *Given $q > 0$ and any $k \in \mathbb{N}$, let $\mathcal{K} \subseteq \left(\mathcal{G}(\mathbb{R}^d), d_{\mathcal{G},q}^{(k)}\right)$ be any compact subspace. Then,*

$$\overline{\mathcal{NN}_k^q(\mathbb{R}^d)|_{\mathcal{K}}} = C(\mathcal{K}, \mathbb{R}).$$

Our universality result resembles the one established in (Azizian and Lelarge, 2021) for message passing neural networks (MPNNs) which proves that MPNNs can universally approximate continuous functions on graphs with bounded size which are less or equally as discriminative as the WL test. Compared with their result, we remark that our universality result applies to the collection of all graphs *with no restriction on their size*. Moreover, we comment that their universal approximation result only works for sets of graphs with labels from a countable set as their proof is based upon the results in Xu et al. (2019); see Remark 38 for a related comment.

5. Reinterpretation of the WL Distance via Markovian Couplings

Observe that the original definition of the WL distance (cf. Definition 12), while well-suited for devising a computation algorithm (cf. Algorithm 1) for the distance, is rather intricate due to its iterative consideration of probability measures on spaces $\mathcal{P}^{\circ k}(Z)$ with increasing complexity (cf. Definition 10). In this section, we adopt a totally different viewpoint to interpret the concept using the language of stochastic processes, which provides a conceptualization independent of the mechanics of the WL test. Specifically, we introduce two complementary approaches: (1) using the concept of *Markovian couplings*, a special type of couplings between Markov chains, and (2) relating the WL distance and the theory of causal Optimal Transport (Lassalle, 2018; Backhoff et al., 2017), a variant of OT specialized for comparing stochastic processes. These interpretations of the WL distance offer valuable insights for future research in this field.

5.1 Markovian Couplings and a Stochastic Process Interpretation of the WL Distance

It turns out that the WL distance tracks how the difference between labels evolves along the trajectories of two coupled Markov chains (see Figure 7b). In this section, we formalize this intuition using stochastic process terminology. Before presenting our interpretation, we introduce some preliminary definitions.

Let $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ be any measure Markov chain. When the state space X is finite, a common description of $\mathcal{X}$ is via a sequence of random variables $\{\boldsymbol{X}_i : \Omega \to X\}_{i\in\mathbb{N}}$ from some probability space $(\Omega, \mathbb{P})$ such that

- $\text{law}(\boldsymbol{X}_0) = \mu_X$ and
- $\mathbb{P}(\boldsymbol{X}_{i+1} = x' | \boldsymbol{X}_i = x) = m_x^X(x')$ for any $x, x' \in X$.

We refer to such a sequence as a *stochastic realization* of the MMC $\mathcal{X}$.

In this section, we restrict our attention to MMCs with *finite* state spaces so that we can fully exploit the language of stochastic realizations. Extensions to general compact state spaces are possible and will be addressed in future work.

Example 8 *A concrete example of a stochastic realization is given by letting* $\Omega := X^{\mathbb{N}}$ *be the path space, and defining the projection maps* $\{\boldsymbol{X}_i : (X^{\mathbb{N}}, \mathbb{P}_X) \to X\}_{i\in\mathbb{N}}$.

In this way, given any two MMCs $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ and $\mathcal{Y} = (Y, m_\bullet^Y, \mu_X)$, one can consider any of their corresponding stochastic realizations $\{\boldsymbol{X}_i : (\Omega_X, \mathbb{P}_X) \to X\}_{i\in\mathbb{N}}$ and $\{\boldsymbol{Y}_i : (\Omega_Y, \mathbb{P}_Y) \to X\}_{i\in\mathbb{N}}$. Then, one can consider couplings $\mathbb{P} \in \mathcal{C}(\mathbb{P}_X, \mathbb{P}_Y) \subseteq \mathcal{P}(\Omega_X \times \Omega_Y)$. This also gives rise to a stochastic process $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i) : (\Omega_X \times \Omega_Y, \mathbb{P}) \to X \times Y\}$. This stochastic process may not be Markovian a priori. If one imposes the Markov property on such a stochastic process, then this leads to the following definition of Markovian couplings which helps to provide an intuitive characterization of the WL distance. We now define the Markovian couplings between two MMCs through stochastic realizations.

Definition 42 (Markovian couplings for stochastic realizations) *Given MMCs* $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ *and* $\mathcal{Y} = (Y, m_\bullet^Y, \mu_Y)$*, let* $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}$ *and* $\{\boldsymbol{Y}_i\}_{i\in\mathbb{N}}$ *be stochastic realizations on the* same *probability space* $(\Omega, \mathbb{P})$*. We call the sequence of random variables* $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ *a* Markovian coupling *between* $\mathcal{X}$ *and* $\mathcal{Y}$*, if the following two properties hold:*

1. *(**Joint Markov Property**) The sequence* $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ *satisfies the Markov property, i.e., for any* $i \in \mathbb{N}$*,* $x_0, \dots, x_i \in X$ *and* $y_0, \dots, y_i \in Y$*, one has that*

$$\begin{aligned}&\mathbb{P}\big((\boldsymbol{X}_i, \boldsymbol{Y}_i) = (x_i, y_i) | \{(\boldsymbol{X}_j, \boldsymbol{Y}_j) = (x_j, y_j)\}_{0\le j\le i-1}\big)\\ &= \mathbb{P}\big((\boldsymbol{X}_i, \boldsymbol{Y}_i) = (x_i, y_i) | (\boldsymbol{X}_{i-1}, \boldsymbol{Y}_{i-1}) = (x_{i-1}, y_{i-1})\big).\end{aligned}$$

2. *(**Transition Coupling**) For any* $i \in \mathbb{N}$*,* $x \in X$ *and* $y \in Y$*, the probability measure* $(\nu_{i+1})_{x,y}$ *constructed below satisfies that* $(\nu_{i+1})_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)$*:*

$$(\nu_{i+1})_{x,y}(x', y') := \mathbb{P}\big((\boldsymbol{X}_{i+1}, \boldsymbol{Y}_{i+1}) = (x', y') | (\boldsymbol{X}_i, \boldsymbol{Y}_i) = (x, y)\big),\ \forall x', y' \in X \times Y.$$

In other words, a Markovian coupling is a Markov chain on $X \times Y$*, whose transition kernels match the individual Markov kernels of* $\mathcal{X}$ *and* $\mathcal{Y}$*. We let* $\mathcal{C}_{\mathcal{M}}(\mathcal{X}, \mathcal{Y})$ *denote the collection of all Markovian couplings* $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ *between* $\mathcal{X}$ *and* $\mathcal{Y}$*.*

In short, a Markovian coupling of two Markov chains $\mathcal{X}$ and $\mathcal{Y}$ is a *time-inhomogeneous* Markov chain on a state space $X \times Y$ that is the product of the state spaces of the two chains and whose transition kernel at each time step is a coupling between the transition kernels of the two chains; see Figure 7a for an illustration.

Remark 43 (Initial distribution) *For a Markovian coupling* $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{X}, \mathcal{Y})$*, at each step* $k \ge 0$*, we have that* $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$ *is a coupling between* $\mathrm{law}(\boldsymbol{X}_k) = \mu_X$ *and* $\mathrm{law}(\boldsymbol{Y}_k) = \mu_Y$ *(where we used the assumption that* μ_X *and* μ_Y *are stationary). In particular,* $(\boldsymbol{X}_0, \boldsymbol{Y}_0) \in \mathcal{C}(\mu_X, \mu_Y)$*. Hence, any Markovian coupling starts with an initial distribution which is itself a coupling between the initial distributions of the two Markov chains.*

Remark 44 (Coupling method) *This definition of Markovian coupling is a generalization of a common coupling technique used for proving convergence of Markov chains (see for example (Levin and Peres, 2017)). Note that whereas we define Markovian couplings for two arbitrary Markov chains, in the context of "coupling technique", the two Markov chains involved have the same transition kernels.*

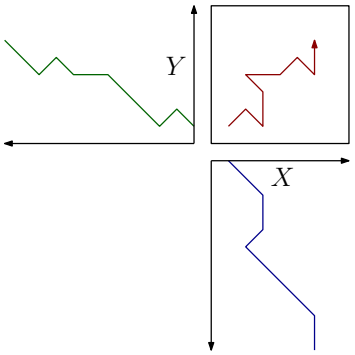
Remark 45 (Existence of Markovian couplings) *Given transition kernels $(\nu_{i+1})_{\bullet,\bullet}$ for all steps $i = 0, \ldots$ and an initial distribution γ, one can always construct a Markovian coupling via the Kolmogorov extension theorem (Kolmogorov and Bharucha-Reid, 2018).*

The following result is our key observation: it turns out that the WL distance is some variant of the Wasserstein distance between probability measures in path spaces with the Markov coupling constraint.

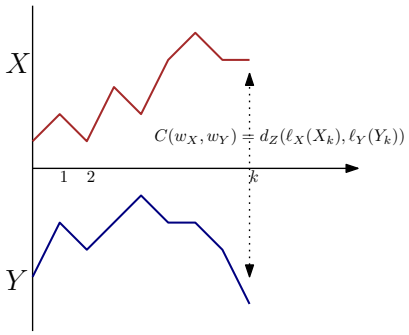
Theorem 46 *For any Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that*

$$d_{\mathrm{WL}}^{(k)}\left((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)\right) = \inf_{\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{X},\mathcal{Y})} \mathbb{E}\, d_Z(\ell_X(\boldsymbol{X}_k), \ell_Y(\boldsymbol{Y}_k)).$$

In this way, we have interpreted the WL distance of depth k as a variant of the Wasserstein distance between the marginal distributions of Markov chains at step k. Here the cost between two states $w_X, w_Y \in \Omega$ is given by $d_Z(\ell_X(\boldsymbol{X}_k(w_X), \ell_Y(\boldsymbol{Y}_k(w_Y))))$ through the label functions. Essentially, the WL distance at depth k is comparing the distribution of the Markov chains at step k; see Figure 7b for an illustration.



(a) Markovian coupling



(b) WL distance interpretation

Figure 7: **Left:** A Markovian coupling between $\mathcal{X}$ and $\mathcal{Y}$ is a Markov chain in the product space $X \times Y$ whose projections onto X and Y have the same law as $\mathcal{X}$ and $\mathcal{Y}$, respectively. **Right:** The WL distance of depth k is a variant of the Wasserstein distance between path distributions, with cost given by $d_Z\Big(\ell_X\big(\boldsymbol{X}_k(w_X)\big),\, \ell_Y\big(\boldsymbol{Y}_k(w_Y)\big)\Big)$ for any paths w_X and w_Y.

5.2 The Connection with Causal Optimal Transport

Now, we comment that the Markovian coupling characterization formula given in the previous section is deeply connected with the notion of Causal Optimal Transport (COT) (Lassalle, 2018). We elucidate this point in this section.

Causal Optimal Transport (COT) provides a framework for comparing stochastic processes by restricting the set of admissible couplings to those that preserve the temporal (or causal) structure. COT has found applications in mathematical finance (Glanzer et al., 2019; Backhoff-Veraguas et al., 2020) and in machine learning (Xu et al., 2020; Xu and Acciaio, 2022; Klemmer et al., 2022). As COT lies in the framework of Optimal Transport, multiple Sinkhorn algorithms have been proposed in the literature (Pichler and Weinhardt, 2022; Eckstein and Pammer, 2022) for accelerating computations. The connection between the WL distance and COT that we study in this section may eventually result in efficient algorithms for computing/approximating the WL distance. We leave this for future study but see the discussion below Theorem 49. Furthermore, we believe that this connection can also be useful for extending the WL distance to the case when input LMMCs have different labeling space Z's.

We first review some basics of COT. Given two (finite) spaces X and Y and an integer $k \in \mathbb{N}$, consider the product spaces X^{k+1} and Y^{k+1} (these product spaces are viewed as spaces of paths in X or Y of length $k+1$)[6]. Let $\alpha \in \mathcal{P}(X^{k+1})$ and $\beta \in \mathcal{P}(Y^{k+1})$. Now we are ready to define the notion of (bi)causal coupling between α and β.

Definition 47 ((bi)causal coupling) *A coupling measure $\pi \in \mathcal{C}(\alpha, \beta)$ is said to be* causal *from α to β if it satisfies*

$$\pi((y_0, \dots, y_l)|(x_0, \dots, x_k)) = \pi((y_0, \dots, y_l)|(x_0, \dots, x_l))$$

for all $l \in \{0, \dots, k\}$ and $(x_0, \dots, x_k) \in X^{k+1}$. Here, the notataion $\pi(\cdot|\cdot)$ denotes conditional probability. This implies that, at time l and given the past $(x_0, \dots, x_l)$ of X, the distribution of y_l does not depend on the future $(x_{l+1}, \dots, x_k)$ of X.

Moreover, π is said to be bicausal *if it is causal both from α to β and from β to α. The set of bicausal couplings between α and β will be denoted by $\mathcal{C}_{\mathrm{bc}}(\alpha, \beta)$.*

Now, let $c : X^{k+1} \times Y^{k+1} \to \mathbb{R}_+$ be any cost function. Then, the *bicausal OT problem* is formulated as follows:

$$d^c(\alpha, \beta) := \inf_{\pi \in \mathcal{C}_{\mathrm{bc}}(\alpha,\beta)} \int_{X^{k+1} \times Y^{k+1}} c(x, y) \pi(dx \times dy).$$

Similarly, one can formulate the causal OT (COT) problem by considering causal couplings. We use the acronym COT to refer to both causal and bicausal OT problems.

It turns out that Markovian couplings naturally give rise to bicausal couplings by restricting paths with infinite time steps to paths with finite time steps. More precisely, given two MMCs $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ and $\mathcal{Y} = (Y, m_\bullet^Y, \mu_X)$, we consider any Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{X}, \mathcal{Y})$ (cf. Definition 42). Then, for any $k \in \mathbb{N}$, we let $\alpha^k := \mathrm{law}((\boldsymbol{X}_0, \dots, \boldsymbol{X}_k)) \in \mathcal{P}(X^{k+1})$ and let $\beta^k := \mathrm{law}((\boldsymbol{Y}_0, \dots, \boldsymbol{Y}_k)) \in \mathcal{P}(Y^{k+1})$. In fact, it is easy to see that α^k and β^k can be expressed explicitly as follows.

$$\alpha^k((x_0, x_1, \dots x_k)) = m^X_{x_{k-1}}(x_k) \cdots m^X_{x_0}(x_1) \mu_X(x_0),$$
$$\beta^k((y_0, y_1, \dots y_k)) = m^Y_{y_{k-1}}(y_k) \cdots m^Y_{y_0}(y_1) \mu_Y(y_0).$$

6. The COT framework can incorporate the setting of both discrete and continuous paths of finite or infinite length. In this paper, we will focus on paths of finite length for simplicity and clarity of our presentation.

Now, we further let

$$\pi^k := \mathrm{law}((\boldsymbol{X}_0, \ldots, \boldsymbol{X}_k, \boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_k)) \in \mathcal{P}(X^{k+1} \times Y^{k+1}).$$

Then, we have that

Lemma 48 *π^k is a bicausal coupling between α^k and β^k.*

Now, we consider the following cost function $c^k : X^{k+1} \times Y^{k+1} \to \mathbb{R}_+$ defined by

$$((x_0, \ldots, x_k), (y_0, \ldots, y_k)) \mapsto d_Z(\ell_X(x_k), \ell_Y(y_k)) \tag{11}$$

and state the following theorem.

Theorem 49 (WL distance as a COT problem) *For any Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that*

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = d^{c^k}(\alpha^k, \beta^k).$$

Theorem 49 shows that the WL distance can be formulated as a COT problem. Consequently, practitioners can likely use causal Sinkhorn algorithms (Pichler and Weinhardt, 2022; Eckstein and Pammer, 2022) to approximate the WL distance efficiently.

Remark 50 (Markov chains in the label space) *In fact, one can also transform a Z-LMMC $(\mathcal{X}, \ell_X)$ into a Markov chain in the label space Z, if we require that $\ell_X : X \to Z$ is injective. In this way, the WL distance can be also interpreted as solving a COT problem in the label space Z. See Section A.2 for more details.*

6. Relationship with the Gromov Wasserstein Distance

Given a LMMC $(\mathcal{X}, \ell_X)$, the label ℓ_X induces the following pseudometric on X: $d_X(x, x') := d_Z(\ell_X(x), \ell_X(x'))$ for $x, x' \in X$. This suggests structures which are closely related to LMMCs: *Markov chain metric spaces* (MCMSs for short). A MCMS is any tuple $(\mathcal{X}, d_X)$ where $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ is a finite MMC and d_X is a proper distance on X. Obviously, endowing a MCMS $(\mathcal{X}, d_X)$ with a label function ℓ_X and forgetting d_X produces a LMMC $(\mathcal{X}, \ell_X)$. We let $\mathcal{M}^{\mathrm{MS}}$ denote the collection of all MCMSs. We now construct a Gromov-Wasserstein type distance between MCMSs and study its relationship with the WL distance.

6.1 The GW Distance between MCMSs

Recall from Equation (2) the definition of the standard GW distance between MMSs. Intuitively, in order to identify a suitable GW-like distance between MCMSs, we would like to alter Equation (2) in order to carry out a comparison between Markov kernels.

Definition 51 *For any $k \geq 1$ and any MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$, we define the k-distortion of any pair consisting of a Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}}$ between $\mathcal{X}$ and $\mathcal{Y}$ and an independent[7] coupling $(\boldsymbol{X}_0', \boldsymbol{Y}_0')$ between μ_X and μ_Y as:*

$$\mathrm{dis}^{(k)}\left(\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}}, (\boldsymbol{X}_0', \boldsymbol{Y}_0')\right) := \mathbb{E}\,|d_X(\boldsymbol{X}_0', \boldsymbol{X}_k) - d_Y(\boldsymbol{Y}_0', \boldsymbol{Y}_k)|.$$

7. Here we implicitly require that the random variable $(\boldsymbol{X}_0', \boldsymbol{Y}_0')$ and the Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}}$ are defined in a same probability space.

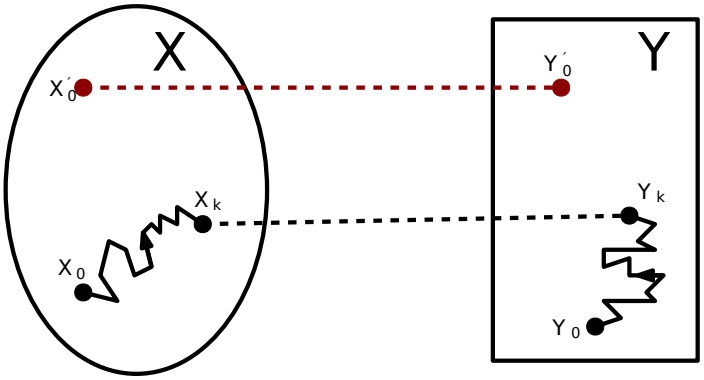


Figure 8: **Illustration of $d_{\mathrm{GW}}^{\mathrm{MCMS}}$** The distortion $\mathrm{dis}^{(k)}$ measures the similarity between the the evolution processes of the two Markov chains from step 0 to step k.

Then, the k-Gromov-Wasserstein distance *between the MCMSs* $(\mathcal{X}, d_X)$ *and* $(\mathcal{Y}, d_Y)$ *is defined by*

$$d_{\mathrm{GW}}^{(k)}\left((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)\right) := \inf_{\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}, (\boldsymbol{X}_0', \boldsymbol{Y}_0')} \mathrm{dis}^{(k)}\left(\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}, (\boldsymbol{X}_0', \boldsymbol{Y}_0')\right), \tag{12}$$

where the infimum is taken over all possible pairs consisting of a Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ *between* $\mathcal{X}$ *and* $\mathcal{Y}$ *and a random variable* $(\boldsymbol{X}_0', \boldsymbol{Y}_0')$ *independent of the Markovian coupling such that* $\mathrm{law}\left((\boldsymbol{X}_0', \boldsymbol{Y}_0')\right) = \mathrm{law}((\boldsymbol{X}_0, \boldsymbol{Y}_0))$.

Finally, we define the MCMS GW distance as follows:

$$d_{\mathrm{GW}}^{\mathrm{MCMS}}\left((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)\right) := \sup_{k\geq 0} d_{\mathrm{GW}}^{(k)}\left((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)\right).$$

See Figure 8 for an illustration of the k-Gromov-Wasserstein distance.

Now we identify a notion of isomorphism between two MCMSs.

Definition 52 *Two MCMSs* $(\mathcal{X}, d_X)$ *and* $(\mathcal{Y}, d_Y)$ *are said to be* isomorphic *if there exists an isometry* $\psi : X \to Y$ *such that* $\psi_{\#}\mu_X = \mu_Y$ *and* $\psi_{\#}m_x^X = m_{\psi(x)}^Y$ *for all* $x \in X$.

Under this notion of isomorphism, we establish the following result.

Proposition 53 $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ *defines a proper distance on the collection* $\mathcal{M}^{\mathrm{MS}}$ *modulo isomorphism of MCMSs.*

Unlike the case for d_{WL}, two MCMSs have zero $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ distance iff they are isomorphic.

6.1.1 A simple example and a lower bound for $d_{\mathrm{GW}}^{\mathrm{MCMS}}$

The simplest MCMS is the one point MCMS denoted by $*$. For any MCMS $(\mathcal{X}, d_X)$, one can explicitly compute the MCMS GW distance between $(\mathcal{X}, d_X)$ and $*$.

Example 9 *For any MCMS $(\mathcal{X}, d_X)$ and for any $k \geq 1$, we have that*

$$\begin{aligned} d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), *) &= \mathbb{E}|d_X(\boldsymbol{X}_0', \boldsymbol{X}_k) - d_*(*, *)| \\ &= \int_X \int_X d_X(x', x)\mu_X(dx')\mu_X(dx), \end{aligned}$$

where the last equality follows from the fact that μ_X is stationary and hence $\mathrm{law}(\boldsymbol{X}_k) = \mu_X$. *This equation implies that*

$$d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), *) = \int_X \int_X d_X(x', x)\mu_X(dx')\mu_X(dx).$$

Similarly to (Mémoli, 2011a, Definition 5.2), we use this distance to the one point space to identify a notion of diameter.

Definition 54 (MCMS diameter) *We let $*$ denote the one point MCMS. For each $k \geq 1$ and a MCMS $(\mathcal{X}, d_X)$, we define its* MCMS diameter *as follows:*

$$\mathrm{diam}_{\mathrm{MCMS}}((\mathcal{X}, d_X)) \coloneqq d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), *).$$

By the triangle inequality, one obtains the following simple lower bound:

Proposition 55 *For all two MCMSs $(\mathcal{X}, d_X)$,$(\mathcal{Y}, d_Y)$ and $k \geq 1$, we have*

$$|\mathrm{diam}_{\mathrm{MCMS}}((\mathcal{X}, d_X)) - \mathrm{diam}_{\mathrm{MCMS}}((\mathcal{Y}, d_Y))| \leq d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \leq d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)).$$

Although the lower bound above is quite coarse, in Section 6.2 we develop more refined lower bounds for $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ utilizing the WL distance. This focus on lower bounds is motivated by the NP-hardness of exactly computing $d_{\mathrm{GW}}^{\mathrm{MCMS}}$, as we detail in Remark 61.

6.1.2 Relationship with the decoupled Gromov-Wasserstein distance

Let $\mathbf{X} = (X, d_X, \mu_X)$ and $\mathbf{Y} = (Y, d_Y, \mu_Y)$ be two MMSs. Recall from Equation (2) that for any coupling $(\boldsymbol{X}, \boldsymbol{Y})$, its ($\ell^1$-)distortion is defined as follows:

$$\mathrm{dis}((\boldsymbol{X}, \boldsymbol{Y})) \coloneqq \mathbb{E}|d_X(\boldsymbol{X}, \bar{\boldsymbol{X}}) - d_Y(\boldsymbol{Y}, \bar{\boldsymbol{Y}})|,$$

where $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}})$ denotes an independent copy of $(\boldsymbol{X}, \boldsymbol{Y})$, i.e., an independent random variable with the same law as $(\boldsymbol{X}, \boldsymbol{Y})$.

On the other hand, when two independent couplings $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}', \boldsymbol{Y}')$ (with possibly distinct laws) are given, we define the (decoupled ℓ^1-)distortion associated to $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}', \boldsymbol{Y}')$ as follows:

$$\mathrm{dis}((\boldsymbol{X}, \boldsymbol{Y}), (\boldsymbol{X}', \boldsymbol{Y}')) \coloneqq \mathbb{E}|d_X(\boldsymbol{X}, \boldsymbol{X}') - d_Y(\boldsymbol{Y}, \boldsymbol{Y}')|.$$

Then, we define the decoupled Gromov-Wasserstein distance as follows:

Definition 56 (Decoupled Gromov-Wasserstein distance) *Suppose two metric measure spaces $\mathbf{X} = (X, d_X, \mu_X)$, $\mathbf{Y} = (Y, d_Y, \mu_Y)$ are given. We define the decoupled Gromov-Wasserstein distance $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y})$ in the following way:*

$$d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) \coloneqq \inf_{\textit{independent } (\boldsymbol{X}, \boldsymbol{Y}), (\boldsymbol{X}', \boldsymbol{Y}') \in \mathcal{C}(\mu_X, \mu_Y)} \mathrm{dis}((\boldsymbol{X}, \boldsymbol{Y}), (\boldsymbol{X}', \boldsymbol{Y}')). \tag{13}$$

This version of the GW distance was implicit in the optimization procedure followed in (Mémoli, 2011a) and has been explicitly considered in (Séjourné et al., 2021; Vayer et al., 2020b) with a measure theoretic formulation. Obviously, $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y}) \leq d_{\mathrm{GW}}(\mathbf{X},\mathbf{Y})$ in general. Furthermore, this inequality is tight as one can see in the following remark.

Remark 57 *We let $\Gamma_{X,Y}(x,y,x',y') \coloneqq |d_X(x,x') - d_Y(y,y')|$ for any $x,x' \in X$ and $y,y' \in Y$. If the kernel $\Gamma_{X,Y} : X \times Y \times X \times Y \to \mathbb{R}$ is negative semi-definite, then one can show that $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y}) = d_{\mathrm{GW}}(\mathbf{X},\mathbf{Y})$ by invoking Séjourné et al. (2021, Theorem 2). More precisely, in terms of our terminology, if $(\boldsymbol{X},\boldsymbol{Y}),(\boldsymbol{X}',\boldsymbol{Y}')$ are the optimal couplings achieving the infimum in the definition of $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y})$, we have that*

$$d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y}) = \mathrm{dis}((\boldsymbol{X},\boldsymbol{Y}),(\boldsymbol{X}',\boldsymbol{Y}')) = \mathrm{dis}((\boldsymbol{X},\boldsymbol{Y})) = \mathrm{dis}((\boldsymbol{X}',\boldsymbol{Y}')) = d_{\mathrm{GW}}(\mathbf{X},\mathbf{Y}).$$

Just like the original version, $d_{\mathrm{GW}}^{\mathrm{bi}}$ also becomes a legitimate metric on the collection of metric measure spaces. This is another original contribution of our work.

Proposition 58 *The decoupled GW distance $d_{\mathrm{GW}}^{\mathrm{bi}}$ is a legitimate metric on $\mathcal{M}^{\mathrm{MS}}$.*

The decoupled GW distance gives a lower bound for the k-MCMS GW distance.

Proposition 59 *Let $(\mathcal{X},d_X)$ and $(\mathcal{Y},d_Y)$ be MCMSs, with underlying MMSs $\mathbf{X}=(X,d_X,\mu_X)$ and $\mathbf{Y}=(Y,d_Y,\mu_Y)$. For every $k \geq 1$,*

$$d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y}) \leq d_{\mathrm{GW}}^{(k)}\big((\mathcal{X},d_X),(\mathcal{Y},d_Y)\big).$$

MMS induced MCMS Given a metric measure space $\mathbf{X}=(X,d_X,\mu_X)$, we produce a MCMS $\mathcal{M}(\mathbf{X}) \coloneqq (\mathcal{X},d_X)$ where $\mathcal{X} \coloneqq (X,m_\bullet^X,\mu_X)$ by letting $m_\bullet^X \coloneqq \mu_X$ be the constant Markov kernel. It is easy to check that μ_X is a stationary distribution w.r.t. $m_\bullet^X$.

Now, one can prove that any k-MCMS GW distance between $\mathcal{M}(\mathbf{X})$ and $\mathcal{M}(\mathbf{Y})$ is the same as the decoupled GW distance between $\mathbf{X}$ and $\mathbf{Y}$ regardless of k.

Proposition 60 *Then, for any two metric measure spaces $\mathbf{X}=(X,d_X,\mu_X)$, $\mathbf{Y}=(Y,d_Y,\mu_Y)$, and $k \geq 1$, we have that*

$$d_{\mathrm{GW}}^{(k)}\big(\mathcal{M}(\mathbf{X}),\mathcal{M}(\mathbf{Y})\big) = d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y}).$$

Remark 61 (Computational complexity) *The quantity $d_{\mathrm{GW}}^{\mathrm{bi}}$ is in general NP-hard to compute (Scetbon et al., 2022), which (via Proposition 60) implies that $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ is also NP-hard to compute.*

6.2 The WL Distance vs. the GW Distance

Given a MCMS $(\mathcal{X},d_X)$, fix any $x \in X$. Then, one can endow $(\mathcal{X},d_X)$ with the label function $d_X(x,\bullet) : X \to \mathbb{R}$. This gives rise to the LMMC $(\mathcal{X},d_X(x,\bullet))$. Then, we have the following lower bound of $d_{\mathrm{GW}}^{(k)}$ in terms of $d_{\mathrm{WL}}^{(k)}$:

Proposition 62 *For each $k \geq 1$ and for any MCMSs $(\mathcal{X},d_X)$ and $(\mathcal{Y},d_Y)$, one has that*

$$\inf_{(\boldsymbol{X}_0,\boldsymbol{Y}_0)\in\mathcal{C}(\mu_X,\mu_Y)} \mathbb{E} d_{\mathrm{WL}}^{(k)}\big((\mathcal{X},d_X(\boldsymbol{X}_0,\bullet)),(\mathcal{Y},d_Y(\boldsymbol{Y}_0,\bullet))\big) \leq d_{\mathrm{GW}}^{(k)}((\mathcal{X},d_X),(\mathcal{Y},d_Y)). \tag{14}$$

Remark 63 *When $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ are induced from MMSs $\mathbf{X}$ and $\mathbf{Y}$, as shown in Proposition 60, the right-hand side of Equation* (14) *coincides with the decoupled GW distance. We point out that the left-hand side is also independent of k and actually coincides with the* third lower bound *(TLB) for the GW distance as defined in (Mémoli, 2011a). See Section B.5.6 for more details. Hence, the proposition above can be viewed as a generalization of the TLB to the setting of MCMS.*

In general, a MCMS $(\mathcal{X}, d_X)$ endowed with any label function $\ell_X : X \to Z$ induces a LMMC $(X, m_\bullet^X, \mu_X, \ell_X)$ which we denote by $(\mathcal{X}, \ell_X)$. If we assign label functions to all MCMSs in a suitable coherent way, we obtain that the WL distance between the induced LMMCs is stable w.r.t. the GW distance between corresponding MCMSs.

Definition 64 (Label invariant of MCMCs) *Given any metric space of labels Z, a Z-valued **label invariant** of MCMSs is a map $\ell_\bullet : \mathcal{M}^{\mathrm{MS}} \to Z^\bullet$ which by definition sends each MCMS $(\mathcal{X}, d_X)$ into a label function $\ell_X : X \to Z$. One such label invariant $\ell_\bullet$ will be said to be* stable *if for all $(\mathcal{X}, d_X), (\mathcal{Y}, d_Y) \in \mathcal{M}^{\mathrm{MS}}$, $k \in \mathbb{N}$, and for any Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ between $\mathcal{X}$ and $\mathcal{Y}$ and any independent coupling $(\boldsymbol{X}'_0, \boldsymbol{Y}'_0)$ between μ_X and μ_Y we have*

$$\mathbb{E}\, d_Z(\ell_X(\boldsymbol{X}_k), \ell_Y(\boldsymbol{Y}_k)) \leq \mathrm{dis}^{(k)}\left(\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}, (\boldsymbol{X}'_0, \boldsymbol{Y}'_0)\right).$$

By $\mathcal{L}(Z)$ we will denote the collection of all stable Z-valued label invariants.

Example 10 *One immediate example of the stable label invariant is the* eccentricity function $\mathrm{ecc}_\bullet$*: for any MCMS $(\mathcal{X}, d_X)$, $\mathrm{ecc}_X(x) \coloneqq \mathbb{E}_{x'\sim\mu_X} d_X(x, x')$ for $x \in X$. This can be seen as follows.*

$$\begin{aligned}\mathbb{E}|\mathrm{ecc}_X(\boldsymbol{X}_k) - \mathrm{ecc}_Y(\boldsymbol{Y}_k)| &= \mathbb{E}|\mathbb{E}_{\boldsymbol{X}'_0\sim\mu_X} d_X(\boldsymbol{X}_k, \boldsymbol{X}'_0) - \mathbb{E}_{\boldsymbol{Y}'_0\sim\mu_X} d_Y(\boldsymbol{Y}_k, \boldsymbol{Y}'_0)| \\ &\leq \mathbb{E}|d_X(\boldsymbol{X}'_0, \boldsymbol{X}_k) - d_Y(\boldsymbol{Y}'_0, \boldsymbol{Y}_k)|.\end{aligned}$$

Then, if one assigns stable labels to MCMSs, one immediately has that the WL distance between the induced LMMCs is stable w.r.t. the GW distance.

Theorem 65 *For every stable label invariant $\ell_\bullet \in \mathcal{L}(Z)$, $k \in \mathbb{N}$ and $(\mathcal{X}, d_X), (\mathcal{Y}, d_Y) \in \mathcal{M}^{\mathrm{MS}}$ we have that*

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \leq d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)).$$

Remark 66 (Sharpness of the GW lower bounds) *The bounds in Proposition 62 and Theorem 65 are sharp. Let $(\mathcal{X}, d_X)$ be any MCMS with at least two points and let $*$ be the one-point MCMS, with label 0. For the eccentricity labels from Example 10, stationarity gives, for every $k \geq 1$,*

$$d_{\mathrm{WL}}^{(k)}\big((\mathcal{X}, \mathrm{ecc}_X), (*, 0)\big) = \int_X \mathrm{ecc}_X(x)\, \mu_X(dx) = d_{\mathrm{GW}}^{(k)}\big((\mathcal{X}, d_X), *\big).$$

The common value is the mean pairwise distance in X and is positive. For each reference point $x \in X$ and every $k \geq 1$, stationarity also gives

$$d_{\mathrm{WL}}^{(k)}\big((\mathcal{X}, d_X(x, \bullet)), (*, 0)\big) = \mathrm{ecc}_X(x).$$

Averaging over $x \sim \mu_X$ gives equality in Proposition 62 with the same positive value. Thus, the coefficient 1 in either bound cannot be improved.

However, either bound can vanish while the GW distance is positive. Let X and Y be the vertex sets of $K_{3,3}$ and the triangular prism, respectively; see Figure 9. Give them their shortest-path distances d_X, d_Y and uniform probability measures μ_X, μ_Y, and denote the resulting MMSs by $\mathbf{X}$ and $\mathbf{Y}$. These spaces are not isometric: an isometry would preserve adjacency, but only the prism contains triangles. Hence, $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) > 0$ by Proposition 58.

For Theorem 65, let $\mathcal{X}$ and $\mathcal{Y}$ be the simple random walks on these graphs. The uniform measures are stationary since both graphs are 3-regular. Every vertex has three vertices at distance 1 and two at distance 2, so

$$\mathrm{ecc}_X(x) = \mathrm{ecc}_Y(y) = \frac{3 \cdot 1 + 2 \cdot 2}{6} = \frac{7}{6} \qquad \text{for all } x \in X,\ y \in Y.$$

These identical constant labels give zero WL distance at every depth. By Proposition 59, for every $k \geq 1$,

$$d_{\mathrm{WL}}^{(k)}\big((\mathcal{X}, \mathrm{ecc}_X), (\mathcal{Y}, \mathrm{ecc}_Y)\big) = 0 < d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) \leq d_{\mathrm{GW}}^{(k)}\big((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)\big).$$

For Proposition 62, keep the same MMSs but replace the random walks by the constant kernels $m_x^X = \mu_X$ and $m_y^Y = \mu_Y$. This gives $\mathcal{M}(\mathbf{X})$ and $\mathcal{M}(\mathbf{Y})$: each chain samples uniformly from the entire vertex set at every step. For any reference vertices $a \in X$ and $b \in Y$, the distance labels $d_X(a, \bullet)$ and $d_Y(b, \bullet)$ have the same distribution,

$$\nu = \frac{1}{6}\delta_0 + \frac{1}{2}\delta_1 + \frac{1}{3}\delta_2.$$

After one WL step, every state in both chains has label ν; subsequent steps give the common labels $\delta_\nu, \delta_{\delta_\nu}, \ldots$. Thus, every k-WL distance on the left-hand side of Proposition 62 is zero for $k \geq 1$, whereas the right-hand side equals $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) > 0$ by Proposition 60. Therefore, neither lower bound admits a finite multiplicative reverse inequality in general.

7. Conclusion and Future Directions

In this paper, we propose the Weisfeiler-Lehman distance – a *quantitative* extension of the Weisfeiler-Lehman test – for measuring the dissimilarity between objects in a fairly general family called *labeled measure Markov chains* (LMMCs). Our approach uncovers intriguing links with graph kernels, graph neural networks (GNNs), causal optimal transport, and the Gromov-Wasserstein distance.

In order to more directly compare the WL test with the WL distance without resorting to relabeling, one interesting future direction would be to redefine the WL distance via positive measures and "unbalanced" (Gromov-)Wasserstein distances (Liero et al., 2018; Séjourné et al., 2021; De Ponti and Mondino, 2022).

Whereas our paper focuses on the application of our WL distance to the graph setting, LMMCs can be used to model not just graphs but also far more general objects such as Riemannian manifolds (equipped with heat kernels) or graphons. Then, our neural network architecture (MCNN) has the potential to be applied to point sets sampled from manifolds too, as well as serving as the limiting object when studying the convergence of GNNs.

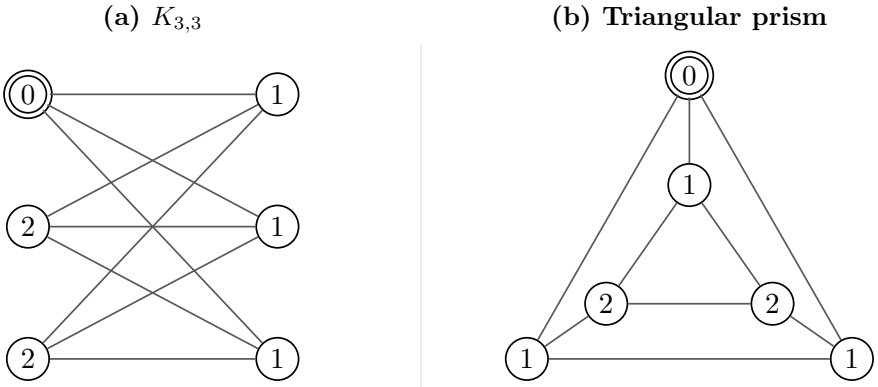


Figure 9: The graphs $K_{3,3}$ and the triangular prism from Remark 66. Vertex annotations indicate shortest-path distance from the double-circled reference vertex; only circles represent vertices. In either graph, every vertex has three vertices at distance 1 and two at distance 2, so its eccentricity is 7/6. These eccentricity labels give zero WL distance, although the graphs are nonisomorphic and have positive GW distance.

It would also be interesting to extend our WL distance to a higher order version that is analogous to the high order k-WL test and k-GNNs. We conjecture that a suitable notion of k-WL distance will converge to the Gromov-Wasserstein distance for MCMSs as k tends to infinity.

Finally, since many of these results are currently restricted to finite spaces, an extension to more general settings, such as compact spaces, is anticipated in future work.

Acknowledgments

This work is partially supported by NSF-DMS-1723003, NSF-CCF-1740761, NSF-RI-1901360, NSF-CCF-1839356, NSF-IIS-2050360, NSF-CCF-2112665, NSF-CCF-2310411.

Appendix A. Subsidiary Details

A.1 Useful Facts about Couplings

Here we collect some useful facts about couplings which will be used in subsequent proofs.

Lemma 67 *Let X, Y be finite metric spaces and let Z be a complete and separable metric space. Let $\varphi_X : X \to Z$ and $\varphi_Y : Y \to Z$ be measurable maps. Consider any $\mu_X \in \mathcal{P}(X)$ and $\mu_Y \in \mathcal{P}(Y)$. Then, we have that*

$$d_{\mathrm{W}}((\varphi_X)_\# \mu_X, (\varphi_Y)_\# \mu_Y) = \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} d_Z(\varphi_X(x), \varphi_Y(y)) \gamma(dx \times dy).$$

Proof The proof is based on the following result.

Lemma 68 *Under the same assumptions as in Lemma 67, if we let $\varphi := \varphi_X \times \varphi_Y$, then we have that*

$$\varphi_\# \mathcal{C}(\mu_X, \mu_Y) = \mathcal{C}\left((\varphi_X)_\# \mu_X, (\varphi_Y)_\# \mu_Y\right).$$

Proof [Proof of Lemma 68] Since X and Y are finite, $\varphi_X(X)$ and $\varphi_Y(Y)$ are discrete sets. Then, the lemma follows directly from Proposition 4.5 in (Schmitzer and Schnörr, 2013). ■

Hence,

$$\begin{aligned} d_{\mathrm{W}}((\varphi_X)_\# \mu_X, (\varphi_Y)_\# \mu_Y) &= \inf_{\gamma \in \mathcal{C}\left((\varphi_X)_\# \mu_X, (\varphi_Y)_\# \mu_Y\right)} \int_{X \times Y} d_Z(z_1, z_2) \gamma(dz_1 \times dz_2) \\ &= \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} d_Z(z_1, z_2) \, \varphi_\# \gamma(dz_1 \times dz_2) \\ &= \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} d_Z(\varphi_X(x), \varphi_Y(y)) \gamma(dx \times dy). \end{aligned}$$

■

The following lemma is a direct consequence of (Villani, 2009, Corollary 5.22)

Lemma 69 *For any complete and separable metric space Z and a continuous cost function $c : Z \times Z \to \mathbb{R}$ such that $\inf c > -\infty$, there exists a measurable map $\varphi : \mathcal{P}(Z) \times \mathcal{P}(Z) \to \mathcal{P}(Z \times Z)$ so that for every $\alpha, \beta \in \mathcal{P}(Z)$, $\varphi(\alpha, \beta)$ is an optimal coupling between α and β.*

A.2 The WL distance as COT in the label space Z

In Theorem 49 we are considering a special COT problem where the two involved state spaces for stochastic processes are different. We note that in the literature, most of the time COT is considering stochastic processes on the same state space. In this section, we then provide an alternative characterization of the WL distance in terms of COT on a same state space under certain conditions.

Given a Z-LMMC $(\mathcal{X}, \ell_X)$, if we require that $\ell_X : X \to Z$ is injective, then $(\mathcal{X}, \ell_X)$ induces a MMC on Z:

$$m_z^{\ell_X} := \begin{cases} (\ell_X)_\# m_x^X, & \exists x \in X \text{ such that } \ell_X(x) = z \\ \delta_z, & \text{otherwise} \end{cases}$$

We further let $\mu_{\ell_X} := (\ell_X)_\# \mu_X$ (note, however, μ_{ℓ_X} may not be fully supported on Z).

Now, given any two Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ and $k \in \mathbb{N}$, if we assume that ℓ_X and ℓ_Y are injective, then these two LMMCs induce two Markov chains on the same state space Z. It is natural to wonder whether one can characterize the WL distance of depth k via a COT problem on Z. Recall notation $\alpha^k \in \mathcal{P}(X^{k+1})$ and $\beta^k \in \mathcal{P}(Y^{k+1})$ from Section 5.2. We now let $\alpha_Z^k := (\ell_X^{k+1})_\# \alpha^k \in \mathcal{P}(Z^{k+1})$ and $\beta_Z^k := (\ell_Y^{k+1})_\# \beta^k \in \mathcal{P}(Z^{k+1})$. Here $\ell_X^{k+1} : X^{k+1} \to Z^{k+1}$ denotes the self product of ℓ_X, and ℓ_Y^{k+1} is similarly defined. Finally, we consider the cost $d_Z^k : Z^{k+1} \times Z^{k+1} \to \mathbb{R}$ defined by $d_Z^k((z_0, \ldots, z_k), (z_0', \ldots, z_k')) := d_Z(z_k, z_k')$. Note that $d_Z^k \circ (\ell_X^{k+1} \times \ell_Y^{k+1}) = c^k$, where c^k is defined through Equation (11). Then, we have the following result:

Theorem 70 *For any Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, if we assume that ℓ_X and ℓ_Y are injective, then we have that*

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = d^{d_Z^k}\left(\alpha_Z^k, \beta_Z^k\right).$$

For the proof of Theorem 70, we establish the following lemma.

Lemma 71 *$(\ell_X^{k+1} \times \ell_Y^{k+1})_\#$ induces a surjective map from $\mathcal{C}_{\mathrm{bc}}(\alpha^k, \beta^k)$ to $\mathcal{C}_{\mathrm{bc}}(\alpha_Z^k, \beta_Z^k)$.*

Given this lemma, and by the fact that $d_Z^k \circ (\ell_X^{k+1} \times \ell_Y^{k+1}) = c^k$, it is easy to check that $d^{c^k}(\alpha^k, \beta^k) = d^{d_Z^k}\left(\alpha_Z^k, \beta_Z^k\right)$. Hence, by emplyoing Theorem 49, we conclude our proof of Theorem 70.

Finally, we provide a proof of Lemma 71.

Proof [Proof of Lemmma 71] Fix an arbitrary $\pi^k \in \mathcal{C}_{\mathrm{bc}}(\alpha^k, \beta^k)$. Let $\pi_Z^k := (\ell_X^{k+1} \times \ell_Y^{k+1})_\# \pi^k$. Then, it is easy to check that π_Z^k is indeed a bicausal coupling between α_Z^k and β_Z^k, so we omit the proof.

Next, fix an arbitrary $\pi_Z^k \in \mathcal{C}_{\mathrm{bc}}(\alpha_Z^k, \beta_Z^k)$. Now, let us define a probability measure π^k on $X^{k+1} \times Y^{k+1}$ in the following way:

$$\pi^k((x_0, \ldots, x_k), (y_0, \ldots, y_k)) := \pi_Z^k((\ell_X(x_0), \ldots, \ell_X(x_k)), (\ell_Y(y_0), \ldots, \ell_Y(y_k)))$$

for all $((x_0, \ldots, x_k), (y_0, \ldots, y_k)) \in X^{k+1} \times Y^{k+1}$. First of all, observe that

$$\pi_Z^k((z_0, \ldots, z_k), (z_0', \ldots, z_k')) = 0$$

if $z_i \notin \ell_X(X)$ for some $i \in \{0, \ldots, k\}$ or $z_j' \notin \ell_Y(Y)$ for some $j \in \{0, \ldots, k\}$ since π_Z^k is a coupling measure between α_Z^k and β_Z^k. This implies that indeed π^k is a coupling measure

between α^k and β^k. Furthermore, since ℓ_X and ℓ_Y are injective, it is easy to check that $\pi_Z^k = (\ell_X^{k+1} \times \ell_Y^{k+1})_{\#}\pi^k$. Finally, note that

$$\begin{aligned}\pi^k((y_0,\dots,y_l)|(x_0,\dots,x_k)) &= \pi_Z^k((\ell_Y(y_0),\dots,\ell_Y(y_l))|(\ell_X(x_0),\dots,\ell_X(x_k))) \\ &= \pi_Z^k((\ell_Y(y_0),\dots,\ell_Y(y_l))|(\ell_X(x_0),\dots,\ell_X(x_l))) \quad (\because \pi_Z^k \text{ is bicausal}) \\ &= \pi^k((y_0,\dots,y_l)|(x_0,\dots,x_l))\end{aligned}$$

for all $l \in \{0,\dots,k\}$, $(y_0,\dots,y_l) \in Y^{l+1}$, and $(x_0,\dots,x_k) \in X^{k+1}$. Hence, π^k is causal from α^k to β^k. In a similar way one can also prove that π^k is causal from β^k to α^k, too. This completes the proof. ■

A.3 More on Lipschitz properties of MP-GNNs

In Section 2.4, we introduced one way of inducing LMMCs from weighted graphs. Below, we introduce a new method for doing so and hence establish the Lipschitz property of a new type of MP-GNNs closely related to GINs.

Given any $\varepsilon \geq 0$ and any finite edge weighted graph $G = (V, E, w)$ endowed with a label function $\ell_G : V \to Z$, we generate a LMMC as follows. We associate to the vertex set V a Markov kernel $m_\bullet^{G,(\varepsilon)}$ as follows: for any $v \in V$,

$$m_v^{G,(\varepsilon)} := \frac{1}{\deg(v) + 1 + \varepsilon}\left((1+\varepsilon)\delta_v + \sum_{v' \in N_G(v)} w_{vv'}\delta_{v'}\right)$$

For each vertex $v \in V$, we let $\deg^\varepsilon(v) := \deg(v) + 1 + \varepsilon$. Then, the following probability measure $\mu_G^\varepsilon := \sum_{v\in V} \frac{\deg^\varepsilon(v)}{\sum_{v'\in V}\deg^\varepsilon(v')}\delta_v$ is a stationary distribution w.r.t. the Markov kernel $m_\bullet^{G,(\varepsilon)}$. Then, we let $\mathcal{X}_{(\varepsilon)}(G) := (V, m_\bullet^{G,(\varepsilon)}, \mu_G^\varepsilon)$.

Now, recall that $\mathcal{G}(Z)$ denotes the collection of all Z-labeled graphs. Given $\varepsilon \geq 0$, let $I_{(\varepsilon)} : \mathcal{G}(Z) \to \mathcal{M}^L(Z)$ denote the map sending a Z-labeled graph G into a Z-LMMC $(\mathcal{X}_{(\varepsilon)}(G), \ell_G)$. Let $\mathcal{G}_{(\varepsilon)}(Z) := I_{(\varepsilon)}(\mathcal{G}(Z)) \subseteq \mathcal{M}^L(Z)$. Now, for any $k \geq 0$, $d_{\mathrm{WL}}^{(k)}$ restricted on $\mathcal{G}_{(\varepsilon)}(Z)$ induces a pseudometric, which we denote by $d_{\mathcal{G},(\varepsilon)}^{(k)}$, on $\mathcal{G}(Z)$.

Now, given any $\varepsilon \geq 0$, we consider the following k-layer "normalized" GIN.

$$\text{Message Passing:}\quad \ell_G^{i+1}(v) = \varphi_{i+1}\left(\frac{1}{\deg(v)+1+\varepsilon}\left((1+\varepsilon)\ell_G^i(v) + \sum_{v'\in N_G(v)} w_{vv'}\ell_G^i(v')\right)\right)$$

$$\text{Readout:}\quad h((G,\ell_G)) := \psi\left(\sum_{v\in V}\frac{\deg^\varepsilon(v)}{\sum_{v'\in V}\deg^\varepsilon(v')}\ell_G^k(v)\right)$$

where $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ and $\psi : \mathbb{R}^{d_k} \to \mathbb{R}$ are MLPs. Note that the only difference between the above MP-GNN and GIN defined in (Xu et al., 2019) is the involvement of normalization terms $\frac{1}{\deg(v)+1+\varepsilon}$ and $\frac{\deg^\varepsilon(v)}{\sum_{v'\in V}\deg^\varepsilon(v')}$.

Lipschitz property of normalized GINs. Now, we establish that the normalized GINs defined above are Lipschitz w.r.t. $d^{(k)}_{\mathcal{G},(\varepsilon)}$.

Theorem 72 *Given a k-layer normalized GIN $h : \mathcal{G}(\mathbb{R}^d) \to \mathbb{R}$ as described above, assume that φ_i is C_i-Lipschitz for $i = 1, \ldots, k$ and that ψ is C-Lipschitz. Then, for any two labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), one has that*

$$|h((G_1, \ell_{G_1})) - h((G_2, \ell_{G_2}))| \leq C \cdot \Pi_{i=1}^k C_i \cdot d^{(k)}_{\mathcal{G},(\varepsilon)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2}))\,.$$

Proof The proof is similar to the one for Corollary 39.

We first prove that

$$d_{\mathrm{W}}\left(\left(\ell^k_{G_1}\right)_{\#}\mu^{\varepsilon}_{G_1}, \left(\ell^k_{G_2}\right)_{\#}\mu^{\varepsilon}_{G_2}\right) \leq \Pi_{i=1}^k C_i \cdot d^{(k)}_{\mathcal{G},(\varepsilon)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2}))\,. \tag{15}$$

Given Equation (15) and the fact that ψ is C-Lipschitz, one has that

$$\begin{aligned}
|h((G_1, \ell_{G_1})) - h((G_2, \ell_{G_2}))| &\leq C \cdot \left\| \sum_{v \in V} \frac{\deg^{\varepsilon}(v)}{\sum_{v' \in V} \deg^{\varepsilon}(v')} \ell^k_{G_1}(v) - \sum_{v \in V} \frac{\deg^{\varepsilon}(v)}{\sum_{v' \in V} \deg^{\varepsilon}(v')} \ell^k_{G_2}(v) \right\| \\
&\leq C \cdot d_{\mathrm{W}}\left(\left(\ell^k_{G_1}\right)_{\#}\mu^{\varepsilon}_{G_1}, \left(\ell^k_{G_2}\right)_{\#}\mu^{\varepsilon}_{G_2}\right) \\
&\leq C \cdot \Pi_{i=1}^k C_i \cdot d^{(k)}_{\mathcal{G},(\varepsilon)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2}))\,,
\end{aligned}$$

where the second inequality follows Lemma 85 by letting $\varphi = \mathrm{id}_{\mathbb{R}^{d_k}}$.

To prove Equation (15), it suffices to prove that for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$,

$$\left\|\ell^k_{G_1}(v_1) - \ell^k_{G_2}(v_2)\right\| \leq \Pi_{i=1}^k C_i \cdot d_{\mathrm{W}}\left(\mathfrak{l}^{(k)}_{(\mathcal{X}_{(\varepsilon)}(G_1), \ell_{G_1})}(v_1), \mathfrak{l}^{(k)}_{(\mathcal{X}_{(\varepsilon)}(G_2), \ell_{G_2})}(v_2)\right).$$

We prove the above inequality by proving the following inequality inductively on $j = 1, \ldots, k$:

$$\left\|\ell^j_{G_1}(v_1) - \ell^j_{G_2}(v_2)\right\| \leq \Pi_{i=1}^j C_i \cdot d_{\mathrm{W}}\left(\mathfrak{l}^{(j)}_{(\mathcal{X}_{(\varepsilon)}(G_1), \ell_{G_1})}(v_1), \mathfrak{l}^{(j)}_{(\mathcal{X}_{(\varepsilon)}(G_2), \ell_{G_2})}(v_2)\right). \tag{16}$$

When $j = 1$, we have that

$$\begin{aligned}
&\left\|\ell^1_{G_1}(v_1) - \ell^1_{G_2}(v_2)\right\| \\
&\leq C_1 \left\| \frac{(1+\varepsilon)\ell_{G_1}(v_1) + \sum_{v_1' \in N_{G_1}(v_1)} w_{v_1 v_1'} \ell_{G_1}(v_1')}{\deg^{\varepsilon}_{G_1}(v_1)} - \frac{(1+\varepsilon)\ell_{G_2}(v_2) + \sum_{v_2' \in N_{G_2}(v_2)} w_{v_2 v_2'} \ell_{G_2}(v_2')}{\deg^{\varepsilon}_{G_2}(v_2)} \right\| \\
&= C_1 \left\| q_{\mathrm{Id}}\left((\ell_{G_1})_{\#} m^{G_1,(\varepsilon)}_{v_1}\right) - q_{\mathrm{Id}}\left((\ell_{G_2})_{\#} m^{G_2,(\varepsilon)}_{v_1}\right) \right\| \\
&\leq C_1 d_{\mathrm{W}}\left((\ell_{G_1})_{\#} m^{G_1,(\varepsilon)}_{v_1}, (\ell_{G_2})_{\#} m^{G_2,(\varepsilon)}_{v_1}\right) = C_1 d_{\mathrm{W}}\left(\mathfrak{l}^{(1)}_{(\mathcal{X}_{(\varepsilon)}(G_1), \ell_{G_1})}(v_1), \mathfrak{l}^{(1)}_{(\mathcal{X}_{(\varepsilon)}(G_2), \ell_{G_2})}(v_2)\right).
\end{aligned}$$

Here $\mathrm{Id} : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}$ is the identity map and thus is 1-Lipschitz.

We now assume that Equation (16) holds for some $j \geq 1$. For $j+1$, we have that

$$
\begin{aligned}
&\left\| \ell_{G_1}^{j+1}(v_1) - \ell_{G_2}^{j+1}(v_2) \right\| \\
&\leq C_{j+1} \left\| \frac{(1+\varepsilon)\ell_{G_1}^{j}(v_1) + \sum_{v_1' \in N_{G_1}(v_1)} w_{v_1 v_1'} \ell_{G_1}^{j}(v_1')}{\deg_{G_1}^{\varepsilon}(v_1)} - \frac{(1+\varepsilon)\ell_{G_2}^{j}(v_2) + \sum_{v_2' \in N_{G_2}(v_2)} w_{v_2 v_2'} \ell_{G_2}^{j}(v_2')}{\deg_{G_2}^{\varepsilon}(v_2)} \right\| \\
&= C_{j+1} \left\| q_{\mathrm{Id}} \left((\ell_{G_1}^{j})_{\#} m_{v_1}^{G_1,(\varepsilon)} \right) - q_{\mathrm{Id}} \left((\ell_{G_2}^{j})_{\#} m_{v_1}^{G_2,(\varepsilon)} \right) \right\| \\
&\leq C_{j+1} d_{\mathrm{W}} \left((\ell_{G_1}^{j})_{\#} m_{v_1}^{G_1,(\varepsilon)}, (\ell_{G_2}^{j})_{\#} m_{v_1}^{G_2,(\varepsilon)} \right) \\
&= C_{j+1} \inf_{\gamma \in \mathcal{C}\left(m_{v_1}^{G_1,(\varepsilon)}, m_{v_2}^{G_2,(\varepsilon)} \right)} \int_{V_{G_1} \times V_{G_2}} \left\| \ell_{G_1}^{j}(v_1') - \ell_{G_2}^{j}(v_2') \right\| \gamma(dv_1' \times dv_2') \\
&\leq C_{j+1} \cdot \inf_{\gamma \in \mathcal{C}\left(m_{v_1}^{G_1,(\varepsilon)}, m_{v_2}^{G_2,(\varepsilon)} \right)} \int_{V_{G_1} \times V_{G_2}} \Pi_{i=1}^{j} C_i \cdot d_{\mathrm{W}} \left(\mathfrak{l}_{(\mathcal{X}_{(\varepsilon)}(G_1), \ell_{G_1})}^{(j)}(v_1'), \mathfrak{l}_{(\mathcal{X}_{(\varepsilon)}(G_2), \ell_{G_2})}^{(j)}(v_2') \right) \gamma(dv_1' \times dv_2') \\
&= \Pi_{i=1}^{j+1} C_i d_{\mathrm{W}} \left(\mathfrak{l}_{(\mathcal{X}_{(\varepsilon)}(G_1), \ell_{G_1})}^{(j+1)}(v_1), \mathfrak{l}_{(\mathcal{X}_{(\varepsilon)}(G_2), \ell_{G_2})}^{(j+1)}(v_2) \right).
\end{aligned}
$$

This concludes the proof. ∎

Appendix B. Proofs

B.1 Proofs from Section 2

B.1.1 Proof of Proposition 9

The "only if" part is obvious. To prove the "if" part, we assume $(\mathcal{X}_q(G_1), \ell_{G_1})$ is isomorphic to $(\mathcal{X}_q(G_2), \ell_{G_2})$. Then, there exists a bijective map $\psi : V_{G_1} \to V_{G_2}$ such that $\psi_{\#} m_v^{G_1,q} = m_{\psi(v)}^{G_2,q}$, $\psi_{\#} \mu_{G_1} = \mu_{G_2}$ and $\ell_{G_1}(v) = \ell_{G_2}(\psi(v))$ for all $v \in V_{G_1}$. Now, by the definition of $m_{\bullet}^{G_1,q}$ and $m_{\bullet}^{G_2,q}$ (see Definition 7), one can easily check that

$$\deg_{G_1}(v) = 0 \Leftrightarrow m_v^{G_1,q}(v) = m_{\psi(v)}^{G_2,q}(\psi(v)) = 1 \Leftrightarrow \deg_{G_2}(\psi(v)) = 0.$$

So, consider the case when $m_v^{G_1,q}(v) < 1$. This implies $\deg_{G_1}(v) > 0$ and $\deg_{G_2}(\psi(v)) > 0$. In this case, again by the definition of $m_{\bullet}^{G_1,q}$ and $m_{\bullet}^{G_2,q}$, one can show that

$$v, v' \in V_{G_1} \text{ are adjacent} \Leftrightarrow m_v^{G_1,q}(v') = m_{\psi(v)}^{G_2,q}(\psi(v')) > 0 \Leftrightarrow \psi(v), \psi(v') \in V_{G_2} \text{ are adjacent.}$$

Hence, G_1 and G_2 are isomorphic as we required.

B.2 Proofs from Section 3

B.2.1 Proof of Proposition 11

Recall that if X is compact, Y is a metric space, and $\psi : X \to Y$ is continuous, then for any Borel probability measure $\alpha \in \mathcal{P}(X)$, $\psi_{\#}\alpha$ has finite 1-moment.

Lemma 73 *Under the same conditions, the pushforward map*

$$\begin{aligned}\psi : \mathcal{P}(X) &\longrightarrow \mathcal{P}(Y)\\ \alpha &\longmapsto \psi_{\#}\alpha\end{aligned}$$

is continuous, where $\mathcal{P}(X)$ is equipped with the weak topology and $\mathcal{P}(Y)$ is equipped with the metric topology induced by the ℓ^1-Wasserstein distance.

Proof Fix an arbitrary sequence $\{\alpha_n\}_{n\geq 1}$ of probability measures in $\mathcal{P}(X)$ that weakly converges to $\alpha \in \mathcal{P}(X)$. Since X is compact and ψ is continuous, we have that for any fixed $y_0 \in Y$,

$$\begin{aligned}\int_Y d_Y(y_0, y)\, \psi_{\#}\alpha_n(dy) &= \int_X d_Y(y_0, \psi(x))\alpha_n(dx)\\ \to \int_X d_Y(y_0, \psi(x))\alpha(dx) &= \int_Y d_Y(y_0, y)\, \psi_{\#}\alpha(dy)\end{aligned}$$

as $n \to \infty$. Furthermore, it is well known that the sequence $\{\psi_{\#}\alpha_n\}_{n\geq 1} \subseteq \mathcal{P}(Z)$ also weakly converges to $\psi_{\#}\alpha$; see (Bogachev, 2018, Section 5.2). Hence, by (Villani, 2003, Theorem 7.12), one can conclude that the sequence $\{\psi_{\#}\alpha_n\}_{n\geq 1}$ converges to $\psi_{\#}\alpha$ with respect to the ℓ^1-Wasserstein distance. ■

Therefore, in order to show that $\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x) = \big(\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)}\big)_{\#} m_x^X$ has a finite 1-moment, it is enough to show that the map $\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)} : X \to \mathcal{P}^{\circ(k-1)}(Z)$ is continuous. We will prove this fact by using the induction on k.

First, for $k = 1$, note that $\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)} = \ell_X$ is a continuous map from X to Z by the assumption.

Next, assume that the claim holds for $k-1$ so that $\mathfrak{l}^{(k-2)}_{(\mathcal{X},\ell_X)} : X \to \mathcal{P}^{\circ(k-2)}(Z)$ is continuous. Note that the map $\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)}$ is a composition of the map $x \mapsto m_x^X$ and $\big(\mathfrak{l}^{(k-2)}_{(\mathcal{X},\ell_X)}\big)_{\#}$. Since the first map is continuous by the assumption and the second map is continuous by Lemma 73, one can conclude that $\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)}$ is also continuous.

B.2.2 Proof of the claim in Example 1

By Lemma 67 we have that

$$\begin{aligned}d^{(1)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) &= d_{\mathrm{W}}(\mathfrak{L}_1((\mathcal{X},\ell_X)), \mathfrak{L}_1((\mathcal{Y},\ell_Y))) = d_{\mathrm{W}}\Big(\big(\mathfrak{l}^{(1)}_{(\mathcal{X},\ell_X)}\big)_{\#}\mu_X, \big(\mathfrak{l}^{(1)}_{(\mathcal{Y},\ell_Y)}\big)_{\#}\mu_Y\Big)\\ &= \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(1)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(1)}_{(\mathcal{Y},\ell_Y)}(y)\big)\, \gamma(dx\times dy)\\ &= \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big((\ell_X)_{\#}m_x^X, (\ell_Y)_{\#}m_y^Y\big)\, \gamma(dx\times dy).\end{aligned}$$

B.2.3 Proof of Proposition 13

First of all, we claim that $d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\big) \leq C_{Z,\ell}(X,Y)$ for all $x \in X$, $y \in Y$, and $k \geq 0$. We prove this claim by using the induction on k. For $k = 0$, this is trivial since

$d_Z\big(\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)}(x),\mathfrak{l}^{(0)}_{(\mathcal{Y},\ell_Y)}(y)\big) = d_Z(\ell_X(x),\ell_Y(y)) \leq C_{Z,\ell}(X,Y)$ by the definition of $C_{Z,\ell}(X,Y)$. Now, assume that the claim holds up to $k-1$. Then, for k, we have that

$$\begin{aligned} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x),\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\big) &= d_{\mathrm{W}}\Big(\big(\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)}\big)_{\#} m_x^X, \big(\mathfrak{l}^{(k-1)}_{(\mathcal{Y},\ell_Y)}\big)_{\#} m_y^Y\Big) \\ &= \inf_{\gamma\in\mathcal{C}(m_x^X,m_y^Y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)}(x'),\mathfrak{l}^{(k-1)}_{(\mathcal{Y},\ell_Y)}(y')\big)\,\gamma(dx'\times dy'); \quad \text{by Lemma 67} \\ &\leq C_{Z,\ell}(X,Y); \quad \text{by the induction.} \end{aligned}$$

Hence, the claim above is true.

Finally, with the claim, one can prove that, for all $k \geq 0$,

$$\begin{aligned} d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) &= d_{\mathrm{W}}(\mathfrak{L}_k((\mathcal{X},\ell_X)),\mathfrak{L}_k((\mathcal{Y},\ell_Y))) = d_{\mathrm{W}}\Big(\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\big)_{\#}\mu_X, \big(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\big)_{\#}\mu_Y\Big) \\ &= \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x),\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\big)\,\gamma(dx\times dy); \quad \text{by Lemma 67} \\ &\leq C_{Z,\ell}(X,Y). \end{aligned}$$

B.2.4 Proof of Theorem 14

The statement (1) follows from Proposition 13 and the definition. It is obvious that when $(\mathcal{X},\ell_X)$ is isomorphic to $(\mathcal{Y},\ell_Y)$, $d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = 0$ for all $k\in\mathbb{N}$ and thus $d_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = 0$. It follows directly from Equation (4) that $d^{(k)}_{\mathrm{WL}}$ satisfies the triangle inequality. Hence, $d_{\mathrm{WL}} := \sup_{k\geq 1} d^{(k)}_{\mathrm{WL}}$ also satisfies the triangle inequality.

B.2.5 Proof of Proposition 15

Fix an arbitrary $k \geq 0$. Observe that

$$\begin{aligned} d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) &= d_{\mathrm{W}}(\mathfrak{L}_k((\mathcal{X},\ell_X)),\mathfrak{L}_k((\mathcal{Y},\ell_Y))) = d_{\mathrm{W}}\Big(\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\big)_{\#}\mu_X, \big(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\big)_{\#}\mu_Y\Big) \\ &= \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x),\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\big)\,\gamma(dx\times dy). \end{aligned}$$

where the last equality holds by Lemma 67. Similarly,

$$\begin{aligned} & d^{(k+1)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) \\ &= \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k+1)}_{(\mathcal{X},\ell_X)}(x),\mathfrak{l}^{(k+1)}_{(\mathcal{Y},\ell_Y)}(y)\big)\,\gamma'(dx\times dy); \quad \text{here } \gamma'\in\mathcal{C}(\mu_X,\mu_Y) \text{ is chosen to be optimal} \\ &= \int_{X\times Y} \left(\inf_{\mu_{x,y}\in\mathcal{C}(m_x^X,m_y^Y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x_1),\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y_1)\big)\,\mu_{x,y}(dx_1\times dy_1) \right) \gamma'(dx\times dy) \\ &= \int_{X\times Y}\int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x_1),\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y_1)\big)\,\nu_{x,y}(dx_1\times dy_1)\gamma'(dx\times dy) \end{aligned}$$

where $\nu_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)$ is a measurable map such that for every $(x, y) \in X \times Y$, $\nu_{x,y} \in \mathcal{C}(m^X_x, m^Y_y)$ is optimal. This map is guaranteed by the fact that $(x, y) \mapsto (m^X_x, m^Y_y)$ is continuous (hence measurable) and Lemma 69.

Furthermore, since μ_X and μ_Y are stationary distributions w.r.t $m^X_\bullet$ and $m^Y_\bullet$ respectively, it is easy to verify that the measure $\int_{X\times Y} \nu_{x,y}\, \gamma'(dx \times dy)$ belongs to $\mathcal{C}(\mu_X, \mu_Y)$.

Therefore, one can conclude that

$$\begin{aligned} d^{(k+1)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) &= \int_{X\times Y} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x_1), \mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y_1)\big)\, \nu_{x,y}(dx_1 \times dy_1)\gamma'(dx \times dy) \\ &\geq \inf_{\gamma \in \mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\big)\, \gamma(dx \times dy) = d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)). \end{aligned}$$

B.2.6 Proof of Lemma 18

We first assume that the WL test cannot distinguish (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), i.e., $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ for all $k = 0, 1, \ldots$. We then prove that $L_k((G_1, \ell^g_{G_1})) = L_k((G_2, \ell^g_{G_2}))$ for all $k = 0, 1, \ldots$. The assumption $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ for all $k = 0, 1, \ldots$ immediately implies that $|V_{G_1}| = |V_{G_2}|$. Then, it suffices to show that for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$

$$\ell^{(k+1)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k+1)}_{(G_2,\ell_{G_2})}(v_2) \Longrightarrow \ell^{(k)}_{(G_1,\ell^g_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell^g_{G_2})}(v_2),\ \forall k = 0, 1, \ldots. \tag{17}$$

We prove Equation (17) by induction on k. When $k = 0$, for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$, if $\ell^1_{(G_1,\ell_{G_1})}(v_1) = \ell^1_{(G_2,\ell_{G_2})}(v_2)$, then

$$(\ell_{G_1}(v_1), \{\!\{\ell_{G_1}(v),\, v \in N_{G_1}(v_1)\}\!\}) = (\ell_{G_2}(v_2), \{\!\{\ell_{G_2}(v),\, v \in N_{G_2}(v_2)\}\!\}).$$

It follows that $\ell_{G_1}(v_1) = \ell_{G_2}(v_2)$ and $\deg_{G_1}(v_1) = \deg_{G_2}(v_2)$. Then, by injectivity of g, one has that $\ell^g_{G_1}(v_1) = \ell^g_{G_2}(v_2)$.

Now, we assume that Equation (17) holds for some $k \geq 0$. For the case of $k + 1$, note that $\ell^{(k+2)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k+2)}_{(G_2,\ell_{G_2})}(v_2)$ implies that

$$\left(\ell^{(k+1)}_{(G_1,\ell_{G_1})}(v_1), \left\{\!\!\left\{\ell^{(k+1)}_{(G_1,\ell_{G_1})}(v),\, v \in N_{G_1}(v_1)\right\}\!\!\right\}\right) = \left(\ell^{(k+1)}_{(G_2,\ell_{G_2})}(v_2), \left\{\!\!\left\{\ell^{(k+1)}_{(G_2,\ell_{G_2})}(v),\, v \in N_{G_2}(v_2)\right\}\!\!\right\}\right).$$

Hence, $\ell^{(k+1)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k+1)}_{(G_2,\ell_{G_2})}(v_2)$ and there exists a bijection $\psi : N_{G_1}(v_1) \to N_{G_2}(v_2)$ such that $\ell^{(k+1)}_{(G_1,\ell_{G_1})}(v) = \ell^{(k+1)}_{(G_2,\ell_{G_2})}(\psi(v))$ for any $v \in N_{G_1}(v_1)$. By the induction assumption, we then have that $\ell^{(k)}_{(G_1,\ell^g_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell^g_{G_2})}(v_2)$ and $\ell^{(k)}_{(G_1,\ell^g_{G_1})}(v) = \ell^{(k)}_{(G_2,\ell^g_{G_2})}(\psi(v))$ for any $v \in N_{G_1}(v_1)$. This implies that

$$\left(\ell^{(k)}_{(G_1,\ell^g_{G_1})}(v_1), \left\{\!\!\left\{\ell^{(k)}_{(G_1,\ell^g_{G_1})}(v),\, v \in N_{G_1}(v_1)\right\}\!\!\right\}\right) = \left(\ell^{(k)}_{(G_2,\ell^g_{G_2})}(v_2), \left\{\!\!\left\{\ell^{(k)}_{(G_2,\ell^g_{G_2})}(v),\, v \in N_{G_2}(v_2)\right\}\!\!\right\}\right)$$

and thus $\ell^{(k+1)}_{(G_1,\ell^g_{G_1})}(v_1) = \ell^{(k+1)}_{(G_2,\ell^g_{G_2})}(v_2)$. Therefore, $L_k\Big((G_1, \ell^g_{G_1})\Big) = L_k\Big((G_2, \ell^g_{G_2})\Big)$ for all $k = 0, 1, \ldots$ and thus the WL test cannot distinguish $(G_1, \ell^g_{G_1})$ and $(G_2, \ell^g_{G_2})$.

Conversely, we assume that the WL test cannot distinguish $(G_1, \ell^g_{G_1})$ and $(G_2, \ell^g_{G_2})$, i.e., $L_k\Big((G_1, \ell^g_{G_1})\Big) = L_k\Big((G_2, \ell^g_{G_2})\Big)$ for all $k = 0, 1, \ldots$. We then prove that $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ for all $k = 0, 1, \ldots$. The proof is similar to the one for the other direction. First, the assumption $L_k\Big((G_1, \ell^g_{G_1})\Big) = L_k\Big((G_2, \ell^g_{G_2})\Big)$ for all $k = 0, 1, \ldots$ implies that $|V_{G_1}| = |V_{G_2}|$. Then, it suffices to show that for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$

$$\ell^{(k)}_{(G_1, \ell^g_{G_1})}(v_1) = \ell^{(k)}_{(G_2, \ell^g_{G_2})}(v_2) \Longrightarrow \ell^{(k)}_{(G_1, \ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2, \ell_{G_2})}(v_2), \ \forall k = 0, 1, \ldots. \tag{18}$$

We prove Equation (18) by induction on k. When $k = 0$, for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$, if $\ell^g_{G_1}(v_1) = \ell^g_{G_2}(v_2)$, then by injectivity of g, we have that $\ell_{G_1}(v_1) = \ell_{G_2}(v_2)$.

Now, we assume that Equation (18) holds for some $k \geq 0$. For the case of $k + 1$, note that $\ell^{(k+1)}_{(G_1, \ell^g_{G_1})}(v_1) = \ell^{(k+1)}_{(G_2, \ell^g_{G_2})}(v_2)$ implies that

$$\left(\ell^{(k)}_{(G_1, \ell^g_{G_1})}(v_1), \left\{\!\!\left\{\ell^{(k)}_{(G_1, \ell^g_{G_1})}(v), \, v \in N_{G_1}(v_1)\right\}\!\!\right\}\right) = \left(\ell^{(k)}_{(G_2, \ell^g_{G_2})}(v_2), \left\{\!\!\left\{\ell^{(k)}_{(G_2, \ell^g_{G_2})}(v), \, v \in N_{G_2}(v_2)\right\}\!\!\right\}\right).$$

By the induction assumption, it is easy to see that

$$\left(\ell^{(k)}_{(G_1, \ell_{G_1})}(v_1), \left\{\!\!\left\{\ell^{(k)}_{(G_1, \ell_{G_1})}(v), \, v \in N_{G_1}(v_1)\right\}\!\!\right\}\right) = \left(\ell^{(k)}_{(G_2, \ell_{G_2})}(v_2), \left\{\!\!\left\{\ell^{(k)}_{(G_2, \ell_{G_2})}(v), \, v \in N_{G_2}(v_2)\right\}\!\!\right\}\right)$$

and thus $\ell^{(k+1)}_{(G_1, \ell_{G_1})}(v_1) = \ell^{(k+1)}_{(G_2, \ell_{G_2})}(v_2)$. Therefore, $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ for all $k = 0, 1, \ldots$ and thus the WL test cannot distinguish (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}).

B.2.7 Proof of Proposition 17

By Lemma 18, we only need to prove that the WL test cannot distinguish $(G_1, \ell^g_{G_1})$ and $(G_2, \ell^g_{G_2})$ iff $d_{\mathrm{WL}}\Big(\big(\mathcal{X}_q(G_1), \ell^g_G\big), \Big(\mathcal{X}_q(G_2), \ell^g_{G_2}\Big)\Big) = 0$. For this purpose, we need to introduce some new notions.

For any metric space Z, we let $\mathrm{MultPow}(Z)$ denote the collection of all finite multisets of Z (including the empty set). We inductively define a family of sets Z^k as follows:

1. $Z^1 \coloneqq Z \times \mathrm{MultPow}(Z)$;
2. for $k \geq 1$, $Z^{k+1} \coloneqq Z^k \times \mathrm{MultPow}(Z^k)$.

Then, we inductively define a family of maps $\varphi^k_q : Z^k \to \mathcal{P}^{\circ k}(Z)$ as follows:

1. define $\varphi^1_q : Z^1 \to \mathcal{P}(Z)$ by

$$(z, A) \in Z \times \mathrm{MultPow}(Z) \mapsto \begin{cases} q\delta_z + \frac{1-q}{|A|} \sum_{z' \in A} \delta_{z'}, & A \neq \emptyset \\ \delta_z, & A = \emptyset \end{cases};$$

2. for $k \geq 1$, define $\varphi^{k+1}_q : Z^{k+1} \to \mathcal{P}^{\circ(k+1)}(Z)$ by

$$(z, A) \in Z^k \times \mathrm{MultPow}(Z^k) \mapsto \begin{cases} q\delta_{\varphi^k_q(z)} + \frac{1-q}{|A|} \sum_{z' \in A} \delta_{\varphi^k_q(z')}, & A \neq \emptyset \\ \delta_{\varphi^k_q(z)}, & A = \emptyset \end{cases}.$$

Lemma 74 *For any labeled graph $(G, \ell_G : V_G \to Z)$ and any $q \in [0,1]$, one has that for any $k \in \mathbb{N}$*

$$\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G),\ell_G)} = \varphi_q^k \circ \ell^{(k)}_{(G,\ell_G)} : V_G \to \mathcal{P}^{\circ k}(Z). \tag{19}$$

Proof [Proof of Lemma 74] We prove by induction on k.

When $k = 1$, for any $v \in V$, if $N_G(v) \neq \emptyset$ we have that

$$\begin{aligned}\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G),\ell_G)}(v) &= (\ell_G)_{\#} m_v^{G,q} = q\,\delta_{\ell_G(v)} + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} \delta_{\ell_G(v')} \\ &= \varphi_q^1\left((\ell_G(v), \{\!\{\ell_G(v') : v' \in N_G(v)\}\!\})\right) = \varphi_q^1(\ell^{(1)}_{(G,\ell_G)}(v)).\end{aligned}$$

If $N_G(v) = \emptyset$ then we have that

$$\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G),\ell_G)}(v) = (\ell_G)_{\#} m_v^{G,q} = \delta_{\ell_G(v)} = \varphi_q^1\left((\ell_G(v), \emptyset)\right) = \varphi_q^1(\ell^{(1)}_{(G,\ell_G)}(v)).$$

Now, we assume that Equation (19) holds for some $k \geq 1$. Then, for $k+1$ and for any $v \in V$, if $N_G(v) \neq \emptyset$, we have that

$$\begin{aligned}\mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G),\ell_G)}(v) &= \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G),\ell_G)}\right)_{\#} m_v^{G,q} = q\,\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G),\ell_G)}(v)} + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G),\ell_G)}(v')} \\ &= q\,\delta_{\varphi_q^k\left(\ell^{(k)}_{(G,\ell_G)}(v)\right)} + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} \delta_{\varphi_q^k\left(\ell^{(k)}_{(G,\ell_G)}(v')\right)} \\ &= \varphi_q^{k+1}\left(\left(\ell^{(k)}_{(G,\ell_G)}(v), \{\!\{\ell^{(k)}_{(G,\ell_G)}(v') : v' \in N_G(v)\}\!\}\right)\right) = \varphi_q^{k+1}\left(\ell^{(k+1)}_{(G,\ell_G)}(v)\right).\end{aligned}$$

If $N_G(v) = \emptyset$ then we have that

$$\begin{aligned}\mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G),\ell_G)}(v) &= \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G),\ell_G)}\right)_{\#} m_v^{G,q} = \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G),\ell_G)}(v)} \\ &= \delta_{\varphi_q^k\left(\ell^{(k)}_{(G,\ell_G)}(v)\right)} = \varphi_q^{k+1}\left(\left(\ell^{(k)}_{(G,\ell_G)}(v), \emptyset\right)\right) = \varphi_q^{k+1}\left(\ell^{(k+1)}_{(G,\ell_G)}(v)\right).\end{aligned}$$

This concludes the proof. ∎

Lemma 75 *Fix any $\frac{1}{2} < q < 1$ and any labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}). Assume that the labels satisfy that for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$, we have that $\ell_{G_1}(v_1) = \ell_{G_2}(v_2)$ implies $\deg(v_1) = \deg(v_2)$. Then, one has that for any $v_1 \in V_{G_1}, v_2 \in V_{G_2}$,*

$$\ell^{(k)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell_{G_2})}(v_2) \;\; \textit{iff} \;\; \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2). \tag{20}$$

Proof [Proof of Lemma 75] By Lemma 74, we have that $\ell^{(k)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell_{G_2})}(v_2)$ implies that $\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$. For the other direction, we prove by induction on k.

When $k = 1$, we first note that $\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = q\,\delta_{\ell_{G_1}(v_1)} + \frac{1-q}{\deg(v_1)} \sum_{v \in N_{G_1}(v_1)} \delta_{\ell_{G_1}(v)}$ if $N_{G_1}(v_1) \neq \emptyset$ and $\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \delta_{\ell_{G_1}(v_1)}$ otherwise. Since $\frac{1}{2} < q < 1$, $\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$ implies that $\delta_{\ell_{G_1}(v_1)} = \delta_{\ell_{G_2}(v_2)}$. Hence $\ell_{G_1}(v_1) = \ell_{G_2}(v_2)$ and thus $\deg_{G_1}(v_1) =$

$\deg_{G_2}(v_2)$. This implies that $N_{G_1}(v_1) = \emptyset$ iff $N_{G_2}(v_2) = \emptyset$. If $N_{G_1}(v_1) = \emptyset$, then obviously, we have that

$$\ell^{(1)}_{(G_1,\ell_{G_1})}(v_1) = (\ell_{G_1}(v_1), \emptyset) = (\ell_{G_2}(v_2), \emptyset) = \ell^{(1)}_{(G_2,\ell_{G_2})}(v_2).$$

If otherwise $N_{G_1}(v_1) \neq \emptyset$, then $\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$ again implies that

$$\frac{1-q}{\deg(v_1)} \sum_{v \in N_{G_1}(v_1)} \delta_{\ell_{G_1}(v)} = \frac{1-q}{\deg(v_2)} \sum_{v \in N_{G_2}(v_2)} \delta_{\ell_{G_2}(v)}.$$

Hence, $\sum_{v \in N_{G_1}(v_1)} \delta_{\ell_{G_1}(v)} = \sum_{v \in N_{G_2}(v_2)} \delta_{\ell_{G_2}(v)}$ and thus

$$\{\!\{\ell_{G_1}(v) : v \in N_{G_1}(v_1)\}\!\} = \{\!\{\ell_{G_2}(v) : v \in N_{G_2}(v_2)\}\!\}.$$

Therefore, $\ell^{(1)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(1)}_{(G_2,\ell_{G_2})}(v_2)$.

Now, we assume that Equation (20) holds for some $k \geq 1$. Note that

$$\mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \begin{cases} q\,\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1)} + \frac{1-q}{\deg(v_1)} \sum_{v \in N_{G_1}(v_1)} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v)}, & N_{G_1}(v_1) \neq \emptyset \\ \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1)}, & N_{G_1}(v_1) = \emptyset \end{cases}$$

Then, for $k+1$, the assumptions $\frac{1}{2} < q < 1$ and $\mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$ imply that $\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1)} = \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)}$. Hence $\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$. By the induction assumption we have that $\ell^{(k)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell_{G_2})}(v_2)$. It is not hard to see that then $\ell^{(1)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(1)}_{(G_2,\ell_{G_2})}(v_2)$ and thus $\deg_{G_1}(v_1) = \deg_{G_2}(v_2)$. Then, similarly as in the case $k = 1$, we have two situations. If $N_{G_1}(v_1) = \emptyset$, then we have that

$$\ell^{(k+1)}_{(G_1,\ell_{G_1})}(v_1) = \left(\ell^{(k)}_{(G_1,\ell_{G_1})}(v_1), \emptyset\right) = \left(\ell^{(k)}_{(G_2,\ell_{G_2})}(v_2), \emptyset\right) = \ell^{(k+1)}_{(G_2,\ell_{G_2})}(v_2).$$

If otherwise $N_{G_1}(v_1) \neq \emptyset$, then $\mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$ again implies that

$$\frac{1-q}{\deg(v_1)} \sum_{v \in N_{G_1}(v_1)} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v)} = \frac{1-q}{\deg(v_2)} \sum_{v \in N_{G_2}(v_2)} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v)}.$$

Hence, $\sum_{v \in N_{G_1}(v_1)} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v)} = \sum_{v \in N_{G_2}(v_2)} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v)}$ and thus

$$\left\{\!\!\left\{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v) : v \in N_{G_1}(v_1)\right\}\!\!\right\} = \left\{\!\!\left\{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v) : v \in N_{G_2}(v_2)\right\}\!\!\right\}.$$

Then, by the induction assumption again, we have that

$$\left\{\!\!\left\{\ell^{(k)}_{(G_1,\ell_{G_1})}(v) : v \in N_{G_1}(v_1)\right\}\!\!\right\} = \left\{\!\!\left\{\ell^{(k)}_{(G_2,\ell_{G_2})}(v) : v \in N_{G_2}(v_2)\right\}\!\!\right\}.$$

Therefore, $\ell^{(k+1)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k+1)}_{(G_2,\ell_{G_2})}(v_2)$. This concludes the proof. ∎

Now, we are ready to prove that

the WL test cannot distinguish $(G_1, \ell^g_{G_1})$ and $(G_2, \ell^g_{G_2})$ iff $d_{\mathrm{WL}}\Big(\big(\mathcal{X}_q(G_1), \ell^g_G\big), \Big(\mathcal{X}_q(G_2), \ell^g_{G_2}\Big)\Big) = 0$.

It suffices to show that for any $k = 0, 1, \ldots,$

$$L_k((G_1, \ell^g_1)) = L_k((G_2, \ell^g_2)) \text{ iff } \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}\right)_{\#} \mu_{G_1} = \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}\right)_{\#} \mu_{G_2} \tag{21}$$

Fix any $k = 0, \ldots$. We first assume that $L_k((G_1, \ell^g_1)) = L_k((G_2, \ell^g_2))$. Then, it is obvious that $|V_{G_1}| = |V_{G_2}|$ and moreover, there exists a bijection $\psi : V_{G_1} \to V_{G_2}$ such that $\ell^{(k)}_{(G_1, \ell^g_{G_1})}(v) = \ell^{(k)}_{(G_2, \ell^g_{G_2})}(\psi(v))$ for any $v \in V_{G_1}$. This implies the following facts:

1. By injectivity of g, $\deg_{G_1}(v) = \deg_{G_2}(\psi(v))$ for any $v \in V_{G_1}$. Hence, $\overline{\deg}_{G_1}(v) = \overline{\deg}_{G_2}(\psi(v))$ for any $v \in V_{G_1}$.
2. By Lemma 74, for any $v \in V_{G_1}$ we have that

$$\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}(v) = \varphi^k_q \circ \ell^{(k)}_{(G_1, \ell^g_{G_1})}(v) = \varphi^k_q \circ \ell^{(k)}_{(G_2, \ell^g_{G_2})}(\psi(v)) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(\psi(v)).$$

Then,

$$\begin{aligned}
\left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}\right)_{\#} \mu_{G_1} &= \sum_{v \in V_{G_1}} \frac{\overline{\deg}_{G_1}(v)}{\sum_{v' \in V_{G_1}} \overline{\deg}_{G_1}(v')} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}(v)} \\
&= \sum_{v \in V_{G_1}} \frac{\overline{\deg}_{G_2}(\psi(v))}{\sum_{v' \in V_{G_1}} \overline{\deg}_{G_2}(\psi(v'))} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(\psi(v))} \\
&= \sum_{v \in V_{G_2}} \frac{\overline{\deg}_{G_2}(v)}{\sum_{v' \in V_{G_2}} \overline{\deg}_{G_2}(v')} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(v)} = \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}\right)_{\#} \mu_{G_2}.
\end{aligned}$$

Conversely, we assume that $\left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}\right)_{\#} \mu_{G_1} = \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}\right)_{\#} \mu_{G_2}$. Then,

$$\sum_{v \in V_{G_1}} \frac{\overline{\deg}_{G_1}(v)}{\sum_{v' \in V_{G_1}} \overline{\deg}_{G_1}(v')} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}(v)} = \sum_{v \in V_{G_2}} \frac{\overline{\deg}_{G_2}(v)}{\sum_{v' \in V_{G_2}} \overline{\deg}_{G_2}(v')} \delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(v)}. \tag{22}$$

Then, for any $v_1 \in V_{G_1}$, there exists $v_2 \in V_{G_2}$ such that $\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}(v_1) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(v_2)$. If $k = 0$, then

$$\ell^g_{G_1}(v_1) = \mathfrak{l}^{(0)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}(v_1) = \mathfrak{l}^{(0)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(v_2) = \ell^g_{G_2}(v_2).$$

Otherwise, we assume that $k > 0$. Since $\frac{1}{2} < q < 1$, we have that $\mathfrak{l}^{(k-1)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}(v_1) = \mathfrak{l}^{(k-1)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(v_2)$. Inductively, we still obtain that

$$\ell^g_{G_1}(v_1) = \mathfrak{l}^{(0)}_{(\mathcal{X}_q(G_1), \ell^g_{G_1})}(v_1) = \mathfrak{l}^{(0)}_{(\mathcal{X}_q(G_2), \ell^g_{G_2})}(v_2) = \ell^g_{G_2}(v_2).$$

Hence, by injectivity of g, we have that $|V_{G_1}| = |V_{G_2}|$, $\deg_{G_1}(v_1) = \deg_{G_2}(v_2)$ and $\overline{\deg}_{G_1}(v_1) = \overline{\deg}_{G_2}(v_2)$. Then, it is easy to see from Equation (22) that

$$\left|\left\{v \in V_{G_1} : \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell^g_{G_1})}(v) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell^g_{G_1})}(v_1)\right\}\right| = \left|\left\{v \in V_{G_2} : \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell^g_{G_2})}(v) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell^g_{G_2})}(v_2)\right\}\right|.$$

It is obvious that $\ell^g_{G_i}$ for $i = 1, 2$ satisfy the condition in Lemma 75. Then, by Lemma 75, we have that $\ell^{(k)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell_{G_2})}(v_2)$ and that

$$\left|\left\{v \in V_{G_1} : \ell^{(k)}_{(G_1,\ell_{G_1})}(v) = \ell^{(k)}_{(G_1,\ell_{G_1})}(v_1)\right\}\right| = \left|\left\{v \in V_{G_2} : \ell^{(k)}_{(G_2,\ell_{G_2})}(v) = \ell^{(k)}_{(G_2,\ell_{G_2})}(v_2)\right\}\right|.$$

Therefore, $L_k\Big((G_1, \ell^g_{G_1})\Big) = L_k\Big((G_2, \ell^g_{G_2})\Big)$.

B.2.8 Proof of Corollary 19

It turns out that one only needs finite steps to determine whether the WL test can distinguish two labeled graphs (Krebs and Verbitsky, 2015). More precisely:

Proposition 76 *For any labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ holds for all $k = 0, \ldots, |V_{G_1}| + |V_{G_2}|$ if and only if $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ holds for all $k \geq 0$.*

Hence, this corollary is a direct consequence of Proposition 17 and Proposition 76.

B.2.9 Proof of Proposition 20

Recall from Equation (4) that the WL distance of depth k is defined as

$$\begin{aligned} d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) &\coloneqq d_{\mathrm{W}}\Big(\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\big)_{\#}\mu_X, \big(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\big)_{\#}\mu_Y\Big) \\ &= \inf_{\gamma \in \mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\big)\,\gamma(dx \times dy). \end{aligned}$$

In order to compute $d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$, we must first compute $d_{\mathrm{W}}(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y))$ for each $x \in X$ and $y \in Y$. To do this, we introduce some notation. For each $i = 1, \ldots, k$, we let C_i denote the $|X| \times |Y|$ matrix such that for each $x \in X$ and $y \in Y$,

$$C_i(x, y) \coloneqq d_{\mathrm{W}}\big(\mathfrak{l}^{(i)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(i)}_{(\mathcal{Y},\ell_Y)}(y)\big).$$

We also let C_0 denote the matrix such that $C_0(x, y) \coloneqq \|\ell_X(x) - \ell_Y(y)\|$ for each $x \in X$ and $y \in Y$. Then, our task is to compute the matrix C_k. For this purpose, we consecutively compute the matrix C_i for $i = 1, \ldots, k$: Given matrix C_{i-1}, since $\mathfrak{l}^{(i)}_{(\mathcal{X},\ell_X)}(x) = \big(\mathfrak{l}^{(i-1)}_{(\mathcal{X},\ell_X)}\big)_{\#} m^X_x$ and $\mathfrak{l}^{(i)}_{(\mathcal{Y},\ell_Y)}(y) = \big(\mathfrak{l}^{(i-1)}_{(\mathcal{Y},\ell_Y)}\big)_{\#} m^Y_y$, computing

$$d_{\mathrm{W}}\big(\mathfrak{l}^{(i)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(i)}_{(\mathcal{Y},\ell_Y)}(y)\big) = \inf_{\gamma \in \mathcal{C}(m^X_x, m^Y_y)} \int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(i-1)}_{(\mathcal{X},\ell_X)}(x'), \mathfrak{l}^{(i-1)}_{(\mathcal{Y},\ell_Y)}(y')\big)\,\gamma(dx' \times dy').$$

is reduced to solving the optimal transport problem with C_{i-1} as the cost matrix and m_x^X and m_y^Y as the source and target distributions, which can be done in $O(n^3 \log(n))$ time (Pele and Werman, 2009). Thus, for each i, computing C_i given that we know C_{i-1}, requires $O(n^2 \cdot n^3 \log(n))$. Finally, we need $O(n^3 \log(n))$ time to compute $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$ based on solving an optimal transport problem with cost matrix C_k and with μ_X and μ_Y being the source and target distributions, respectively.

Therefore, the total time needed to compute $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$ is

$$k \cdot O(n^5 \log(n)) + O(n^3 \log(n)) = O(k\, n^5 \log(n)).$$

B.2.10 Proof of Proposition 21

We need the following lemma:

Lemma 77 *Given any $k \geq 0$ and any Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that*

$$\begin{aligned} &d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \\ &= \int_{X \times Y} \cdots \int_{X \times Y} d_Z(\ell_X(x_k), \ell_Y(y_k))\, (\nu_k)_{x_{k-1}, y_{k-1}}(dx_k \times dy_k) \cdots (\nu_1)_{x,y}(dx_1 \times dy_1)\gamma(dx \times dy) \end{aligned}$$

for some $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and measurable maps $(\nu_i)_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)$ such that $(\nu_i)_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)$ $\forall (x, y) \in X \times Y$ for $i = 1, \ldots, k$.

Proof First of all, we claim that for all $k \geq 0$ and $(x, y) \in X \times Y$ we have that

$$\begin{aligned} &d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}(x), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}(y)\big) \\ &= \int_{X \times Y} \cdots \int_{X \times Y} d_Z(\ell_X(x_k), \ell_Y(y_k))\, (\nu_k)_{x_{k-1}, y_{k-1}}(dx_k \times dy_k) \cdots (\nu_1)_{x,y}(dx_1 \times dy_1). \end{aligned}$$

where each $(\nu_i)_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)$ is a measurable map such that for every $(x, y) \in X \times Y$, $(\nu_i)_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)$ for $i = 1, \ldots, k$.

We prove the above claim by induction on k. The case $k = 0$ trivially since $d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(0)}(x), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(0)}(y)\big) = d_X(x, y)$. Now, assume that the claim holds up to $(k-1)$ for some $k \geq 1$. Then, by Lemma 67 for any $(x, y) \in X \times Y$ we have that

$$d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}(x), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}(y)\big) = \inf_{\mu_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)} \int_{X \times Y} d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k-1)}(x_1), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k-1)}(y_1)\big)\, \mu_{x,y}(dx_1 \times dy_1).$$

Since $(x, y) \mapsto (m_x^X, m_y^Y)$ is continuous (hence measurable) by our definition of Markov chains, by Lemma 69 we have that there exists a measurable map $(\nu_1)_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)$ such that for every $(x, y) \in X \times Y$, $(\nu_1)_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)$ is optimal, i.e.,

$$d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}(x), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}(y)\big) = \int_{X \times Y} d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k-1)}(x_1), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k-1)}(y_1)\big)\, (\nu_1)_{x,y}(dx_1 \times dy_1). \tag{23}$$

Hence, by the inductive assumption one can finish the proof of the claim.
Finally, with the claim above, we have the following formulas.

$$
\begin{aligned}
&d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = d_{\mathrm{W}}\Big(\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\big)_{\#}\mu_X, \big(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\big)_{\#}\mu_Y\Big)\\
=&\int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x), \mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\big)\,\gamma(dx\times dy); \quad \text{here } \gamma\in\mathcal{C}(\mu_X,\mu_Y) \text{ is chosen to be optimal}\\
=&\int_{X\times Y}\cdots\int_{X\times Y} d_Z(\ell_X(x_k),\ell_Y(y_k))\,(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy).
\end{aligned}
$$

where the last equality holds by the claim. Here, each $(\nu_i)_{\bullet,\bullet}\in\mathcal{C}(m_\bullet^X, m_\bullet^Y)$ for $i=1,\dots,k$ is optimal in the sense of Equation (23). ■

We are ready to prove Proposition 21. By Lemma 77, we know that

$$
\begin{aligned}
&d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y))\\
=&\int_{X\times Y}\cdots\int_{X\times Y} d_Z(\ell_X(x_k),\ell_Y(y_k))\,(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)
\end{aligned}
$$

where γ and ν_i's are as above. Furthermore, it is easy to verify that $(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)$ is a coupling between $m_x^{X,\otimes k}$ and $m_y^{X,\otimes k}$ by induction on k. Hence, we have that

$$
\begin{aligned}
&d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y))\\
=&\int_{X\times Y}\cdots\int_{X\times Y} d_Z(\ell_X(x_k),\ell_Y(y_k))\,(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)\\
\geq&\int_{X\times Y} d_{\mathrm{W}}\left((\ell_X)_{\#}m_x^{X,\otimes k},(\ell_Y)_{\#}m_y^{Y,\otimes k}\right)\gamma(dx\times dy); \quad \text{by Lemma 67}\\
\geq& d_{\mathrm{WLLB}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)).
\end{aligned}
$$

B.2.11 Proof of Proposition 22

Given two finite LMMCs,

$$(\mathcal{X},\ell_X: X\to\mathbb{R}) \text{ where } \mathcal{X}=\left(X=\{x_1,x_2,...,x_n\},m_\bullet^X,\mu_X\right)$$

and

$$(\mathcal{Y},\ell_Y: Y\to\mathbb{R}) \text{ where } \mathcal{Y}=\left(Y=\{y_1,y_2,...,y_m\},m_\bullet^Y,\mu_Y\right)$$

We represent their Markov kernels as two transition matrices, $M_{\mathcal{X}}$ and $M_{\mathcal{Y}}$, respectively. Then, the k-step Markov kernels $m_\bullet^{X,\otimes k}$ and $m_\bullet^{Y,\otimes k}$ are expressed as matrices $M_{\mathcal{X}}^k$ and $M_{\mathcal{Y}}^k$, respectively. Assume that $n=\max(|X|,|Y|)$. Then computing the k-step Markov kernels of $\mathcal{X}$ and $\mathcal{Y}$ will require $O(n^3\log(k))$ time where $O(n^3)$ is time needed for matrix multiplication.

Then since $(\ell_X)_{\#}m_x^{X,\otimes k}$ and $(\ell_Y)_{\#}m_y^{Y,\otimes k}$ are both distributions in $\mathbb{R}$, by (Vallender, 1974), $d_{\mathrm{W}}\left((\ell_X)_{\#}m_x^{X,\otimes k}, (\ell_Y)_{\#}m_y^{Y,\otimes k}\right)$ can be computed in $O(n)$ time for each $x \in X$ and $y \in Y$.

Finally, computing $d_{\mathrm{WLLB}}^{(k)}$ can be formulated as finding the optimal transport cost where each (x, y) entry of the cost matrix is defined as $d_{\mathrm{W}}\left((\ell_X)_{\#}m_x^{X,\otimes k}, (\ell_Y)_{\#}m_y^{Y,\otimes k}\right)$ and the source and target distributions are μ_X, μ_Y respectively. Therefore, the overall time complexity is

$$O(n^3 \log(k)) + O(n^3 \log(n)) = O(n^3 \log(kn)).$$

B.2.12 Proof of Proposition 25

The proof of Proposition 25 is based on the following basic fact about the Wasserstein distance:

Lemma 78 *Let Z be a complete and separable metric space. Endow $Z \times Z$ with any product metric $d_{Z\times Z}$ such that*

$$d_{Z\times Z}((z_1, z_2), (z_3, z_4)) \leq d_Z(z_1, z_3) + d_Z(z_2, z_4) \ \ \forall z_1, z_2, z_3, z_4 \in Z.$$

For example, one can let

$$d_{Z\times Z}((z_1, z_2), (z_3, z_4)) \coloneqq \sqrt{(d_Z(z_1, z_3))^2 + (d_Z(z_2, z_4))^2}.$$

Given any complete and separable metric space X, for any $i = 1, 2$ and any measurable maps $f_i, g_i : X \to Z$, if we let $h_i \coloneqq (f_i, g_i) : X \to Z \times Z$, then for any $\mu_1, \mu_2 \in \mathcal{P}(X)$

$$d_{\mathrm{W}}((h_1)_{\#}\mu_1, (h_2)_{\#}\mu_2) \leq d_{\mathrm{W}}((f_1)_{\#}\mu_1, (f_2)_{\#}\mu_2) + d_{\mathrm{W}}((g_1)_{\#}\mu_1, (g_2)_{\#}\mu_2)\,.$$

Proof [Proof of Lemma 78] For any $\gamma_f, \gamma_g \in \mathcal{C}(\mu_1, \mu_2)$, we define a probability measure $\nu \in \mathcal{P}(Z \times Z \times Z \times Z)$ as follows: for any measurable $A, A', B, B' \subseteq Z$

$$\nu(A \times A' \times B \times B') \coloneqq (f_1 \times f_2)_{\#}\gamma_f(A \times B) \cdot (g_1 \times g_2)_{\#}\gamma_g(A' \times B').$$

It is easy to show that $\nu \in \mathcal{C}\left((h_1)_{\#}\mu_1, (h_2)_{\#}\mu_2\right)$. Then,

$$\begin{aligned}
d_{\mathrm{W}}((h_1)_{\#}\mu_1, (h_2)_{\#}\mu_2) &\leq \int_{Z\times Z}\int_{Z\times Z} d_{Z\times Z}((z_1, z_2), (z_3, z_4))\, \nu(dz_1 \times dz_2 \times dz_3 \times dz_4) \\
&\leq \int_{Z\times Z}\int_{Z\times Z} (d_Z(z_1, z_3) + d_Z(z_2, z_4))\ \nu(dz_1 \times dz_2 \times dz_3 \times dz_4) \\
&\leq \int_{Z\times Z} d_Z(z_1, z_3)\,(f_1 \times f_2)_{\#}\gamma_f(dz_1 \times dz_3) + \int_{Z\times Z} d_Z(z_2, z_4)\,(g_1 \times g_2)_{\#}\gamma_g(dz_2 \times dz_4)
\end{aligned}$$

By Lemma 67, infimizing over all $\gamma_f, \gamma_g \in \mathcal{C}(\mu_1, \mu_2)$, we obtain the conclusion. ■

Now we are ready to prove Proposition 25.

Since $L_G^k := (\ell_G^0, \ldots, \ell_G^k)$, by inductively applying Lemma 78, we have that

$$\hat{D}^{(k)}(G_1, G_2) = d_{\mathrm{W}}\left(\left(L_{G_1}^k\right)_{\#}\mu_{G_1}, \left(L_{G_2}^k\right)_{\#}\mu_{G_2}\right) \leq \sum_{i=1}^{k} d_{\mathrm{W}}\left(\left(\ell_{G_1}^i\right)_{\#}\mu_{G_1}, \left(\ell_{G_2}^i\right)_{\#}\mu_{G_2}\right).$$

Choose $\varphi_j := \mathrm{id} : \mathbb{R}^d \to \mathbb{R}^d$ to be the identity map for each $j = 1, \ldots, k$, then using notation from Section B.3.1, we have that

$$\ell_{G_i}^j = \ell_{G_i}^{(\varphi,j)}, \quad \forall i = 1, 2 \text{ and } \forall j = 1, \ldots, k.$$

Then, by Equation (29), we conclude that

$$\begin{aligned}\hat{D}^{(k)}(G_1, G_2) &\leq \sum_{i=1}^{k} d_{\mathrm{W}}\left(\left(\mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}\right)_{\#}\mu_{G_1}, \left(\mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}\right)_{\#}\mu_{G_2}\right) \\ &= \sum_{i=1}^{k} d_{\mathrm{WL}}^{(i)}((\mathcal{X}_q(G_1), \ell_{G_1}), (\mathcal{X}_q(G_2), \ell_{G_2}))\,.\end{aligned}$$

By Proposition 15, we have that

$$\hat{D}^{(k)}(G_1, G_2) \leq k \cdot d_{\mathrm{WL}}^{(k)}((\mathcal{X}_q(G_1), \ell_{G_1}), (\mathcal{X}_q(G_2), \ell_{G_2}))\,.$$

B.3 Proofs from Section 4

B.3.1 Proof of Theorem 29

Item 1 follows directly from Theorem 32.

Next, we prove item 2. The proof is based on the following two basic results:

Lemma 79 *For any $d \in \mathbb{N}$ and any $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$, if $\alpha \neq \beta$, then there exists a Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\int_{\mathbb{R}^d} \varphi(x)\alpha(dx) \neq \int_{\mathbb{R}^d} \varphi(x)\beta(dx).$$

Proof [Proof of Lemma 79] By Kantorovich duality (see for example Remark 6.5 in (Villani, 2009)),

$$d_{\mathrm{W}}(\alpha, \beta) = \sup\left\{\left|\int_{\mathbb{R}^d} \varphi(x)\alpha(dx) - \int_{\mathbb{R}^d} \varphi(x)\beta(dx)\right| : \varphi : \mathbb{R}^d \to \mathbb{R} \text{ is 1-Lipschitz.}\right\}$$

Since $\alpha \neq \beta$, we have that $d_{\mathrm{W}}(\alpha, \beta) > 0$, and thus there exists a 1-Lipschitz $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $\int_{\mathbb{R}^d} \varphi(x)\alpha(dx) \neq \int_{\mathbb{R}^d} \varphi(x)\beta(dx)$. ■

Lemma 80 *For any $d \in \mathbb{N}$ and any $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$, if $\alpha \neq \beta$ and they are both compactly supported, then there exists a MLP $\varphi : \mathbb{R}^d \to \mathbb{R}$, with one hidden layer and with a Lipschitz non-polynomial activation function, such that*

$$\int_{\mathbb{R}^d} \varphi(x)\alpha(dx) \neq \int_{\mathbb{R}^d} \varphi(x)\beta(dx).$$

Proof [Proof of Lemma 80] By Lemma 79, there exists a Lipschitz $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $\int_{\mathbb{R}^d} \psi(x)\alpha(dx) \neq \int_{\mathbb{R}^d} \psi(x)\beta(dx)$. Let $\varepsilon \coloneqq \left|\int_{\mathbb{R}^d} \psi(x)\alpha(dx) - \int_{\mathbb{R}^d} \psi(x)\beta(dx)\right| > 0$. Let $K \coloneqq \operatorname{supp}(\alpha) \cup \operatorname{supp}(\beta)$. Then, by the classic universal approximation theorem (Pinkus, 1999, Theorem 3.1), there exists a MLP $\varphi : \mathbb{R}^d \to \mathbb{R}$, with one hidden layer and with any Lipschitz non-polynomial activation function, such that $\sup_{x\in K} \|\varphi(x) - \psi(x)\| < \frac{\varepsilon}{4}$. Therefore,

$$\left|\int_{\mathbb{R}^d} \varphi(x)\alpha(dx) - \int_{\mathbb{R}^d} \psi(x)\alpha(dx)\right| \leq \int_{\mathbb{R}^d} |\varphi(x) - \psi(x)|\, \alpha(dx) = \int_K |\varphi(x) - \psi(x)|\, \alpha(dx) \leq \frac{\varepsilon}{4}$$

and similarly,

$$\left|\int_{\mathbb{R}^d} \varphi(x)\beta(dx) - \int_{\mathbb{R}^d} \psi(x)\beta(dx)\right| \leq \frac{\varepsilon}{4}.$$

Hence, $\int_{\mathbb{R}^d} \varphi(x)\alpha(dx) \neq \int_{\mathbb{R}^d} \varphi(x)\beta(dx)$. ■

Now, given any $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ such that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) > 0$, we have that

$$d_{\mathrm{W}}\left(\left(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\right)_{\#}\mu_X, \left(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\right)_{\#}\mu_Y\right) = d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) > 0.$$

We introduce some notation. Given a MCNN_k $h \coloneqq \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1}$ and any $(\mathcal{X}, \ell_X) \in \mathcal{M}^L(Z)$, we let

$$\left(\mathcal{X}, \ell_X^{(\varphi,i)}\right) \coloneqq F_{\varphi_i} \circ \cdots \circ F_{\varphi_1}((\mathcal{X}, \ell_X)) \tag{24}$$

Then, we prove that for each $i = 1, \ldots, k$, there exists a MLP $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ for suitable dimensions d_{i-1} and d_i such that

$$\forall x \in X,\, y \in Y,\ \ell_X^{(\varphi,i)}(x) = \ell_Y^{(\varphi,i)}(y) \text{ iff } \mathfrak{l}^{(i)}_{(\mathcal{X},\ell_X)}(x) = \mathfrak{l}^{(i)}_{(\mathcal{Y},\ell_Y)}(y). \tag{25}$$

Given Equation (25), it is obvious that $\left(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\right)_{\#}\mu_X \neq \left(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\right)_{\#}\mu_Y$ implies that $\left(\ell_X^{(\varphi,k)}\right)_{\#}\mu_X \neq \left(\ell_Y^{(\varphi,k)}\right)_{\#}\mu_Y$. Then, by Lemma 80 and by the fact that X and Y are finite, there exists a MLP $\varphi_{k+1} : \mathbb{R}^{d_k} \to \mathbb{R}$ such that

$$\int_{\mathbb{R}^d} \varphi_{k+1}(t)\left(\ell_X^{(\varphi,k)}\right)_{\#}\mu_X(dt) \neq \int_{\mathbb{R}^d} \varphi_{k+1}(t)\left(\ell_Y^{(\varphi,k)}\right)_{\#}\mu_Y(dt).$$

Then, by the classic universal approximation theorem (Pinkus, 1999, Theorem 3.1) again, there exists a MLP $\psi : \mathbb{R} \to \mathbb{R}$ which distinguishes the two numbers above. Hence, we have

$$\begin{aligned}
\psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1}(\mathcal{X}) &= \psi\left(\int_{\mathbb{R}^d} \varphi_{k+1}(t)\left(\ell_X^{(\varphi,k)}\right)_{\#}\mu_X(dt)\right) \\
&\neq \psi\left(\int_{\mathbb{R}^d} \varphi_{k+1}(t)\left(\ell_Y^{(\varphi,k)}\right)_{\#}\mu_Y(dt)\right) \\
&= \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1}(\mathcal{Y}).
\end{aligned}$$

To conclude the proof, we prove Equation (25) by induction on $i = 1, \ldots, k$. When $i = 1$, let

$$A_1 \coloneqq \left\{(x, y) \in X \times Y : (\ell_X)_{\#}m_x^X \neq (\ell_Y)_{\#}m_y^Y\right\}.$$

Since X and Y are finite, A_1 is a finite set. We enumerate elements in A_1 and write $A_1 = \{(x_1, y_1), \ldots, (x_{d_1}, y_{d_1})\}$. By Lemma 80, for each $j = 1, \ldots, d_1$, there exists a MLP $\varphi_1^j : \mathbb{R}^d \to \mathbb{R}$ such that

$$\int_{\mathbb{R}^d} \varphi_1^j(t)(\ell_X)_\# m_{x_j}^X(dt) \neq \int_{\mathbb{R}^d} \varphi_1^j(t)(\ell_Y)_\# m_{y_j}^Y(dt).$$

We then let $\varphi_1 \coloneqq \left(\varphi_1^1, \varphi_1^2, \ldots, \varphi_1^{d_1}\right) : \mathbb{R}^d \to \mathbb{R}^{d_1}$. It is not hard to see that φ_1 is still a MLP with one hidden layer and it satisfies that

$$\forall x \in X,\, y \in Y,\ (\ell_X)_\# m_x^X = (\ell_Y)_\# m_y^Y \text{ iff } \int_{\mathbb{R}^d} \varphi_1(t)(\ell_X)_\# m_x^X(dt) = \int_{\mathbb{R}^d} \varphi_1(t)(\ell_Y)_\# m_y^Y(dt).$$

Equivalent speaking,

$$\forall x \in X,\, y \in Y,\ \mathfrak{l}_{(\mathcal{X},\ell_X)}^{(1)}(x) = \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(1)}(y) \text{ iff } \ell_X^{(\varphi,1)}(x) = \ell_Y^{(\varphi,1)}(y).$$

Now, we assume that Equation (25) holds for some $i \geq 1$. For $i+1$, we let

$$A_{i+1} \coloneqq \left\{(x, y) \in X \times Y :\ \left(\ell_X^{(\varphi,i)}\right)_\# m_x^X \neq \left(\ell_Y^{(\varphi,i)}\right)_\# m_y^Y\right\}.$$

Since X and Y are finite, A_{i+1} is a finite set. We enumerate elements in A_{i+1} and write $A_{i+1} = \{(x_1, y_1), \ldots, (x_{d_{i+1}}, y_{d_{i+1}})\}$. By Lemma 80, for each $j = 1, \ldots, d_{i+1}$, there exists a MLP $\varphi_{i+1}^j : \mathbb{R}^{d_i} \to \mathbb{R}$ such that

$$\int_{\mathbb{R}^{d_i}} \varphi_{i+1}^j(t) \left(\ell_X^{(\varphi,i)}\right)_\# m_{x_j}^X(dt) \neq \int_{\mathbb{R}^{d_i}} \varphi_{i+1}^j(t) \left(\ell_Y^{(\varphi,i)}\right)_\# m_{y_j}^Y(dt).$$

We then let $\varphi_{i+1} \coloneqq \left(\varphi_{i+1}^1, \varphi_{i+1}^2, \ldots, \varphi_{i+1}^{d_{i+1}}\right) : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$. Again φ_{i+1} is a MLP with one hidden layer and it satisfies that $\forall x \in X$ and $y \in Y$,

$$\left(\ell_X^{(\varphi,i)}\right)_\# m_x^X = \left(\ell_Y^{(\varphi,i)}\right)_\# m_y^Y \text{ iff } \int_{\mathbb{R}^{d_i}} \varphi_{i+1}(t) \left(\ell_X^{(\varphi,i)}\right)_\# m_x^X(dt) = \int_{\mathbb{R}^{d_i}} \varphi_{i+1}(t) \left(\ell_Y^{(\varphi,i)}\right)_\# m_y^Y(dt).$$

Equivalent speaking,

$$\forall x \in X,\, y \in Y,\ \left(\ell_X^{(\varphi,i)}\right)_\# m_x^X = \left(\ell_Y^{(\varphi,i)}\right)_\# m_y^Y \text{ iff } \ell_X^{(\varphi,i+1)}(x) = \ell_Y^{(\varphi,i+1)}(y).$$

By the induction assumption, $\forall x \in X$, $y \in Y$ we have that

$$\ell_X^{(\varphi,i)}(x) = \ell_Y^{(\varphi,i)}(y) \text{ iff } \mathfrak{l}_{(\mathcal{X},\ell_X)}^{(i)}(x) = \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(i)}(y).$$

This implies that

$$\left(\ell_X^{(\varphi,i)}\right)_\# m_x^X = \left(\ell_Y^{(\varphi,i)}\right)_\# m_y^Y \text{ iff } \left(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(i)}\right)_\# m_x^X = \left(\mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(i)}\right)_\# m_y^Y \text{ iff } \mathfrak{l}_{(\mathcal{X},\ell_X)}^{(i+1)}(x) = \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(i+1)}(y).$$

Therefore,

$$\ell_X^{(\varphi,i+1)}(x) = \ell_Y^{(\varphi,i+1)}(y) \text{ iff } \mathfrak{l}_{(\mathcal{X},\ell_X)}^{(i+1)}(x) = \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(i+1)}(y)$$

and we thus conclude the proof.

B.3.2 Proof of Theorem 31

The proof of the theorem is based on the following several lemmas.

Lemma 81 *Consider the union*

$$C_P(Z) \coloneqq \bigcup_{(\mathcal{X},\ell_X)\in\mathcal{M}_P^L(Z)} \left(\{(\ell_X)_\# \mu_X\} \cup \left\{(\ell_X)_\# m_x^X : x \in X\right\}\right).$$

Then, $C_P(Z)$ is a countable subset of $\mathcal{P}(\mathbb{R}^d)$.

Proof Every probability measure $\alpha \in C_P(Z)$ can be expressed as

$$\alpha = \sum_{z\in Z} \alpha(z)\delta_z, \text{ where } \alpha(z) \text{ is a sum of finitely many elements in } P.$$

Then, by the fact that both Z and P is countable, we have that $C_P(Z)$ is countable. ■

Lemma 82 *Fix any $d \in \mathbb{N}$. For any Lipschitz $\varphi : \mathbb{R}^d \to \mathbb{R}$, any countable subset $Z \subseteq \mathbb{R}^d$ and any countable subset $P \subseteq [0,1]$, there exists a countable subset $Z' \subseteq \mathbb{R}$ such that*

$$F_\varphi(\mathcal{M}_P^L(Z)) \subseteq \mathcal{M}_P^L(Z') \text{ and } S_\varphi(\mathcal{M}_P^L(Z)) \subseteq Z'. \tag{26}$$

Proof Let $Z' \coloneqq q_\varphi(C_P(Z)) \subseteq \mathbb{R}$. Since $C_P(Z)$ is countable, Z' is also countable. It is obvious that Z' satisfies Equation (26) which concludes the proof. ■

Lemma 83 *For any countable subset $C \subseteq \mathcal{P}(\mathbb{R}^d)$, there exists a Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that the restriction of $q_\varphi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ to C is injective.*

Proof Consider the following vector space:

$$\mathrm{Lip}_0(\mathbb{R}^d) \coloneqq \{f : \mathbb{R}^d \to \mathbb{R} |\, f \text{ is Lipschitz and } f(0) = 0\}.$$

When equipped with the norm $\|f\| \coloneqq \sup_{x\neq y} \frac{|f(x)-f(y)|}{\|x-y\|}$, $\mathrm{Lip}_0(\mathbb{R}^d)$ is a Banach space (Weaver, 1995). Consider the following collection of signed measures $D \coloneqq \{\alpha - \beta : \alpha, \beta \in C \text{ and } \alpha \neq \beta\}$. Since C is countable, D is also countable. Each $\mu \in D$ gives rise to a linear functional $\psi_\mu : \mathrm{Lip}_0(\mathbb{R}^d) \to \mathbb{R}$ defined by $f \mapsto \int_{\mathbb{R}^d} f(x)\mu(dx)$. This functional is well-defined since every probability measure in $\mathcal{P}(\mathbb{R}^d)$ is assumed to have finite 1-moment (see Section 2.3 for this assumption on $\mathcal{P}(\mathbb{R}^d)$). Moreover, it is easy to check that ψ_μ is bounded and non-zero (see also Lemma 79 for proving that ψ_μ is non-zero). Then, $\ker(\psi_\mu)$ is nowhere dense since it is a proper closed subspace of $\mathrm{Lip}_0(\mathbb{R}^d)$. By Baire category theorem, we have that $\cup_{\mu\in D} \ker(\psi_\mu) \neq \mathrm{Lip}_0(\mathbb{R}^d)$. Then, choose an arbitrary $\varphi \in \mathrm{Lip}_0(\mathbb{R}^d) \setminus \cup_{\mu\in D} \ker(\psi_\mu)$ and it follows that $q_\varphi|_C$ is injective. ■

Lemma 84 *For any countable subset $S \subseteq \mathbb{R}$, there exists a Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ such that the restriction of $\psi : \mathbb{R} \to \mathbb{R}$ to S is injective.*

Proof Consider the isometric embedding $E : \mathbb{R} \to (\mathcal{P}(\mathbb{R}), d_{\mathrm{W}}^{\mathbb{R}})$ sending each $t \in \mathbb{R}$ to the Dirac delta measure δ_t. By Lemma 83, there exists a Lipschitz map $\varphi : \mathbb{R} \to \mathbb{R}$ such that $q_\varphi|_{E(S)}$ is injective. We let $\psi \coloneqq q_\varphi \circ E : \mathbb{R} \to \mathbb{R}$. Then, it is obvious that ψ is Lipschitz and $\psi|_S$ is injective. ■

Proof [Proof of Theorem 31] We let $Z_0 \coloneqq Z$ and $C_1 \coloneqq C_P(Z_0)$. By Lemma 83, there exists a Lipschitz map $\varphi_1 : \mathbb{R}^d \to \mathbb{R}$ such that $q_{\varphi_1}|_{C_1}$ is injective. By Lemma 82, there exists a countable subset $Z_1 \subseteq \mathbb{R}$ such that

$$F_{\varphi_1}(\mathcal{M}_P^L(Z_0)) \subseteq \mathcal{M}_P^L(Z_1).$$

Let $C_2 \coloneqq C_P(Z_1)$. By Lemma 83 again, there exists a Lipschitz map $\varphi_2 : \mathbb{R} \to \mathbb{R}$ such that $q_{\varphi_2}|_{C_2}$ is injective. Then, inductively, for each $i = 1, \ldots, k$, there exist a countable subset $Z_i \subseteq \mathbb{R}$ and a Lipschitz map $\varphi_i : \mathbb{R} \to \mathbb{R}$ so that q_{φ_i} is injective when restricted to $C_i \coloneqq C_P(Z_{i-1})$ and $F_{\varphi_i}(\mathcal{M}_P^L(Z_{i-1})) \subseteq \mathcal{M}_P^L(Z_i)$. Similarly, there exist a countable subset $S \subseteq \mathbb{R}$ and a Lipschitz map $\varphi_{k+1} : \mathbb{R} \to \mathbb{R}$ such that $q_{\varphi_{k+1}}$ is injective when restricted to $C_{k+1} \coloneqq C_P(Z_k)$ and $S_{\varphi_{k+1}}(\mathcal{M}_P^L(Z_k)) \subseteq S$ and, moreover, by Lemma 84 there exists a Lipschitz map $\psi : \mathbb{R} \to \mathbb{R}$ which is injective when restricted to S. We then let $h \coloneqq \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} : \mathcal{M}^L(\mathbb{R}^d) \to \mathbb{R}$.

Pick any $(\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y) \in \mathcal{M}_P^L(Z)$. We prove that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) > 0$ iff $h((\mathcal{X}, \ell_X)) \neq h((\mathcal{Y}, \ell_Y))$. The proof is similar to the one for proving Theorem 29. For each $i = 1, \ldots, k$, we let

$$\left(\mathcal{X}, \ell_X^{(\varphi,i)}\right) \coloneqq F_{\varphi_i} \circ \cdots \circ F_{\varphi_1}((\mathcal{X}, \ell_X)). \tag{27}$$

We similarly define $\left(\mathcal{Y}, \ell_Y^{(\varphi,i)}\right)$. Then, we prove that for each $i = 1, \ldots, k$,

$$\forall x \in X,\, y \in Y,\ \ell_X^{(\varphi,i)}(x) = \ell_Y^{(\varphi,i)}(y) \text{ iff } \mathfrak{l}_{(\mathcal{X},\ell_X)}^{(i)}(x) = \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(i)}(y). \tag{28}$$

Given Equation (28), it is obvious that $\left(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(k)}\right)_\# \mu_X \neq \left(\mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(k)}\right)_\# \mu_Y$ iff $\left(\ell_X^{(\varphi,k)}\right)_\# \mu_X \neq \left(\ell_Y^{(\varphi,k)}\right)_\# \mu_Y$. Note that $\left(\ell_X^{(\varphi,k)}\right)_\# \mu_X, \left(\ell_Y^{(\varphi,k)}\right)_\# \mu_Y \in C_{k+1}$. Since $q_{\varphi_{k+1}}$ is injective on C_{k+1} and ψ is injective on S, we have that $\left(\ell_X^{(\varphi,k)}\right)_\# \mu_X \neq \left(\ell_Y^{(\varphi,k)}\right)_\# \mu_Y$ iff $h((\mathcal{X}, \ell_X)) \neq h((\mathcal{Y}, \ell_Y))$. Therefore, $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) > 0$ iff $h((\mathcal{X}, \ell_X)) \neq h((\mathcal{Y}, \ell_Y))$.

To conclude the proof, we prove Equation (28) by induction on $i = 1, \ldots, k$. When $i = 1$, since $(\ell_X)_\# m_x^X, (\ell_Y)_\# m_y^Y \in C_1$, $\forall x \in X$, $y \in Y$, by injectivity of q_{φ_1} on C_1, we have that

$$\forall x \in X,\, y \in Y,\ (\ell_X)_\# m_x^X = (\ell_Y)_\# m_y^Y \text{ iff } q_{\varphi_1}((\ell_X)_\# m_x^X) = q_{\varphi_1}((\ell_Y)_\# m_y^Y).$$

Equivalent speaking,

$$\forall x \in X,\, y \in Y,\ \mathfrak{l}_{(\mathcal{X},\ell_X)}^{(1)}(x) = \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(1)}(y) \text{ iff } \ell_X^{(\varphi,1)}(x) = \ell_Y^{(\varphi,1)}(y).$$

Now, we assume that Equation (28) holds for some $i \geq 1$. For $i + 1$, since

$$\left(\ell_X^{(\varphi,i)}\right)_\# m_x^X, \left(\ell_Y^{(\varphi,i)}\right)_\# m_y^Y \in C_{i+1},\ \forall x \in X,\, y \in Y,$$

by injectivity of $q_{\varphi_{i+1}}$ on C_{i+1}, we have that

$$\forall x \in X, y \in Y,\ \left(\ell_X^{(\varphi,i)}\right)_{\#} m_x^X = \left(\ell_Y^{(\varphi,i)}\right)_{\#} m_y^Y \text{ iff } q_{\varphi_{i+1}}\left(\left(\ell_X^{(\varphi,i)}\right)_{\#} m_x^X\right) = q_{\varphi_{i+1}}\left(\left(\ell_Y^{(\varphi,i)}\right)_{\#} m_y^Y\right).$$

Equivalent speaking,

$$\forall x \in X,\, y \in Y,\ \left(\ell_X^{(\varphi,i)}\right)_{\#} m_x^X = \left(\ell_Y^{(\varphi,i)}\right)_{\#} m_y^Y \text{ iff } \ell_X^{(\varphi,i+1)}(x) = \ell_Y^{(\varphi,i+1)}(y).$$

By the induction assumption, $\forall x \in X,\, y \in Y$ we have that

$$\ell_X^{(\varphi,i)}(x) = \ell_Y^{(\varphi,i)}(y) \text{ iff } \mathfrak{l}^{(i)}_{(\mathcal{X},\ell_X)}(x) = \mathfrak{l}^{(i)}_{(\mathcal{Y},\ell_Y)}(y).$$

This implies that

$$\left(\ell_X^{(\varphi,i)}\right)_{\#} m_x^X = \left(\ell_Y^{(\varphi,i)}\right)_{\#} m_y^Y \text{ iff } \left(\mathfrak{l}^{(i)}_{(\mathcal{X},\ell_X)}\right)_{\#} m_x^X = \left(\mathfrak{l}^{(i)}_{(\mathcal{Y},\ell_Y)}\right)_{\#} m_y^Y \text{ iff } \mathfrak{l}^{(i+1)}_{(\mathcal{X},\ell_X)}(x) = \mathfrak{l}^{(i+1)}_{(\mathcal{Y},\ell_Y)}(y).$$

Therefore,

$$\ell_X^{(\varphi,i+1)}(x) = \ell_Y^{(\varphi,i+1)}(y) \text{ iff } \mathfrak{l}^{(i+1)}_{(\mathcal{X},\ell_X)}(x) = \mathfrak{l}^{(i+1)}_{(\mathcal{Y},\ell_Y)}(y)$$

and we thus conclude the proof. ■

B.3.3 Proof of Theorem 32

We need the following lemma:

Lemma 85 *For any C-Lipschitz function $\varphi : \mathbb{R}^i \to \mathbb{R}^j$, we have that the map $q_\varphi : \mathcal{P}(\mathbb{R}^i) \to \mathbb{R}^j$ is C-Lipschitz.*

Proof [Proof of Lemma 85] For any $\alpha, \beta \in \mathcal{P}(\mathbb{R}^i)$, pick any $\gamma \in \mathcal{C}(\alpha, \beta)$. Then, we have

$$\begin{aligned}|q_\varphi(\alpha) - q_\varphi(\beta)| &= \left|\int_{\mathbb{R}^i} \varphi(x)\alpha(dx) - \int_{\mathbb{R}^i} \varphi(x)\beta(dx)\right| = \left|\int_{\mathbb{R}^i \times \mathbb{R}^i} (\varphi(x) - \varphi(y))\, \gamma(dx \times dy)\right| \\ &\leq \int_{\mathbb{R}^i \times \mathbb{R}^i} |\varphi(x) - \varphi(y)|\, \gamma(dx \times dy) \leq C \cdot \int_{\mathbb{R}^i \times \mathbb{R}^i} |x - y|\, \gamma(dx \times dy).\end{aligned}$$

Since $\gamma \in \mathcal{C}(\alpha, \beta)$ is arbitrary, we have that

$$|q_\varphi(\alpha) - q_\varphi(\beta)| \leq C \cdot d_{\mathrm{W}}(\alpha, \beta).$$

Hence q_φ is C-Lipschitz. ■

Given a MCNN$_k$ $h \coloneqq \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1}$ and any $(\mathcal{X}, \ell_X) \in \mathcal{M}^L(Z)$, recall the notation from Equation (24):

$$\left(\mathcal{X}, \ell_X^{(\varphi,i)}\right) \coloneqq F_{\varphi_i} \circ \cdots \circ F_{\varphi_1}((\mathcal{X}, \ell_X))$$

We assume that for $i = 1, \ldots, k$, φ_i is a C_i-Lipschitz MLP for some $C_i > 0$. Then, by Lemma 85, we have that q_{φ_i} is a C_i-Lipschitz map for $i = 1, \ldots, k$.

We then prove the following inequality:

$$d_{\mathrm{W}}\Big(\big(\ell_X^{(\varphi,k)}\big)_{\#}\mu_X, \big(\ell_Y^{(\varphi,k)}\big)_{\#}\mu_Y\Big) \leq \Pi_{i=1}^{k} C_i \cdot d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)). \tag{29}$$

Given Equation (29), by Lemma 85 and the fact that ψ is C-Lipschitz, one has that

$$\begin{aligned}
|h((\mathcal{X},\ell_X)) - h((\mathcal{Y},\ell_Y))| &\leq C \cdot \left\| q_{\varphi_{k+1}}\Big(\big(\ell_X^{(\varphi,k)}\big)_{\#}\mu_X\Big) - q_{\varphi_{k+1}}\Big(\big(\ell_Y^{(\varphi,k)}\big)_{\#}\mu_Y\Big)\right\| \\
&\leq C \cdot C_{k+1} d_{\mathrm{W}}\Big(\big(\ell_X^{(\varphi,k)}\big)_{\#}\mu_X, \big(\ell_Y^{(\varphi,k)}\big)_{\#}\mu_Y\Big) \\
&\leq C \cdot \Pi_{i=1}^{k+1} C_i \cdot d_{\mathrm{WL}}^{(k)}((X,\ell_X),(Y,\ell_Y))\,.
\end{aligned}$$

To prove Equation (29), it suffices to prove that for any $x \in X$ and $y \in Y$ (see Lemma 67),

$$\big\|\ell_X^{(\varphi,k)}(x) - \ell_Y^{(\varphi,k)}(y)\big\| \leq \Pi_{i=1}^{k} C_i \cdot d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(k)}(x), \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(k)}(y)\big)\,.$$

We prove the above inequality by proving the following inequality inductively on $j = 1,\ldots,k$:

$$\big\|\ell_X^{(\varphi,j)}(x) - \ell_Y^{(\varphi,j)}(y)\big\| \leq \Pi_{i=1}^{j} C_i \cdot d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(j)}(x), \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(j)}(y)\big)\,. \tag{30}$$

When $j = 0$, we have that $\ell_X^{(\varphi,0)} = \ell_X = \mathfrak{l}_{(\mathcal{X},\ell_X)}^{(0)}$ and $\ell_Y^{(\varphi,0)} = \ell_Y = \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(0)}$. Therefore, Equation (30) obviously holds (we let $\Pi_{i=1}^{0} C_i \coloneqq 1$). We now assume that Equation (30) holds for some $j \geq 0$. For $j+1$, we have that

$$\begin{aligned}
&\Pi_{i=1}^{j+1} C_i \cdot d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(j+1)}(x), \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(j+1)}(y)\big) \\
&= \Pi_{i=1}^{j+1} C_i \cdot d_{\mathrm{W}}\Big(\big(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(j)}\big)_{\#} m_x^X, \big(\mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(j)}\big)_{\#} m_y^Y\Big) \\
&= C_{j+1} \cdot \inf_{\gamma \in \mathcal{C}(m_x^X, m_y^Y)} \int_{X\times Y} \Pi_{i=1}^{j} C_i \cdot d_{\mathrm{W}}\big(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(j)}(x'), \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(j)}(y')\big)\,\gamma(dx'\times dy') \\
&\geq C_{j+1} \inf_{\gamma \in \mathcal{C}(m_x^X, m_y^Y)} \int_{X\times Y} \big\|\ell_X^{(\varphi,j)}(x') - \ell_Y^{(\varphi,j)}(y')\big\|\,\gamma(dx'\times dy') \\
&= C_{j+1} \cdot d_{\mathrm{W}}\Big(\big(\ell_X^{(\varphi,j)}\big)_{\#} m_x^X, \big(\ell_Y^{(\varphi,j)}\big)_{\#} m_y^Y\Big) \\
&\geq \left\| q_{\varphi_{j+1}}\Big(\big(\ell_X^{(\varphi,j)}\big)_{\#} m_x^X\Big) - q_{\varphi_{j+1}}\Big(\big(\ell_Y^{(\varphi,j)}\big)_{\#} m_y^Y\Big)\right\| \\
&= \big\|\ell_X^{(\varphi,j+1)}(x) - \ell_Y^{(\varphi,j+1)}(y)\big\|\,.
\end{aligned}$$

B.3.4 Proof of Theorem 33

The main technique we utilize to prove the theorem is the following Stone-Weierstrass theorem.

Lemma 86 (Stone-Weierstrass) *Let X be a compact Hausdorff space. Let $\mathcal{F} \subseteq C(X,\mathbb{R})$ be a subalgebra containing the constant function 1. If moreover $\mathcal{F}$ separates points, then $\mathcal{F}$ is dense in $C(X,\mathbb{R})$.*

Ideally, one would try to let $X \coloneqq \mathcal{K}$ to apply the Stone-Weierstrass theorem. However, notice that $\mathcal{M}_k^L(\mathbb{R}^d)$ is a pseudometric space and may not be a Hausdorff space. To handle this subtlety, we consider the following variant of the Stone-Weierstrass theorem.

Lemma 87 (Variant of Stone-Weierstrass) *Let (X, d) be a compact pseudometric space. Let $\mathcal{F} \subseteq C(X, \mathbb{R})$ be a subalgebra containing the constant function 1. If moreover $\mathcal{F}$ separates points, then $\overline{\mathcal{F}} = C(X, \mathbb{R})$.*

Proof [Proof of Lemma 87] In general, a pseudometric space canonically induces a metric space by identifying points at 0 distance (see Burago et al. (2001, Proposition 1.1.5)). We let $\widehat{X}$ denote the metric space induced from X with the metric $\hat{d}([x], [y]) := d(x, y)$ for any $[x], [y] \in \widehat{X}$.

Let Q denote this identification map (also called the *quotient* map): $Q : (X, d) \to (\widehat{X}, d)$ sending each $x \in X$ to its equivalence class $[x] \in \widehat{X}$. Then, consider the map $Q_c : C(\widehat{X}, \mathbb{R}) \to C(X, \mathbb{R})$ defined by sending $\hat{f} : \widehat{X} \to \mathbb{R}$ to $f := \hat{f} \circ Q : X \to \mathbb{R}$. It is easy to check that Q_c is an isometric embedding w.r.t. the sup metrics on $C(\widehat{X}, \mathbb{R})$ and $C(X, \mathbb{R})$ and that Q_c has its isometric inverse $(Q_c)^{-1} : C(X, \mathbb{R}) \to C(\widehat{X}, \mathbb{R})$ defined as follows: for any $f \in C(X, \mathbb{R})$, we let $\hat{f} : \widehat{X} \to \mathbb{R}$ be defined such that for any equivalence class $[x] \in \widehat{X}$, $\hat{f}([x]) := f(x)$ (it is obvious that $\hat{f}$ does not depend on the choice of the representative x). Hence, Q_c gives rise to an isometry from $C(\widehat{X}, \mathbb{R})$ to $C(X, \mathbb{R})$.

Note that $Q_c(\mathcal{F})$ is a subalgebra of $C(\widehat{X}, \mathbb{R})$ containing the constant function 1. If $\mathcal{F}$ separates points, then $Q_c(\mathcal{F})$ separates points on $\widehat{X}$. By the Stone-Weierstrass theorem, we have that $\overline{Q_c(\mathcal{F})} = C(\widehat{X}, \mathbb{R})$. Since Q_c is an isometry, we have that $\overline{\mathcal{F}} = C(X, \mathbb{R})$. ■

We now start by proving a relaxation of Theorem 33. Let

$$\mathcal{NN}_k^c(\mathbb{R}^d) \coloneqq \{\psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} : \forall \text{ MLPs } \varphi_i, i = 1, \ldots, k+1 \text{ and continuous } \psi.\}.$$

Note that $\mathcal{NN}_k(\mathbb{R}^d) \subseteq \mathcal{NN}_k^c(\mathbb{R}^d)$ since the latter relaxes the MLP assumption for ψ in $\mathcal{NN}_k(\mathbb{R}^d)$. Then, we prove the following lemma:

Lemma 88 *For any $k \in \mathbb{N}$, let $\mathcal{K} \subseteq \mathcal{M}_k^L(\mathbb{R}^d)$ be any compact subspace. Then, $\overline{\mathcal{NN}_k^c(\mathbb{R}^d)|_{\mathcal{K}}} = C(\mathcal{K}, \mathbb{R})$.*

We prove the lemma by applying Lemma 87 to $\mathcal{K}$.

$\mathcal{NN}_k^c(\mathbb{R}^d)|_{\mathcal{K}}$ **Contains 1.** Given any choice of φ_is, we let $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$ be the constant map 1. Then, the corresponding function $h = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} \equiv 1 \in \mathcal{NN}_k^c(\mathbb{R}^d)$.

$\mathcal{NN}_k^c(\mathbb{R}^d)|_{\mathcal{K}}$ **Separates Points.** This follows from item 2 in Theorem 29.

$\mathcal{NN}_k^c(\mathbb{R}^d)|_{\mathcal{K}}$ **Is a Subalgebra.** By Equation (29), Lemma 85 and the continuity of ψ, we have that $\mathcal{NN}_k^c(\mathbb{R}^d) \subseteq C(\mathcal{K}, \mathbb{R})$. Next, we show that $\mathcal{NN}_k^c(\mathbb{R}^d)$ is, in fact, a subalgebra of $C(\mathcal{K}, \mathbb{R})$. Given any constant c and function $h = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} \in \mathcal{NN}_k^c(\mathbb{R}^d)$, we have that

$$c \cdot h = c \cdot \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} = (c \cdot \psi) \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} \in \mathcal{NN}_k^c(\mathbb{R}^d).$$

Then, we show that the sum and the product of any $h_1 = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1}$ and $h_2 = \tilde{\psi} \circ S_{\tilde{\varphi}_{k+1}} \circ F_{\tilde{\varphi}_k} \circ \cdots F_{\tilde{\varphi}_1}$ belongs to $\mathcal{NN}_k^c(\mathbb{R}^d)$. We define

$$\Phi_1 \coloneqq (\varphi_1, \tilde{\varphi}_1) : \mathbb{R}^d \to \mathbb{R}^{d_1} \times \mathbb{R}^{\tilde{d}_1},$$

and for each $2 \leq i \leq k+1$, we define

$$\Phi_i = \varphi_i \times \tilde{\varphi}_i : \mathbb{R}^{d_{i-1}} \times \mathbb{R}^{\tilde{d}_{i-1}} \to \mathbb{R}^{d_i} \times \mathbb{R}^{\tilde{d}_i}.$$

It is not hard to see that for each $i = 1, \ldots, k+1$, Φ_i is also a single hidden layer MLP with the activation function σ. Indeed, if we write $\varphi_i(x) = C_i \sigma * (W_i x + b_i)$ for any $x \in \mathbb{R}^{d_i}$, and $\varphi_i(\tilde{x}) = \tilde{C}_i \sigma * \left(\tilde{W}_i \tilde{x} + \tilde{b}_i\right)$ for any $\tilde{x} \in \mathbb{R}^{\tilde{d}_i}$, then for any $\begin{pmatrix} x \\ \tilde{x} \end{pmatrix} \in \mathbb{R}^{d_i} \times \mathbb{R}^{\tilde{d}_i}$, we have that

$$\Phi_1(x) = \begin{pmatrix} C_1 \\ \tilde{C}_1 \end{pmatrix} \sigma * \left(\begin{pmatrix} W_1 \\ \tilde{W}_1 \end{pmatrix} x + \begin{pmatrix} b_1 \\ \tilde{b}_1 \end{pmatrix} \right)$$

and

$$\Phi_i \left(\begin{pmatrix} x \\ \tilde{x} \end{pmatrix} \right) = \begin{pmatrix} C_i & 0 \\ 0 & \tilde{C}_i \end{pmatrix} \sigma * \left(\begin{pmatrix} W_i & 0 \\ 0 & \tilde{W}_i \end{pmatrix} \begin{pmatrix} x \\ \tilde{x} \end{pmatrix} + \begin{pmatrix} b_i \\ \tilde{b}_i \end{pmatrix} \right), \quad \forall i > 1.$$

We let $P : \mathbb{R}^{d_{k+1}} \times \mathbb{R}^{\tilde{d}_{k+1}} \to \mathbb{R}^{d_{k+1}}$ and $\tilde{P} : \mathbb{R}^{d_{k+1}} \times \mathbb{R}^{\tilde{d}_{k+1}} \to \mathbb{R}^{\tilde{d}_{k+1}}$ denote projection maps. Then, we can rewrite h_1 and h_2 as follows.

Claim 89 *$h_1 = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} = \psi \circ P \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1}$ and $h_2 = \tilde{\psi} \circ S_{\tilde{\varphi}_{k+1}} \circ F_{\tilde{\varphi}_k} \circ \cdots F_{\tilde{\varphi}_1} = \tilde{\psi} \circ \tilde{P} \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1}$.*

Proof [Proof of Claim 89]

We define

$$\left(\mathcal{X}_q(G), \ell_G^{(\varphi,i)}\right) \coloneqq F_{\varphi_i} \circ \cdots \circ F_{\varphi_1}((\mathcal{X}_q(G), \ell_G)) \tag{31}$$

Then, we first prove inductively on $i = 1, \ldots, k$ that for any $\mathcal{X} \in \mathcal{K}$

$$\ell_X^{(\Phi,i)}(x) = \left(\ell_X^{(\varphi,i)}(x), \ell_X^{(\tilde{\varphi},i)}(x)\right), \quad \forall x \in X. \tag{32}$$

When $i = 1$,

$$\ell_X^{(\Phi,1)}(x) = q_{\Phi_1}((\ell_X)_\# m_x^X)) = (q_{\varphi_1}((\ell_X)_\# m_x^X), q_{\tilde{\varphi}_1}((\ell_X)_\# m_x^X)) = \left(\ell_X^{(\varphi,1)}(x), \ell_X^{(\tilde{\varphi},1)}(x)\right).$$

Now, we assume that Equation (32) holds for some $i \geq 1$. Then, for $i+1$, we have that

$$\begin{aligned}
\ell_X^{(\Phi,i+1)}(x) &= q_{\Phi_{i+1}}\left(\left(\ell_X^{(\Phi,i)}\right)_\# m_x^X\right) = q_{\Phi_{i+1}}\left(\left(\left(\ell_X^{(\varphi,i)}, \ell_X^{(\tilde{\varphi},i)}\right)\right)_\# m_x^X\right) \\
&= q_{\Phi_{i+1}}\left(\left(\ell_X^{(\varphi,i)}\right)_\# m_x^X \otimes \left(\ell_X^{(\tilde{\varphi},i)}\right)_\# m_x^X\right) \\
&= \left(q_{\varphi_{i+1}}\left(\left(\ell_X^{(\varphi,i)}\right)_\# m_x^X\right), q_{\tilde{\varphi}_{i+1}}\left(\left(\ell_X^{(\tilde{\varphi},i)}\right)_\# m_x^X\right)\right) \\
&= \left(\ell_X^{(\varphi,i+1)}(x), \ell_X^{(\tilde{\varphi},i+1)}(x)\right),
\end{aligned}$$

which concludes the proof of Equation (32).

Similarly,

$$\begin{aligned}
S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1}((\mathcal{X}, \ell_X)) &= q_{\Phi_{k+1}}\left(\left(\ell_X^{(\Phi,k)}\right)_\# \mu_X\right) = q_{\Phi_{k+1}}\left(\left(\left(\ell_X^{(\varphi,k)}, \ell_X^{(\tilde{\varphi},k)}\right)\right)_\# \mu_X\right) \\
&= q_{\Phi_{k+1}}\left(\left(\ell_X^{(\varphi,k)}\right)_\# \mu_X \otimes \left(\ell_X^{(\tilde{\varphi},k)}\right)_\# \mu_X\right) \\
&= \left(q_{\varphi_{k+1}}\left(\left(\ell_X^{(\varphi,k)}\right)_\# \mu_X\right), q_{\tilde{\varphi}_{k+1}}\left(\left(\ell_X^{(\tilde{\varphi},k)}\right)_\# \mu_X\right)\right) \\
&= \left(S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1}((\mathcal{X}, \ell_X)), S_{\tilde{\varphi}_{k+1}} \circ F_{\tilde{\varphi}_k} \circ \cdots F_{\tilde{\varphi}_1}((\mathcal{X}, \ell_X))\right).
\end{aligned}$$

Therefore, $\psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} = \psi \circ P \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1}$ and similarly, $\tilde{\psi} \circ S_{\tilde{\varphi}_{k+1}} \circ F_{\tilde{\varphi}_k} \circ \cdots F_{\tilde{\varphi}_1} = \tilde{\psi} \circ \tilde{P} \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1}$. ■

Given these claims, we then have that

$$\psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} + \tilde{\psi} \circ S_{\tilde{\varphi}_{k+1}} \circ F_{\tilde{\varphi}_k} \circ \cdots F_{\tilde{\varphi}_1} = (\psi \circ P + \tilde{\psi} \circ \tilde{P}) \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1} \in \mathcal{MN}_k^c(\mathbb{R}^d)$$

and

$$\psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} \times \tilde{\psi} \circ S_{\tilde{\varphi}_{k+1}} \circ F_{\tilde{\varphi}_k} \circ \cdots F_{\tilde{\varphi}_1} = (\psi \circ P \times \tilde{\psi} \circ \tilde{P}) \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1} \in \mathcal{MN}_k^c(\mathbb{R}^d).$$

This concludes the proof of Claim 88. Now we finish proving Theorem 33 by showing that $\mathcal{MN}_k(\mathbb{R}^d)|_{\mathcal{K}}$ is dense in $\mathcal{MN}_k^c(\mathbb{R}^d)|_{\mathcal{K}}$. Choose any $h = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} \in \mathcal{MN}_k^c(\mathbb{R}^d)$. By Equation (29) and Lemma 85 again, we have that the map $S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} : \mathcal{M}_k^L(\mathbb{R}^d) \to \mathbb{R}^{d_{k+1}}$ is continuous. Hence, $\mathfrak{K} \coloneqq S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1}(\mathcal{K})$ is a compact subspace of $\mathbb{R}^{d_{k+1}}$. Now, for any $\varepsilon > 0$, by the universal approximation theorem for MLP (Pinkus, 1999, Theorem 3.1), there exists a single hidden layer MLP $\tilde{\psi} : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$ such that $\sup_{x \in \mathfrak{K}} \left\| \tilde{\psi}(x) - \psi(x) \right\| < \varepsilon$. If we let $\tilde{h} \coloneqq \tilde{\psi} \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1}$, then $\tilde{h} \in \mathcal{MN}_k(\mathbb{R}^d)$ and moreover,

$$\sup_{(\mathcal{X}, \ell_X) \in \mathcal{K}} \left| \tilde{h}((\mathcal{X}, \ell_X)) - h((\mathcal{X}, \ell_X)) \right| < \varepsilon.$$

This implies that $\mathcal{MN}_k(\mathbb{R}^d)|_{\mathcal{K}}$ is dense in $\mathcal{MN}_k^c(\mathbb{R}^d)|_{\mathcal{K}}$. This concludes the proof.

B.3.5 Proof of Proposition 35

We now carry out calculations for the first layer given a graph induced LMMC as input. Let $(G, \ell_G) \in \mathcal{G}(\mathbb{R}^d)$. Consider the corresponding LMMC $(\mathcal{X}_q(G), \ell_G)$ defined in Section 2.4. Then, by applying H to $(\mathcal{X}_q(G), \ell_G)$, one has that for any $v \in V$ (without loss of generality, we assume $\deg(v) > 0$),

$$\begin{aligned}
\ell_G^{\varphi_1}(v) = q_{\varphi_1}((\ell_G)_{\#} m_v^{G,q}) &= q_{\varphi_1}\left(q\delta_{\ell_G(v)} + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} w_{vv'} \delta_{\ell_G(v')} \right) \\
&= \int_{\mathbb{R}^d} \varphi_1(v'') \left(q\delta_{\ell_G(v)} + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} w_{vv'} \delta_{\ell_G(v')} \right) (dv'') \\
&= q\varphi_1(\ell_G(v)) + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} w_{vv'} \varphi_1(\ell_G(v')).
\end{aligned}$$

Hence the label map $\ell_G^{\varphi_1} : V \to \mathbb{R}^{d_1}$ agrees with the one ℓ_G^1 obtained via the message passing rule specified in Section 4.2. In this way, it is straightforward to verify that any MP-GNN defined in Section Section 4.2 is a restriction of an MCNN to the collection of graph induced LMMCs $\mathcal{G}_q(\mathbb{R}^d)$, i.e., $h = H \circ I_q$.

B.4 Proofs from Section 5

B.4.1 Proof of Theorem 46

By Lemma 77, we have that

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y))$$
$$= \int_{X\times Y}\cdots\int_{X\times Y} d_Z(\ell_X(x_k),\ell_Y(y_k))\,(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)$$

for some $\gamma\in\mathcal{C}(\mu_X,\mu_Y)$ and measurable maps $(\nu_i)_{\bullet,\bullet}: X\times Y\to\mathcal{P}(X\times Y)$ such that $(\nu_i)_{x,y}\in\mathcal{C}(m_x^X,m_y^Y)\ \forall (x,y)\in X\times Y$ for $i=1,\dots,k$.

Furthermore, as pointed out in Remark 45, one can construct a Markovian coupling $\{(\overline{\boldsymbol{X}}_i,\overline{\boldsymbol{Y}}_i)\}_{i\in\mathbb{N}}\in\mathcal{C}_{\mathcal{M}}(\mathcal{X},\mathcal{Y})$ such that

$$(\nu_{i+1})_{x_i,y_i}(x_{i+1},y_{i+1})=\mathbb{P}\big((\overline{\boldsymbol{X}}_{i+1},\overline{\boldsymbol{Y}}_{i+1})=(x_{i+1},y_{i+1})|(\overline{\boldsymbol{X}}_i,\overline{\boldsymbol{Y}}_i)=(x_i,y_i)\big),$$

and $\mathrm{law}((\overline{\boldsymbol{X}}_0,\overline{\boldsymbol{Y}}_0))=\gamma$. Hence, we have that

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y))$$
$$= \int_{X\times Y}\cdots\int_{X\times Y} d_Z(\ell_X(x_k),\ell_Y(y_k))\,(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)$$
$$=\mathbb{E}\,d_Z(\ell_X(\overline{\boldsymbol{X}}_k),\ell_Y(\overline{\boldsymbol{Y}}_k))\geq \inf_{\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\}_{i\in\mathbb{N}}\in\mathcal{C}_{\mathcal{M}}(\mathcal{X},\mathcal{Y})}\mathbb{E}\,d_Z(\ell_X(\boldsymbol{X}_k),\ell_Y(\boldsymbol{Y}_k)).$$

For the other direction, fix an arbitrary $\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\}_{i\in\mathbb{N}}\in\mathcal{C}_{\mathcal{M}}(\mathcal{X},\mathcal{Y})$. Let

$$(\nu_{i+1})_{x_i,y_i}(x_{i+1},y_{i+1}):=\mathbb{P}\big((\boldsymbol{X}_{i+1},\boldsymbol{Y}_{i+1})=(x_{i+1},y_{i+1})|(\boldsymbol{X}_i,\boldsymbol{Y}_i)=(x_i,y_i)\big),$$

and $\mathrm{law}((\boldsymbol{X}_0,\boldsymbol{Y}_0))=\gamma$. Then,

$$\mathbb{E}\,d_Z(\ell_X(\boldsymbol{X}_k),\ell_Y(\boldsymbol{Y}_k))$$
$$= \int_{X\times Y}\cdots\int_{X\times Y} d_{\mathrm{W}}(\ell_X(x_k),\ell_Y(y_k))\,(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)$$
$$= \int_{X\times Y}\cdots\int_{X\times Y} d_{\mathrm{W}}\left(\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)}(x_k),\mathfrak{l}^{(0)}_{(\mathcal{Y},\ell_Y)}(y_k)\right)(\nu_k)_{x_{k-1},y_{k-1}}(dx_k\times dy_k)\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)$$
$$\geq \int_{X\times Y}\cdots\int_{X\times Y} d_{\mathrm{W}}\left(\mathfrak{l}^{(1)}_{(\mathcal{X},\ell_X)}(x_{k-1}),\mathfrak{l}^{(1)}_{(\mathcal{Y},\ell_Y)}(y_{k-1})\right)(\nu_{k-1})_{x_{k-2},y_{k-2}}(dx_{k-1}\times dy_{k-1})\cdots(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)$$
$$\cdots$$
$$\geq \int_{X\times Y}\int_{X\times Y} d_{\mathrm{W}}\left(\mathfrak{l}^{(k-1)}_{(\mathcal{X},\ell_X)}(x_1),\mathfrak{l}^{(k-1)}_{(\mathcal{Y},\ell_Y)}(y_1)\right)(\nu_1)_{x,y}(dx_1\times dy_1)\gamma(dx\times dy)$$
$$\geq \int_{X\times Y} d_{\mathrm{W}}\left(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}(x),\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}(y)\right)\gamma(dx\times dy)\geq d_{\mathrm{W}}\left(\left(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\right)_{\#}\mu_X,\left(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\right)_{\#}\mu_Y\right)=d_{\mathrm{WL}}^{(k)}\left((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)\right).$$

where the inequalities hold by Lemma 67. Hence, we have that

$$d_{\mathrm{WL}}^{(k)}\left((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)\right)\leq \inf_{\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\}_{i\in\mathbb{N}}\in\mathcal{C}_{\mathcal{M}}(\mathcal{X},\mathcal{Y})}\mathbb{E}\,d_Z(\ell_X(\boldsymbol{X}_k),\ell_Y(\boldsymbol{Y}_k)).$$

This completes the proof.

B.4.2 Proof of Lemma 48

Note that it is easy to check that π^k is indeed a coupling measure between α^k and β^k from the definition. Hence, we only need to verify the bicausality of π^k. Observe that

$$\begin{aligned}
&\pi^k((y_0,\dots,y_l)|(x_0,\dots,x_k))\\
&= \mathbb{P}\big(\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_k = x_k\big)\\
&= \frac{\mathbb{P}\big(\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_k = x_k,\boldsymbol{Y}_0 = y_0,\dots,Y_l = y_l\big)}{\mathbb{P}\big(\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_k = x_k\big)}\\
&= \frac{\mathbb{P}\big(\boldsymbol{X}_{l+1} = x_{l+1},\dots,\boldsymbol{X}_k = x_k|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l,\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l\big)}{\mathbb{P}\big(\boldsymbol{X}_{l+1} = x_{l+1},\dots,\boldsymbol{X}_k = x_k|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l\big)}\\
&\quad\times \frac{\mathbb{P}\big(\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l,\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l\big)}{\mathbb{P}\big(\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l\big)}.
\end{aligned}$$

Moreover, since $\mathbb{P}$ is a Markovian coupling, it is easy to check that

$$\begin{aligned}
&\mathbb{P}\big(\boldsymbol{X}_{l+1} = x_{l+1},\dots,\boldsymbol{X}_k = x_k|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l,\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l\big)\\
&= \mathbb{P}\big(\boldsymbol{X}_k = x_k|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_{k-1} = x_{k-1},\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l\big)\\
&\quad\times \mathbb{P}\big(\boldsymbol{X}_{k-1} = x_{k-1}|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_{k-2} = x_{k-2},\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l\big)\\
&\quad\cdots\times \mathbb{P}\big(\boldsymbol{X}_{l+1} = x_{l+1}|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l,\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l\big)\\
&= m^X_{x_l}(x_{l+1})\cdots m^X_{x_{k-1}}(x_k)\\
&= \mathbb{P}\big(\boldsymbol{X}_{l+1} = x_{l+1},\dots,\boldsymbol{X}_k = x_k|\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l\big).
\end{aligned}$$

Therefore, one can conclude that

$$\begin{aligned}
\pi^k((y_0,\dots,y_l)|(x_0,\dots,x_k)) &= \frac{\mathbb{P}\big(\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l,\boldsymbol{Y}_0 = y_0,\dots,\boldsymbol{Y}_l = y_l\big)}{\mathbb{P}\big(\boldsymbol{X}_0 = x_0,\dots,\boldsymbol{X}_l = x_l\big)}\\
&= \pi^k((y_0,\dots,y_l)|(x_0,\dots,x_l)).
\end{aligned}$$

This implies that π^k is causal from α^k to β^k. In a similar way, one can also prove that π^k is causal from β^k to α^k. This completes the proof.

B.4.3 Proof of Theorem 49

In order to prove Theorem 49, we first note that d^{c^k} can be recursively computed. For each $i = 0,\dots,k$, we define $V_i : X^{i+1}\times Y^{i+1}\to\mathbb{R}$ recursively as follows:

$$\begin{aligned}
&V_k(x_0,\dots,x_k,y_0,\dots,y_k) := d_Z(\ell_X(x_k),\ell_Y(y_k)),\text{ and}\\
&V_{i-1}(x_0,\dots,x_{i-1},y_0,\dots,y_{i-1}) := \inf_{\nu^i\in\mathcal{C}(\alpha^k_{x_0,\dots,x_{i-1}},\beta^k_{y_0,\dots,y_{i-1}})}\mathbb{E}_{\nu^i}\,V_i(\boldsymbol{X}_0,\dots,\boldsymbol{X}_{i-1},\boldsymbol{X}_i,\boldsymbol{Y}_0,\dots,\boldsymbol{Y}_{i-1},\boldsymbol{Y}_i)
\end{aligned}$$

where $\alpha^k_{x_0,\dots,x_{i-1}}$ (resp. $\beta^k_{y_0,\dots,y_{i-1}}$) is the probability measure on X (resp. Y) defined as the conditional probability measure $\alpha^k_{x_0,\dots,x_{i-1}}(x_i) := \alpha^k(x_i|x_0,\dots,x_{i-1})$ (resp. $\beta^k_{y_0,\dots,y_{i-1}}(y_i) := \beta^k(y_i|y_0,\dots,y_{i-1})$)[8].

8. It might happen that the event $(\boldsymbol{X}_0,\dots,\boldsymbol{X}_{i-1}) = (x_0,\dots,x_{i-1})$ is null, then the conditional probability is not defined. In this case, we simply let $\alpha^k_{x_0,\dots,x_{i-1}}$ be any probability measure on X. We adopt the

Then, as a direct consequence of Proposition 5.2 in (Backhoff et al., 2017), one has the following equality

$$d^{c^k}(\alpha^k, \beta^k) = \inf_{\gamma^0 \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} V_0(x_0, y_0)\, \gamma^0(dx_0 \times dy_0). \tag{33}$$

It turns out that $V_i(x_0, \ldots, x_i, y_0, \ldots, y_i)$ is independent of $(x_0, \ldots, x_{i-1}, y_0, \ldots, y_{i-1})$. In order to show that this is indeed the case, for each $i = 0, \ldots, k$, we define $W_i : X \times Y \to \mathbb{R}$ in a way similar to how we defined V_i:

$$\begin{aligned} &W_k(x_k, y_k) := d_Z(\ell_X(x_k), \ell_Y(y_k)), \text{ and} \\ &W_{i-1}(x_{i-1}, y_{i-1}) := \inf_{\nu^i \in \mathcal{C}(\alpha^k_{x_{i-1}}, \beta^k_{y_{i-1}})} \mathbb{E}_{\nu^i} W_i(\boldsymbol{X}_i, \boldsymbol{Y}_i) \end{aligned}$$

for each $i = 0, \ldots, k-1$ where $\alpha^k_{x_{i-1}}$ (resp. $\beta^k_{y_{i-1}}$) is the probability measure on X (resp. Y) defined as the conditional probability measure $\alpha^k_{x_{i-1}}(x_i) := \alpha^k(x_i | x_{i-1})$ (resp. $\beta^k_{y_{i-1}}(y_i) := \beta^k(y_i | y_{i-1})$).

Lemma 90 *For all $i \in \{0, \ldots, k-1, k\}$ and $(x_0, \ldots, x_i, y_0, \ldots, y_i) \in X^{i+1} \times Y^{i+1}$, we have the following equality:*

$$V_i(x_0, \ldots, x_i, y_0, \ldots, y_i) = W_i(x_i, y_i).$$

Proof The proof is by (backward) induction. First, observe that $\alpha^k_{x_0, \ldots, x_{i-1}} = \alpha^k_{x_{i-1}} = m^X_{x_{i-1}}$ and $\beta^k_{y_0, \ldots, y_{i-1}} = \beta^k_{y_{i-1}} = m^Y_{y_{i-1}}$ since both of α^k and β^k are Markovian. The remaining steps are straightforward, so we omit it. ∎

Now we are ready to prove Theorem 49.

Proof [Proof of Theorem 49] First, fix an arbitrary Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{X}, \mathcal{Y})$. Then, by Lemma 48, we know that $\pi^k := \text{law}((\boldsymbol{X}_0, \ldots, \boldsymbol{X}_k, \boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_k)) \in \mathcal{P}(X^{k+1} \times Y^{k+1})$ is a bicausal coupling between α^k and β^k. Therefore,

$$\mathbb{E}\, d_Z(\ell_X(\boldsymbol{X}_k), \ell_Y(\boldsymbol{Y}_k)) = \int_{X^{k+1} \times Y^{k+1}} c^k(x, y)\, \pi^k(dx \times dy) \geq d^{c^k}(\alpha^k, \beta^k).$$

Since the Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}}$ is chosen arbitrarily, by Theorem 46, we have $d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \geq d^{c^k}(\alpha^k, \beta^k)$.

Now, let us prove the other direction. By Equation (33) and Lemma 90, one can choose an optimal coupling $\gamma^0 \in \mathcal{C}(\mu_X, \mu_Y)$ such that

$$d^{c^k}(\alpha^k, \beta^k) = \int_{X \times Y} W_0(x_0, y_0)\, \gamma^0(dx_0 \times dy_0).$$

Next, for arbitrary $(x_0, y_0) \in X \times Y$, one can choose an optimal coupling $(\nu_1)_{x_0, y_0} \in \mathcal{C}(m^X_{x_0}, m^Y_{y_0})$ such that

$$W_0(x_0, y_0) = \int_{X \times Y} W_1(x_1, y_1)\, (\nu_1)_{x_0, y_0}(dx_1 \times dy_1).$$

same convention for β^k. An alternative and more rigorous way of dealing with this is to define $\alpha^k_{x_0, \ldots, x_{i-1}}$ as the disintegration of α^k w.r.t. $(x_0, \ldots, x_{i-1})$ (cf. (Ambrosio et al., 2005, Section 5.3)). However, for simplicity of presentation, we avoid such a definition.

We repeat this process inductively to obtain the following maps $(\nu_1)_{\bullet,\bullet}, \ldots, (\nu_k)_{\bullet,\bullet}$ such that $(\nu_i)_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)$, for all $(x,y) \in X \times Y$ and for $i = 1, \ldots, k$. Furthermore,

$$W_{i-1}(x_{i-1}, y_{i-1}) = \int_{X\times Y} W_i(x_i, y_i)\,(\nu_i)_{x_{i-1},y_{i-1}}(dx_i \times dy_i)$$

for all $(x_{i-1}, y_{i-1}) \in X \times Y$. Now, the following equality holds:

$$d^{c^k}(\alpha^k, \beta^k) = \int_{X\times Y} \cdots \int_{X\times Y} d_Z(\ell_X(x_k), \ell_Y(y_k))\,(\nu_k)_{x_{k-1},y_{k-1}}(dx_k \times dy_k) \cdots (\nu_1)_{x_0,y_0}(dx_1 \times dy_1)\gamma^0(dx_0 \times dy_0).$$

Also, recall that there exists a Markovian coupling induced by γ^0 and these maps $(\nu_1)_{\bullet,\bullet}, \cdots, (\nu_k)_{\bullet,\bullet}$ by employing Kolmogorov extension theorem (cf. Remark 45). Therefore, again by Theorem 46, we have $d^{c^k}(\alpha^k, \beta^k) \geq d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$. This completes the proof. ■

B.5 Proofs from Section 6

B.5.1 Proof of Proposition 53

The proof is rather lengthy, and we start with some preliminary definitions and lemmas.

Lemma 91 *For any $k \geq 1$ and any MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$, consider any pair consisting of a Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ between $\mathcal{X}$ and $\mathcal{Y}$. Let γ denote the law of $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ and γ_k denote the law of $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$. Let $(\nu_{i+1})_{\bullet,\bullet}$ denote the transition kernel of the Markovian coupling for $i = 0, \ldots$. Then, we have that*

$$\gamma_k(dx_k \times dy_k) = \int (\nu_k)_{x_{k-1},y_{k-1}}(dx_k \times dy_k) \cdots \int (\nu_1)_{x_0,y_0}(dx_1 \times dy_1)\gamma(dx_0 \times dy_0).$$

Lemma 92 *Let $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ be two finite MCMCs. Then, for each $k \geq 1$, there exists an optimal Markovian coupling such that the infimum in Equation (12) is achieved.*

Proof Consider any Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ between $\mathcal{X}$ and $\mathcal{Y}$ and an independent coupling (X_0', Y_0') between μ_X and μ_Y. Let γ_k denote the law of $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$ and let γ denote the law of (X_0', Y_0'). Then, we have that

$$\mathrm{dis}^{(k)}\left(\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}, (X_0', Y_0')\right) = \int |d_X(x_0', x_k) - d_Y(y_0', y_k)|\gamma'(dx_0' \times dy_0')\gamma_k(dx_k \times dy_k).$$

By Lemma 91, the left hand side is simply a continuous function of $(\gamma', \gamma, (\nu_{i+1})_{i=0,\ldots})$. Notice that the following set is compact:

$$\mathcal{C}(\mu_X, \mu_Y) \times (\Pi_{x\in X, y\in Y}\mathcal{C}(m_x^X, m_y^Y))^{\mathbb{N}}$$

Therefore, the infimum can be achieved and hence the optimal Markovian coupling exists. ■

Now we start to prove Proposition 53. First of all, $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ is obviously symmetric. Next, we prove that $d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = 0$ happens if and only if $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ are isomorphic.

When are $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ are isomorphic, without loss of generality, we simply assume that $(\mathcal{Y}, d_Y) = (\mathcal{X}, d_X)$. In this case, consider the (diagonal) Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{X}_i)\}_{i\in\mathbb{N}}$. Then, for each $k \in \mathbb{N}$

$$\begin{aligned} d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) &\leq \mathrm{dis}^{(k)}\big(\{(\boldsymbol{X}_i, \boldsymbol{X}_i)\}_{i\in\mathbb{N}}, (\boldsymbol{X}_0', \boldsymbol{X}_0')\big) \\ &= \mathbb{E}|d_X(\boldsymbol{X}_0', \boldsymbol{X}_k) - d_X(\boldsymbol{X}_0', \boldsymbol{X}_k)| = 0. \end{aligned}$$

Hence, $d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = 0$,

Now, we assume that $d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = 0$ for some MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$. Then, $d_{\mathrm{GW}}^{(1)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = 0$. By Lemma 92, there exist optimal $(\nu_1)_{x,y} \in \mathcal{C}\big(m_x^X, m_y^Y\big)$ for all $x \in X$ and $y \in Y$ and $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ such that

$$\int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x, x') - d_Y(y, y')|(\nu_1)_{x'',y''}(dx' \times dy')\gamma(dx'' \times dy'')\gamma(dx \times dy) = 0. \quad (34)$$

We define the average of ν under γ as the probability measure $\gamma' \in \mathcal{P}(X \times Y)$ such that for any measurable $A \subseteq X \times Y$,

$$\gamma'(A) := \int_{X\times Y} \nu_{x,y}(A)\, \gamma(dx \times dy).$$

Note that $\gamma' \in \mathcal{C}(\mu_X, \mu_Y)$. Since X and Y are finite, we rewrite Equation (34) as sums:

$$\sum_{(x,y)\in X\times Y} \sum_{(x',y')\in X\times Y} |d_X(x, x') - d_Y(y, y')|\, \gamma'(x', y')\, \gamma(x, y) = 0.$$

By Claim 94, there exists an isometry $\phi : X \to Y$ such that

$$\{(x, \phi(x)) : x \in X\} = \mathrm{supp}(\gamma) = \mathrm{supp}(\gamma').$$

Since $\gamma, \gamma' \in \mathcal{C}(\mu_X, \mu_Y)$, this immediately implies that for any Borel subset $A \subseteq X$, one has

$$\mu_X(A) = \gamma(A \times Y) = \gamma(A \times \phi(A)) = \gamma(X \times \phi(A)) = \mu_Y(\phi(A)).$$

Hence, $\phi_\# \mu_X = \mu_Y$. Moreover, we have that

$$\gamma = \gamma' = \sum_{x\in X} \mu_X(x)\delta_{(x,\phi(x))}.$$

Then, by the definition of γ', we have that

$$\begin{aligned} \sum_{x\in X} \mu_X(x)\delta_{(x,\phi(x))} = \gamma' &= \sum_{x_1\in X, y_1\in Y} \gamma(x_1, y_1)\nu_{x_1,y_1} = \sum_{x\in X, y\in Y} \left(\sum_{x_1\in X, y_1\in Y} \gamma(x_1, y_1)\nu_{x_1,y_1}(x, y) \right) \delta_{(x,y)} \\ &= \sum_{x\in X, y\in Y} \left(\sum_{x_1\in X} \gamma(x_1, \phi(x_1))\nu_{x_1,\phi(x_1)}(x, y) \right) \delta_{(x,y)} \\ &= \sum_{x\in X, y\in Y} \left(\sum_{x_1\in X} \mu_X(x_1)\nu_{x_1,\phi(x_1)}(x, y) \right) \delta_{(x,y)}. \end{aligned}$$

By comparing coefficients for the Dirac delta measures above, one has that

$$\nu_{x_1,\phi(x_1)}(x,y) = \begin{cases} 0, & \text{if } y \neq \phi(x) \\ \nu_{x_1,\phi(x_1)}(x,y), & \text{if } y = \phi(x) \end{cases}$$

for any $x_1 \in X$. This means that $\{(x,\phi(x)) : x \in X\} \supseteq \mathrm{supp}[\nu_{x_1,\phi(x_1)}]$. Since $\nu_{x_1,\phi(x_1)} \in \mathcal{C}\left(m^X_{x_1}, m^Y_{\phi(x_1)}\right)$, for any Borel subset $A \subseteq X$, one has that

$$m^X_{x_1}(A) = \nu_{x_1,\phi(x_1)}(A \times Y) = \nu_{x_1,\phi(x_1)}(A \times \phi(A)) = \nu_{x_1,\phi(x_1)}(X \times \phi(A)) = m^Y_{\phi(x_1)}(\phi(A)).$$

By an argument similar to the one for proving $\phi_\# \mu_X = \mu_Y$, we have that

$$\phi_\# m^X_{x_1} = m^Y_{\phi(x_1)}, \ \forall x_1 \in X.$$

Therefore, $(\mathcal{X}, d_X)$ is isomorphic to $(\mathcal{Y}, d_Y)$.

Finally, we prove that $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ satisfies the triangle inequality. It suffices to prove that for each $k \in \mathbb{N}$, $d_{\mathrm{GW}}^{(k)}$ satisfies the triangle inequality. Fix arbitrary three MCMSs $(\mathcal{X}, d_X)$, $(\mathcal{Y}, d_Y)$, and $(\mathcal{Z}, d_Z)$. Then, we have the following gluing lemma for Markovian couplings.

Lemma 93 (Gluing lemma for Markovian couplings (Brugère et al., 2024, Lemma 31)) *Let $(\boldsymbol{X}_t, \boldsymbol{Z}^1_t)_{t\in\mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{X}, \mathcal{Z})$ and $(\boldsymbol{Z}^2_t, \boldsymbol{Y}_t)_{t\in\mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{Z}, \mathcal{Y})$. Then, there exists a (time inhomogeneous) Markov chain $(\boldsymbol{X}'_t, \boldsymbol{Z}'_t, \boldsymbol{Y}'_t)_{t\in\mathbb{N}}$ so that $(\boldsymbol{X}'_t, \boldsymbol{Z}'_t)_{t\in\mathbb{N}} \sim (\boldsymbol{X}_t, \boldsymbol{Z}^1_t)_{t\in\mathbb{N}}$, $(\boldsymbol{Z}'_t, \boldsymbol{Y}'_t)_{t\in\mathbb{N}} \sim (\boldsymbol{Z}^2_t, \boldsymbol{Y}_t)_{t\in\mathbb{N}}$ and furthermore $(\boldsymbol{X}'_t, \boldsymbol{Y}'_t)_{t\in\mathbb{N}}$ is a Markovian coupling between $\mathcal{X}$ and $\mathcal{Y}$.*

We let $\Gamma_{X,Y}$ denote the function sending $(x,y,x',y') \in X \times Y \times X \times Y$ to $|d_X(x,x') - d_Y(y,y')|$. Then, for any $x, x' \in X$, $y, y' \in Y$, and $z, z' \in Z$, we obviously have that

$$\Gamma_{X,Y}(x,y,x',y') \leq \Gamma_{X,Z}(x,z,x',z') + \Gamma_{Z,Y}(z,y,z',y').$$

Now, consider any Markovian couplings $\{(\boldsymbol{X}_t, \boldsymbol{Z}^1_t)\}_{t\in\mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{X}, \mathcal{Z})$ and $\{(\boldsymbol{Z}^2_t, \boldsymbol{Y}_t)\}_{t\in\mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{Z}, \mathcal{Y})$. By the Gluing lemma above, there exists a Markov chain $(\boldsymbol{X}'_t, \boldsymbol{Z}'_t, \boldsymbol{Y}'_t)_{t\in\mathbb{N}}$ so that $(\boldsymbol{X}'_t, \boldsymbol{Z}'_t)_{t\in\mathbb{N}} \sim (\boldsymbol{X}_t, \boldsymbol{Z}^1_t)_{t\in\mathbb{N}}$, $(\boldsymbol{Z}'_t, \boldsymbol{Y}'_t)_{t\in\mathbb{N}} \sim (\boldsymbol{Z}^2_t, \boldsymbol{Y}_t)_{t\in\mathbb{N}}$ and that $(\boldsymbol{X}'_t, \boldsymbol{Y}'_t)_{t\in\mathbb{N}} \in \mathcal{C}_{\mathcal{M}}(\mathcal{X}, \mathcal{Y})$. Similarly, by the gluing lemma for usual couplings, for any couplings $(\boldsymbol{X}'_0, \boldsymbol{Z}^{1\prime}_0)$ and $(\boldsymbol{Z}^{2\prime}_0, \boldsymbol{Y}'_0)$ there exists a random variable $(\boldsymbol{X}'_0, \boldsymbol{Z}'_0, \boldsymbol{Y}'_0)$ whose corresponding marginals coincide with the two couplings and that $(\boldsymbol{X}'_0, \boldsymbol{Y}'_0)$ is a coupling between μ_X and μ_Y. Therefore,

$$\begin{aligned} d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) &\leq \mathbb{E}\Gamma_{X,Y}(\boldsymbol{X}'_0, \boldsymbol{Y}'_0, \boldsymbol{X}_k, \boldsymbol{Y}_k) \\ &\leq \mathbb{E}\Gamma_{X,Z}(\boldsymbol{X}'_0, \boldsymbol{Z}'_0, \boldsymbol{X}'_k, \boldsymbol{Z}'_k) + \mathbb{E}\Gamma_{X,Y}(\boldsymbol{Z}'_0, \boldsymbol{Y}'_0, \boldsymbol{Z}'_k, \boldsymbol{Y}'_k) \\ &= \mathbb{E}\Gamma_{X,Z}(\boldsymbol{X}'_0, \boldsymbol{Z}^{1\prime}_0, \boldsymbol{X}_k, \boldsymbol{Z}^1_k) + \mathbb{E}\Gamma_{X,Y}(\boldsymbol{Z}^{2\prime}_0, \boldsymbol{Y}'_0, \boldsymbol{Z}^2_k, \boldsymbol{Y}_k) \end{aligned}$$

Since the choice of couplings are arbitrary, by taking the infimum one concludes that

$$d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \leq d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Z}, d_Z)) + d_{\mathrm{GW}}^{(k)}((\mathcal{Z}, d_Z), (\mathcal{Y}, d_Y))$$

as we required. Then, we have that $d_{\mathrm{GW}}^{\mathrm{MCMS}} \coloneqq \sup_{k\geq 0} d_{\mathrm{GW}}^{(k)}$ satisfies the triangle inequality.

B.5.2 Proof of Proposition 58

To prove this result, we consider the equivalent measure theoretic formulation as follows:

$$d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y}) \coloneqq \inf_{\gamma,\gamma'\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y}\int_{X\times Y} |d_X(x,x')-d_Y(y,y')|\gamma'(dx'\times dy')\gamma(dx\times dy).$$

Symmetry is obvious. We need to prove the triangle inequality plus the fact that $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y})=0$ happens if and only if $\mathbf{X}$ and $\mathbf{Y}$ are isomorphic. The "if" part is trivial. For the other direction we proceed as follows. Suppose that $d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y})=0$. By the compactness of $\mathcal{C}(\mu_X,\mu_Y)$ with respect to the weak topology (see Villani (2003, p.49)), there must be optimal couplings $\gamma,\gamma'\in\mathcal{C}(\mu_X,\mu_Y)$ such that

$$\sum_{(x,y)\in X\times Y}\sum_{(x',y')\in X\times Y} |d_X(x,x')-d_Y(y,y')|\,\gamma'(x',y')\,\gamma(x,y)=0. \tag{35}$$

Claim 94 *There exists an isometry $\phi: X\to Y$ such that*

$$\{(x,\phi(x)) : x\in X\} = \mathrm{supp}(\gamma) = \mathrm{supp}(\gamma').$$

Proof [Proof of Claim 94] By Equation (35), we have that

$$d_X(x,x') = d_Y(y,y') \tag{36}$$

for any $(x,y)\in\mathrm{supp}(\gamma)$ and $(x',y')\in\mathrm{supp}(\gamma')$.

Fix an arbitrary $x\in X$. Then, since both μ_X and μ_Y are fully supported, there must exist $y,y'\in Y$ such that $(x,y)\in\mathrm{supp}(\gamma)$ and $(x,y')\in\mathrm{supp}(\gamma')$. Then, $y=y'$ by Equation (36). Now, if there exists $y''\in Y$ such that $(x,y'')\in\mathrm{supp}(\gamma)$, then similarly, we have that $y''=y'$ and thus $y''=y$. In other words, for each $x\in X$, there exists a unique $y\in Y$ such that $(x,y)\in\mathrm{supp}(\gamma)$. Similarly, this same $y\in Y$ is unique such that $(x,y)\in\mathrm{supp}(\gamma')$. Hence, we define $\phi: X\to Y$ by letting $\phi(x)$ be the unique $y\in Y$ such that $(x,\phi(x))\in\mathrm{supp}(\gamma)$. It is obvious that ϕ is bijective and satisfies that $\{(x,\phi(x)) : x\in X\} = \mathrm{supp}(\gamma) = \mathrm{supp}(\gamma')$. By Equation (36), we conclude that ϕ is an isometry. ■

Based on the claim above, consider an arbitrary Borel subset $A\subseteq X$. Then,

$$\mu_X(A) = \gamma(A\times Y) = \gamma(A\times Y\cap A\times\phi(A)) = \gamma(A\times\phi(A)) = \mu_Y(\phi(A)).$$

Hence, ϕ is a isomorphism between $\mathbf{X}$ and $\mathbf{Y}$.

Finally, for the triangle inequality, fix finite metric measure space $\mathbf{X}$,$\mathbf{Y}$, and $\mathbf{Z}$. Notice first that for all $x,x'\in X$, $y,y'\in Y$, and $z,z'\in Z$,

$$\Gamma_{X,Y}(x,y,x',y') \le \Gamma_{X,Z}(z,x,z',x') + \Gamma_{Z,Y}(y,z,y',z').$$

Next, fix arbitrary coupling measures $\gamma_1,\gamma_1'\in\mathcal{C}(\mu_X,\mu_Z)$ and $\gamma_2,\gamma_2'\in\mathcal{C}(\mu_Z,\mu_Y)$. By the Gluing Lemma (see Villani (2003, Lemma 7.6)), there exist probability measures $\pi,\pi'\in$

$\mathcal{P}(X \times Y \times Z)$ with marginals γ_1, γ_1' on $X \times Z$ and γ_2, γ_2' on $Z \times Y$. Let γ_3, γ_3' be the marginal of π, π' on $X \times Y$. Then, by the triangle inequality of L^1 norm,

$$\begin{aligned} d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) &\leq \|\Gamma_{X,Y}\|_{L^1(\gamma_3 \otimes \gamma_3')} \\ &= \|\Gamma_{X,Y}\|_{L^1(\pi \otimes \pi')} \\ &\leq \|\Gamma_{X,Z}\|_{L^1(\pi \otimes \pi')} + \|\Gamma_{Z,Y}\|_{L^1(\pi \otimes \pi')} \\ &= \|\Gamma_{X,Z}\|_{L^1(\gamma_1 \otimes \gamma_1')} + \|\Gamma_{Z,Y}\|_{L^1(\gamma_2 \otimes \gamma_2')}. \end{aligned}$$

Since $\gamma_1, \gamma_1', \gamma_2, \gamma_2'$ are arbitrary, by taking the infimum one can conclude

$$d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) \leq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Z}) + d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{Z}, \mathbf{Y}).$$

B.5.3 Proof of Proposition 59

Fix $k \geq 1$ and any MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$. Consider any Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}}$ between $\mathcal{X}$ and $\mathcal{Y}$, and a coupling $(\boldsymbol{X}_0', \boldsymbol{Y}_0')$ between μ_X and μ_Y that is independent of this Markovian coupling and has the same law as $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$, as in Definition 51. By stationarity, $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$ also has marginals μ_X and μ_Y. Thus, $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$ and $(\boldsymbol{X}_0', \boldsymbol{Y}_0')$ are two independent couplings between μ_X and μ_Y. Hence, by the definition of the decoupled GW distance,

$$\mathbb{E}\big|d_X(\boldsymbol{X}_0', \boldsymbol{X}_k) - d_Y(\boldsymbol{Y}_0', \boldsymbol{Y}_k)\big| \geq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}).$$

Taking the infimum over all such choices in Equation (12) proves the result.

B.5.4 Proof of Proposition 60

Fix arbitrary independent couplings $(\boldsymbol{X}, \boldsymbol{Y}), (\boldsymbol{X}', \boldsymbol{Y}') \in \mathcal{C}(\mu_X, \mu_Y)$. Because of the definitions of $\mathcal{M}(\mathbf{X})$ and $\mathcal{M}(\mathbf{Y})$, one can establish a Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}}$ and an independent coupling $(\boldsymbol{X}_0', \boldsymbol{Y}_0')$ satisfying that

1. $\mathrm{law}(\boldsymbol{X}_0, \boldsymbol{Y}_0) = \mathrm{law}(\boldsymbol{X}_0', \boldsymbol{Y}_0') = \mathrm{law}(\boldsymbol{X}', \boldsymbol{Y}')$;
2. $\mathrm{law}(\boldsymbol{X}_{k+1}, \boldsymbol{Y}_{k+1} | \boldsymbol{X}_k, \boldsymbol{Y}_k) = \mathrm{law}(\boldsymbol{X}, \boldsymbol{Y})$.

Also, it is easy to verify that

$$\mathrm{law}(\boldsymbol{X}_k, \boldsymbol{Y}_k) = \mathrm{law}(\boldsymbol{X}, \boldsymbol{Y}) \quad \text{for all } k = 1, 2, \ldots. \tag{37}$$

Hence,

$$\begin{aligned} d_{\mathrm{GW}}^{(k)}(\mathcal{M}(\mathbf{X}), \mathcal{M}(\mathbf{Y})) &\leq \mathbb{E}|d_X(\boldsymbol{X}_0', \boldsymbol{X}_k) - d_Y(\boldsymbol{Y}_0', \boldsymbol{Y}_k)| \\ &= \mathbb{E}|d_X(\boldsymbol{X}', \boldsymbol{X}) - d_Y(\boldsymbol{Y}', \boldsymbol{Y})|. \end{aligned}$$

Since $(\boldsymbol{X}, \boldsymbol{Y}), (\boldsymbol{X}', \boldsymbol{Y}')$ are arbitrary, one concludes that

$$d_{\mathrm{GW}}^{(k)}(\mathcal{M}(\mathbf{X}), \mathcal{M}(\mathbf{Y})) \leq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}).$$

For the reverse direction, fix an arbitrary Markovian coupling $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i \in \mathbb{N}}$ and an independent coupling $(\boldsymbol{X}_0', \boldsymbol{Y}_0') \in \mathcal{C}(\mu_X, \mu_Y)$. Let $(\boldsymbol{X}', \boldsymbol{Y}') := (\boldsymbol{X}_0', \boldsymbol{Y}_0')$ and $(\boldsymbol{X}, \boldsymbol{Y}) := (\boldsymbol{X}_k, \boldsymbol{Y}_k)$. Then,

$$\mathbb{E}|d_X(\boldsymbol{X}_0', \boldsymbol{X}_k) - d_Y(\boldsymbol{Y}_0', \boldsymbol{Y}_k)| = \mathbb{E}|d_X(\boldsymbol{X}, \boldsymbol{X}') - d_Y(\boldsymbol{Y}, \boldsymbol{Y}')| \geq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}).$$

Since the choices of $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i\in\mathbb{N}}$ and $(\boldsymbol{X}_0', \boldsymbol{Y}_0')$ are arbitrary, one concludes that

$$d_{\mathrm{GW}}^{(k)}(\mathcal{M}(\mathbf{X}), \mathcal{M}(\mathbf{Y})) \geq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}).$$

Hence, $d_{\mathrm{GW}}^{(k)}(\mathcal{M}(\mathbf{X}), \mathcal{M}(\mathbf{Y})) = d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y})$ as we required.

B.5.5 Proof of Proposition 62

For any $(\boldsymbol{X}_0, \boldsymbol{Y}_0) \in \mathcal{C}(\mu_X, \mu_Y)$, let $\{(\boldsymbol{X}_i', \boldsymbol{Y}_i')\}_{i\in\mathbb{N}}$ denote an independent Markovian coupling between $\mathcal{X}$ and $\mathcal{Y}$. Then,

$$\mathbb{E} d_{\mathrm{WL}}^{(k)}\big((\mathcal{X}, d_X(\boldsymbol{X}_0, \bullet)), (\mathcal{Y}, d_Y(\boldsymbol{Y}_0, \bullet))\big) \leq \mathbb{E}\big|d_X(\boldsymbol{X}_0, \boldsymbol{X}_k') - d_Y(\boldsymbol{Y}_0, \boldsymbol{Y}_k')\big|.$$

One concludes the proof by infimizing both sides.

B.5.6 Proof of the statement in Remark 63

We first recall the third lower bound (TLB) from (Mémoli, 2011a):

$$\mathrm{TLB}(\mathbf{X}, \mathbf{Y}) \coloneqq \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} \left(\inf_{\gamma'\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} |d_X(x,x') - d_Y(y,y')| \gamma'(dx' \times dy') \right) \gamma(dx \times dy).$$

where we omit the $\frac{1}{2}$ factor from (Mémoli, 2011a) for simplicity of presentation.

For simplicity, we let $\ell_X^x \coloneqq d_X(x, \bullet)$. Notice that

$$\inf_{\gamma'\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} |d_X(x,x') - d_Y(y,y')| \gamma'(dx' \times dy') = d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)).$$

We hence have that

$$\mathrm{TLB}(\mathbf{X}, \mathbf{Y}) = \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) \gamma(dx \times dy).$$

For any $k \in \mathbb{N}$, we show the following equation using the lemma below.

$$\mathrm{TLB}(\mathbf{X}, \mathbf{Y}) = \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) \gamma(dx \times dy).$$

Lemma 95 *For any $k \in \mathbb{N}$ we have that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) = d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y))$.*

Proof By Equation (37), we have that

$$\begin{aligned}
&d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) = \inf_{\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\}_{i\in\mathbb{N}}} \mathbb{E}|d_X(x, \boldsymbol{X}_k) - d_Y(y, \boldsymbol{Y}_k)| \\
&\leq \inf_{(\boldsymbol{X}_k,\boldsymbol{Y}_k)\in\mathcal{C}(\mu_X,\mu_Y)} \mathbb{E}|d_X(x, \boldsymbol{X}_k) - d_Y(y, \boldsymbol{Y}_k)| \\
&= d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)).
\end{aligned}$$

By Proposition 15, we conclude that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) = d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y))$. ■

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Shreya Arya, Arnab Auddy, Ranthony A Clark, Sunhyuk Lim, Facundo Memoli, and Daniel Packer. The Gromov–Wasserstein distance between spheres. *Foundations of Computational Mathematics*, pages 1–56, 2024.

Waiss Azizian and Marc Lelarge. Expressive power of invariant and equivariant graph neural networks. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=lxHgXYN4bwl`.

László Babai and Ludik Kucera. Canonical labelling of graphs in linear average time. In *20th Annual Symposium on Foundations of Computer Science (sfcs 1979)*, pages 39–46. IEEE, 1979.

László Babai and Eugene M Luks. Canonical labeling of graphs. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 171–183, 1983.

Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Manu Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24(3):601–632, 2020.

Martin Bauer, Facundo Mémoli, Tom Needham, and Mao Nishino. The Z-Gromov-Wasserstein distance. *Journal of Machine Learning Research*, 26(291):1–57, 2025.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Vladimir Igorevich Bogachev. *Weak convergence of measures*. American Mathematical Society Providence, 2018.

Jan Böker, Ron Levie, Ningyuan Huang, Soledad Villar, and Christopher Morris. Fine-grained expressivity of graph neural networks. *Advances in neural information processing systems*, 36:46658–46700, 2023.

Tristan Brugère, Zhengchao Wan, and Yusu Wang. Distances for markov chains, and their differentiation. In *International Conference on Algorithmic Learning Theory*, pages 282–336. PMLR, 2024.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

Samantha Chen, Sunhyuk Lim, Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Weisfeiler-lehman meets gromov-wasserstein. In *International Conference on Machine Learning*, pages 3371–3416. PMLR, 2022.

Samantha Chen, Sunhyuk Lim, Facundo Mémoli, Zhengchao Wan, and Yusu Wang. The Weisfeiler-Lehman distance: Reinterpretation and connection with GNNs. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pages 404–425. PMLR, 2023. URL `https://proceedings.mlr.press/v221/chen23a.html`.

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, 2019.

Ching-Yao Chuang and Stefanie Jegelka. Tree mover's distance: Bridging graph metrics and stability of graph neural networks. *Advances in Neural Information Processing Systems*, 35:2944–2957, 2022.

Giuseppe Da Prato and Jerzy Zabczyk. *Ergodicity for infinite dimensional systems*, volume 229. Cambridge university press, 1996.

Nicolo De Ponti and Andrea Mondino. Entropy-transport distances between unbalanced metric measure spaces. *Probability Theory and Related Fields*, 184(1):159–208, 2022.

Frank Den Hollander. Probability theory: The coupling method. *Lecture notes available online (http://websites. math. leidenuniv. nl/probability/lecturenotes/CouplingLectures.pdf)*, 2012.

Stephan Eckstein and Gudmund Pammer. Computational methods for adapted optimal transport. *arXiv preprint arXiv:2203.05005*, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Martin Glanzer, Georg Ch Pflug, and Alois Pichler. Incorporating statistical model error into the calculation of acceptability prices of contingent claims. *Mathematical Programming*, 174(1):499–524, 2019.

Konstantin Klemmer, Tianlin Xu, Beatrice Acciaio, and Daniel B Neill. SPATE-GAN: Improved generative modeling of dynamic spatio-temporal patterns with an autoregressive embedding loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4523–4531, 2022.

Andreĭ Nikolaevich Kolmogorov and Albert T Bharucha-Reid. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

Andreas Krebs and Oleg Verbitsky. Universal covers, color refinement, and two-variable counting logic: Lower bounds for the depth. In *2015 30th Annual ACM/IEEE Symposium on Logic in Computer Science*, pages 689–700. IEEE, 2015.

Nils M Kriege, Pierre-Louis Giscard, and Richard Wilson. On valid optimal assignment kernels and applications to graph classification. *Advances in Neural Information Processing Systems*, 29:1623–1631, 2016.

Rémi Lassalle. Causal transport plans and their Monge–Kantorovich problems. *Stochastic Analysis and Applications*, 36(3):452–484, 2018.

AA Lehman and B Weisfeiler. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya*, 2(9):12–16, 1968.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Gaëlle Loosli, Stéphane Canu, and Cheng Soon Ong. Learning SVM in kreĭn spaces. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1204–1216, 2015.

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. ISBN 978-3-905673-51-7. doi: 10.2312/SPBG/SPBG07/081-090.

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.

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

Facundo Mémoli. A spectral notion of Gromov-Wasserstein distance and related methods. *Applied and Computational Harmonic Analysis*, 30(3):363–401, 2011b.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL `www.graphlearning.io`.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

Alois Pichler and Michael Weinhardt. The nested sinkhorn divergence to learn the nested distance. *Computational Management Science*, 19(2):269–293, 2022.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8:143–195, 1999.

Laurent Saloff-Coste. Lectures on finite Markov chains. In *Lectures on probability theory and statistics*, pages 301–413. Springer, 1997.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80, 2008.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov-Wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

Bernhard Schmitzer and Christoph Schnörr. Modelling convex shape priors and matching based on the Gromov-Wasserstein distance. *Journal of mathematical imaging and vision*, 46(1):143–159, 2013.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced Gromov-Wasserstein distance: Conic formulation and relaxation. In *35th Conference on Neural Information Processing Systems*, 2021.

Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Karl-Theodor Sturm. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*, volume 290. American Mathematical Society, 2023.

Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. Wasserstein Weisfeiler-Lehman graph kernels. *Advances in Neural Information Processing Systems*, 32:6439–6449, 2019.

SS Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

Antonis Vasileiou, Ben Finkelshtein, Floris Geerts, Ron Levie, and Christopher Morris. Covered Forest: Fine-grained generalization analysis of graph neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 60984–61034. PMLR, 2025. URL `https://proceedings.mlr.press/v267/vasileiou25a.html`.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020a.

Titouan Vayer, Ievgen Redko, Rémi Flamary, and Nicolas Courty. Co-optimal transport. *Advances in Neural Information Processing Systems*, 33, 2020b.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Max-K von Renesse and Karl-Theodor Sturm. Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on pure and applied mathematics*, 58(7): 923–940, 2005.

Nik Weaver. Order completeness in Lipschitz algebras. *Journal of Functional Analysis*, 130 (1):118–130, 1995.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Tianlin Xu and Beatrice Acciaio. Conditional COT-GAN for video prediction with kernel smoothing. *Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems*, 2022.

Tianlin Xu, Li Kevin Wenliang, Michael Munn, and Beatrice Acciaio. COT-GAN: Generating sequential data via causal optimal transport. *Advances in Neural Information Processing Systems*, 33:8798–8809, 2020.