The Weisfeiler-Lehman Distance: Reinterpretation and Connection with GNNs

Samantha Chen [*1] Sunhyuk Lim [*2] Facundo Mémoli [*3] Zhengchao Wan [*4] Yusu Wang [*1 4]

Abstract

In this paper, we present a novel interpretation of the Weisfeiler-Lehman (WL) distance introduced by (Chen et al., 2022) using concepts from stochastic processes. The WL distance compares graphs with node features, has the same discriminative power as the classic Weisfeiler-Lehman graph isomorphism test and has deep connections to the Gromov-Wasserstein distance. Our interpretation connects the WL distance to the literature on distances for stochastic processes, which also makes the interpretation of the distance more accessible and intuitive. We further explore the connections between the WL distance and certain Message Passing Neural Networks, and discuss the implications of the WL distance for understanding the Lipschitz property and the universal approximation results for these networks.

1. Introduction

The Weisfeiler-Lehman (WL) test, a classic graph isomorphism test (Lehman & Weisfeiler, 1968) which has recently gained renewed interest as a tool for analyzing Message Passing Graph Neural Networks (MP-GNNs) (Xu et al., 2018; Azizian et al., 2020). Recently, Chen et al. (2022) introduced the Weisfeiler-Lehman (WL) distance between labeled measure Markov chains (LMMCs). The WL distance has the same power as the WL test in distinguishing non-isomorphic graphs and it is more discrimative than a certain WL based graph kernel (Togninalli et al., 2019). Moreover, Chen et al. (2022) unveiled interesting connections between the WL distance and both a certain neural network architecture on Markov chains and a variant of the Gromov-Wasserstein distance (Mémoli, 2011).

Although the WL distance possesses nice theoretical properties and good empirical performance in graph classification tasks, the original formulation of the WL distance is complicated and can be hard to decipher. In this work, we identify a novel characterization of the WL distance using concepts from stochastic processes. This new characterization eventually provides an alternative, more intuitive reformulation of the WL distance. Via this reformulation, we further identify connections between the WL distance and the so-called Causal Optimal Transport (COT), a branch of the Optimal Transport theory specifically taylored for comparing stochastic processes (Lassalle, 2018).

Finally, we recall that Chen et al. (2022) introduced a certain neural network structure, called Markov Chain Neural Networks (MCNNs), for Markov chains and utilized the WL distance to explain theoretical properties of MCNNs. It was mentioned that MCNNs will reduce to a special type of MP-GNNs when restricted to Markov chains induced by graphs. In this work, we proceed further along this line and explicitly clarify how the WL distance can be used to understand properties of MP-GNNs. In particular, inspired by the analysis of (Chen et al., 2022; Chuang & Jegelka, 2022), we establish the Lipschitz property and a universal approximation result for such MP-GNNs.

Related work. Our work builds on recent developments in graph similarity measures, particularly the WL distance introduced by (Chen et al., 2022). Another relevant work is the Tree Mover's (TM) distance, introduced in (Chuang & Jegelka, 2022), which compares labeled graphs using Wasserstein distance and has similar discriminative power to the WL distance. However, in contrast to the combinatorial computation tree structure used in the TM distance (which currently cannot yet handle weighted graphs), the WL distance benefits from a more flexible Markov chain formulation. This formulation enables the WL distance to compare weighted graphs, potentially handle continuous objects, such as heat kernels on Riemannian manifolds, and facilitates the development of differentiable distances for comparing labeled graphs, as demonstrated in (Brugère et al., 2023). Additionally, Toth et al. (2022) introduced the hypo-elliptic graph Laplacian and a corresponding diffu-

[*]Equal contribution [1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA [2]Max Planck Institute for Mathematics in the Sciences, Leipzig, Saxony, Germany [3]Department of Mathematics and Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA [4]Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, California, USA. Correspondence to: Zhengchao Wan <zcwan@ucsd.edu>.

Proceedings of the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023.

sion model that captures the evolution of random walks on graphs. This approach shares similarities with the WL test and the WL distance by focusing on entire walk trajectories instead of individual steps. Exploring the relationship between these concepts in future research would be intriguing.

2. Preliminaries

2.1. Probability measures and Optimal Transport

Let Z be any metric space. We let $\mathcal{P}(Z)$ denotes the space of all Borel probability measures on Z with finite 1-moment[1]. We let $\mathcal{P}^{\circ 1}(Z) := \mathcal{P}(Z)$ and for each $k = 1, \ldots$, we inductively let $\mathcal{P}^{\circ(k+1)}(Z) := \mathcal{P}(\mathcal{P}^{\circ k}(Z))$.

Given any measurable map $f : X \to Y$ and a probability measure $\alpha \in \mathcal{P}(X)$, we let $f_\# \alpha \in \mathcal{P}(Y)$ denote the *pushforward* of α, i.e., for any measurable $B \subseteq Y$, $f_\# \alpha(B) := \alpha(f^{-1}B)$.

For any probability measures α, β on Z, the (ℓ_1) Wasserstein distance between them is defined as follows:

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

Here $\mathcal{C}(\alpha, \beta)$ denotes the set of all couplings between α and β, i.e., the set of all probability measure $\pi \in \mathcal{P}(Z \times Z)$ such that marginals of π are α and β, respectively.

A note on notation for probability measures. We will mostly deal with finite sets in this paper. Given a finite set X and $\alpha \in \mathcal{P}(X)$, we let $\alpha(x) := \alpha(\{x\})$ for any $x \in X$.

An alternative description of the Wasserstein distance. A random variable X with values in a complete and separable metric space Z is any measurable map from a probability space $(\Omega, \mathbb{P})$ to Z. We let $\mathrm{law}(X) := X_\#\mathbb{P}$ denote the law of X. Now, we define the notion of coupling in terms of random variables.

Definition 2.1 (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 $(X, Y) : (\Omega, \mathbb{P}) \to Z \times Z$ a coupling between α and β if $\mathrm{law}(X) = \alpha$ and $\mathrm{law}(Y) = \beta$.

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

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:

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

where the infimum is taken over all couplings (X, Y) between α and β.

2.2. Markov chains

Let X be a finite set. We denote by $m_\bullet^X : X \to \mathcal{P}(X)$ a Markov transition kernel on X. Let $\mu_X \in \mathcal{P}(X)$ be a *stationary* distribution w.r.t. $m_\bullet^X$. Then, we call the tuple $\mathcal{X} := (X, m_\bullet^X, \mu_X)$ a measure Markov chain (MMC).

Due to the Kolmogorov extension theorem (Kolmogorov & Bharucha-Reid, 2018) (see also (Durrett, 2019, Theorem 2.1.14)), an equivalent way of describing a measure Markov chain $\mathcal{X} := (X, m_\bullet^X, \mu_X)$ is to view it as a probability measure $\mathbb{P}_X$ on the path space $X^{\mathbb{N}} = X \times X \times \cdots$, i.e., $X^{\mathbb{N}} = \{w = (x_i)_{i=0}^\infty : x_i \in X\}$[2]: If we let $X_i : X^{\mathbb{N}} \to X$ denote the projection map to the i-th component for $i \in \mathbb{N}$, then $\mathbb{P}_X$ is required to satisfy that

- $\mathbb{P}_X(X_{i+1}(w) = x' | X_i(w) = x) = m_x^X(x')$ for any $x, x' \in X$ and
- $(X_0)_\#\mathbb{P}_X = \mu_X$.

As X_i can be viewed as a random variable on the probability space $(X^{\mathbb{N}}, \mathbb{P}_X)$ with values in X, the second condition can be also rewritten as $\mathrm{law}(X_0) = \mu_X$.

Now given any metric space Z, consider any map $\ell_X : X \to Z$. Then, we call the tuple $(\mathcal{X}, \ell_X)$ a Z-labeled measure Markov chain (($Z-$)LMMC).

2.3. The Weisfeiler-Lehman Distance

In (Chen et al., 2022), a notion of (pseudo-)distance is proposed for labeled measure Markov chains (LMMCs). The motivation comes from the classical Weisfeiler-Lehman (WL) graph isomorphism test, which compares two graphs by iteratively testing whether certain aggregated node-label summaries of the two input graphs are the same. The WL distance of (Chen et al., 2022) introduced a measure-theoretic treatment of the node labels via Markov kernels, and it essentially "metrized" the WL-test procedure into a distance measure that is compatible with WL test. We briefly introduce the concept below; see the original paper (Chen et al., 2022) for details.

Consider a $Z-$LMMC $(\mathcal{X}, \ell_X)$. We recursively define a sequence of maps $\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)} : X \to \mathcal{P}^{\circ k}(Z)$ for $k \in \mathbb{N}$. First of all, we let $\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)} := \ell_X$. Then,

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

[1] This is equivalent to saying that for any $\alpha \in \mathcal{P}(Z)$ and any $z_0 \in Z$, one has that $\int_Z d_Z(z, z_0)\alpha(dz) < \infty$.

[2] Here $\mathbb{N}$ denotes the set of all non negative integers.

Finally, we let

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

Now, given two $Z-$LMMCs $(\mathcal{X},\ell_X)$ and $(\mathcal{Y},\ell_Y)$ and any $k\in\mathbb{N}$, the Weisfeiler-Lehman distance of depth k is defined as follows:

$$d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) \coloneqq d_{\mathrm{W}}(\mathfrak{L}_k((\mathcal{X},\ell_X)),\mathfrak{L}_k((\mathcal{Y},\ell_Y))). \tag{2}$$

3. Reinterpretation of the WL distance

The original definition of the WL distance, while well-suited for devising a computation algorithm for the distance, is rather intricate due to its iterative consideration of probability measures on spaces $\mathcal{P}^{\circ k}(Z)$ with increasing complexity. However, a better understanding of the concept can be achieved by using the language of stochastic processes. In this section, we will elaborate on two approaches: (1) using the concept of *Markovian couplings*, a special type of couplings between Markov chains, and (2) exploring the connection between 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.

3.1. Recall of a characterization of the WL distance

We first recall a characterization of the WL distance from (Chen et al., 2022). Given any two MMCs $\mathcal{X}$ and $\mathcal{Y}$, one can inductively define the notion of k-step coupling between $m^X_\bullet$ and $m^Y_\bullet$ as follows:

$k=1$: A *1-step coupling* between $m^X_\bullet$ and $m^Y_\bullet$ is defined to be any measurable map

$$\nu^{(1)}_{\bullet,\bullet}: X\times Y\to\mathcal{P}(X\times Y)$$

such that $\nu^{(1)}_{x,y}\in\mathcal{C}(m^X_x,m^Y_y)$ for any $x\in X$ and $y\in Y$.

$k\geq 2$: A measurable map $\nu^{(k)}_{\bullet,\bullet}: X\times Y\to\mathcal{P}(X\times Y)$ is called a *k-step coupling* between $m^X_\bullet$ and $m^Y_\bullet$ if there exist a $(k-1)$-step coupling $\nu^{(k-1)}_{\bullet,\bullet}$ and a 1-step coupling $\nu^{(1)}_{\bullet,\bullet}$ such that for any $x\in X$ and $y\in Y$, one has

$$\nu^{(k)}_{x,y}=\int_{X\times Y}\nu^{(k-1)}_{x',y'}\cdot\nu^{(1)}_{x,y}(dx'\times dy')$$

Let $\mathcal{C}^{(k)}(m^X_\bullet,m^Y_\bullet)$ denote the collection of all k-step couplings between $m^X_\bullet$ and $m^Y_\bullet$. Furthermore, for any $\gamma\in\mathcal{C}(\mu_X,\mu_Y)$ and any $\nu^{(k)}_{\bullet,\bullet}\in\mathcal{C}^{(k)}(m^X_\bullet,m^Y_\bullet)$, denote by

$$\nu^{(k)}_{\bullet,\bullet}\odot\gamma \coloneqq \int_{X\times Y}\nu^{(k)}_{x',y'}\cdot\gamma(dx'\times dy').$$

It is shown in (Chen et al., 2022) that $\nu^{(k)}_{\bullet,\bullet}\odot\gamma\in\mathcal{C}(\mu_X,\mu_Y)$. We then let $\mathcal{C}^{(k)}(\mu_X,\mu_Y)$ denote the collection of all such couplings. Then, it turns out the WL distance can be characterized as follows[3].

Theorem 3.1 ((Chen et al., 2022, Theorem A.7)). *Given $k\in\mathbb{N}$ and any two Z-LMMCs $(\mathcal{X},\ell_X)$ and $(\mathcal{Y},\ell_Y)$, one has that*

$$d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = \inf_{\gamma^{(k)}\in\mathcal{C}^{(k)}(\mu_X,\mu_Y)}\mathbb{E}_{\gamma^{(k)}}d_Z(\ell_X(X_0),\ell_Y(Y_0)).$$

3.2. Markovian couplings and a stochastic process interpretation of the WL distance

Readers familiar with stochastic processes may recognize the construction of k-step couplings in the previous section. In fact, we observe that these couplings can essentially be derived from the so-called Markovian couplings (which we will introduce later). Based on this observation, in this section, we provide a clean and intuitive interpretation of the WL distance as a variant of the Wasserstein distance between distributions of paths by highlighting this connection.

Recall that any measure Markov chain $(X,m^X_\bullet,\mu_X)$ can be equivalently described as a probability measure $\mathbb{P}_X$ on the path space $X^{\mathbb{N}}$ such that $(X_0)_\#\mathbb{P}_X=\mu_X$ and $\mathbb{P}_X(X_{i+1}=x'|X_i=x)=m^X_x(x')$ for any $x,x'\in X$. In this way, given any two MMCs $\mathcal{X}=(X,m^X_\bullet,\mu_X)$ and $\mathcal{Y}=(Y,m^Y_\bullet,\mu_X)$, one can consider couplings $\mathbb{P}\in\mathcal{C}(\mathbb{P}_X,\mathbb{P}_Y)\subseteq\mathcal{P}(X^{\mathbb{N}}\times Y^{\mathbb{N}})$ between their corresponding path distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. There is a canonical identification between $X^{\mathbb{N}}\times Y^{\mathbb{N}}$ and $(X\times Y)^{\mathbb{N}}$ sending $((x_j)_{j\in\mathbb{N}},(y_j)_{j\in\mathbb{N}})$ to $(x_j,y_j)_{j\in\mathbb{N}}$. Then, $\mathbb{P}$ can be also considered as a probability measure on $(X\times Y)^{\mathbb{N}}$, i.e., a distribution of paths inside the space $X\times Y$. As this distribution is a coupling between two Markov chains, it is natural to also impose the Markov property on such couplings. This naturally leads to the following definition of Markovian couplings which helps to provide an intuitive characterization of the WL distance. We first introduce some notation. For any $i\in\mathbb{N}$, we let $X_i: X^{\mathbb{N}}\times Y^{\mathbb{N}}\to X$ sending $((x_j)_{j\in\mathbb{N}},(y_j)_{j\in\mathbb{N}})$ to x_i and we similarly define Y_i. Now, we introduce the notion of Markovian couplings.

Definition 3.2 (Markovian couplings). A coupling $\mathbb{P}\in\mathcal{C}(\mathbb{P}_X,\mathbb{P}_Y)$ is called a *Markovian coupling*, if the following conditions are satisfied:

1. The sequence $\{(X_i,Y_i)\}_{i\in\mathbb{N}}$, regarded as random variables valued in $X\times Y$ on the probability measure space $(X^{\mathbb{N}}\times Y^{\mathbb{N}},\mathbb{P})$, satisfies the *Markov property*, i.e., for any

[3] In Theorem A.7 of (Chen et al., 2022), it is assumed that the measures equipped in LMMCs are stationary with respect to their corresponding Markov kernels. However, we note that the statement in Theorem A.7 still holds if we remove the requirement measures to be stationary.

$i \in \mathbb{N}$, $x_0, \ldots, x_i \in X$ and $y_0, \ldots, y_i \in Y$, one has that

$$\mathbb{P}\big((X_i, Y_i) = (x_i, y_i) | \{(X_j, Y_j) = (x_j, y_j)\}_{0 \le j \le i-1}\big) = \mathbb{P}\big((X_i, Y_i) = (x_i, y_i) | (X_{i-1}, Y_{i-1}) = (x_{i-1}, y_{i-1})\big).$$

2. For any $i \in \mathbb{N}$, $x \in X$ and $y \in Y$, if we construct a probability measure $(\nu_{i+1})_{x,y} \in \mathcal{P}(X \times Y)$ as follows

$$(\nu_{i+1})_{x,y}(x', y') := \mathbb{P}\big((X_{i+1}, Y_{i+1}) = (x', y') | (X_i, Y_i) = (x, y)\big)$$

$\forall\, x' \in X$, $y' \in Y$, then $(\nu_{i+1})_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)$. i.e., $(\nu_{i+1})_{\bullet,\bullet}$ is a 1-step coupling between $m_\bullet^X$ and $m_\bullet^Y$.[4]

We let $\mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$ denote the collection of all Markovian couplings between $\mathbb{P}_X$ and $\mathbb{P}_Y$.

In short, a Markovian coupling of two Markov chains $\mathbb{P}_X$ and $\mathbb{P}_Y$ 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 1 for an illustration.

Remark 3.3 (Initial distribution). Note that coupling $\mathbb{P} \in \mathcal{C}(\mathbb{P}_X, \mathbb{P}_Y)$ satisfies that $(X_0, Y_0)_\# \mathbb{P} \in \mathcal{C}(\mu_X, \mu_Y)$. Hence, any Markovian coupling also starts with an initial distribution which is itself a coupling between the initial distributions of the two Markov chains.

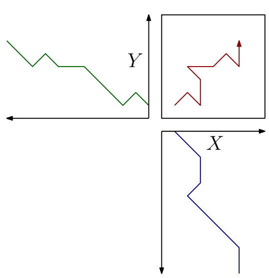

Figure 1: A Markovian coupling between $\mathcal{X}$ and $\mathcal{Y}$ is a Markov chain in the product space $X \times Y$ so that its projections on to X and Y have the same law as $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Remark 3.4 (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 & 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.

[4] When the event $\{(X_i, Y_i) = (x, y)\}$ is null, the conditional probability is not defined and hence we simply let $(\nu_{i+1})_{x,y}$ be the product measure $m_x^X \otimes m_y^Y \in \mathcal{C}(m_x^X, m_y^Y)$.

Remark 3.5 (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 & Bharucha-Reid, 2018).

As promised at the beginning of this section, below we provide an intuitive explanation of k-step couplings. See Appendix A.1 for the proof of this result.

Proposition 3.6. *For any coupling $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and any k-step coupling $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}(m_\bullet^X, m_\bullet^Y)$, there exists a Markovian coupling $\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$ such that*

$$\mathrm{law}((X_k, Y_k)) = \nu_{\bullet,\bullet}^{(k)} \odot \gamma \text{ and } \mathrm{law}((X_0, Y_0)) = \gamma.$$

Conversely, for any Markovian coupling $\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$, $\gamma := \mathrm{law}((X_0, Y_0))$ is a coupling between μ_X and μ_Y and furthermore, there exists k-step coupling $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ such that $\mathrm{law}((X_k, Y_k)) = \nu_{\bullet,\bullet}^{(k)} \odot \gamma$.

This proposition implies that any k-step coupling, just as indicated by the name, is the distribution of two coupled random walks at exactly step k. The direct implication of this result is that we provide a characterization of the WL distance of depth k as follows. We note that the following theorem is an immediate consequence of Theorem 3.1 and Proposition 3.6.

Theorem 3.7. *For any Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that*

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = \inf_{\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)} \mathbb{E}_{\mathbb{P}}\, d_Z(\ell_X(X_k), \ell_Y(Y_k)).$$

In this way, we have interpreted the WL distance of depth k as a variant of the Wasserstein distance between the distributions of paths corresponding to the input Markov chains. Here the cost between two paths $w_X \in X^{\mathbb{N}}$ and $w_Y \in Y^{\mathbb{N}}$ is given by $d_Z(\ell_X(X_k(w_X)), \ell_Y(Y_k(w_Y)))$ through the label functions. Essentially, the WL distance at depth k is comparing the distribution of the Markov chains at step k; see Figure 2 for an illustration.

3.2.1. In terms of the language of random variables

A Markov chain can of course be realized on probability spaces beyond the path space. This flexibility can be useful in situations where some random variables need to be independent of the Markov chain. We consider this slightly more generalized setting in this section and reformulate the definition of Markovian couplings and also Theorem 3.7. It is worth noting that skipping this section will not impede understanding of the other parts of the paper.

Note that a common way for describing a measure Markov chain $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ is through a sequence of random

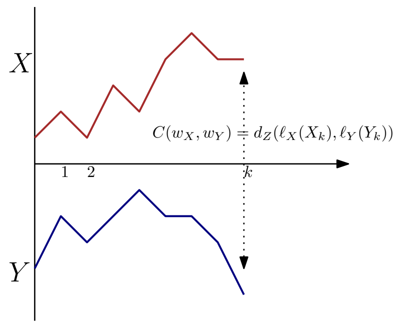

Figure 2: The WL distance of depth k is almost like the Wasserstein distance between path distributions where the cost $C(w_X, w_Y)$ between paths is given by the distance between labels of the two random walks at step k: $d_Z(\ell_X(X_k(w_X), \ell_Y(Y_k(w_Y)))$.

variables $\{X_i : \Omega \to X\}_{i\in\mathbb{N}}$ from some probability space $(\Omega, \mathbb{P})$ such that $\mathrm{law}(X_0) = \mu_X$ and $\mathbb{P}(X_{i+1} = x' | X_i = x) = m_x^X(x')$ for any $x, x' \in X$. Just to distinguish from our terminology of measure Markov chains $\mathcal{X}$, we call any such sequence of random variables as a *stochastic realization* of the MMC $\mathcal{X}$. In this way, the projection maps $\{X_i : (X^{\mathbb{N}}, \mathbb{P}_X) \to X\}_{i\in\mathbb{N}}$ on the path space give a concrete example of stochastic realization.

We now generalize the Markovian couplings from Definition 3.2 to the case of general stochastic realizations in a way similar to Definition 2.1.

Definition 3.8 (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 $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ be stochastic realizations on the *same* probability space $(\Omega, \mathbb{P})$. We call the sequence of random variables $\{(X_i, Y_i)\}_{i\in\mathbb{N}}$ a *Markovian coupling* between $\mathcal{X}$ and $\mathcal{Y}$, if the following properties hold:

1. The sequence of random variables $\{(X_i, Y_i)\}_{i\in\mathbb{N}}$ satisfies the Markov property, i.e., for any $i \in \mathbb{N}$, $x_0, \ldots, x_i \in X$ and $y_0, \ldots, y_i \in Y$, one has that
$$\mathbb{P}\big((X_i, Y_i) = (x_i, y_i) | \{(X_j, Y_j) = (x_j, y_j)\}_{0\le j\le i-1}\big) = \mathbb{P}\big((X_i, Y_i) = (x_i, y_i) | (X_{i-1}, Y_{i-1}) = (x_{i-1}, y_{i-1})\big).$$
2. For any $i \in \mathbb{N}$, $x \in X$ and $y \in Y$, if we construct a probability measure in $\mathcal{P}(X \times Y)$ as follows
$$(\nu_{i+1})_{x,y}(x', y') := \mathbb{P}\big((X_{i+1}, Y_{i+1}) = (x', y') | (X_i, Y_i) = (x, y)\big),$$
then $(\nu_{i+1})_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)$, i.e., $(\nu_{i+1})_{\bullet,\bullet}$ is a 1-step coupling between $m_\bullet^X$ and $m_\bullet^Y$.

Note that the above definition is almost identical to the definition of Markovian couplings between path distribution (see Definition 3.2). Hence, it is not surprising that Theorem 3.7 can be rephrased in terms of stochastic realizations; see also the similarity between Equation (1) and the corollary below:

Corollary 3.9. *For any Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that*
$$d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = \inf_{\{(X_i, Y_i)\}_{i\in\mathbb{N}}} \mathbb{E}\, d_Z(\ell_X(X_k), \ell_Y(Y_k))$$
where the infimum is taken over all possible Markovian couplings between $\mathcal{X}$ and $\mathcal{Y}$.

3.3. The Connection with Causal Optimal Transport

Finally, 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.

Note that COT has already found applications in mathematical finance (Glanzer et al., 2019; Backhoff-Veraguas et al., 2020) and in machine learning (Xu et al., 2020; Xu & 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 & Weinhardt, 2022; Eckstein & 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. 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 $k+1$)[5]. 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 3.10 ((bi)causal coupling)**.** A coupling measure $\pi \in \mathcal{C}(\alpha, \beta)$ is said to be *causal* from α to β if it satisfies
$$\pi((y_0, \ldots, y_l) | (x_0, \ldots, x_k)) = \pi((y_0, \ldots, y_l) | (x_0, \ldots, x_l))$$
for all $l \in \{0, \ldots, k\}$ and $(x_0, \ldots, x_k) \in X^{k+1}$. Here, the notataion $\pi(\cdot|\cdot)$ denotes conditional probability. This implies that, at time l and given the past $(x_0, \ldots, x_l)$ of X, the distribution of y_l does not depend on the future $(x_{l+1}, \ldots, x_k)$ 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 $\mathcal{C}_{\mathrm{bc}}(\alpha, \beta)$.

[5]The framework of COT 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, 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 from $\mathbb{N}$ 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 $\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$ (cf. Definition 3.2). Then, for any $k \in \mathbb{N}$, we let $\alpha^k := \mathrm{law}((X_0,\ldots,X_k)) \in \mathcal{P}(X^{k+1})$ and let $\beta^k := \mathrm{law}((Y_0,\ldots,Y_k)) \in \mathcal{P}(Y^{k+1})$. Recall that here $\mathrm{law}((X_0,\ldots,X_k)) := (X_0,\ldots,X_k)_\#\mathbb{P}$ and $\mathrm{law}((Y_0,\ldots,Y_k))$ is similarly defined. In fact, it is easy to see that α^k and β^k can be expressed explicitly as follows.

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

Now, we further let

$$\pi^k := \mathrm{law}((X_0,\ldots,X_k,Y_0,\ldots,Y_k)) \in \mathcal{P}(X^{k+1}\times Y^{k+1}).$$

Then, we have that

Lemma 3.11. *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

$$((x_0,\ldots,x_k),(y_0,\ldots,y_k)) \mapsto d_Z(\ell_X(x_k),\ell_Y(y_k)) \quad (3)$$

and state the following theorem.

Theorem 3.12. *For any Z-LMMCs $(\mathcal{X},\ell_X)$ and $(\mathcal{Y},\ell_Y)$, we have that*

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

In order to prove Theorem 3.12, we first note that d^{c^k} can be recursively computed. For each $i = 0,\ldots,k$, we define $V_i : X^{i+1}\times Y^{i+1} \to \mathbb{R}$ recursively as follows:

$$V_k(x_0,\ldots,x_k,y_0,\ldots,y_k) := d_Z(\ell_X(x_k),\ell_Y(y_k)), \text{ and}$$
$$V_{i-1}(x_0,\ldots,x_{i-1},y_0,\ldots,y_{i-1}) :=$$
$$\inf_{\nu^i\in\mathcal{C}(\alpha^k_{x_0,\ldots,x_{i-1}},\beta^k_{y_0,\ldots,y_{i-1}})} \mathbb{E}_{\nu^i} V_i(X_0,\ldots,X_{i-1},Y_0,\ldots,Y_{i-1})$$

where $\alpha^k_{x_0,\ldots,x_{i-1}}$ (resp. $\beta^k_{y_0,\ldots,y_{i-1}}$) is the probability measure on X (resp. Y) defined as the conditional probability measure $\alpha^k_{x_0,\ldots,x_{i-1}}(x_i) := \alpha^k(x_i|x_0,\ldots,x_{i-1})$ (resp. $\beta^k_{y_0,\ldots,y_{i-1}}(y_i) := \beta^k(y_i|y_0,\ldots,y_{i-1})$)[6].

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

$$d^{c^k}(\alpha^k,\beta^k) = \inf_{\gamma^0\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} V_0(x_0,y_0)\,\gamma^0(dx_0\times dy_0). \quad (4)$$

As we are dealing with Markov chains, it is expected that $V_i(x_0,\ldots,x_{i-1},y_0,\ldots,y_{i-1})$ is independent of the past, i.e., independent of $(x_0,\ldots,x_{i-2},y_0,\ldots,y_{i-2})$. In order to show that this is indeed the case, for each $i = 0,\ldots,k-1$, we define $W_i : X\times Y \to \mathbb{R}$ in a way similar to how we defined V_i:

$$W_k(x_k,y_k) := d_Z(\ell_X(x_k),\ell_Y(y_k)), \text{ and}$$
$$W_{i-1}(x_{i-1},y_{i-1}) := \inf_{\nu^i\in\mathcal{C}(\alpha^k_{x_{i-1}},\beta^k_{y_{i-1}})} \mathbb{E}_{\nu^i} W_i(X_i,Y_i)$$

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

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

Finally, as a direct consequence of Theorem 3.7, Lemma 3.11, and Lemma 3.13, one can prove Theorem 3.12. See Appendix A.5 for a complete proof.

Remark 3.14 (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 Appendix A.6 for more details.

4. Implications to Message Passing GNNs

In this section, we will provide some results on the use of the WL distance for studying the universality and the Lipschitz property of message passing GNNs. In (Chen et al., 2022), the authors introduced a neural network framework,

[6] It might happen that the event $(X_0,\ldots,X_{i-1}) = (x_0,\ldots,x_{i-1})$ is null, then the conditional probability is not defined. In this case, we simply let $\alpha^k_{x_0,\ldots,x_{i-1}}$ be any probability measure on X. We adopt the 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.

named Markov chain neural network (MCNN), on the collection of all LMMCs and established that this framework is universal w.r.t. the WL distance on this collection. It is briefly touched upon in (Chen et al., 2022) that a MCNN will reduce to a standard Message Passing Graph Neural Network (MP-GNN) when the input LMMCs are induced from labeled graphs. In this section, we will restrict ourselves to graph induced LMMCs and study properties of MP-GNNs via the use of the WL distances. In particular, we show that a special yet common type of MP-GNNs (1) has the same power as the WL distance in distinguishing labeled graphs, (2) is Lipschitz w.r.t. change of input labeled graphs through the lens of the WL distance and (3) satisfies the universal approximation property, i.e., any continuous function defined on a compact space of labeled graphs can be approximated by such MP-GNNs.

4.1. Graph induced LMMCs

We first introduce some terminology related to graphs. In this paper, we consider finite edge weighted graph $G = (V, E, w : E \to (0, \infty))$ where V denotes the vertex set, E denotes the edge set and w denotes the edge weight function. For each $v \in V$, we define its degree as $\deg(v) := \sum_{v' \in V} w_{vv'}$. In order to distinguish between multiple graphs, we sometimes include the graph symbol G in subscripts when referring to these notions, such as in the case of V_G or w_G. This helps to clarify which graph we are referring to in cases where multiple graphs are involved.

Now, given a finite edge weighted graph G endowed with a label function $\ell_G : V \to Z$, one can generate a LMMC as follows. Given $q \in (0, 1)$, we associate to the vertex set V a Markov kernel $m_\bullet^{G,q}$ as follows: for any $v \in V$,

$$m_v^{G,q} := \begin{cases} q\delta_v + \frac{1-q}{\deg(v)} \sum_{v' \in N_G(v)} w_{vv'} \delta_{v'}, & \deg(v) > 0; \\ \delta_v, & \deg(v) = 0. \end{cases}$$

For each vertex $v \in V$, we introduce the following modified notion of degree

$$\overline{\deg}(v) := \begin{cases} \deg(v) & \deg(v) > 0; \\ 1 & \deg(v) = 0. \end{cases} \tag{5}$$

Based on $\overline{\deg}$, we introduce the following probability measure on V: $\mu_G := \sum_{v \in V} \frac{\overline{\deg}(v)}{\sum_{v' \in V} \overline{\deg}(v')} \delta_v$. Then for any $q \in (0, 1)$, μ_G is a stationary distribution w.r.t. the Markov kernel $m_\bullet^{G,q}$. Then, we let $\mathcal{X}_q(G) := (V, m_\bullet^{G,q}, \mu_G)$ and we say that the LMMC $(\mathcal{X}_q(G), \ell_G)$ is induced by the labeled graph (G, ℓ_G).

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

4.2. Message Passing Graph Neural Networks

Given $q > 0$ we consider the following special type of k-layer MP-GNNs.

Message Passing: $\ell_G^{i+1}(v) :=$

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

Readout:

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

where $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ and $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$ are MLPs.

Remark 4.1. Any such MP-GNN $h : \mathcal{G}(Z) \to \mathbb{R}$ arises as the restriction of a Markov Chain Neural Network (MCNN) $H : \mathcal{M}^L(Z) \to \mathbb{R}$ (see (Chen et al., 2022, Section 4)) to the collection $\mathcal{G}_q(Z)$. For more details, see Appendix B.1.

We use $\mathcal{NN}_k^q(\mathbb{R}^d)$ to denote the collection of all such MP-GNNs with k layers.

Discriminative Power of MP-GNNs In addition to the Lipschitz property, we also establish that $\mathcal{NN}_k^q(\mathbb{R}^d)$ has the same discriminative power as the WL distance.

Proposition 4.2. *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 $h((G_1, \ell_{G_1})) \neq h((G_2, \ell_{G_2}))$.*

Proof. This follows directly from Remark 4.1 and (Chen et al., 2022, Proposition 4.1). □

In fact, we establish a fact stronger than Proposition 4.2 under a setting similar to (but more flexible than) the one used in (Xu et al., 2018). Choose any *countable* subset $Z \subseteq \mathbb{R}^d$ and any *countable* subset $P \subseteq \mathbb{R}$. Let $\mathcal{G}_P(Z)$ denote the collection of all Z-labeled weighted graphs so that their edge weights are contained in P.

Now, we will establish that a very restricted set of MP-GNNs is sufficient to have the same discriminative power as the WL distance and, in this way, we establish a much stronger result than Proposition 4.2. We let $\mathcal{NN}_k^{q,1}(\mathbb{R}^d)$ denote the collection of maps $h \in \mathcal{NN}_k^{q}(\mathbb{R}^d)$ where $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ satisfies that $d_i = 1$ for each $i = 1, \ldots, k+1$ (note $d_0 = d$). Then, we establish the following main result:

Theorem 4.3. *For each $k \geq 0$, there exists $h \in \mathcal{NN}_k^{q,1}(\mathbb{R}^d)$ such that for any $(G_1, \ell_{G_1}), (G_2, \ell_{G_2}) \in \mathcal{G}_P(Z)$, $d_{\mathcal{G},q}^{(k)}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) > 0$ iff $h((G_1, \ell_{G_1})) \neq h((G_2, \ell_{G_2}))$.*

Notice that whereas in Proposition 4.2 the existence of h may depend on the choice of labeled graphs, h in Theorem 4.3 is universal for all pairs of labeled graphs. Furthermore, while maps φ_i involved in intermediate layers of h in Proposition 4.2 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 Theorem 4.3. In this way, the latter result is stronger than the previous one.

Remark 4.4. (Xu et al., 2018) established a result similar to Theorem 4.3: for any $k \geq 0$, there exists a MP-GNN h which has the same discriminative power as the k-step WL test when restricted to the set of graphs whose labels are from a common *countable* set. We remark that in order to show the existence of such h, however, the MP-GNN h they constructed utilizes aggregation functions defined on a *countable* set which may not be able to be continuously extended to a "continuous" domain such as $\mathbb{R}^d$. In contrast, each function involved in constructing $h \in \mathcal{NN}_k^{q,1}(\mathbb{R}^d)$ is a continuous map between Euclidean spaces.

Lipschitz property of MP-GNNs In addition to the study of zero sets, one can generalize the results above in a quantitative manner. More specifically, we establish that the specific MP-GNNs defined in this section are Lipschitz w.r.t. the WL distance. This in particular indicates that MP-GNNs are stable w.r.t. small perturbations of graphs in the sense of the WL distance.

Theorem 4.5. *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*

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

Remark 4.6. We note that the WL distance could potentially be used to study the Lipschitz property of other types of MP-GNNs. See Appendix B.4 for such a study of a normalized version of Graph Isomorphism Network.

Universal approximation Based on the Lipschitz property, we finally establish the universal approximation property of MP-GNNs. For this purpose, we introduce some notation. Given any $k \in \mathbb{N}$ and any subset $\mathcal{K} \subseteq \left(\mathcal{G}(\mathbb{R}^d), d_{\mathcal{G},q}^{(k)}\right)$, we let $C(\mathcal{K}, \mathbb{R})$ denote the set of all continuous functions $f : \mathcal{K} \to \mathbb{R}$. We further let $\mathcal{NN}_k^{q}(\mathbb{R}^d)|_{\mathcal{K}}$ denote the collection of all functions $h|_{\mathcal{K}}$ where $h \in \mathcal{NN}_k^{q}(\mathbb{R}^d)$, i.e.,

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

Then, we state our main result as follows.

Theorem 4.7 (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, $\overline{\mathcal{NN}_k^{q}(\mathbb{R}^d)|_{\mathcal{K}}} = C(\mathcal{K}, \mathbb{R})$.*

The theorem above follows directly from the universal approximation result for MCNNs (Chen et al., 2022, Theorem 4.3); see Appendix B.5 for the proof.

5. Conclusion and Future Directions

In this paper, we further investigate the WL distance, proposed in (Chen et al., 2022), for comparing LMMCs and establish that the WL distance of depth k can be seen as a variant of the ℓ_1-Wasserstein distance comparing distributions of trajectories from random walks on the label space. We further identify connections between the WL distance and causal optimal transport (COT), suggesting potential applications in computing/approximating COT distances and extending the WL distance to handle different label spaces. These avenues offer interesting future research directions.

Given that the WL distance is compatible with the WL-graph isomorphism test, which has been connected to the expressiveness of message-passing graph neural networks (MP-GNNs) (Xu et al., 2018; Azizian et al., 2020), it is natural to equip the space of graphs with a (pseudo-)metric structure induced by the WL distance. As already observed in (Chen et al., 2022) and further detailed in this paper, MP-GNN can universally approximate continuous functions defined on the space of graphs w.r.t. the WL distance. In fact, a more refined result is obtained (Theorem 4.7) which shows universal approximation result for the family of k-layer MP-GNNs. Furthermore, similar to the work of (Chuang & Jegelka, 2022), we show that the WL distance can also be used to study stability and Lipschitz property of sub-families of MP-GNNs. We also remark that the WL distance can serve as a suitable choice of metric for the space of graphs to study questions such as the generalization power of message passing GNNs as done in (Chuang & Jegelka, 2022).

Acknowledgement. This work is partially supported by the NSF through grants IIS-1901360, IIS-2050360, CCF-1740761, CCF-2112665, CCF-2310412, DMS-1547357, and by the BSF under grant 2020124.

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Azizian, W. et al. Expressive power of invariant and equivariant graph neural networks. In *International Conference on Learning Representations*, 2020.

Backhoff, J., Beiglbock, M., Lin, Y., and Zalashko, A. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., and Eder, M. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24(3):601–632, 2020.

Brugère, T., Wan, Z., and Wang, Y. Distances for Markov chains, and their differentiation. *arXiv preprint arXiv:2302.08621*, 2023.

Burago, D., Burago, Y., and Ivanov, S. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

Chen, S., Lim, S., Mémoli, F., Wan, Z., and Wang, Y. Weisfeiler-Lehman meets Gromov-Wasserstein. In *International Conference on Machine Learning (ICML)*, pp. 3371–3416. PMLR, 2022.

Chuang, C.-Y. and Jegelka, S. Tree mover's distance: Bridging graph metrics and stability of graph neural networks. *arXiv preprint arXiv:2210.01906*, 2022.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Eckstein, S. and Pammer, G. Computational methods for adapted optimal transport. *arXiv preprint arXiv:2203.05005*, 2022.

Glanzer, M., Pflug, G. C., and Pichler, A. Incorporating statistical model error into the calculation of acceptability prices of contingent claims. *Mathematical Programming*, 174(1):499–524, 2019.

Klemmer, K., Xu, T., Acciaio, B., and Neill, D. B. 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, pp. 4523–4531, 2022.

Kolmogorov, A. N. and Bharucha-Reid, A. T. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

Lassalle, R. Causal transport plans and their Monge–Kantorovich problems. *Stochastic Analysis and Applications*, 36(3):452–484, 2018.

Lehman, A. and Weisfeiler, B. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya*, 2(9): 12–16, 1968.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Mémoli, F. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Pichler, A. and Weinhardt, M. The nested sinkhorn divergence to learn the nested distance. *Computational Management Science*, 19(2):269–293, 2022.

Togninalli, M., Ghisu, E., Llinares-López, F., Rieck, B., and Borgwardt, K. Wasserstein Weisfeiler-Lehman graph kernels. *Advances in Neural Information Processing Systems*, 32:6439–6449, 2019.

Toth, C., Lee, D., Hacker, C., and Oberhauser, H. Capturing graphs with hypo-elliptic diffusions. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

Weaver, N. Order completeness in Lipschitz algebras. *Journal of Functional Analysis*, 130(1):118–130, 1995.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

Xu, T. and Acciaio, B. Conditional COT-GAN for video prediction with kernel smoothing. *Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems*, 2022.

Xu, T., Wenliang, L. K., Munn, M., and Acciaio, B. COT-GAN: Generating sequential data via causal optimal transport. *Advances in Neural Information Processing Systems*, 33:8798–8809, 2020.

A. Details from Section 3

A.1. Proof of Proposition 3.6

To facilitate the proof below, we consider the canonical identification $\iota : (X \times Y)^{\mathbb{N}} \to X^{\mathbb{N}} \times Y^{\mathbb{N}}$ sending $(x_j, y_j)_{j\in\mathbb{N}}$ to $((x_j)_{j\in\mathbb{N}}, (y_j)_{j\in\mathbb{N}})$. This is obviously a homeomorphism given the product topology. By slight abuse of notation, we let X_i denote both the projection map $X_i : X^{\mathbb{N}} \times Y^{\mathbb{N}} \to X$ to the i-th component for $i \in \mathbb{N}$ and also the projection map $X_i \circ \iota : (X \times Y)^{\mathbb{N}} \to X$. We adopt a similar convention for Y_i.

For any k-step coupling $\gamma^{(k)} \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)$, there exist $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and $(\nu_i)_{\bullet,\bullet} \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$ for $i = 1, \ldots, k$ such that

$$\gamma^{(k)} = \int\limits_{X\times Y} \cdots \int\limits_{X\times Y} (\nu_k)_{x_{k-1},y_{k-1}} (\nu_{k-1})_{x_{k-2},y_{k-2}}(dx_{k-1} \times dy_{k-1}) \cdots (\nu_1)_{x_0,y_0}(dx_1 \times dy_1)\gamma(dx_0 \times dy_0).$$

We further let $(\nu_i)_{\bullet,\bullet} := (\nu_k)_{\bullet,\bullet}$ for all $i > k$. Now, by the Kolmogorov extension theorem, there exists $\mathbb{Q} \in \mathcal{P}((X \times Y)^{\mathbb{N}})$ such that

1. $(\nu_{i+1})_{x,y}(x', y') = \mathbb{Q}\big((X_{i+1}, Y_{i+1}) = (x', y')|(X_i, Y_i) = (x, y)\big)$, for any $x, x' \in X$, $y, y' \in Y$ and any $i = 0, \ldots$.
2. $\mathbb{Q}((X_i, Y_i) \in B_i, 0 \le i \le n) = \int_{B_0} \gamma(dx_0 \times dy_0) \cdots \int_{B_n} (\nu_n)_{x_{n-1},y_{n-1}}(dx_n \times dy_n)$ for any $n \in \mathbb{N}$ and any measurable $B_i \subseteq X \times Y$ for $i = 0, \ldots, n$.

Now, we let $\mathbb{P} := \iota_\#\mathbb{Q} \in \mathcal{P}(X^{\mathbb{N}} \times Y^{\mathbb{N}})$. Then, it is straightforward to check that

1. $(\nu_{i+1})_{x,y}(x', y') = \mathbb{P}\big((X_{i+1}, Y_{i+1}) = (x', y')|(X_i, Y_i) = (x, y)\big)$, for any $x, x' \in X$, $y, y' \in Y$ and any $i = 0, \ldots$.
2. $\mathbb{P}((X_i, Y_i) \in B_i, 0 \le i \le n) = \int_{B_0} \gamma(dx_0 \times dy_0) \cdots \int_{B_n} (\nu_n)_{x_{n-1},y_{n-1}}(dx_n \times dy_n)$ for any $n \in \mathbb{N}$ and any measurable $B_i \subseteq X \times Y$ for $i = 0, \ldots, n$.

In this way, it is easy to check that $\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$. Finally, we have that

$$\begin{aligned}&\mathrm{law}((X_k, Y_k)) = (X_k, Y_k)_\#\mathbb{P}\\ =&\int\limits_{X\times Y} \cdots \int\limits_{X\times Y} (\nu_k)_{x_{k-1},y_{k-1}} (\nu_{k-1})_{x_{k-2},y_{k-2}}(dx_{k-1} \times dy_{k-1}) \cdots (\nu_1)_{x_0,y_0}(dx_1 \times dy_1)\gamma(dx_0 \times dy_0)\\ =&\gamma^{(k)}.\end{aligned}$$

For the other direction, given a Markovian coupling $\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$, we let

$$(\nu_{i+1})_{x,y}(x', y') := \mathbb{P}\big((X_{i+1}, Y_{i+1}) = (x', y')|(X_i, Y_i) = (x, y)\big)$$

for any $i \in \mathbb{N}$, $x, x' \in X$ and $y, y' \in Y$.[7] Then, by definition of the Markovian coupling, $(\nu_{i+1})_{\bullet,\bullet}$ is a 1-step coupling between $m_\bullet^X$ and $m_\bullet^Y$ for each $i = 0, 1, \ldots$. Let $\gamma := \mathrm{law}((X_0, Y_0)) \in \mathcal{C}(\mu_X, \mu_Y)$. Then, if we let

$$\nu^{(k)}_{x_0,y_0} := \int\limits_{X\times Y} \cdots \int\limits_{X\times Y} (\nu_k)_{x_{k-1},y_{k-1}} (\nu_{k-1})_{x_{k-2},y_{k-2}}(dx_{k-1} \times dy_{k-1}) \cdots (\nu_1)_{x_0,y_0}(dx_1 \times dy_1)$$

for any $x_0 \in X$ and $y_0 \in Y$, then $\nu^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ and thus

$$\mathrm{law}((X_k, Y_k)) = (X_k, Y_k)_\#\mathbb{P} = \nu^{(k)}_{\bullet,\bullet} \odot \gamma.$$

[7] Recall from Definition 3.2 that when the event $\{(X_i, Y_i) = (x, y)\}$ is null, we let $(\nu_{i+1})_{x,y} := m_x^X \otimes m_y^Y$.

A.2. Proof of Theorem 3.7

Fix an arbitrary k-step coupling $\gamma^{(k)} = \nu^{(k)}_{\bullet,\bullet} \odot \gamma \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)$. Then, by Proposition 3.6, there exists a Markovian coupling $\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$ such that

$$\mathrm{law}((X_k, Y_k)) = \gamma^{(k)} \text{ and } \mathrm{law}((X_0, Y_0)) = \gamma.$$

This implies that

$$\begin{aligned}
&\int_{X\times Y} d_Z(\ell_X(x), \ell_Y(y))\gamma^{(k)}(dx \times dy) = \int_{X\times Y} d_Z(\ell_X(x), \ell_Y(y))(X_k, Y_k)_\# \mathbb{P}(dx \times dy) \\
&= \int_{X^{\mathbb{N}}\times Y^{\mathbb{N}}} d_Z(\ell_X(X_k(w_X)), \ell_Y(Y_k(w_Y)))\mathbb{P}(dw_X \times dw_Y) \\
&\geq \inf_{\mathbb{P}\in\mathcal{C}_{\mathcal{M}}(\mathbb{P}_X,\mathbb{P}_Y)} \int_{X^{\mathbb{N}}\times Y^{\mathbb{N}}} d_Z(\ell_X(X_k(w_X)), \ell_Y(Y_k(w_Y)))\mathbb{P}(dw_X \times dw_Y).
\end{aligned}$$

Since $\gamma^{(k)}$ is arbitrary, we have that

$$\begin{aligned}
&\inf_{\gamma^{(k)}\in\mathcal{C}^{(k)}(\mu_X,\mu_Y)} \int_{X\times Y} d_Z(\ell_X(x), \ell_Y(y))\gamma^{(k)}(dx \times dy) \\
&\geq \inf_{\mathbb{P}\in\mathcal{C}_{\mathcal{M}}(\mathbb{P}_X,\mathbb{P}_Y)} \int_{X^{\mathbb{N}}\times Y^{\mathbb{N}}} d_Z(\ell_X(X_k(w_X)), \ell_Y(Y_k(w_Y)))\mathbb{P}(dw_X \times dw_Y).
\end{aligned}$$

Conversely, fix an arbitrary Markovian coupling $\mathbb{P} \in \mathcal{C}_{\mathcal{M}}(\mathbb{P}_X, \mathbb{P}_Y)$. Then, by Proposition 3.6, $\gamma^{(k)} := \mathrm{law}((X_k, Y_k))$ is a k-step coupling between μ_X and μ_Y. Therefore,

$$\begin{aligned}
&\int_{X^{\mathbb{N}}\times Y^{\mathbb{N}}} d_Z(\ell_X(X_k(w_X)), \ell_Y(Y_k(w_Y)))\mathbb{P}(dw_X \times dw_Y) = \int_{X\times Y} d_Z(\ell_X(x), \ell_Y(y))\gamma^{(k)}(dx \times dy) \\
&\geq \inf_{\gamma^{(k)}\in\mathcal{C}^{(k)}(\mu_X,\mu_Y)} \int_{X\times Y} d_Z(\ell_X(x), \ell_Y(y))\gamma^{(k)}(dx \times dy).
\end{aligned}$$

Since $\mathbb{P}$ is arbitrary,

$$\begin{aligned}
&\inf_{\gamma^{(k)}\in\mathcal{C}^{(k)}(\mu_X,\mu_Y)} \int_{X\times Y} d_Z(\ell_X(x), \ell_Y(y))\gamma^{(k)}(dx \times dy) \\
&\leq \inf_{\mathbb{P}\in\mathcal{C}_{\mathcal{M}}(\mathbb{P}_X,\mathbb{P}_Y)} \int_{X^{\mathbb{N}}\times Y^{\mathbb{N}}} d_Z(\ell_X(X_k(w_X)), \ell_Y(Y_k(w_Y)))\mathbb{P}(dw_X \times dw_Y).
\end{aligned}$$

Finally, by Theorem 3.1, one can conclude that

$$\begin{aligned}
d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) &= \inf_{\gamma^{(k)}\in\mathcal{C}^{(k)}(\mu_X,\mu_Y)} \int_{X\times Y} d_Z(\ell_X(x), \ell_Y(y))\gamma^{(k)}(dx \times dy) \\
&= \inf_{\mathbb{P}\in\mathcal{C}_{\mathcal{M}}(\mathbb{P}_X,\mathbb{P}_Y)} \int_{X^{\mathbb{N}}\times Y^{\mathbb{N}}} d_Z(\ell_X(X_k(w_X)), \ell_Y(Y_k(w_Y)))\mathbb{P}(dw_X \times dw_Y).
\end{aligned}$$

A.3. Proof of Lemma 3.11

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

$$\begin{aligned}
&\pi^k((y_0,\ldots,y_l)|(x_0,\ldots,x_k))\\
&=\mathbb{P}\big(Y_0=y_0,\ldots,Y_l=y_l|X_0=x_0,\ldots,X_k=x_k\big)\\
&=\frac{\mathbb{P}\big(X_0=x_0,\ldots,X_k=x_k,Y_0=y_0,\ldots,Y_l=y_l\big)}{\mathbb{P}\big(X_0=x_0,\ldots,X_k=x_k\big)}\\
&=\frac{\mathbb{P}\big(X_{l+1}=x_{l+1},\ldots,X_k=x_k|X_0=x_0,\ldots,X_l=x_l,Y_0=y_0,\ldots,Y_l=y_l\big)}{\mathbb{P}\big(X_{l+1}=x_{l+1},\ldots,X_k=x_k|X_0=x_0,\ldots,X_l=x_l\big)}\\
&\quad\times\frac{\mathbb{P}\big(X_0=x_0,\ldots,X_l=x_l,Y_0=y_0,\ldots,Y_l=y_l\big)}{\mathbb{P}\big(X_0=x_0,\ldots,X_l=x_l\big)}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\big(X_{l+1}=x_{l+1},\ldots,X_k=x_k|X_0=x_0,\ldots,X_l=x_l,Y_0=y_0,\ldots,Y_l=y_l\big)\\
&=\mathbb{P}\big(X_k=x_k|X_0=x_0,\ldots,X_{k-1}=x_{k-1},Y_0=y_0,\ldots,Y_l=y_l\big)\\
&\quad\times\mathbb{P}\big(X_{k-1}=x_{k-1}|X_0=x_0,\ldots,X_{k-2}=x_{k-2},Y_0=y_0,\ldots,Y_l=y_l\big)\\
&\quad\cdots\times\mathbb{P}\big(X_{l+1}=x_{l+1}|X_0=x_0,\ldots,X_l=x_l,Y_0=y_0,\ldots,Y_l=y_l\big)\\
&=m^X_{x_l}(x_{l+1})\cdots m^X_{x_{k-1}}(x_k)\\
&=\mathbb{P}\big(X_{l+1}=x_{l+1},\ldots,X_k=x_k|X_0=x_0,\ldots,X_l=x_l\big).
\end{aligned}$$

Therefore, one can conclude that

$$\begin{aligned}
\pi^k((y_0,\ldots,y_l)|(x_0,\ldots,x_k))&=\frac{\mathbb{P}\big(X_0=x_0,\ldots,X_l=x_l,Y_0=y_0,\ldots,Y_l=y_l\big)}{\mathbb{P}\big(X_0=x_0,\ldots,X_l=x_l\big)}\\
&=\pi^k((y_0,\ldots,y_l)|(x_0,\ldots,x_l)).
\end{aligned}$$

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.

A.4. Proof of Lemma 3.13

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.

A.5. Proof of Theorem 3.12

First, fix an arbitrary Markovian coupling $\mathbb{P}\in\mathcal{C}_{\mathcal{M}}(\mathbb{P}_X,\mathbb{P}_Y)$. Then, by Lemma 3.11, we know that $\pi^k:=\mathrm{law}((X_0,\ldots,X_k,Y_0,\ldots,Y_k))\in\mathcal{P}(X^{k+1}\times Y^{k+1})$ is a bicausal coupling between α^k and β^k. Therefore,

$$\int_{X^{\mathbb{N}}\times Y^{\mathbb{N}}} d_Z(\ell_X(X_k(w_X)),\ell_Y(Y_k(w_Y)))\,\mathbb{P}(dw_X\times dw_Y)=\int_{X^{k+1}\times Y^{k+1}} c^k(x,y)\,\pi^k(dx\times dy)\geq d^{c^k}(\alpha^k,\beta^k).$$

Since the Markovian coupling $\mathbb{P}$ is chosen arbitrarily, by Theorem 3.7, 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 (4) and Lemma 3.13, 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

$$W_0(x_0,y_0)=\int_{X\times Y} W_1(x_1,y_1)\,\gamma^1_{x_0,y_0}(dx_1\times dy_1).$$

This implies that $\nu^1_{\bullet,\bullet}$ is a 1-step coupling between $m^X_\bullet$ and $m^Y_\bullet$. We repeat this process inductively. Then, we have the following 1-step couplings $\nu^1_{\bullet,\bullet}, \nu^2_{\bullet,\bullet}, \ldots, \nu^k_{\bullet,\bullet}$ between $m^X_\bullet$ and $m^Y_\bullet$ such that

$$W_{i-1}(x_{i-1}, y_{i-1}) = \int_{X\times Y} W_i(x_i, y_i)\, \nu^i_{x_{i-1},y_{i-1}}(dx_i \times dy_i)$$

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 1-step couplings $\nu^1_{\bullet,\bullet}, \nu^2_{\bullet,\bullet}, \ldots, \nu^k_{\bullet,\bullet}$ by employing Kolmogorov extension theorem (cf. Remark 3.5). Therefore, again by Theorem 3.7, we have $d^{c^k}(\alpha^k, \beta^k) \geq d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$. This completes the proof.

A.6. The WL distance as COT in the label space Z

In Theorem 3.12 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:

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

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 3.3. We now let $\alpha^k_Z := (\ell^{k+1}_X)_\# \alpha^k \in \mathcal{P}(Z^{k+1})$ and $\beta^k_Z := (\ell^{k+1}_Y)_\# \beta^k \in \mathcal{P}(Z^{k+1})$. Here $\ell^{k+1}_X : X^{k+1} \to Z^{k+1}$ denotes the self product of ℓ_X, and ℓ^{k+1}_Y is similarly defined. Finally, we consider the cost $d^k_Z : Z^{k+1} \times Z^{k+1} \to \mathbb{R}$ defined by $d^k_Z((z_0, \ldots, z_k), (z'_0, \ldots, z'_k)) := d_Z(z_k, z'_k)$. Note that $d^k_Z \circ (\ell^{k+1}_X \times \ell^{k+1}_Y) = c^k$, where c^k is defined through Equation (3). Then, we have the following result:

Theorem A.1. *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*

$$d^{(k)}_{\mathrm{WL}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = d^{d^k_Z}\left(\alpha^k_Z, \beta^k_Z\right).$$

For the proof of Theorem A.1, we establish the following lemma.

Lemma A.2. *$(\ell^{k+1}_X \times \ell^{k+1}_Y)_\#$ induces a surjective map from $\mathcal{C}_{\mathrm{bc}}(\alpha^k, \beta^k)$ to $\mathcal{C}_{\mathrm{bc}}(\alpha^k_Z, \beta^k_Z)$.*

Given this lemma, and by the fact that $d^k_Z \circ (\ell^{k+1}_X \times \ell^{k+1}_Y) = c^k$, it is easy to check that $d^{c^k}(\alpha^k, \beta^k) = d^{d^k_Z}\left(\alpha^k_Z, \beta^k_Z\right)$. Hence, by emplyoing Theorem 3.12, we conclude our proof of Theorem A.1.

Finally, we provide a proof of Lemma A.2.

Proof of Lemma A.2. Fix an arbitrary $\pi^k \in \mathcal{C}_{\mathrm{bc}}(\alpha^k, \beta^k)$. Let $\pi^k_Z := (\ell^{k+1}_X \times \ell^{k+1}_Y)_\# \pi^k$. Then, it is easy to check that π^k_Z is indeed a bicausal coupling between α^k_Z and β^k_Z, so we omit the proof.

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

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

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

$$\pi^k_Z((z_0, \ldots, z_k), (z'_0, \ldots, z'_k)) = 0$$

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

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

for all $l \in \{0, \ldots, k\}$, $(y_0, \ldots, y_l) \in Y^{l+1}$, and $(x_0, \ldots, x_k) \in X^{k+1}$. 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.

□

B. Details from Section 4

B.1. Details about Markov chain neural networks (MCNNs)

We first briefly recall the definition of MCNNs from (Chen et al., 2022).

Given any Lipschitz map $\varphi : \mathbb{R}^i \to \mathbb{R}^j$, define the map $q_\varphi : \mathcal{P}(\mathbb{R}^i) \to \mathbb{R}^j$ by sending α to $\int_{\mathbb{R}^i} \varphi(x)\alpha(dx)$.

One layer of MCNN We define one layer of MCNN as follows: $F_\varphi : \mathcal{M}^L(\mathbb{R}^i) \to \mathcal{M}^L(\mathbb{R}^j)$ sends $(\mathcal{X}, \ell_X : X \to \mathbb{R}^i)$ to $(\mathcal{X}, \ell_X^\varphi : X \to \mathbb{R}^j)$ where $\ell_X^\varphi(x) := q_\varphi((\ell_X)_\# m_x^X)$ for each $x \in X$.

Readout layer We define a readout layer for MCNN as follows: we first let $S_\varphi : \mathcal{M}^L(\mathbb{R}^i) \to \mathbb{R}^j$ be defined via $(\mathcal{X}, \ell_X) \mapsto q_\varphi((\ell_X)_\# \mu_X)$; then we consider any Lipschitz $\psi : \mathbb{R}^j \to \mathbb{R}$ and call $\psi \circ S_\varphi : \mathcal{M}^L(\mathbb{R}^i) \to \mathbb{R}$ a readout layer.

k-layer MCNNs Now, a k-layer MCNN is defined as follows. Given a sequence of MLPs $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ for $i = 1, \ldots, k+1$ and a MLP $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$, the map of the following form is called a k-layer MCNN:

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

We let $\mathcal{NN}_k(\mathbb{R}^d)$ denote the collection of all k-layer MCNNs.

MCNNs are MP-GNNs 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 in Section 4.2. Let $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ for $i = 1, \ldots, k+1$ and $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$ be the maps involved in defining h. Then, we consider the following k-layer MCNN:

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

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 4.1. 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 4.2 is a restriction of an MCNN to the collection of graph induced LMMCs $\mathcal{G}_q(\mathbb{R}^d)$, i.e.,

$$h = H \circ I_q. \tag{6}$$

Here recall that $I_q : \mathcal{G}(\mathbb{R}^d) \to \mathcal{M}^L(\mathbb{R}^d)$ sends a labeled graph to a LMMC and I_q has image $\mathcal{G}_q(\mathbb{R}^d)$.

Universal approximation property of MCNNs Given any $k \in \mathbb{N}$ and any subset $\mathcal{K} \subseteq \left(\mathcal{G}(\mathbb{R}^d), d_{\mathcal{G},q}^{(k)}\right)$, we let $\mathcal{NN}_k^q(\mathbb{R}^d)|_{\mathcal{K}}$ denote the collection of restrictions of functions $h \in \mathcal{NN}_k(\mathbb{R}^d)$ to $\mathcal{K}$, i.e.,

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

The following theorem is a slight variant of (Chen et al., 2022, Theorem 4.3): Notice that $d_{\mathrm{WL}}^{(k)}$ is a pseudo-distance on $\mathcal{M}^L(\mathbb{R}^d)$. Hence in (Chen et al., 2022, Theorem 4.3), all pseudometric spaces are first transformed to metric spaces by identifying points at 0 distance (cf. (Burago et al., 2001, Proposition 1.1.5)). In the following result, we remove this subtlety and deal with the pseudo-distance topology directly.

Theorem B.1. *For any $k \in \mathbb{N}$ and any compact subset $\mathcal{K} \subseteq \left(\mathcal{M}^L(\mathbb{R}^d), d_{\mathrm{WL}}^{(k)}\right)$, one has that $\overline{\mathcal{NN}_k(\mathbb{R}^d)|_{\mathcal{K}}} = C(\mathcal{K}, \mathbb{R})$.*

Proof. We let $\mathcal{M}_k^L(\mathbb{R}^d)$ denote the space obtained by identifying points at 0 distance (w.r.t $d_{\mathrm{WL}}^{(k)}$) in $\mathcal{M}^L(\mathbb{R}^d)$. Then, $d_{\mathrm{WL}}^{(k)}$ become a metric on $\mathcal{M}_k^L(\mathbb{R}^d)$. Let Q denote this identification map (also called the *quotient* map):

$$Q : \left(\mathcal{M}^L(\mathbb{R}^d), d_{\mathrm{WL}}^{(k)}\right) \to \left(\mathcal{M}_k^L(\mathbb{R}^d), d_{\mathrm{WL}}^{(k)}\right). \tag{7}$$

Let $\mathcal{K}_Q := Q(\mathcal{K})$. Then, $\mathcal{K}_Q$ is compact in $\mathcal{M}_k^L(\mathbb{R}^d)$. Consider the map $Q_c : C(\mathcal{K}_Q, \mathbb{R}) \to C(\mathcal{K}, \mathbb{R})$ defined by sending f_Q to $f_Q \circ Q$. It is easy to check that Q_c is an isometric embedding w.r.t. the sup metrics on $C(\mathcal{K}_Q, \mathbb{R})$ and $C(\mathcal{K}, \mathbb{R})$. It is easy to check that Q_c has its isometric inverse $(Q_c)^{-1} : C(\mathcal{K}, \mathbb{R}) \to C(\mathcal{K}_Q, \mathbb{R})$ defined as follows, which is also an isometric embedding: for any $f \in C(\mathcal{K}, \mathbb{R})$, we let $f_Q : \mathcal{K}_Q \to \mathbb{R}$ be defined such that for any equivalence class $[(\mathcal{X}, \ell_X)] \in \mathcal{K}_Q$, $f_Q([(\mathcal{X}, \ell_X)]) := f((\mathcal{X}, \ell_X))$ (it is obvious that f_Q does not depend on the choice of the representative $(\mathcal{X}, \ell_X)$.). Hence, Q_c gives rise to an isometry from $C(\mathcal{K}_Q, \mathbb{R})$ to $C(\mathcal{K}, \mathbb{R})$.

Now, by (Chen et al., 2022, Theorem 4.3), one has that

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

Since Q_c is an isometry, we have that $\overline{\mathcal{NN}_k(\mathbb{R}^d)|_{\mathcal{K}}} = C(\mathcal{K}, \mathbb{R})$. ☐

B.2. Proof of Theorem 4.3

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

Lemma B.2. *We let $P_S := \{\sum_{i=1}^n p_i : p_1, \ldots, p_n \in P\}$ denote the collection of finite sums of elements in P. Then, for a fixed $q \in (0,1)$, let*

$$\begin{aligned} Q :=& \{1\} \cup \left\{ \frac{\overline{\deg}(v)}{\sum_{v' \in V} \overline{\deg}(v')} : G \text{ has edge weights in } P,\, v \in V \right\} \\ & \cup \left\{ mq + s\frac{(1-q)}{\deg(v)} : G \text{ has edge weights in } P,\, v \in V \text{ is such that } \deg(v) > 0, m \in \mathbb{N},\, s \in P_S \right\}. \end{aligned}$$

Then, Q is a countable set.

Proof. Note that P_S is still a countable set. Since graphs have edge weights from a countable set P, every $\deg(v)$ (and $\overline{\deg}(v)$) encountered in the definition above is from a countable set. Hence, Q is countable. ☐

Lemma B.3. *Consider the union*

$$C_P(Z) := \bigcup_{(G,\ell_G) \in \mathcal{G}_P(Z)} \left(\{(\ell_G)_\# \mu_G\} \cup \left\{ (\ell_G)_\# m_v^{G,q} : \, v \in V \right\} \right) \subseteq \mathcal{P}(\mathbb{R}^d)$$

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

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

Then, by the fact that both Z and Q are countable (cf. Lemma B.2), we have that $C_P(Z)$ is countable. $\square$

Recall notation from Appendix B.1. Then, we have that

Lemma B.4. *Fix any $d \in \mathbb{N}$. Then, for any Lipschitz $\varphi : \mathbb{R}^d \to \mathbb{R}$, any countable subset $Z \subseteq \mathbb{R}^d$ and any countable subset $P \subseteq \mathbb{R}$, there exists a countable subset $Z' \subseteq \mathbb{R}$ such that*

$$F_\varphi(\mathcal{G}_P(Z)) \subseteq \mathcal{G}_P(Z') \text{ and } S_\varphi(\mathcal{G}_P(Z)) \subseteq Z'. \tag{8}$$

Proof. Let $Z' := q_\varphi(C_P(Z)) \subseteq \mathbb{R}$. Since $C_P(Z)$ is countable, Z' is also countable. It is obvious that Z' satisfies Equation (8) which concludes the proof. $\square$

Lemma B.5. *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:

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

When equipped with the norm $\|f\| := \sup_{x \neq y} \frac{|f(x)-f(y)|}{\|x-y\|}$, $\mathrm{Lip}_0(\mathbb{R}^d)$ is a Banach space (Weaver, 1995). Consider the following collection of signed measures $D := \{\alpha - \beta : \alpha, \beta \in C \text{ and } \alpha \neq \beta\}$. 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 2.1 for this assumption on $\mathcal{P}(\mathbb{R}^d)$). Moreover, it is easy to check that ψ_μ is bounded and 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. $\square$

Proof of Theorem 4.3. We let $Z_0 := Z$ and $C_1 := C_P(Z_0)$. By Lemma B.5, there exists a Lipschitz map $\varphi_1 : \mathbb{R}^d \to \mathbb{R}$ such that $q_{\varphi_1}|_{C_1}$ is injective. By Lemma B.4, there exists a countable subset $Z_1 \subseteq \mathbb{R}$ such that

$$F_{\varphi_1}(\mathcal{G}_P(Z_0)) \subseteq \mathcal{G}_P(Z_1).$$

Let $C_2 := C_P(Z_1)$. By Lemma B.5 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 $C_i := C_P(Z_{i-1})$ and $F_{\varphi_i}(\mathcal{G}_P(Z_{i-1})) \subseteq \mathcal{G}_P(Z_i)$. 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 $C_{k+1} := C_P(Z_k)$ and $S_{\varphi_{k+1}}(\mathcal{G}_P(Z_k)) \subseteq S$. We then let $\psi : \mathbb{R} \to \mathbb{R}$ denote the identity map and let $h := \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} : \mathcal{G}(\mathbb{R}^d) \to \mathbb{R}$.

Pick any $(G_1, \ell_{G_1}), (G_2, \ell_{G_2}) \in \mathcal{G}_P(Z)$. We prove that $d^{(k)}_{\mathcal{G},q}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) > 0$ iff $h((G_1, \ell_{G_1})) \neq h((G_2, \ell_{G_2}))$.

For each $i = 1, \ldots, k$, we let

$$\left(G_1, \ell^{(\varphi,i)}_{G_1}\right) := F_{\varphi_i} \circ \cdots \circ F_{\varphi_1}((G_1, \ell_{G_1})). \tag{9}$$

We similarly define $\left(G_2, \ell^{(\varphi,i)}_{G_2}\right)$. Then, we prove that for each $i = 1, \ldots, k$,

$$\forall v_1 \in V_{G_1},\, v_2 \in V_{G_2},\; \ell^{(\varphi,i)}_{G_1}(v_1) = \ell^{(\varphi,i)}_{G_2}(v_2) \text{ iff } \mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2) \tag{10}$$

Given Equation (10), it is obvious that

$$\left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}\right)_\# \mu_{G_1} \neq \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}\right)_\# \mu_{G_2} \text{ iff } \left(\ell^{(\varphi,k)}_{G_1}\right)_\# \mu_{G_1} \neq \left(\ell^{(\varphi,k)}_{G_2}\right)_\# \mu_{G_2}.$$

Note that $\left(\ell_{G_1}^{(\varphi,k)}\right)_\# \mu_{G_1}, \left(\ell_{G_2}^{(\varphi,k)}\right)_\# \mu_{G_2} \in C_{k+1}$. Since $q_{\varphi_{k+1}}$ is injective on C_{k+1} and ψ is injective on S, we have that $\left(\ell_{G_1}^{(\varphi,k)}\right)_\# \mu_{G_1} \neq \left(\ell_{G_2}^{(\varphi,k)}\right)_\# \mu_{G_2}$ iff $h((G_1,\ell_{G_1})) \neq h((G_2,\ell_{G_2}))$. Therefore, $d_{\mathcal{G},q}^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2})) > 0$ iff $h((G_1,\ell_{G_1})) \neq h((G_2,\ell_{G_2}))$.

To conclude the proof, we prove Equation (10) by induction on $i = 1,\ldots,k$. When $i = 1$, since $(\ell_{G_1})_\# m_{v_1}^{G_1,q}, (\ell_{G_2})_\# m_{v_2}^{G_2,q} \in C_1$, $\forall v_1 \in V_{G_1}$, $v_2 \in V_{G_2}$, by injectivity of q_{φ_1} on C_1, we have that

$$\forall v_1 \in V_{G_1},\, v_2 \in V_{G_2},\; (\ell_{G_1})_\# m_{v_1}^{G_1,q} = (\ell_{G_2})_\# m_{v_2}^{G_2,q} \text{ iff } q_{\varphi_1}\big((\ell_{G_1})_\# m_{v_1}^{G_1,q}\big) = q_{\varphi_1}\big((\ell_{G_2})_\# m_{v_2}^{G_2,q}\big).$$

Equivalent speaking,

$$\forall v_1 \in V_{G_1},\, v_2 \in V_{G_2},\; \mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2) \text{ iff } \ell_{G_1}^{(\varphi,1)}(x) = \ell_{G_2}^{(\varphi,1)}(v_2).$$

Now, we assume that Equation (10) holds for some $i \geq 1$. For $i+1$, since

$$\left(\ell_{G_1}^{(\varphi,i)}\right)_\# m_{v_1}^{G_1,q}, \left(\ell_{G_2}^{(\varphi,i)}\right)_\# m_{v_2}^{G_2,q} \in C_{i+1},\, \forall v_1 \in V_{G_1},\, v_2 \in V_{G_2},$$

by injectivity of $q_{\varphi_{i+1}}$ on C_{i+1}, we have that $\forall v_1 \in V_{G_1}, v_2 \in V_{G_2}$,

$$\left(\ell_{G_1}^{(\varphi,i)}\right)_\# m_{v_1}^{G_1,q} = \left(\ell_{G_2}^{(\varphi,i)}\right)_\# m_{v_2}^{G_2,q} \text{ iff } q_{\varphi_{i+1}}\left(\left(\ell_{G_1}^{(\varphi,i)}\right)_\# m_{v_1}^{G_1,q}\right) = q_{\varphi_{i+1}}\left(\left(\ell_{G_2}^{(\varphi,i)}\right)_\# m_{v_2}^{G_2,q}\right).$$

Equivalent speaking,

$$\forall v_1 \in V_{G_1},\, y \in Y,\; \left(\ell_{G_1}^{(\varphi,i)}\right)_\# m_{v_1}^{G_1,q} = \left(\ell_{G_2}^{(\varphi,i)}\right)_\# m_{v_2}^{G_2,q} \text{ iff } \ell_{G_1}^{(\varphi,i+1)}(x) = \ell_{G_2}^{(\varphi,i+1)}(y).$$

By the induction assumption, $\forall v_1 \in V_{G_1}$, $v_2 \in V_2$ we have that

$$\ell_{G_1}^{(\varphi,i)}(v_1) = \ell_{G_2}^{(\varphi,i)}(v_2) \text{ iff } \mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2).$$

This implies that

$$\begin{aligned} &\left(\ell_{G_1}^{(\varphi,i)}\right)_\# m_{v_1}^{G_1,q} = \left(\ell_{G_2}^{(\varphi,i)}\right)_\# m_{v_2}^{G_2,q} \\ \text{iff } &\left(\mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}\right)_\# m_{v_1}^{G_1,q} = \left(\mathfrak{l}^{(i)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}\right)_\# m_{v_2}^{G_2,q} \\ \text{iff } &\mathfrak{l}^{(i+1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(i+1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2). \end{aligned}$$

Therefore,

$$\ell_{G_1}^{(\varphi,i+1)}(v_1) = \ell_{G_2}^{(\varphi,i+1)}(v_2) \text{ iff } \mathfrak{l}^{(i+1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(i+1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$$

and we thus conclude the proof. □

B.3. Proof of Theorem 4.5

We need the following basic fact.

Lemma B.6 ((Chen et al., 2022, Lemma B.4))**.** *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.*

Recall notations from Appendix B.1. Given a k-layer MCNN $H := \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1}$, recall from Equation (6) that $h := H \circ I_q$ gives rise to a MP-GNN. For any $(G,\ell_G) \in \mathcal{G}(Z)$, consider its induced LMMC $(\mathcal{X}_q(G),\ell_G)$. Following notation in Equation (9) we let

$$\left(\mathcal{X}_q(G), \ell_G^{(\varphi,i)}\right) := F_{\varphi_i} \circ \cdots \circ F_{\varphi_1}((\mathcal{X}_q(G),\ell_G)) \tag{11}$$

Now, we assume that for $i = 1,\ldots,k$, φ_i is a C_i-Lipschitz MLP for some $C_i > 0$. Then, by Lemma B.6, we have that q_{φ_i} is a C_i-Lipschitz map for $i = 1,\ldots,k$.

Then, we prove that

$$d_{\mathrm{W}}\left(\left(\ell_{G_1}^{(\varphi,k)}\right)_{\#}\mu_{G_1},\left(\ell_{G_2}^{(\varphi,k)}\right)_{\#}\mu_{G_2}\right)\leq \Pi_{i=1}^{k}C_i\cdot d_{\mathcal{G},q}^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2})). \tag{12}$$

Given Equation (12), by Lemma B.6 again and the fact that ψ is C-Lipschitz, one has that

$$\begin{aligned}|h((G_1,\ell_{G_1}))-h((G_2,\ell_{G_2}))|&\leq C\cdot\left\|q_{\varphi_{k+1}}\left(\left(\ell_{G_1}^{(\varphi,k)}\right)_{\#}\mu_{G_1}\right)-q_{\varphi_{k+1}}\left(\left(\ell_{G_2}^{(\varphi,k)}\right)_{\#}\mu_{G_2}\right)\right\|\\&\leq C\cdot C_{k+1}d_{\mathrm{W}}\left(\left(\ell_{G_1}^{(\varphi,k)}\right)_{\#}\mu_{G_1},\left(\ell_{G_2}^{(\varphi,k)}\right)_{\#}\mu_{G_2}\right)\\&\leq C\cdot\Pi_{i=1}^{k+1}C_i\cdot d_{\mathcal{G},q}^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2})).\end{aligned}$$

The following proof of Equation (12) is adapted from the proof of (Chen et al., 2022, Equation (19)): it suffices to prove that for any $v_1\in V_{G_1}$ and $v_2\in V_{G_2}$,

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

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

$$\left\|\ell_{G_1}^{(\varphi,j)}(v_1)-\ell_{G_2}^{(\varphi,j)}(v_2)\right\|\leq\Pi_{i=1}^{j}C_i\cdot d_{\mathrm{W}}\left(\mathfrak{l}^{(j)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1),\mathfrak{l}^{(j)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)\right). \tag{13}$$

When $j=1$, we have that

$$\begin{aligned}\left\|\ell_{G_1}^{(\varphi,1)}(v_1)-\ell_{G_2}^{(\varphi,1)}(v_2)\right\|&=\left\|q_{\varphi_1}\left((\ell_{G_1})_{\#}m_{v_1}^{G_1,q}\right)-q_{\varphi_j}\left((\ell_{G_2})_{\#}m_{v_2}^{G_2,q}\right)\right\|\\&\leq C_1 d_{\mathrm{W}}\left((\ell_{G_1})_{\#}m_{v_1}^{G_1,q},(\ell_{G_2})_{\#}m_{v_2}^{G_2,q}\right)\\&=C_1 d_{\mathrm{W}}\left(\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1),\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)\right).\end{aligned}$$

We now assume that Equation (13) holds for some $j\geq 1$. For $j+1$, we have that

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

B.4. More on Lipschitz properties

In Section 4.1, 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 pseudo-distance, which we denote by $d_{\mathcal{G},(\varepsilon)}^{(k)}$, on $\mathcal{G}(Z)$. Note the subtle differences in notation introduced in Section 4.1.

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 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. $d_{\mathcal{G},(\varepsilon)}^{(k)}$.

Theorem B.7. *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*

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

Proof. The proof is similar to the one for Theorem 4.5.

We first prove that

$$d_{\mathrm{W}}\left(\left(\ell_{G_1}^k\right)_\# \mu_{G_1}^\varepsilon, \left(\ell_{G_2}^k\right)_\# \mu_{G_2}^\varepsilon\right) \leq \Pi_{i=1}^k C_i \cdot d_{\mathcal{G},(\varepsilon)}^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2})). \tag{14}$$

Given Equation (14) and the fact that ψ is C-Lipschitz, one has that

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

where the second inequality follows Lemma B.6 by letting $\varphi = \mathrm{id}_{\mathbb{R}^{d_k}}$.

To prove Equation (14), it suffices to prove that for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$,

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

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

$$\left\|\ell_{G_1}^j(v_1) - \ell_{G_2}^j(v_2)\right\| \leq \Pi_{i=1}^j C_i \cdot d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}_{(\varepsilon)}(G_1),\ell_{G_1})}^{(j)}(v_1), \mathfrak{l}_{(\mathcal{X}_{(\varepsilon)}(G_2),\ell_{G_2})}^{(j)}(v_2)\right). \tag{15}$$

When $j=1$, we have that

$$\begin{aligned}
&\left\|\ell^1_{G_1}(v_1)-\ell^1_{G_2}(v_2)\right\|\\
&\leq C_1\left\|\frac{(1+\varepsilon)\ell_{G_1}(v_1)+\sum_{v_1'\in N_{G_1}(v_1)}w_{v_1v_1'}\ell_{G_1}(v_1')}{\deg^\varepsilon_{G_1}(v_1)}-\frac{(1+\varepsilon)\ell_{G_2}(v_2)+\sum_{v_2'\in N_{G_2}(v_2)}w_{v_2v_2'}\ell_{G_2}(v_2')}{\deg^\varepsilon_{G_2}(v_2)}\right\|\\
&=C_1\left\|q_{\mathrm{Id}}\left((\ell_{G_1})_\# m_{v_1}^{G_1,(\varepsilon)}\right)-q_{\mathrm{Id}}\left((\ell_{G_2})_\# m_{v_1}^{G_2,(\varepsilon)}\right)\right\|\\
&\leq C_1 d_{\mathrm{W}}\left((\ell_{G_1})_\# m_{v_1}^{G_1,(\varepsilon)},(\ell_{G_2})_\# m_{v_1}^{G_2,(\varepsilon)}\right)=C_1 d_{\mathrm{W}}\left(\mathfrak{l}^{(1)}_{(\mathcal{X}_{(\varepsilon)}(G_1),\ell_{G_1})}(v_1),\mathfrak{l}^{(1)}_{(\mathcal{X}_{(\varepsilon)}(G_2),\ell_{G_2})}(v_2)\right).
\end{aligned}$$

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 (15) holds for some $j\geq 1$. For $j+1$, we have that

$$\begin{aligned}
&\left\|\ell^{j+1}_{G_1}(v_1)-\ell^{j+1}_{G_2}(v_2)\right\|\\
&\leq C_{j+1}\left\|\frac{(1+\varepsilon)\ell^j_{G_1}(v_1)+\sum_{v_1'\in N_{G_1}(v_1)}w_{v_1v_1'}\ell^j_{G_1}(v_1')}{\deg^\varepsilon_{G_1}(v_1)}-\frac{(1+\varepsilon)\ell^j_{G_2}(v_2)+\sum_{v_2'\in N_{G_2}(v_2)}w_{v_2v_2'}\ell^j_{G_2}(v_2')}{\deg^\varepsilon_{G_2}(v_2)}\right\|\\
&=C_{j+1}\left\|q_{\mathrm{Id}}\left((\ell^j_{G_1})_\# m_{v_1}^{G_1,(\varepsilon)}\right)-q_{\mathrm{Id}}\left((\ell^j_{G_2})_\# m_{v_1}^{G_2,(\varepsilon)}\right)\right\|\\
&\leq C_{j+1} d_{\mathrm{W}}\left((\ell^j_{G_1})_\# m_{v_1}^{G_1,(\varepsilon)},(\ell^j_{G_2})_\# m_{v_1}^{G_2,(\varepsilon)}\right)\\
&=C_{j+1}\inf_{\gamma\in\mathcal{C}\left(m_{v_1}^{G_1,(\varepsilon)},m_{v_2}^{G_2,(\varepsilon)}\right)}\int_{V_{G_1}\times V_{G_2}}\left\|\ell^j_{G_1}(v_1')-\ell^j_{G_2}(v_2')\right\|\gamma(dv_1'\times dv_2')\\
&\leq C_{j+1}\cdot\inf_{\gamma\in\mathcal{C}\left(m_{v_1}^{G_1,(\varepsilon)},m_{v_2}^{G_2,(\varepsilon)}\right)}\int_{V_{G_1}\times V_{G_2}}\Pi_{i=1}^j C_i\cdot d_{\mathrm{W}}\left(\mathfrak{l}^{(j)}_{(\mathcal{X}_{(\varepsilon)}(G_1),\ell_{G_1})}(v_1'),\mathfrak{l}^{(j)}_{(\mathcal{X}_{(\varepsilon)}(G_2),\ell_{G_2})}(v_2')\right)\gamma(dv_1'\times dv_2')\\
&=\Pi_{i=1}^{j+1}C_i d_{\mathrm{W}}\left(\mathfrak{l}^{(j+1)}_{(\mathcal{X}_{(\varepsilon)}(G_1),\ell_{G_1})}(v_1),\mathfrak{l}^{(j+1)}_{(\mathcal{X}_{(\varepsilon)}(G_2),\ell_{G_2})}(v_2)\right).
\end{aligned}$$

This concludes the proof.

□

B.5. Proof of Theorem 4.7

Recall the map $I_q:\left(\mathcal{G}(\mathbb{R}^d),d^{(k)}_{\mathcal{G},q}\right)\to\left(\mathcal{G}_q(\mathbb{R}^d),d^{(k)}_{\mathrm{WL}}\right)$. This map is continuous by the definition of $d^{(k)}_{\mathcal{G},q}$. Hence $\mathcal{K}_q:=I_q(\mathcal{K})$ is compact. Then, we define the map $J_q:C(\mathcal{K}_q,\mathbb{R})\to C(\mathcal{K},\mathbb{R})$ sending f_q to $f:=f_q\circ I_q$.

Claim 1. $J_q:C(\mathcal{K}_q,\mathbb{R})\to C(\mathcal{K},\mathbb{R})$ is a homeomorphism.

By Theorem B.1, we know that $\overline{\mathcal{NN}_k(\mathbb{R}^d)|_{\mathcal{K}_q}}=C(\mathcal{K}_q,\mathbb{R})$. Here $\mathcal{NN}_k(\mathbb{R}^d)$ denotes the collection of k-layer MCNNs (see Appendix B.1) and $\mathcal{NN}_k(\mathbb{R}^d)|_{\mathcal{K}_q}$ refers to the collection of restrictions of functions in $\mathcal{NN}_k(\mathbb{R}^d)$ to $\mathcal{K}_q$. By Equation (6), we have that $\mathcal{NN}^q_k(\mathbb{R}^d)|_{\mathcal{K}}=J_q(\mathcal{NN}_k(\mathbb{R}^d)|_{\mathcal{K}_q})$. Hence, by Claim 1 we have that

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

Proof of Claim 1. We show that, in fact, J_q is an isometry w.r.t. the sup metric on $C(\mathcal{K}_q,\mathbb{R})$ and $C(\mathcal{K},\mathbb{R})$. Given continuous $f_q,g_q:\mathcal{K}_q\to\mathbb{R}$, one has that

$$\begin{aligned}
&\sup_{(\mathcal{X}_q(G),\ell_G)\in\mathcal{K}_q}|f_q((\mathcal{X}_q(G),\ell_G))-g_q((\mathcal{X}_q(G),\ell_G))|\\
=&\sup_{(\mathcal{X}_q(G),\ell_G)\in\mathcal{K}_q}|f((G,\ell_G))-g((G,\ell_G))|\\
=&\sup_{(G,\ell_G)\in\mathcal{K}}|f((G,\ell_G))-g((G,\ell_G))|.
\end{aligned}$$

Hence, J_q is an isometric embedding. The inverse $(J_q)^{-1}$ of J_q is identified as follows: for any $f \in C(\mathcal{K}, \mathbb{R})$, we let $f_q : \mathcal{K}_q \to \mathbb{R}$ be defined such that for any $(\mathcal{X}_q(G), \ell_G) \in \mathcal{K}_q$, $f_q((\mathcal{X}_q(G), \ell_G)) := f((G, \ell_G))$ (f_q is well-defined: if $(\mathcal{X}_q(G_1), \ell_{G_1}) = (\mathcal{X}_q(G_2), \ell_{G_2})$, then $d^{(k)}_{\mathcal{G},q}((G_1, \ell_{G_1}), (G_2, \ell_{G_2})) = 0$ and thus $f((\mathcal{X}_q(G_1), \ell_{G_1})) = f((\mathcal{X}_q(G_2), \ell_{G_2}))$.). It is easy to check that $(J_q)^{-1}$ is the inverse of J_q and is also an isometric embedding. This finishes the proof. $\square$