Weisfeiler-Lehman Meets Gromov-Wasserstein

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

Abstract

The Weisfeiler-Lehman (WL) test is a classical procedure for graph isomorphism testing. The WL test has also been widely used both for designing graph kernels and for analyzing graph neural networks. In this paper, we propose the *Weisfeiler-Lehman (WL) distance*, a notion of distance between *labeled measure Markov chains* (LMMCs), of which labeled graphs are special cases. The WL distance is polynomial time computable and is also compatible with the WL test in the sense that the former is positive if and only if the WL test can distinguish the two involved graphs. The WL distance captures and compares subtle structures of the underlying LMMCs and, as a consequence of this, it is more discriminating than the distance between graphs used for defining the state-of-the-art Wasserstein Weisfeiler-Lehman graph kernel. Inspired by the structure of the WL distance we identify a neural network architecture on LMMCs which turns out to be universal w.r.t. continuous functions defined on the space of all LMMCs (which includes all graphs) endowed with the WL distance. Finally, the WL distance turns out to be stable w.r.t. a natural variant of the Gromov-Wasserstein (GW) distance for comparing metric Markov chains that we identify. Hence, the WL distance can also be construed as a polynomial time lower bound for the GW distance which is in general NP-hard to compute.

[*]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 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

The Weisfeiler-Lehman (WL) test (Lehman & Weisfeiler, 1968) is a classical procedure which provides a polynomial time proxy for testing graph isomorphism. It is efficient and can distinguish most pairs of graphs in linear time (Babai & Kucera, 1979; Babai & Luks, 1983). The WL test has a close relationship with graph neural networks (GNNs), both in the design of GNN architectures and in terms of characterizing their expressive power. For example, Xu et al. (2018) showed that graph isomorphism networks (GINs) have the same discriminative power as the WL test in distinguishing whether two graphs are isomorphic or not. Recently, Azizian et al. (2020) showed that message passing graph neural networks (MPNNs) are universal with respect to the continuous functions defined on the set of graphs (with the topology induced by a specific variant of the graph edit distance) that have equivalent to or less discriminative power than the WL test.

However, the WL test is only suitable for testing graph isomorphism and cannot directly *quantitatively* compare graphs. This state of affairs naturally suggests identifying a distance function between graphs so that two graphs have positive distance iff they can be distinguished by the WL test. We note that there have been WL-inspired graph kernels which can quantitatively compare graphs (Shervashidze et al., 2011; Togninalli et al., 2019). However, these either cannot handle continuous node features naturally or they do not have the same discriminative power as the WL test.

New work and connections to related work. Our work provides novel connections between the WL test, GNNs and the Gromov-Wasserstein distance. The central object we define in this paper is a distance between graphs *which has the same discriminative power as the WL test*. We do this by combining ideas inherent to the WL test with optimal transport (OT) (Villani, 2009). We call this distance the *Weisfeiler-Lehman (WL) distance*. We show that two graphs are at zero WL distance if and only if they cannot be distinguished by the WL test. Moreover, the WL distance can be computed in polynomial time. Furthermore, our WL distance is actually defined on a more general and flexible type of objects called the *labeled measure Markov chains* (LMMCs), of which labeled graphs (i.e, graph with node features) are special cases. LMMCs can naturally model

the interaction between graphs and their node labels (node features), and thus provides a new OT based perspective for comparing node labeled graphs. Besides graphs, the LMMC framework also encompasses continuous objects such as Riemannian manifolds and graphons. It is worth noting that the idea of combining OT and Markov chains/heat kernels has long been used to study notions of curvature of geometric objects, e.g., graphs or Riemannian manifolds (von Renesse & Sturm, 2005; Ollivier, 2009).

Our definition of the WL distance is able to capture and compare subtle geometric and combinatorial structures from the underlying LMMCs. This allows us to establish various lower bounds for the WL distance which are not just useful in practical computations but also clarify its discriminating power relative to existing approaches. In particular, we carry out experiments which demonstrate the effectiveness of our WL distance (and its lower bounds) in graph comparison tasks.[1] Furthermore, based on the hierarchy inherent to the WL distance, we are able to identify a neural network architecture on the collection of all LMMCs, which we call MCNNs (for Markov chain NNs). We show that MCNNs have the same discriminative power as the WL test when applied to graphs; while at the same time, they have the desired universal approximation property w.r.t. continuous functions defined on the space of all LMMCs (including the space of graphs) equipped with the WL distance. It turns out that from MCNNs, one can recover Weisfeiler-Lehman graph kernels (Togninalli et al., 2019) and in particular, we show that a slight variant of a key pseudo-distance between graphs defined in (Togninalli et al., 2019) serves as a lower bound for our WL distance. This indicates that the WL distance has a stronger discriminating ability than WWL graph kernels (see Appendix A.3 for an infinite family of examples).

Finally, we observe that our formulation of the WL distance resembles the Gromov-Wasserstein (GW) distance (Mémoli, 2007; 2011; Peyré et al., 2016; Sturm, 2012; Vayer et al., 2020; Chowdhury & Mémoli, 2019) which is a OT-based distance between *metric measure spaces* and has been recently widely used in shape matching and machine learning. We hence identify a special variant of the GW distance between *Markov chain metric spaces* (MCMSs) (including all graphs). Our version of the GW distance implements a certain *multiscale comparison* of MCMSs, it vanishes only when the two MCMSs are isomorphic, but leads to NP-hard problems. Interestingly, it turns out that the poly-time computable WL distance is not only stable w.r.t. (i.e., upper bounded by) this variant of the GW distance, but can also be construed as a variant of the *third lower bound* (TLB) of this GW distance, as in (Mémoli, 2011).

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

Proofs of results and details can be found in the Appendix.

2. Preliminaries

2.1. The Weisfeiler-Lehman Test

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

Definition 1 (Weisfeiler-Lehman hierarchy)**.** *Given any labeled graph (G, ℓ_G), we consider the following hierarchy of multisets, which we call the Weisfeiler-Lehman hierarchy:*

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

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

...

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

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

Here, $\{\!\{\cdot\}\!\}$ denotes multisets. In the literature, $\ell^{(k)}_{(G,\ell_G)}(v)$ is usually often mapped to a common space of labels such as $\mathbb{N}$ through a hash function, a step which we do not require in this paper. We induce, at each step k, a multiset

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

Definition 2 (Weisfeiler-Lehman test)**.** *For each integer $k \geq 0$, we compare $L_k((G_1, \ell_{G_1}))$ with $L_k((G_2, \ell_{G_2}))$. If $\exists k \geq 0$ so that $L_k((G_1, \ell_{G_1})) \neq L_k((G_2, \ell_{G_2}))$ then we conclude that the two label graphs are non-isomorphic; otherwise we say that the two labeled graphs pass the WL test and that the two graphs are "possibly isomorphic".*

2.2. Probability Measures and Optimal Transport

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

Pushforward Maps. Given two measurable spaces X and Y and a measurable map $\psi : X \to Y$, the *pushforward* map induced by ψ is the map $\psi_\# : \mathcal{P}(X) \to \mathcal{P}(Y)$ sending α to $\psi_\#\alpha$ where for any measurable $B \subseteq Y$, $\psi_\#\alpha(B) :=$

$\alpha\left(\psi^{-1}(B)\right)$. In the case when X is finite and Y is a metric space, $\psi_{\#}\alpha$ obviously has finite 1-moment and is thus an element of $\mathcal{P}(Y)$.

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

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

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

By (Villani, 2009, Proposition 2.1), the infimum above is always achieved by some $\gamma \in \mathcal{C}(\alpha, \beta)$ which we call an *optimal coupling* between α and β.

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

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

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

The Gromov-Wasserstein Distance. We call a triple $\mathbf{X} = (X, d_X, \mu_X)$ a *metric measure space* (MMS) if (X, d_X) is a metric space and μ_X is a (Borel) probability measure on X with full support. Given any $\mathbf{X} = (X, d_X, \mu_X)$ and $\mathbf{Y} = (Y, d_Y, \mu_Y)$, for any coupling $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$, we define its distortion by

$$\mathrm{dis}(\gamma) \coloneqq \int_{X\times Y}\int_{X\times Y} |d_X(x, x') - d_Y(y, y')|\gamma(dx \times dy)\gamma(dx' \times dy').$$

Then, the (ℓ^1-)Gromov-Wasserstein (GW) distance between $\mathbf{X} = (X, d_X, \mu_X)$ and $\mathbf{Y} = (Y, d_Y, \mu_Y)$ is defined as follows (Mémoli, 2011)

$$d_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) \coloneqq \inf_{\gamma \in \mathcal{C}(\mu_X,\mu_Y)} \mathrm{dis}(\gamma), \tag{1}$$

where we omit the usual $\frac{1}{2}$ factor for simplicity.

2.3. Markov Chains

Given a *finite* set X, we call any map $m_\bullet^X : X \to \mathcal{P}(X)$ a *Markov kernel* on X. Of course Markov kernels can be represented as transition matrices but we adopt this more flexible language. A probability measure $\mu_X \in \mathcal{P}(X)$ is called a *stationary distribution w.r.t.* $m_\bullet^X$ if for every measurable subset $A \subseteq X$ we have:

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

The existence of stationary distributions is guaranteed by the Perron-Frobenius Theorem (Saloff-Coste, 1997). A *measure Markov chain (MMC)* is any tuple $\mathcal{X} = (X, m_\bullet^X, \mu_X)$ where X is a finite set, $m_\bullet^X$ is a Markov kernel on X and μ_X is a fully supported stationary distribution w.r.t. $m_\bullet^X$.

Definition 3 (Labeled measure Markov chain)**.** *Given any metric space Z, which we refer to as the* metric space of labels, *a Z-*labeled measure Markov chain *((Z-)LMMC for short) is a tuple $(\mathcal{X}, \ell_X)$ where $\mathcal{X}$ is a MMC and $\ell_X : X \to Z$ is a continuous map. For technical reasons, throughout this paper, we assume that the metric space of labels Z is* complete *and* separable[2]*. We let $\mathcal{M}^L(Z)$ denote the collection of all Z-LMMCs.*

The following definition of isomorphism between LMMCs is similar to that of labeled graph isomorphism.

Definition 4. *Two Z-LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ are said to be* isomorphic *if there exists a bijective map $\psi : X \to Y$ such that $\ell_X(x) = \ell_Y(\psi(x))$ and $\psi_{\#}m_x^X = m_{\psi(x)}^Y$ for all $x \in X$ and $\psi_{\#}\mu_X = \mu_Y$.*

Labeled Graphs as LMMCs. Any labeled graph induces a family of LMMCs which we explain as follows.

Definition 5 (q-Markov chains on graphs)**.** *For any graph G and parameter $q \in [0, 1)$, we define the q-Markov chain $m_\bullet^{G,q}$ associated to G as follows: for any $v \in V_G$,*

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

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

$$\mu_G \coloneqq \sum_{v \in V_G} \frac{\overline{\deg}_G(v)}{\sum_{v' \in V_G} \overline{\deg}_G(v')} \delta_v$$

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

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

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

also let $m_\bullet^G := m_\bullet^{G,q}$ and let $\mathcal{X}(G) := \mathcal{X}_0(G)$. One has the following desirable property for graph induced LMMCs.

Proposition 2.1. *For any $q \in [0,1)$,(G_1, ℓ_{G_1}) is isomorphic to (G_2, ℓ_{G_1}) as labeled graphs iff $(\mathcal{X}_q(G_1), \ell_{G_1})$ is isomorphic to $(\mathcal{X}_q(G_2), \ell_{G_2})$ as LMMCs.*

3. The WL Distance

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

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

where $m(s)$ denotes the multiplicity of s in S. This point of view permits reinterpreting the multisets appearing in the WL hierarchy (see Definition 1) through the language of probability measures, which will eventually lead us to a distance between graphs.

Definition 6 (Weisfeiler-Lehman measure hierarchy)**.** *Given any Z-LMMC $(\mathcal{X}, \ell_X)$, we let $\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)} := \ell_X$ and produce the following label functions whose codomains span a certain hierarchy of probability measures:*

Step 1 *For each $x \in X$, we have $(\mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)})_\# m_x^X \in \mathcal{P}(Z)$. Hence we in fact have the function*

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

...

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

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

We then induce at each step k a probability measure

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

$\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}$ should be compared to $\ell^{(k)}_{(G,\ell_G)}$ and $\mathfrak{L}_k((\mathcal{X},\ell_X))$ should be compared to $L_k((G,\ell_G))$ from the WL hierarchy (cf. Definition 1). See Figure 1 for an illustration of the WL measure hierarchy of a graph induced LMMC and its comparison with the corresponding WL hierarchy. We will show later that, up to certain change of labels, the WL measure hierarchy for a graph induced LMMC captures all the information contained in the WL hierarchy of the original graph (see Proposition 3.3).

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

Definition 7 (Weisfeiler-Lehman distance)**.** *For each integer $k \geq 0$ and any metric space of labels Z we define the* Weisfeiler-Lehman (WL) distance of depth k *between the Z-LMMCs $(\mathcal{X},\ell_X)$ and $(\mathcal{Y},\ell_Y)$ as*

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

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

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

Example 1. *We write down explicit formulas for $d^{(k)}_{\mathrm{WL}}$ when $k = 0$ and 1. When $k = 0$, it is easy to see that*

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

which agrees with the Wasserstein distance between the global *label distributions $(\ell_X)_\#\mu_X$ and $(\ell_Y)_\#\mu_Y$ (cf. a similar concept for MMSs (Mémoli, 2011)).*

When $k = 1$, we have that

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

implementing the comparison of local *label distributions.*

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

Proposition 3.1. *Let $k \geq 0$ be any integer. Given any two Z-LMMCs $(\mathcal{X},\ell_X)$ and $(\mathcal{Y},\ell_Y)$, we have that $d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) \leq d^{(k+1)}_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y))$.*

As a first step towards understanding the WL distance, we show that d_{WL} is a pseudo-distance. We discuss its relationship with the WL test in Proposition 3.3 below.

Proposition 3.2. *d_{WL} (resp. $d^{(k)}_{\mathrm{WL}}$ for $k \geq 0$) defines a pseudo-distance[3] on the collection $\mathcal{M}^L(Z)$.*

3.1. Comparison with the WL Test

Given the apparent similarity between the WL hierarchy and the WL measure hierarchy, it should not be surprising that (as we will show later in Proposition 3.3), on graphs, d_{WL} essentially has the same discriminative power as the WL test, i.e., those pairs of graphs which can be distinguished by the WL test are the same as those with $d_{\mathrm{WL}} > 0$. The WL distance therefore should be interpreted as a quantification of the *degree* to which the graphs fail to pass the WL test.

[3] By pseudo-distance, we mean that d_{WL} (resp. $d^{(k)}_{\mathrm{WL}}$) is symmetric and satisfies the triangle inequality, but non-isomorphic LMMCs can have zero d_{WL} (resp. $d^{(k)}_{\mathrm{WL}}$) distance.

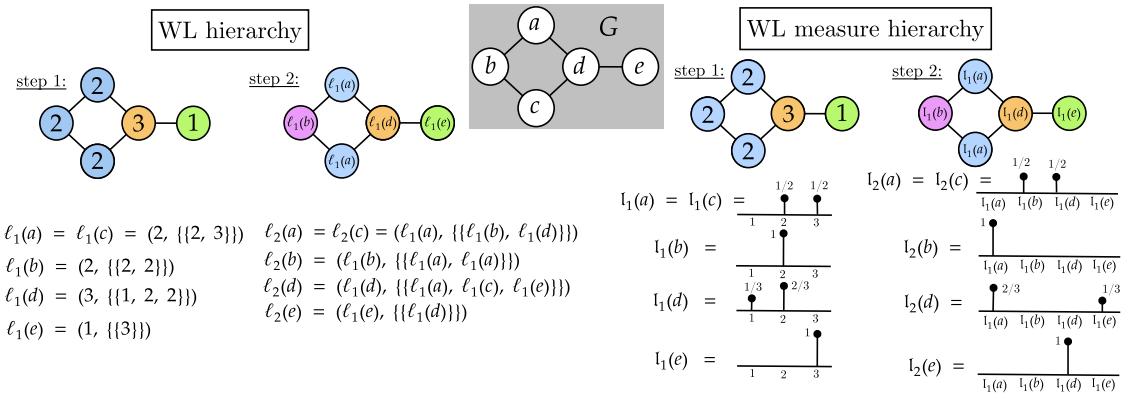

Figure 1. **Illustration of the WL (measure) hierarchy.** The graph G shown in the middle of the figure is assigned the degree label ℓ_G. We explicitly present two steps of the WL hierarchy of (G, ℓ_G) (on the left) and of the WL measure hierarchy of $(\mathcal{X}(G), \ell_G)$ (on the right). Every probability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

However, there is an apparent loss of information in the WL measure hierarchy due to the normalization inherent in probability measures. This could result in certain cases when $d_{\mathrm{WL}}((\mathcal{X}_q(G_1), \ell_{G_1}), (\mathcal{X}_q(G_2), \ell_{G_2})) = 0$ but (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) are distinguished by the WL test. See the examples in Appendix A.5. However, as we explain next, with appropriate label transformations, the discriminative power of d_{WL} is the same as that of the WL test.

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

Now, given any labeled graph $(G, \ell_G : V_G \to Z)$, we generate a new label function $\ell^g_G \coloneqq g(\ell_G, \deg_G(\bullet), |V_G|) : V_G \to Z_1$. Intuitively, this is understood as relabeling G via the map g. Modulo this change of label function, we establish that d_{WL} has the *same* discriminative power as the WL test.

Proposition 3.3. *For any $q \in (\frac{1}{2}, 1)$, the WL test distinguishes two labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) iff $d_{\mathrm{WL}}\big((\mathcal{X}_q(G_1), \ell^g_{G_1}), (\mathcal{X}_q(G_2), \ell^g_{G_2})\big) > 0$.*

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

Lemma 3.4. *The WL test distinguishes (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) iff it distinguishes $(G_1, \ell^g_{G_1})$ and $(G_2, \ell^g_{G_2})$.*

By Proposition 3.3 and a convergence result pertaining to the WL test (Krebs & Verbitsky, 2015), one has the following convergence result for $d^{(k)}_{\mathrm{WL}}$ (which implies that to determine whether two graph induced LMMCs satisfy $d_{\mathrm{WL}} = 0$ one only needs to inspect $d^{(k)}_{\mathrm{WL}}$ for finitely many k.)

Corollary 3.5. *For any $q \in (\frac{1}{2}, 1)$ and any two labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), if $d^{(k)}_{\mathrm{WL}}\big((\mathcal{X}_q(G_1), \ell^g_{G_1}), (\mathcal{X}_q(G_2), \ell^g_{G_2})\big) = 0$ holds for each $k = 0, \ldots, |V_{G_1}| + |V_{G_2}|$, we then have that $d_{\mathrm{WL}}\big((\mathcal{X}_q(G_1), \ell^g_{G_1}), (\mathcal{X}_q(G_2), \ell^g_{G_2})\big) = 0$.*

Results from (Babai & Kucera, 1979) imply that the WL test can certify isomorphism of random graphs (with degree labels) with high probability. Then, immediately by Proposition 3.3, we have that with high probability d_{WL} generates positive distance for non-isomorphic random-graph-induced LMMCs.

3.2. A Lower Bound for $d^{(k)}_{\mathrm{WL}}$

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

$$m^{X,\otimes k}_x(A) \coloneqq \int_X m^{X,\otimes(k-1)}_{x'}(A)\, m^X_x(dx').$$

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

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

$$d^{(k)}_{\mathrm{WLLB}}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \coloneqq$$
$$\inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X\times Y} d_{\mathrm{W}}\big((\ell_X)_\# m^{X,\otimes k}_x, (\ell_Y)_\# m^{Y,\otimes k}_y\big)\gamma(dx \times dy).$$

Notice that d_{W} above takes place in $\mathcal{P}(Z)$ whereas d_{W} in Equation (2) for defining $d^{(k)}_{\mathrm{WL}}$ takes place in $\mathcal{P}^{\circ k}(Z)$. Of

course, we have that $d_{\mathrm{WLLB}}^{(1)} = d_{\mathrm{WL}}^{(1)}$. It turns out that for each $k \geq 1$, $d_{\mathrm{WLLB}}^{(k)}$ is a lower bound for $d_{\mathrm{WL}}^{(k)}$.

Proposition 3.6. *For any $(\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y) \in \mathcal{M}^L(Z)$ and any integer $k \geq 1$ we have that $d_{\mathrm{WLLB}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \leq d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$.*

For any fixed $k \in \mathbb{N}$ and any two finite $\mathbb{R}$-LMMCs, $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, if we let $n := \max(|X|, |Y|)$, then computing $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$ can be done in $O(n^5 \log(n)\, k)$ time whereas the total time complexity for computing $d_{\mathrm{WLLB}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y))$ is $O(n^3 \log(nk))$. Hence it is far more efficient to compute $d_{\mathrm{WLLB}}^{(k)}$ than $d_{\mathrm{WL}}^{(k)}$. Details can be found in Appendix A.7.

4. WL Distance Inspired Neural Networks

We now focus on the case when the metric space of labels Z is Euclidean, i.e., $Z = \mathbb{R}^d$ and define a family of real functions on $\mathcal{M}^L(\mathbb{R}^d)$ called *Markov chain neural networks (*MCNN*s)*. We both study the discriminative power and establish a universality result for this family of functions.

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

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

(2) $S_\varphi : \mathcal{M}^L(\mathbb{R}^i) \to \mathbb{R}^j$ sending $(\mathcal{X}, \ell_X)$ to $q_\varphi((\ell_X)_\# \mu_X)$.

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

Then, for any sequence of MLPs $\varphi_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ for $i = 1, \ldots, k+1$, and any MLP $\psi : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$, we define a map of the following form, which we call a k-layer *Markov chain neural network (*MCNN$_k$*)*:

$$\psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} : \mathcal{M}^L(\mathbb{R}^d) \to \mathbb{R}. \quad (3)$$

Note the resemblance between our MCNNs and message passing neural networks (MPNNs) for graphs (Gilmer et al., 2017): Specifically, $(\ell_X)_\# m_x^X$ is analogous to the AGGREGATION operation, q_ϕ is analogous to the UPDATE operation, and $\psi \circ S_\phi$ corresponds to the readout function that appears in the context of MPNN.

Example 2 (Relation with WWL graph kernels)**.** *MCNNs recover the framework of Wasserstein Weisfeiler-Lehman (WWL) graph kernels w.r.t.* continuous attributes *(Togninalli et al., 2019): Consider any labeled graph $(G, \ell_G : V_G \to \mathbb{R}^d)$ and any $q \in [0, 1)$. Let $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be any continuous map. Applying q_φ to $(\mathcal{X}_q(G), \ell_G)$ (see Definition 5), then for any $v \in V_G$ such that $N_G(v) \neq \emptyset$, we have*

$$\begin{aligned}\ell_G^\varphi(v) &= \int_{\mathbb{R}^d} \varphi(t)\,(\ell_G)_\# m_v^{G,q}(dt) \\ &= q\,\varphi(\ell_G(v)) + \frac{1-q}{\deg_G(v)} \sum_{v' \in N_G(v)} \varphi(\ell_G(v')).\end{aligned}$$

Notice that if we further let $q = \frac{1}{2}$ and $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be the identity map id *(by relaxing the assumption that ψ is a MLP), then we have*

$$\ell_G^{\mathrm{id}}(v) = \frac{1}{2}\left(\ell_G(v) + \frac{1}{\deg_G(v)} \sum_{v' \in N_G(v)} \ell_G(v')\right). \quad (4)$$

This is exactly how labels are updated in the WWL graph kernel framework. A slight modification of the ground distance computation in the WWL framework generates a lower bound for $d_{\mathrm{WL}}^{(k)}$, which implies that the WL distance is more capable at discriminating labeled graphs than WWL graph kernels. This is confirmed by the examples in Appendix A.3.

We let $\mathcal{NN}_k(\mathbb{R}^d)$ denote the collection of all MCNN$_k$. Below, we show that $\mathcal{NN}_k(\mathbb{R}^d)$ has the same discriminative power as the WL distance.

Proposition 4.1. *Given any $(\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y) \in \mathcal{M}^L(\mathbb{R}^d)$,*

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

Recall from Corollary 3.5 that given any $q \in (\frac{1}{2}, 1)$, any injective map g and any labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), we need at most $2n$ steps to determine whether $d_{\mathrm{WL}}((\mathcal{X}_q(G_1), \ell_G^g), (\mathcal{X}_q(G_2), \ell_{G_2}^g)) = 0$, where $n = \max(|V_{G_1}|, |V_{G_2}|)$. Consequently, MCNNs have the same discriminative power as the WL test:

Corollary 4.2. *For any $\frac{1}{2} < q < 1$, the WL test distinguishes the labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) iff there exists $h \in \mathcal{NN}_{2n}(\mathbb{R}^d)$ for which $h\big((\mathcal{X}_q(G_1), \ell_{G_1}^g)\big) \neq h\big((\mathcal{X}_q(G_2), \ell_{G_2}^g)\big)$.*

Since MPNNs also have the same discriminative power as the WL test (Xu et al., 2018), we know that our MCNNs can separate all pairs of graphs that MPNNs can separate.

Next, we establish a universal approximation theorem for MCNNs. We first introduce some notation. In general, a

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

pseudometric space canonically induces a metric space by identifying points at 0 distance (see Burago et al. (2001, Proposition 1.1.5)). We let $\mathcal{M}_k^L(\mathbb{R}^d)$ denote the metric space induced by the pseudometric space $(\mathcal{M}^L(\mathbb{R}^d), d_{\mathrm{WL}}^{(k)})$. As a direct consequence of Proposition 4.1, every $h \in \mathcal{NN}_k(\mathbb{R}^d)$ induces a real function (which we still denote by h) in $\mathcal{M}_k^L(\mathbb{R}^d)$. Then, our MCNNs are actually universal w.r.t. continuous functions defined on $\mathcal{M}_k^L(\mathbb{R}^d)$ (see the proof in Appendix B.3.2).

Theorem 4.3. *For any $k \in \mathbb{N}$, let $\mathcal{K} \subseteq \mathcal{M}_k^L(\mathbb{R}^d)$ be any compact subspace. Then*[5]*, $\overline{\mathcal{NN}_k(\mathbb{R}^d)} = C(\mathcal{K}, \mathbb{R})$.*

Our universality result resembles the one established in (Azizian et al., 2020) for message passing neural networks (MPNNs) which proves that MPNNs can universally approximate continuous functions on graphs with bounded size which are less or equally as discriminative as the WL test. Compared with their result, we remark that our universality result applies to the collection of all LMMCs (and hence all graphs) *with no restriction on their size*. Moreover, although for simplicity LMMCs are restricted to finite spaces throughout the paper, our MCNNs and universality result can potentially be extended to more general LMMCs including continuous objects such as manifolds and graphons.

5. Relationship with the GW Distance

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

5.1. The GW Distance between MCMSs

Recall from Equation (1) the definition of the standard GW distance between MMSs. Intuitively, in order to identify a suitable GW-like distance between MCMSs, we would like to incorporate a comparison between Markov kernels into Equation (1). Towards this goal, for each $k \in \mathbb{N}$, we consider a special type of maps $\nu_{\bullet,\bullet}^{(k)} : X \times Y \to \mathcal{P}(X \times Y)$ which are defined similarly to how we define k-step Markov kernels and satisfy that for any $x \in X$ and $y \in Y$, $\nu_{x,y}^{(k)}$ is a coupling between the k-step Markov kernels $m_x^{X,\otimes k}$ and $m_y^{Y,\otimes k}$; see Appendix A.2 for the precise definition. We refer to $\nu_{\bullet,\bullet}^{(k)}$ as a "k-step coupling" between k-step Markov kernels. We let $\mathcal{C}^{(k)}\left(m_\bullet^X, m_\bullet^Y\right)$ denote the collection of all such k-step couplings $\nu_{\bullet,\bullet}^{(k)}$.

Definition 8. *For any $k \geq 1$ and any MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$, we define the k-distortion of any pair $(\gamma, \nu_{\bullet,\bullet}^{(k)})$ where $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\left(m_\bullet^X, m_\bullet^Y\right)$ as:*

$$\mathrm{dis}^{(k)}\left(\gamma, \nu_{\bullet,\bullet}^{(k)}\right) \coloneqq \int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')| \, \nu_{x'',y''}^{(k)}(dx' \times dy')\, \gamma(dx'' \times dy'')\, \gamma(dx \times dy).$$

*This notion of distortion implements a multiscale reweighting of the coupling γ through the k-step coupling $\nu_{\bullet,\bullet}^{(k)}$. Then, the k-*Gromov-Wasserstein distance *between the MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ is defined by*

$$d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \coloneqq \inf_{\substack{\gamma \in \mathcal{C}(\mu_X, \mu_Y) \\ \nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\left(m_\bullet^X, m_\bullet^Y\right)}} \mathrm{dis}^{(k)}\left(\gamma, \nu_{\bullet,\bullet}^{(k)}\right).$$

We then define the (absolute) ***Gromov-Wasserstein distance*** *between MCMSs by*

$$d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \coloneqq \sup_k d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)).$$

Proposition 5.1. *$d_{\mathrm{GW}}^{\mathrm{MCMS}}$ defines a proper*[6] *distance on the collection $\mathcal{M}^{\mathrm{MS}}$ modulo isomorphism of MCMSs.*

Example 3 (MMS induced MCMS)**.** *Given a metric measure space $\mathbf{X} = (X, d_X, \mu_X)$, we produce a MCMS $\mathcal{M}(\mathbf{X}) \coloneqq (\mathcal{X}, d_X)$ where $\mathcal{X} \coloneqq (X, m_\bullet^X, \mu_X)$ by letting $m_\bullet^X \coloneqq \mu_X$ be the constant Markov kernel. It is easy to check that μ_X is a stationary distribution w.r.t. $m_\bullet^X$. Then, for any two metric measure spaces $\mathbf{X} = (X, d_X, \mu_X)$, $\mathbf{Y} = (Y, d_Y, \mu_Y)$, and $k \geq 1$, we have that*

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

where $d_{\mathrm{GW}}^{\mathrm{bi}}$ denotes a "decoupled" version of the Gromov-Wasserstein distance between metric measure spaces (see Appendix A.4) which is of course independent of k.

$d_{\mathrm{GW}}^{\mathrm{bi}}$ is in general NP-hard to compute (Scetbon et al., 2021), which (via Example 3) implies that $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ is also NP-hard to compute. See Appendix A.6 for a basic computable lower bound estimate of $d_{\mathrm{GW}}^{\mathrm{MCMS}}$. In the next section, we establish more sophisticated lower bounds for $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ involving the WL distance.

[5] For simplicity of notation, we still use $\mathcal{NN}_k(\mathbb{R}^d)$ to denote induced functions on $\mathcal{M}_k^L(\mathbb{R}^d)$ with domain restricted to $\mathcal{K}$.

[6] Unlike the case for d_{WL}, two MCMSs have zero $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ distance iff they are isomorphic. The precise definition of isomorphism between MCMSs is postponed to Definition 15 in Appendix B.4.

Table 1. 1-Nearest Neighbor classification accuracy.

METHOD	MUTAG	PROTEINS	PTC-FM	PTC-MR	IMDB-B	IMDB-M	COX2
$d_{\mathrm{WL}}^{(k)}$	**92.1 ± 6.3**	63.0 ± 3.5	62.2 ± 8.5	56.2 ± 6.3	**70.0 ± 4.3**	**41.3 ± 4.8**	76.1 ± 5.5
$d_{\mathrm{WLLB}}^{(k)}$	87.3 ± 1.9	**66.2 ± 2.2**	**62.5 ± 8.5**	**57.8 ± 6.8**	69.9 ± 2.5	40.6 ± 3.8	**81.2 ± 5.3**
WWL	85.1 ± 6.5	64.7 ± 2.8	58.2 ± 8.5	54.3 ± 7.9	65.0 ± 3.3	40.0 ± 3.3	76.1 ± 5.6

Table 2. SVM classification accuracy.

METHOD	MUTAG	PROTEINS	PTC-FM	PTC-MR	IMDB-B	IMDB-M	COX2
$d_{\mathrm{WL}}^{(k)}$	89.9 ± 6.4	72.6 ± 3.1	**63.6 ± 7.0**	57.9 ± 7.9	**76.2 ± 4.1**	51.6 ± 4.0	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(k)}$	**90.0 ± 5.6**	70.5 ± 1.0	63.3 ± 5.6	59.0 ± 8.3	75.1 ± 2.2	**52.0 ± 1.8**	78.1 ± 0.8
WWL	85.3 ± 7.3	**72.9 ± 3.6**	62.2 ± 6.1	**63.0 ± 7.4**	70.8 ± 5.4	50.0 ± 5.3	78.2 ± 0.8
WL	85.5± 1.6	71.6 ± 0.6	56.6 ± 2.1	56.2 ± 2.0	72.4 ± 0.7	50.9 ± 0.4	78.4 ± 1.1
WL-OA	86.3 ± 2.1	72.6 ± 0.7	58.4 ± 2.0	54.2 ± 1.6	73.0 ± 1.1	50.2 ± 1.1	**78.8 ± 1.3**

5.2. The WL Distance vs. the GW Distance

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

Proposition 5.2. *For each $k \geq 1$ and for any MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$, one has that*

$$d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \geq \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} d_{\mathrm{WL}}^{(k)}\big((\mathcal{X}, d_X(x, \bullet)), (\mathcal{Y}, d_Y(y, \bullet))\big) \gamma(dx \times dy).$$

Remark 5.3. *When $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ are induced from MMSs $\mathbf{X}$ and $\mathbf{Y}$, as shown in Example 3, the left-hand side of the above inequality coincides with the decoupled GW distance. We point out that the right-hand side is also independent of k and actually coincides with the* third lower bound *(TLB) for the GW distance as defined in (Mémoli, 2011). See Appendix B.4.4 for more details. Hence, the proposition above can be viewed as a generalization of the TLB to the setting of MCMS.*

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

Definition 9 (Label invariant of MCMCs)**.** *Given any metric space of labels Z, a Z-valued **label invariant** of MCMSs is a map $\ell_\bullet : \mathcal{M}^{\mathrm{MS}} \to Z^\bullet$ which by definition sends each MCMS $(\mathcal{X}, d_X)$ into a label function $\ell_X : X \to Z$. One such label invariant $\ell_\bullet$ will be said to be* stable *if for all $(\mathcal{X}, d_X), (\mathcal{Y}, d_Y) \in \mathcal{M}^{\mathrm{MS}}$, $k \in \mathbb{N}$, and for any $\gamma \in \mathcal{C}(\mu_X, \mu_Y), \nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ we have*

$$\int_{X \times Y} \int_{X \times Y} d_Z(\ell_X(x'), \ell_Y(y'))\, \nu_{x,y}^{(k)}(dx' \times dy') \gamma(dx \times dy) \leq \mathrm{dis}^{(k)}\big(\gamma, \nu_{\bullet,\bullet}^{(k)}\big).$$

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

Example 4. *One immediate example of the stable label invariant is the* eccentricity function $\mathrm{ecc}_\bullet$ *(see Appendix B.4.5 for a proof): for any MCMS $(\mathcal{X}, d_X)$, $\mathrm{ecc}_X(x) := \int_X d_X(x, x')\, \mu_X(dx')$ for $x \in X$.*

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

Proposition 5.4. *For every stable label invariant $\ell_\bullet \in \mathcal{L}(Z)$, $k \in \mathbb{N}$ and $(\mathcal{X}, d_X), (\mathcal{Y}, d_Y) \in \mathcal{M}^{\mathrm{MS}}$ we have that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \leq d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y))$.*

6. Experimental Results

We provide some results showing the effectiveness of our WL distance in terms of comparing graphs. We conduct both 1-NN and SVM graph classification experiments and evaluate the performance of both our lower bound, $d_{\mathrm{WLLB}}^{(k)}$, and our WL distance, $d_{\mathrm{WL}}^{(k)}$, against the WWL kernel/distance (Togninalli et al., 2019), the WL kernel and the WL optimal assignment (WL-OA) (Kriege et al., 2016) kernel. We note that the WWL kernel of (Togninalli et al., 2019) is a state-of-the-art graph kernel. We use the degree label for both $d_{\mathrm{WL}}^{(k)}$ and $d_{\mathrm{WLLB}}^{(k)}$. More details on the experimental setup and extra experiments can be found in Appendix A.8.

1-NN Classification. In this case, both $d_{\mathrm{WLLB}}^{(k)}$ and $d_{\mathrm{WL}}^{(k)}$ slightly outperform the WWL distance on all datasets we tested; see Table 1. Overall, the classification accuracies of $d_{\mathrm{WLLB}}^{(k)}$ and $d_{\mathrm{WL}}^{(k)}$ were close to those of the WWL distance. These results illustrate the close relationship between $d_{\mathrm{WL}}^{(k)}$ and the WWL distance that was outlined in Section 4.

SVM Classification. See Table 2. First, we observe that our lower bound kernel slightly outperforms the WWL kernel for MUTAG, IMDB-B, and IMDB-M. For the other datasets, $d_{\mathrm{WLLB}}^{(k)}$ had comparable classification accuracy with the other methods, coming within one to two percent of WWL and WWL-OA. The $d_{\mathrm{WL}}^{(k)}$ kernel had similar classification accuracy to $d_{\mathrm{WLLB}}^{(k)}$ but only outperformed the $d_{\mathrm{WLLB}}^{(k)}$ kernel on PTC-FM, PROTEINS, and IMDB-B.

Note that our lower bound distance d_{WLLB} performs similarly to our WL distance d_{WL}, but is more efficient to compute. See Appendix A.8.3 for the runtime comparison.

7. Conclusion and Future Directions

In this paper, we proposed the WL distance – a quantitative extension of the WL test – for measuring the dissimilarity between objects in a fairly general family called LMMCs. The WL distance possesses interesting connections with graph kernels, GNNs and the GW distance. In order to more directly compare the WL test with the WL distance without resorting to relabeling, one future direction is to redefine the WL distance via positive measures and "unbalanced" (Gromov-)Wasserstein distances (Liero et al., 2018; Séjourné et al., 2021; De Ponti & Mondino, 2020). Whereas our paper focuses on the application of our WL distance to the graph setting, LMMCs can be used to model not just graphs but also far more general objects such as Riemannian manifolds (equipped with heat kernels) or graphons. Then, our neural network architecture (MCNN) has the potential to be applied to point sets sampled from manifolds too, as well as serving as the limiting object when studying the convergence of GNNs. It is also interesting to extend our WL distance to a higher order version that is analogous to the high order k-WL test and k-GNNs. We conjecture that a suitable notion of k-WL distance will converge to the Gromov-Wasserstein distance for MCMSs as k tends to infinity. Finally, we point out that our WL distance can actually be regarded as a certain instance of optimal transport with special couplings (see Appendix A.2). It is of interest to explore the relation between our WL distance and some other distances between stochastic processes defined via optimal transport with special couplings (Backhoff et al., 2017; Moulos, 2021; O'Connor et al., 2022).

Acknowledgements

This work is partially supported by NSF-DMS-1723003, NSF-CCF-1740761, NSF-RI-1901360, NSF-CCF-1839356, NSF-IIS-2050360, NSF-CCF-2112665 and BSF-2020124.

References

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

Babai, L. and Kucera, L. Canonical labelling of graphs in linear average time. In *20th Annual Symposium on Foundations of Computer Science (sfcs 1979)*, pp. 39–46. IEEE, 1979.

Babai, L. and Luks, E. M. Canonical labeling of graphs. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pp. 171–183, 1983.

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.

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

Chowdhury, S. and Mémoli, F. The Gromov-Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4): 757–787, 2019.

De Ponti, N. and Mondino, A. Entropy-transport distances between unbalanced metric measure spaces. *arXiv preprint arXiv:2009.10636*, 2020.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Krebs, A. and Verbitsky, O. Universal covers, color refinement, and two-variable counting logic: Lower bounds for the depth. In *2015 30th Annual ACM/IEEE Symposium on Logic in Computer Science*, pp. 689–700. IEEE, 2015.

Kriege, N. M., Giscard, P.-L., and Wilson, R. On valid optimal assignment kernels and applications to graph classification. *Advances in Neural Information Processing Systems*, 29:1623–1631, 2016.

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.

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Loosli, G., Canu, S., and Ong, C. S. Learning svm in kreĭn spaces. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1204–1216, 2015.

Luss, R. and d'Aspremont, A. Support vector machine classification with indefinite kernels. *Mathematical Programming Computation*, 1(2):97–118, 2009.

Mémoli, F. On the use of Gromov-Hausdorff distances for shape comparison. In Botsch, M., Pajarola, R., Chen, B., and Zwicker, M. (eds.), *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007. ISBN 978-3-905673-51-7. doi: 10.2312/SPBG/SPBG07/081-090.

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

Morris, C., Kriege, N. M., Bause, F., Kersting, K., Mutzel, P., and Neumann, M. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL www.graphlearning.io.

Moulos, V. Bicausal optimal transport for Markov chains via dynamic programming. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 1688–1693. IEEE, 2021.

Ollivier, Y. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

O'Connor, K., McGoff, K., and Nobel, A. B. Optimal transport for stationary Markov chains via policy iteration. *Journal of Machine Learning Research*, 23(45): 1–52, 2022.

Pele, O. and Werman, M. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pp. 460–467. IEEE, 2009.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672. PMLR, 2016.

Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8:143–195, 1999.

Saloff-Coste, L. Lectures on finite Markov chains. In *Lectures on probability theory and statistics*, pp. 301–413. Springer, 1997.

Scetbon, M., Peyré, G., and Cuturi, M. Linear-time Gromov-Wasserstein distances using low rank couplings and costs. *arXiv preprint arXiv:2106.01128*, 2021.

Schmitzer, B. and Schnörr, C. Modelling convex shape priors and matching based on the Gromov-Wasserstein distance. *Journal of mathematical imaging and vision*, 46(1):143–159, 2013.

Séjourné, T., Vialard, F.-X., and Peyré, G. The unbalanced Gromov-Wasserstein distance: Conic formulation and relaxation. In *35th Conference on Neural Information Processing Systems*, 2021.

Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Sturm, K.-T. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

Titouan, V., Redko, I., Flamary, R., and Courty, N. Co-optimal transport. *Advances in Neural Information Processing Systems*, 33, 2020.

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.

Vallender, S. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

von Renesse, M.-K. and Sturm, K.-T. Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on pure and applied mathematics*, 58(7): 923–940, 2005.

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

A. Extra Details

A.1. Useful Facts about Couplings

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

Lemma A.1. *Let X, Y be finite metric spaces and let Z be a complete and separable metric space. Let $\varphi_X : X \to Z$ and $\varphi_Y : Y \to Z$ be measurable maps. Consider any $\mu_X \in \mathcal{P}(X)$ and $\mu_Y \in \mathcal{P}(Y)$. Then, we have that*

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

Proof. The proof is based on the following result.

Lemma A.2. *Let X, Y be finite metric spaces and let Z be a metric space of labels. Let $\varphi_X : X \to Z$ and $\varphi_Y : Y \to Z$ be measurable maps. Consider any $\mu_X \in \mathcal{P}(X)$ and $\mu_Y \in \mathcal{P}(Y)$. If we let $\varphi \coloneqq \varphi_X \times \varphi_Y$, then we have that*

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

Proof of Lemma A.2. Since X and Y are finite, $\varphi_X(X)$ and $\varphi_Y(Y)$ are discrete sets. Then, the lemma follows directly from Proposition 4.5 in (Schmitzer & Schnörr, 2013). □

Hence,

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

□

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

Lemma A.3. *For any complete and separable metric space Z, there exists a measurable map $\varphi : \mathcal{P}(Z) \times \mathcal{P}(Z) \to \mathcal{P}(Z \times Z)$ so that for every $\alpha, \beta \in \mathcal{P}(Z)$, $\varphi(\alpha, \beta)$ is an optimal coupling between α and β.*

A.2. A Characterization of the WL Distance

Recall from Definition 7 that the WL distance of depth k is the Wasserstein distance between the local distributions of labels generated at the kth step of the WL hierarchy. In this section, we prove that in fact $d_{\mathrm{WL}}^{(k)}$ can be characterized through a novel variant of the Wasserstein distance between the distributions of initial labels (i.e., $(\ell_X)_{\#}\mu_X$ and $(\ell_Y)_{\#}\mu_Y$).

A.2.1. k-STEP COUPLINGS

We introduce a convenient notation which will be used in the sequel.

Definition 10. *Suppose a measurable space Z, a probability measure $\gamma \in \mathcal{P}(Z)$, and a measurable map $\nu_\bullet : Z \longrightarrow \mathcal{P}(Z)$ are given. Then, we define the average of $\nu_\bullet$ under γ, denoted by $\nu_\bullet \odot \gamma$, which is still a probability measure on Z:*

$$\text{For any measurable } A \subseteq Z,\ \nu_\bullet \odot \gamma(A) \coloneqq \int_Z \nu_z(A)\, \gamma(dz).$$

The operation $\odot$ will be useful for constructing a special type of couplings between Markov kernels.

Given two MMCs $\mathcal{X}$ and $\mathcal{Y}$, we introduce the following notion of **k-step coupling between Markov kernels $m_\bullet^X$ and $m_\bullet^Y$**:

- $k = 1$: A 1-step coupling between $m_\bullet^X$ and $m_\bullet^Y$ is any measurable map
$$\nu_{\bullet,\bullet}^{(1)} : X \times Y \to \mathcal{P}(X \times Y)$$
such that $\nu_{x,y}^{(1)} \in \mathcal{C}(m_x^X, m_y^Y)$ for any $x \in X$ and $y \in Y$.
- $k \geq 2$: We say a map
$$\nu_{\bullet,\bullet}^{(k)} : X \times Y \to \mathcal{P}(X \times Y)$$
is a k**-step coupling** between the Markov kernels $m_\bullet^X$ and $m_\bullet^Y$ if there exist a $(k-1)$-step coupling $\nu_{\bullet,\bullet}^{(k-1)}$ and a 1-step coupling $\nu_{\bullet,\bullet}^{(1)}$ such that
$$\nu_{x,y}^{(k)} = \nu_{\bullet,\bullet}^{(k-1)} \odot \nu_{x,y}^{(1)}, \ \forall x \in X, \ y \in Y.$$

Lemma A.4. *For any k-step coupling $\nu_{\bullet,\bullet}^{(k)}$, one has that $\nu_{x,y}^{(k)} \in \mathcal{C}\big(m_x^{X,\otimes k}, m_y^{Y,\otimes k}\big)$ for every $x \in X$ and $y \in Y$.*

Proof. We prove the statement by induction on k. When $k = 1$, the statement is trivially true. Assume that the statement is true for some $k \geq 1$. Then, for any measurable $A \subseteq X$, we have that
$$\begin{aligned}
\nu_{x,y}^{(k+1)}(A \times Y) &= \int_{X \times Y} \nu_{x',y'}^{(k)}(A \times Y)\, \nu_{x,y}^{(1)}(dx' \times dy') \\
&= \int_{X \times Y} m_{x'}^{X,\otimes k}(A)\, \nu_{x,y}^{(1)}(dx' \times dy') \\
&= \int_X m_{x'}^{X,\otimes k}(A)\, m_x^X(dx') \\
&= m_x^{X,\otimes(k+1)}(A).
\end{aligned}$$
Similarly, for any measurable $B \subseteq Y$, we have that
$$\nu_{x,y}^{(k+1)}(X \times B) = m_y^{Y,\otimes(k+1)}(B).$$
Hence $\nu_{x,y}^{(k+1)} \in \mathcal{C}\left(m_x^{X,\otimes(k+1)}, m_y^{Y,\otimes(k+1)}\right)$ and thus we conclude the proof. □

Henceforth, we denote by $\mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ the collection of all k-step couplings between Markov kernels $m_\bullet^X$ and $m_\bullet^Y$.

Definition 11. *Given any $k \in \mathbb{N}$, any coupling $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and any k-step coupling $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$, we define a probability measure on $X \times Y$ as follows:*
$$\mu^{(k)} \coloneqq \nu_{\bullet,\bullet}^{(k)} \odot \gamma. \tag{5}$$
*We call $\mu^{(k)}$ defined as above a k**-step coupling between** μ_X **and** μ_Y. We let $\mathcal{C}^{(k)}(\mu_X, \mu_Y)$ denote the collection of all k-step couplings between μ_X and μ_Y and we also let $\mathcal{C}^{(0)}(\mu_X, \mu_Y) \coloneqq \mathcal{C}(\mu_X, \mu_Y)$.*

As defined above, k-step couplings are indeed couplings.

Lemma A.5. *Any $\mu^{(k)} \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)$ is a coupling between μ_X and μ_Y.*

Proof. For any measurable $A \subseteq X$, we have that
$$\begin{aligned}
\mu^{(k)}(A \times Y) &= \int_{X \times Y} \nu_{x,y}^{(k)}(A \times Y)\, \gamma(dx \times dy) \\
&= \int_{X \times Y} m_x^{X,\otimes k}(A)\, \gamma(dx \times dy) \\
&= \int_X m_x^{X,\otimes k}(A)\, \mu_X(dx) \\
&= \mu_X(A).
\end{aligned}$$
Similarly, for any measurable $B \subseteq Y$, we have that $\mu^{(k)}(X \times B) = \mu_Y(B)$. Therefore, $\mu^{(k)} \in \mathcal{C}(\mu_X, \mu_Y)$. □

Lemma A.6. *We have the following hierarchy of k-step couplings:*

$$\mathcal{C}^{(0)}(\mu_X, \mu_Y) \supseteq \mathcal{C}^{(1)}(\mu_X, \mu_Y) \supseteq \mathcal{C}^{(2)}(\mu_X, \mu_Y) \supseteq \cdots$$

Proof. We prove the following inclusion by induction on $k = 0, 1, \ldots$:

$$\mathcal{C}^{(k)}(\mu_X, \mu_Y) \supseteq \mathcal{C}^{(k+1)}(\mu_X, \mu_Y) \tag{6}$$

When $k = 0$, we only need to check that any $\gamma^{(1)} \in \mathcal{C}^{(1)}(\mu_X, \mu_Y)$ is a coupling between μ_X and μ_Y. Assume that

$$\gamma^{(1)} = \nu^{(1)}_{\bullet,\bullet} \odot \gamma$$

for some $\nu^{(1)}_{\bullet,\bullet} \in \mathcal{C}^{(1)}\big(m^X_\bullet, m^Y_\bullet\big)$ and $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$. For any measurable set $A \subseteq X$, we have that

$$\begin{aligned}
\gamma^{(1)}(A \times Y) &= \int_{X\times Y} \nu^{(1)}_{x,y}(A \times Y)\, \gamma(dx \times dy) \\
&= \int_{X\times Y} m^X_x(A)\, \gamma(dx \times dy) \\
&= \int_X m^X_x(A)\, \mu_X(dx) \\
&= \mu_X(A).
\end{aligned}$$

Similarly, for any measurable $B \subseteq Y$ we have that

$$\gamma^{(1)}(X \times B) = \mu_Y(B).$$

Hence, $\gamma^{(1)} \in \mathcal{C}(\mu_X, \mu_Y)$.

Now, assume that Equation (6) holds for some $k \geq 0$. For any $\gamma^{(k+2)} \in \mathcal{C}^{(k+2)}(\mu_X, \mu_Y)$, we assume that

$$\gamma^{(k+2)} = \nu^{(k+2)}_{\bullet,\bullet} \odot \gamma,$$

where $\nu^{(k+2)}_{\bullet,\bullet} \in \mathcal{C}^{(k+2)}\big(m^X_\bullet, m^Y_\bullet\big)$ and $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$. Then, there exist $\nu^{(k+1)}_{\bullet,\bullet} \in \mathcal{C}^{(k+1)}\big(m^X_\bullet, m^Y_\bullet\big)$ and $\nu^{(1)}_{\bullet,\bullet} \in \mathcal{C}^{(1)}\big(m^X_\bullet, m^Y_\bullet\big)$ such that

$$\nu^{(k+2)}_{x,y} = \nu^{(k+1)}_{\bullet,\bullet} \odot \nu^{(1)}_{x,y}, \ \forall x \in X, \ y \in Y.$$

Hence,

$$\begin{aligned}
\gamma^{(k+2)} &= \int_{X\times Y} \int_{X\times Y} \nu^{(k+1)}_{x',y'} \nu^{(1)}_{x,y}(dx' \times dy')\, \gamma(dx \times dy) \\
&= \int_{X\times Y} \nu^{(k+1)}_{x',y'} \int_{X\times Y} \nu^{(1)}_{x,y}(dx' \times dy')\, \gamma(dx \times dy) \\
&= \int_{X\times Y} \nu^{(k+1)}_{x',y'}\, \gamma^{(1)}(dx' \times dy').
\end{aligned}$$

Here that $\gamma^{(1)} \coloneqq \nu^{(1)}_{\bullet,\bullet} \odot \gamma$ belongs to $\mathcal{C}(\mu_X, \mu_Y)$ follows from the case $k = 0$. Hence, by the induction assumption, $\gamma^{(k+2)} \in \mathcal{C}^{(k+1)}(\mu_X, \mu_Y)$ which concludes the proof. □

A.2.2. A Characterization of $d^{(k)}_{\mathrm{WL}}$ via k-step Couplings

Now, we characterize $d^{(k)}_{\mathrm{WL}}$ via k-step couplings defined in the previous section.

Theorem A.7. *Given any integer $k \geq 0$ and any two Z-LMMC $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$, we have that*

$$d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = \inf_{\gamma^{(k)} \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)} \int_{X \times Y} d_Z(\ell_X(x), \ell_Y(y)) \gamma^{(k)}(dx \times dy).$$

Proof of Theorem A.7. The case $k = 0$ holds trivially. Now, for any $k \geq 1$ and for any $x \in X$ and $y \in Y$, by Lemma A.1 we have that

$$d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}(x), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}(y)\right) = \inf_{\nu_{x,y} \in \mathcal{C}(m_x^X, m_y^Y)} \int_{X \times Y} d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k-1)}(x'), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k-1)}(y')\right) \nu_{x,y}(dx' \times dy').$$

Since $(x, y) \mapsto (m_x^X, m_y^Y)$ is measurable by definition of Markov kernels, by Lemma A.3 we have that there exists a measurable map $\nu_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)$ such that for every $x \in X$ and $y \in Y$, $\nu_{x,y}$ is optimal, i.e.,

$$d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}(x), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}(y)\right) = \int_{X \times Y} d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k-1)}(x'), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k-1)}(y')\right) \nu_{x,y}(dx' \times dy'). \tag{7}$$

Hence, we have the following formulas.

$$\begin{aligned}
& d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) \\
=& d_{\mathrm{W}}\left(\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}\right)_{\#} \mu_X, \left(\mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}\right)_{\#} \mu_Y\right) \\
=& \int_{X \times Y} d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}(x), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}(y)\right) \gamma(dx \times dy), \quad \text{here } \gamma \in \mathcal{C}(\mu_X, \mu_Y) \text{ is chosen to be optimal} \\
=& \int_{X \times Y} \int_{X \times Y} d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k-1)}(x_1), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k-1)}(y_1)\right) (\nu_1)_{x,y}(dx_1 \times dy_1) \gamma(dx \times dy), \\
=& \int_{X \times Y} \cdots \int_{X \times Y} d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(0)}(x_k), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(0)}(y_k)\right) (\nu_k)_{x_{k-1}, y_{k-1}}(dx_k \times dy_k) \cdots (\nu_1)_{x,y}(dx_1 \times dy_1) \gamma(dx \times dy).
\end{aligned}$$

Here, each $(\nu_i)_{\bullet,\bullet} \in \mathcal{C}^{(1)}(m_\bullet^X, m_\bullet^Y)$ for $i = 1, \ldots, k$ is optimal in the sense of Equation (7).

From the above equations, we identify a probability measure on $X \times Y$ for every $x \in X$ and $y \in Y$ as follows

$$\nu_{x,y}^{(k)} \coloneqq \int_{X \times Y} \cdots \int_{X \times Y} (\nu_k)_{x_{k-1}, y_{k-1}} (\nu_{k-1})_{x_{k-2}, y_{k-2}}(dx_{k-1} \times dy_{k-1}) \cdots (\nu_1)_{x,y}(dx_1 \times dy_1).$$

It is obvious that $\nu_{\bullet,\bullet}^{(k)}$ is a k-step coupling and thus

$$\begin{aligned}
d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) =& d_{\mathrm{W}}\left(\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(k)}\right)_{\#} \mu_X, \left(\mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(k)}\right)_{\#} \mu_Y\right) \\
=& \int_{X \times Y} \int_{X \times Y} d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X}, \ell_X)}^{(0)}(x'), \mathfrak{l}_{(\mathcal{Y}, \ell_Y)}^{(0)}(y')\right) \nu_{x,y}^{(k)}(dx' \times dy') \gamma(dx \times dy) \\
\geq& \inf_{\gamma^{(k)} \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)} \int_{X \times Y} d_{\mathrm{W}}(\ell_X(x), \ell_Y(y)) \, \gamma^{(k)}(dx \times dy),
\end{aligned}$$

Conversely, we have that given any $\gamma^{(k)} \coloneqq \nu_{\bullet,\bullet}^{(k)} \odot \gamma \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)$, where $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\left(m_\bullet^X, m_\bullet^Y\right)$ can be written for every $x \in X$ and $y \in Y$ as follows

$$\nu_{x,y}^{(k)} \coloneqq \int_{X \times Y} \cdots \int_{X \times Y} (\nu_k)_{x_{k-1}, y_{k-1}} (\nu_{k-1})_{x_{k-2}, y_{k-2}}(dx_{k-1} \times dy_{k-1}) \cdots (\nu_1)_{x,y}(dx_1 \times dy_1),$$

the following inequalities hold:

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

Infimizing over all γ and $\nu^{(k)}_{\bullet,\bullet}$, one concludes the proof. □

A.3. The Wasserstein Weisfeiler-Lehman Graph Kernel and its Relationship with d_{WL}

The Wasserstein Weisfeiler-Lehman graph kernel deals with graphs with either categorical or "continuous" (i.e., Euclidean) labels (Togninalli et al., 2019). We only consider WWL graph kernel w.r.t. continuous labels since categorical labels usually don't come equipped with a distance. Even if we artificially defined a distance on a set of categorical labels (e.g. using Euclidean distance via one-hot encoding), we point out that there wouldn't be a clear relation between WWL graph kernel and the WL distance since at each step of WWL graph kernel a hash function is invoked to map into a single label a pair consisting of a label and a multiset of labels. Now, we describe the WWL framework w.r.t. continuous labels as follows. For technical reasons, we assume that all graphs involved in this section are such that all their connected components have cardinality at least 2 (i.e. no graph contains an isolated vertex).

Given a labeled graph $(G,\ell_G:V_G\to\mathbb{R}^d)$, the label function is updated for a fixed number k of iterations according to the equation below for $i=0,\ldots,k-1$, where $\ell^0_G\coloneqq\ell_G$.

$$
\forall v\in V_G,\quad \ell^{i+1}_G(v)\coloneqq\frac{1}{2}\left(\ell^i_G(v)+\frac{1}{\deg_G(v)}\sum_{v'\in N_G(v)}\ell^i_G(v')\right).
$$

Then, for each $i=0,\ldots,k$ there is a label function $\ell^{(i)}_{(G,\ell_G)}:V_G\to\mathbb{R}^d$. Define the stacked label function L^k_G as follows:

$$
L^k_G\coloneqq\left(\ell^0_G,\ldots,\ell^k_G\right):V_G\to\mathbb{R}^{d\times(k+1)}.
$$

Now, given any two labeled graphs (G_1,ℓ_{G_1}) and (G_2,ℓ_{G_2}), Togninalli et al. (2019) first computed $L^k_{G_1}$ and $L^k_{G_2}$, then computed the Wasserstein distance between their induced distributions (which we call the WWL distance) and finally, built a kernel upon this Wasserstein distance. If we let λ_{G_i} denote the uniform measure on V_{G_i}, then we express their WWL distance via pushforward of uniform measures as follows.

$$
D^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2}))\coloneqq d_{\mathrm{W}}\left(\left(L^k_{G_1}\right)_{\#}\lambda_{G_1},\left(L^k_{G_2}\right)_{\#}\lambda_{G_2}\right). \tag{8}
$$

Now, if we instead of uniform measures consider the stationary distributions μ_{G_1} and μ_{G_2} w.r.t. $m^{G_1,\frac{1}{2}}_{\bullet}$ and $m^{G_2,\frac{1}{2}}_{\bullet}$,

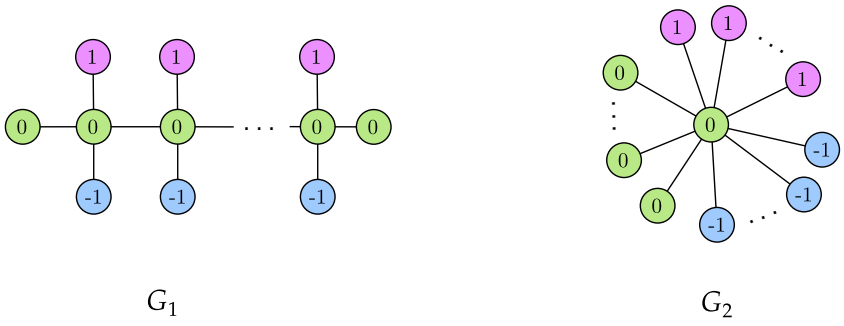

Figure 2. In this figure we show two labeled graphs and each of them has $3n+2$ vertices. Each of the graphs has n vertices with label 1, n vertices with label -1, and $n+2$ vertices with label 0.

respectively, we define the following variant of $D^{(k)}$:

$$\hat{D}^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2})) \coloneqq d_{\mathrm{W}}\left(\left(L^k_{G_1}\right)_{\#}\mu_{G_1},\left(L^k_{G_2}\right)_{\#}\mu_{G_2}\right), \tag{9}$$

which is the distance which we will relate to our WL distance next. In fact, we then prove that $\hat{D}^{(k)}((G_1,\ell_{G_1}),(G_2,\ell_{G_2}))$ actually provides a lower bound for $d^{(k)}_{\mathrm{WL}}((\mathcal{X}_q(G_1),\ell_{G_1}),(\mathcal{X}_q(G_2),\ell_{G_2}))$ when $q=\frac{1}{2}$.

Proposition A.8. *For any two labeled graphs $(G_1,\ell_{G_1}:V_{G_1}\to\mathbb{R}^d)$ and $(G_2,\ell_{G_2}:V_{G_2}\to\mathbb{R}^d)$, one has that for $q=\frac{1}{2}$ and any $k\in\mathbb{N}$,*

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

The proposition will be proved after we provide some examples and remarks.

Example 5 ($d^{(k)}_{\mathrm{WL}}$ is more discriminating than $\hat{D}^{(k)}$)**.** *In this example, we construct a family of pairs of graphs so that $\hat{D}^{(k)}$ between the pairs is always zero but d_{WL} between the pairs is positive. For any $n\geq 2$, consider the two $(3n+2)$-point labeled graphs shown in Figure 2. It is easy to see that for each $i=1,2$, and any $v\in V_{G_i}$,*

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

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

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

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

For proving that $d_{\mathrm{WL}}>0$, we first analyze $\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1),\ell_{G_1})}(v_1)$ for any $v_1\in V_{G_1}$.

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

3. *If v_1 is either the leftmost or the rightmost vertex, then*

$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1),\ell_{G_1})}(v_1) = \delta_0.$$

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

1. *If $\ell_{G_2}(v_2) = \pm 1$, then*

$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}(v_2) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\pm 1}.$$

2. *If v_2 is the center vertex, then*

$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}(v_2) = \frac{2n+1}{3n+1}\delta_0 + \frac{n}{2(3n+1)}\delta_1 + \frac{n}{2(3n+1)}\delta_{-1}.$$

3. *If $\ell_{G_2}(v_2) = 0$ and v_2 is not the center vertex, then*

$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2),\ell_{G_2})}(v_2) = \delta_0.$$

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

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

Remark A.9. *Our WL distance formulation is flexible and we can of course relax it by allowing the comparison of measure Markov chains with general reference probability measures which are not necessarily stationary. In that case, we can directly compare $D^{(k)}$ defined in Equation (8) with $d^{(k)}_{\mathrm{WL}}$. More precisely, for any graph G, we can replace the stationary distribution inherent to $\mathcal{X}_q(G)$ with the uniform measure and hence obtain a new measure Markov chain $\mathcal{X}^{\mathrm{u}}_q(G)$. Then, the same proof technique used for proving Proposition A.8 can be used for proving that*

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

Moreover, we can show that $d^{(k)}_{\mathrm{WL}}$ is strictly more discriminating than $D^{(k)}$ via the same pairs of graphs as in Example 5.

The proof of Proposition A.8 is based on the following basic fact about the Wasserstein distance:

Lemma A.10. *Let Z be a complete and separable metric space. Endow $Z \times Z$ with any product metric $d_{Z\times Z}$ such that*

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

For example, one can let

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

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

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

Proof of Lemma A.10. For any $\gamma_f, \gamma_g \in \mathcal{C}(\mu_1,\mu_2)$, we define a probability measure $\nu \in \mathcal{P}(Z\times Z\times Z\times Z)$ as follows: for any measurable $A, A', B, B' \subseteq Z$

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

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

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

By Lemma A.1, infimizing over all $\gamma_f, \gamma_g \in \mathcal{C}(\mu_1,\mu_2)$, we obtain the conclusion. □

Proof of Proposition A.8. Since $L^k_G \coloneqq (\ell^0_G, \ldots, \ell^k_G)$, by inductively applying Lemma A.10, we have that

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

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

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

Then, by Equation (19), we conclude that

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

By Proposition 3.1, we have that

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

□

A.4. A Decoupled Version of the Gromov-Wasserstein Distance

For simplicity, in this section we will assume that the cardinality of all underlying spaces to always be finite.

The Gromov-Wasserstein distance d_{GW} was proposed as a measure of dissimilarity between two metric measure spaces; see Section 2.2 for its definition and also (Mémoli, 2011) for its more general version involving a parameter $p \in [1, \infty]$. Note that one can define a variant of the standard GW distance by considering two coupling measures γ, γ' *independently*, and use $\gamma \otimes \gamma'$ instead of $\gamma \otimes \gamma$ in Equation (1). This version of the GW distance was implicit in the optimization procedure followed in (Mémoli, 2011) and has been explicitly considered in (Séjourné et al., 2021; Titouan et al., 2020), and this is closely connected to our GW distance between MCMSs (see Definition 8) as shown in Example 3.

Here we give the definition of this "decoupled" variant of the GW distance.

Definition 12 (Decoupled Gromov-Wasserstein distance)**.** *Suppose two metric measure spaces $\mathbf{X} = (X, d_X, \mu_X)$, $\mathbf{Y} = (Y, d_Y, \mu_Y)$ are given. We define the decoupled Gromov-Wasserstein distance $d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y})$ in the following way:*

$$d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) \coloneqq \inf_{\gamma, \gamma' \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X\times Y} \int_{X\times Y} |d_X(x, x') - d_Y(y, y')| \gamma'(dx' \times dy') \gamma(dx \times dy).$$

Obviously, $d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) \leq d_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y})$ in general. Furthermore, this inequality is actually tight as one can see in the following remark.

Remark A.11. *We let $\Gamma_{X,Y}(x, y, x', y') \coloneqq |d_X(x, x') - d_Y(y, y')|$ for any $x, x' \in X$ and $y, y' \in Y$. If the kernel $\Gamma_{X,Y} : X \times Y \times X \times Y \to \mathbb{R}$ is negative semi-definite, then one can show that $d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) = d_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y})$ by invoking Séjourné et al. (2021, Theorem 4). More precisely, if γ, γ' are the optimal coupling measures achieving the infimum in the definition of $d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y})$, then both γ and γ' are optimal for d_{GW}, i.e.,*

$$d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) = \|\Gamma_{X,Y}\|_{L^1(\gamma\otimes\gamma')} = \|\Gamma_{X,Y}\|_{L^1(\gamma\otimes\gamma)} = \|\Gamma_{X,Y}\|_{L^1(\gamma'\otimes\gamma')} = d_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}).$$

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

Proposition A.12. *The decoupled Gromov Wasserstein distance $d^{\mathrm{bi}}_{\mathrm{GW}}$ is a legitimate metric on $\mathcal{M}^{\mathrm{MS}}$.*

Proof. Symmetry is obvious. We need to prove the triangle inequality plus the fact that $d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) = 0$ happens if and only if $\mathbf{X}$ and $\mathbf{Y}$ are isomorphic. The "if" part is trivial. For the other direction we proceed as follows. Suppose that $d^{\mathrm{bi}}_{\mathrm{GW}}(\mathbf{X}, \mathbf{Y}) = 0$. By Lemma B.11 and the compactness of $\mathcal{C}(\mu_X, \mu_Y)$ for the weak topology (see Villani (2003, p.49)), there must be optimal couplings $\gamma, \gamma' \in \mathcal{C}(\mu_X, \mu_Y)$ such that

$$\sum_{(x,y)\in X\times Y} \sum_{(x',y')\in X\times Y} |d_X(x, x') - d_Y(y, y')| \, \gamma'(x', y') \, \gamma(x, y) = 0. \tag{10}$$

Claim 1. *There exists an isometry $\phi : X \to Y$ such that*

$$\{(x, \phi(x)) : x \in X\} = \mathrm{supp}(\gamma) = \mathrm{supp}(\gamma').$$

Proof of Claim 1. By Equation (10), we have that

$$d_X(x, x') = d_Y(y, y') \tag{11}$$

for any $(x, y) \in \mathrm{supp}(\gamma)$ and $(x', y') \in \mathrm{supp}(\gamma')$.

Fix an arbitrary $x \in X$. Then, since both μ_X and μ_Y are fully supported and X, Y are finite, there must exist $y, y' \in Y$ such that $(x, y) \in \mathrm{supp}(\gamma)$ and $(x, y') \in \mathrm{supp}(\gamma')$. Then, $y = y'$ by Equation (11). Now, if there exists $y'' \in Y$ such that $(x, y'') \in \mathrm{supp}(\gamma)$, then similarly, we have that $y'' = y'$ and thus $y'' = y$. In other words, for each $x \in X$, there exists a unique $y \in Y$ such that $(x, y) \in \mathrm{supp}(\gamma)$. Similarly, this same $y \in Y$ is unique such that $(x, y) \in \mathrm{supp}(\gamma')$. Hence, we define $\phi : X \to Y$ by letting $\phi(x)$ be the unique $y \in Y$ such that $(x, \phi(x)) \in \mathrm{supp}(\gamma)$. It is obvious that ϕ is bijective and satisfies that $\{(x, \phi(x)) : x \in X\} = \mathrm{supp}(\gamma) = \mathrm{supp}(\gamma')$. By Equation (11), we conclude that ϕ is an isometry. □

Based on the claim above, consider an arbitrary Borel subset $A \subseteq X$. Then,

$$\mu_X(A) = \gamma(A \times Y) = \gamma(A \times Y \cap A \times \phi(A)) = \gamma(A \times \phi(A)) = \mu_Y(\phi(A)).$$

Hence, ϕ is a isomorphism between $\mathbf{X}$ and $\mathbf{Y}$.

Finally, for the triangle inequality, fix finite metric measure space $\mathbf{X}$,$\mathbf{Y}$, and $\mathbf{Z}$. Notice first that for all $x, x' \in X$, $y, y' \in Y$, and $z, z' \in Z$,

$$\Gamma_{X,Y}(x, y, x', y') \leq \Gamma_{X,Z}(z, x, z', x') + \Gamma_{Z,Y}(y, z, y', z').$$

Next, fix arbitrary coupling measures $\gamma_1, \gamma_1' \in \mathcal{C}(\mu_X, \mu_Z)$ and $\gamma_2, \gamma_2' \in \mathcal{C}(\mu_Z, \mu_Y)$. By the Gluing Lemma (see Villani (2003, Lemma 7.6)), there exist probability measures $\pi, \pi' \in \mathcal{P}(X \times Y \times Z)$ with marginals γ_1, γ_1' on $X \times Z$ and γ_2, γ_2' on $Z \times Y$. Let γ_3, γ_3' be the marginal of π, π' on $X \times Y$. Then, by the triangle inequality of L^1 norm,

$$\begin{aligned} d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) &\leq \|\Gamma_{X,Y}\|_{L^1(\gamma_3 \otimes \gamma_3')} \\ &= \|\Gamma_{X,Y}\|_{L^1(\pi \otimes \pi')} \\ &\leq \|\Gamma_{X,Z}\|_{L^1(\pi \otimes \pi')} + \|\Gamma_{Z,Y}\|_{L^1(\pi \otimes \pi')} \\ &= \|\Gamma_{X,Z}\|_{L^1(\gamma_1 \otimes \gamma_1')} + \|\Gamma_{Z,Y}\|_{L^1(\gamma_2 \otimes \gamma_2')}. \end{aligned}$$

Since the choice of $\gamma_1, \gamma_1', \gamma_2, \gamma_2'$ are arbitrary, by taking the infimum one can conclude

$$d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Y}) \leq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X}, \mathbf{Z}) + d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{Z}, \mathbf{Y}).$$

□

A.5. Examples When d_{WL} Fails to Separate Graphs

Example 6 (Constant labels)**.** *Let G_1 be a claw and G_2 be a path with four nodes; see Figure 3. Let the label functions ℓ_{G_i} for $i = 1, 2$ be constant and equal to 1 for both graphs.*

In the first step of the WL test, we find

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

and

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

Since $L_1((G_1, \ell_{G_1})) \neq L_1((G_2, \ell_{G_2}))$, (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) are recognized as non-isomorphic by the WL test. Notice that within the first step, the WL test collects degree information and comparing $L_1((G_1, \ell_{G_1}))$ and $L_1((G_2, \ell_{G_2}))$ is equivalent to comparing the multisets of degrees w.r.t. G_1 and G_2. However, for $d_{\mathrm{WL}}((\mathcal{X}(G_1), \ell_{G_1}), (\mathcal{X}(G_2), \ell_{G_2}))$ (abbreviated to $d_{\mathrm{WL}}(\mathcal{X}(G_1), \mathcal{X}(G_2))$ in Figure 3), because of the normalization inherent to the Markov chains $m_\bullet^{G_1}$ and $m_\bullet^{G_2}$, each step inside the hierarchy pertaining to the WL distance cannot collect degree information when the labels are constant.

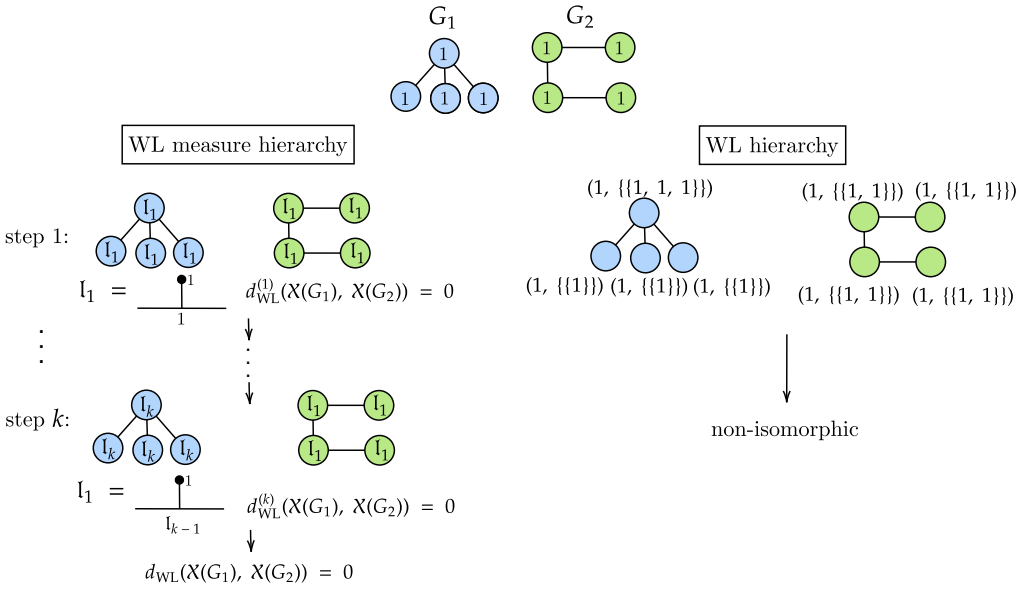

Figure 3. **Illustration of Example 6**. Notice that if we start from constant labels, one step of the WL hierarchy will collect degree information for the vertices. In contrast, because of the normalization of the Markov kernel, a single step of the WL measure hierarchy with constant labels will not be able to accumulate the same information.

Example 7 (Degree label)**.** *Let G_1 be a two-point graph consisting of a single edge with the vertex set $\{v_1, v_2\}$. Let G_2 be a four-point graph consisting of two disjoint edges denoted by $\{u_1, u_2\}$ and $\{u_3, u_4\}$; see Figure 4. For each $i = 1, 2$, let ℓ_{G_i} be the degree label function for both graphs.*

In the first step of the WL test,

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

and

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

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

In the case of $d_{\mathrm{WL}}((\mathcal{X}(G_1), \ell_{G_1}), (\mathcal{X}(G_2), \ell_{G_2}))$, notice that $(\ell_{G_1})_{\#} m_x^{G_1}(z) = 1$ if $z = 1$ and 0 otherwise for both $x = v_1$ and $x = v_2$. Hence, $\mathfrak{l}^{(1)}_{(\mathcal{X}(G_1), \ell_{G_1})}(v_1) = \mathfrak{l}^{(1)}_{(\mathcal{X}(G_1), \ell_{G_1})}(v_2)$. Similarly, for G_2,

$$\mathfrak{l}^{(1)}_{(\mathcal{X}(G_2), \ell_{G_2})}(u_1) = \cdots = \mathfrak{l}^{(1)}_{(\mathcal{X}(G_2), \ell_{G_2})}(u_4) = \mathfrak{l}^{(1)}_{(\mathcal{X}(G_1), \ell_{G_1})}(v_1).$$

It is not hard to show inductively that for each $k \in \mathbb{N}$,

$$\mathfrak{l}^{(k)}_{(\mathcal{X}(G_1), \ell_{G_1})}(v_1) = \mathfrak{l}^{(k)}_{(\mathcal{X}(G_1), \ell_{G_1})}(v_2) = \mathfrak{l}^{(k)}_{(\mathcal{X}(G_2), \ell_{G_2})}(u_1) = \cdots = \mathfrak{l}^{(k)}_{(\mathcal{X}(G_2), \ell_{G_2})}(u_4).$$

Then, for each $k \in \mathbb{N}$,

$$\mathfrak{L}_1((\mathcal{X}(G_1), \ell_{G_1})) = \left(\mathfrak{l}^{(k)}_{(\mathcal{X}(G_1), \ell_{G_1})}\right)_{\#} \mu_{G_1} = \left(\mathfrak{l}^{(k)}_{(\mathcal{X}(G_2), \ell_{G_2})}\right)_{\#} \mu_{G_2} = \mathfrak{L}_1((\mathcal{X}(G_2), \ell_{G_2}))$$

and thus $d^{(k)}_{\mathrm{WL}}((\mathcal{X}(G_1), \ell_{G_1}), (\mathcal{X}(G_2), \ell_{G_2})) = 0$ which implies that $d_{\mathrm{WL}}((\mathcal{X}(G_1), \ell_{G_1}), (\mathcal{X}(G_2), \ell_{G_2})) = 0$.

Notice that the standard WL test with degree labels is able to capture (and therefore compare) information about the number of nodes in the graph. On the other hand, $(\mathcal{X}(G_1), \ell_{G_1})$ and $(\mathcal{X}(G_2), \ell_{G_2})$ cannot be distinguished by the WL distance because of the normalization of the reference measures, μ_{G_1} and μ_{G_2}.

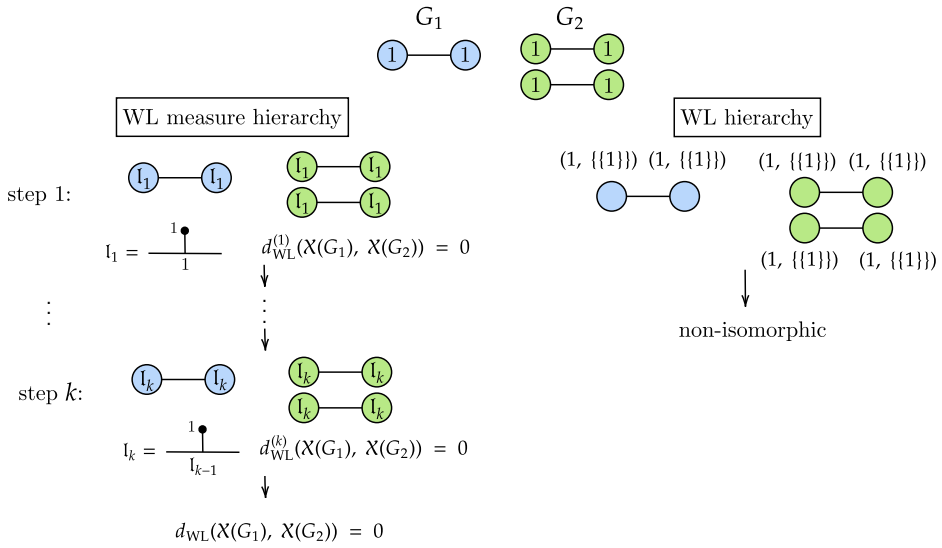

Figure 4. **Illustration of Example 7**. One step of the WL hierarchy with degree labels can distinguish graphs of different sizes whereas the normalization of μ_{G_1} and μ_{G_2} does not allow graph size to be distinguished when the label function is degree.

A.6. A Basic Lower Bound for $d_{\mathrm{GW}}^{\mathrm{MCMS}}$

One can produce some basic lower bounds for $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ by invoking the notion of *diameter* for MCMSs which we define below. We first introduce the one point MCMS.

Example 8. *The one point MCMS is the tuple* $* := (\{*\}, (0), \delta_*, \delta_*)$.

Definition 13 (MCMS diameter)**.** *For each* $k \geq 1$ *and a MCMS* $(\mathcal{X}, d_X)$*, we define*

$$\mathrm{diam}_{\mathrm{MCMS}}^{(k)}((\mathcal{X}, d_X)) := d_{\mathrm{GW}}^{(k)}((\mathcal{X}, \ell_X), *),$$

and

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

Notice that $\mathcal{C}(\mu_X, \mu_*) = \{\mu_X \otimes \delta_*\}$ and $\mathcal{C}^{(k)}\left(m_\bullet^X, \delta_*\right) = \left\{m_\bullet^{X,\otimes k} \otimes \delta_*\right\}$. Then, it turns out that the diameter of $\mathcal{X}$ is independent of k:

$$\begin{aligned}\mathrm{diam}_{\mathrm{MCMS}}^{(k)}((\mathcal{X}, d_X)) = d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), *) &= \int_X \int_X \int_X d_X(x, x')\, m_{x''}^{X,\otimes k}(dx')\mu_X(dx'')\mu_X(dx) \\ &= \int_X \int_X d_X(x, x')\mu_X(dx')\mu_X(dx)\end{aligned}$$

By the triangle inequality, one can prove the following result.

Proposition A.13. *For all two MCMSs* $(\mathcal{X}, d_X)$*,*$(\mathcal{Y}, d_Y)$ *and* $k \geq 1$*, we have*

$$\left|\int_X \int_X d_X(x, x')\mu_X(dx')\mu_X(dx) - \int_Y \int_Y d_Y(y, y')\mu_Y(dy')\mu_Y(dy)\right| \leq d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)),$$

and

$$\left|\int_X \int_X d_X(x, x')\mu_X(dx')\mu_X(dx) - \int_Y \int_Y d_Y(y, y')\mu_Y(dy')\mu_Y(dy)\right| \leq d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)).$$

A.7. More Details on the Complexity of Computing the WL Distance

In the following subsections, we provide an algorithm for computing the WL distance and complexity analysis for both computing the WL distance and its lower bound defined in Section 3.2.

A.7.1. Computation of the WL Distance

In this section, we devise an algorithm (with pseudocode in Algorithm 1) for computing $d_{\mathrm{WL}}^{(k)}$ and establish the following complexity analysis.

Proposition A.14. *For any fixed $k \in \mathbb{N}$, computing $d_{\mathrm{WL}}^{(k)}$ between any LMMCs $(\mathcal{X}, \ell_X)$ and $(\mathcal{Y}, \ell_Y)$ can be achieved in time at most $O(k\, n^5 \log(n))$ where $n = \max(|X|, |Y|)$.*

Recall from Equation (2) that the WL distance of depth k is defined as

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

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

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

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

$$d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(i)}(x), \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(i)}(y)\right) = \inf_{\gamma \in \mathcal{C}(m_x^X, m_y^Y)} \int_{X \times Y} d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(i-1)}(x), \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(i-1)}(y)\right) \gamma(dx \times dy).$$

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

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

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

For any $n \in \mathbb{N}$, $d_{\mathrm{WL}}^{(2n)}$ generates a distance between graph induced LMMCs with size bounded by n. By Corollary 3.5, $d_{\mathrm{WL}}^{(2n)}$ has the same discriminating power as the WL test in separating graphs with size bounded by n. Now, given labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}) so that $\max(|V_{G_1}|, |V_{G_2}|) \leq n$, computing $d_{\mathrm{WL}}^{(2n)}((\mathcal{X}_q(G_1), \ell_{G_1}), (\mathcal{X}_q(G_2), \ell_{G_2}))$ takes time at most $O(n^6 \log(n))$.

A.7.2. Computation of the Lower Bound Distance

Recall from Section 3.2 that the WL lower bound distance was defined as

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

Given two finite LMMCs,

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

and

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

we represent their Markov kernels as two transition matrices, $M_{\mathcal{X}}$ and $M_{\mathcal{Y}}$, respectively. Then, k-Markov kernels $m_\bullet^{X, \otimes k}$ and $m_\bullet^{Y, \otimes k}$ are expressed as matrices $M_{\mathcal{X}}^k$ and $M_{\mathcal{Y}}^k$, respectively. Assume that $n \geq m$. Then computing the k-Markov

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

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

kernels of $\mathcal{X}$ and $\mathcal{Y}$ will require $O(n^3 \log(k))$ time where $O(n^3)$ is time needed for matrix multiplication. Then since $(\ell_X)_\# m_x^{X,\otimes k}$ and $(\ell_Y)_\# m_y^{Y,\otimes k}$ are both distributions in $\mathbb{R}$, by (Vallender, 1974), $d_{\mathrm{W}}\left((\ell_X)_\# m_x^{X,\otimes k}, (\ell_Y)_\# m_y^{Y,\otimes k}\right)$ can be computed in $O(n)$ time for each $x \in X$ and $y \in Y$. Finally, computing $d_{\mathrm{WLLB}}^{(k)}$ can be formulated as finding the optimal transport cost where each entry of the cost matrix is defined as $d_{\mathrm{W}}\left((\ell_X)_\# m_x^{X,\otimes k}, (\ell_Y)_\# m_y^{Y,\otimes k}\right)$ and the source and target distributions are μ_X, μ_Y respectively. Recall from the previous section that μ_X and μ_Y are normalized degree distributions for $\mathcal{X}$ and $\mathcal{Y}$. Therefore, the overall time complexity is

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

A.8. Experiments

A.8.1. Experimental Setup

We use several publicly available graph benchmark datasets from TUDatasets (Morris et al., 2020) and evaluate the performance of our WL distance $d_{\mathrm{WL}}^{(k)}$ distance as well as $d_{\mathrm{WLLB}}^{(k)}$ (lower bound of our d_{WL} which is more efficient to compute) through two types of graph classification experiments compared with several representative methods. Note that for all of our experiments, we use $q = 0.6$ to transform every graph G into the Markov chain $\mathcal{X}_q(G)$.

We use 1-Nearest Neighbors classifier in the first graph classification experiment and for the second experiment, we use support vector machines (SVM). For the first experiment, we compare classification accuracies with the WWL distance (Togninalli et al., 2019) (see Equation (8)). For the second graph classification task, we run an SVM using the indefinite kernel matrices $\exp\left(-\gamma d_{\mathrm{WLLB}}^{(k)}\right)$ and $\exp\left(-\gamma d_{\mathrm{WL}}^{(k)}\right)$, which are seen as noisy observations of the true positive semi-definite kernels (Luss & d'Aspremont, 2009). We also evaluate the classification accuracies of d_{WLLB} and $d_{\mathrm{WL}}^{(k)}$ using KSVM (Loosli et al., 2015) for their indefinite kernel matrices. Additionally, for the SVM method, we cross validate the parameter $C \in \{10^{-3}, \ldots, 10^3\}$ and the parameter $\gamma \in \{10^{-3}, \ldots, 10^3\}$. We compare classification accuracies with the WWL kernel (Togninalli et al., 2019), the WL kernel (Shervashidze et al., 2011), and the Weisfeiler-Lehman optimal assignment kernel (WL-OA) (Kriege et al., 2016). Note that we only use WWL distance in the 1-NN graph classification experiment since the WL-OA and WL kernels are not defined in terms of a distance unlike the WWL kernel.

In addition to the full accuracies for $k \in \{1, 2, 3, 4\}$ for $d_{\mathrm{WL}}^{(k)}$ and $d_{\mathrm{WLLB}}^{(k)}$ with the degree label, call this f_1, we also evaluate $d_{\mathrm{WL}}^{(k)}$ and $d_{\mathrm{WLLB}}^{(k)}$ with the label function $f_2(G, v) = \frac{1}{|V_G|} + \deg_G(v)$ for any graph G and vertex $v \in V_G$. Note that f_2 is a relabeling of any constant label function, which assigns a constant c to each vertex, via the injective map $g : \{c\} \times \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ sending (c, n_1, n_2) to $n_1 + \frac{1}{n_2}$ as described in Section 3.1. So under f_2, $d_{\mathrm{WL}}^{(k)}$ is as discriminative as the k-step WL test. Thus, we also evaluate the performance of the WWL distance/kernel, WL, and WL-OA kernels using only degree label. Additionally, we report the best accuracies for WWL, WL, and WL-OA for iterations $1, \ldots, 4$.

Table 3. 1-Nearest Neighbor classification accuracy. Let $f_1(G,v) = \deg_G(v)$, $f_2(G,v) = \frac{1}{|V_G|} + \deg_G(v)$

METHOD	MUTAG	PROTEINS	PTC-FM	PTC-MR	IMDB-B	IMDB-M	COX2
$d_{\mathrm{WL}}^{(1)}, f_1$	90.5 ± 6.5	61.8 ± 4.3	60.0 ± 8.5	53.9 ± 7.1	70.1 ± 4.7	41.1 ± 3.9	73.8 ± 3.6
$d_{\mathrm{WL}}^{(2)}, f_1$	92.1 ± 6.3	60.8 ± 4.4	62.2 ± 7.4	56.2 ± 6.3	69.9 ± 4.2	41.1 ± 4.7	74.2 ± 4.5
$d_{\mathrm{WL}}^{(3)}, f_1$	91.1 ± 4.3	60.8 ± 3.5	59.4 ± 8.2	54.0 ± 7.7	69.4 ± 3.9	41.0 ± 4.8	74.2 ± 3.9
$d_{\mathrm{WL}}^{(4)}, f_1$	90.1 ± 4.8	63.0 ± 3.8	59.1 ± 8.3	54.2 ± 6.8	70.2 ± 4.3	41.3 ± 4.8	76.1 ± 5.5
$d_{\mathrm{WL}}^{(1)}, f_2$	91.6 ± 7.1	63.3 ± 4.4	57.5 ± 6.0	51.9 ± 9.3	$\mathbf{71.4 \pm 4.5}$	40.6 ± 5.3	72.5 ± 4.5
$d_{\mathrm{WL}}^{(2)}, f_2$	91.1 ± 5.8	62.4 ± 3.4	58.2 ± 8.2	56.2 ± 7.6	70.4 ± 4.5	$\mathbf{41.6 \pm 4.3}$	74.0 ± 4.7
$d_{\mathrm{WL}}^{(3)}, f_2$	91.5 ± 5.8	63.4 ± 3.9	58.5 ± 7.9	53.4 ± 8.4	$\mathbf{71.4 \pm 5.9}$	40.6 ± 4.3	74.6 ± 4.4
$d_{\mathrm{WL}}^{(4)}, f_2$	$\mathbf{92.6 \pm 4.8}$	63.3 ± 4.9	58.5 ± 8.0	54.8 ± 7.9	71.2 ± 5.1	40.7 ± 4.8	75.9 ± 4.9
$d_{\mathrm{WLLB}}^{(1)}, f_1$	87.3 ± 1.9	64.0 ± 2.3	$\mathbf{62.5 \pm 8.5}$	57.4 ± 6.8	69.0 ± 3.9	40.6 ± 3.8	75.1 ± 3.8
$d_{\mathrm{WLLB}}^{(2)}, f_1$	86.8 ± 3.7	$\mathbf{66.2 \pm 2.2}$	60.0 ± 8.1	53.4 ± 6.4	69.4 ± 3.2	40.1 ± 3.6	75.1 ± 3.8
$d_{\mathrm{WLLB}}^{(3)}, f_1$	85.2 ± 3.5	64.6 ± 2.2	58.0 ± 1.1	54.5 ± 9.1	69.8 ± 3.3	40.1 ± 3.9	$\mathbf{81.2 \pm 5.3}$
$d_{\mathrm{WLLB}}^{(4)}, f_1$	84.7 ± 3.1	65.4 ± 2.3	58.0 ± 1.1	52.0 ± 9.1	69.9 ± 2.5	40.1 ± 3.6	80.4 ± 2.3
$d_{\mathrm{WLLB}}^{(1)}, f_2$	87.3 ± 2.5	64.7 ± 1.4	$\mathbf{62.5 \pm 7.4}$	$\mathbf{57.8 \pm 6.8}$	69.0 ± 3.9	40.4 ± 3.6	75.5 ± 3.7
$d_{\mathrm{WLLB}}^{(2)}, f_2$	86.3 ± 3.6	65.6 ± 2.2	60.0 ± 8.1	53.4 ± 6.4	69.2 ± 3.2	40.2 ± 3.6	77.0 ± 4.9
$d_{\mathrm{WLLB}}^{(3)}, f_2$	85.3 ± 3.6	64.3 ± 1.1	58.0 ± 10.8	54.7 ± 9.1	69.7 ± 3.1	40.1 ± 3.9	80.4 ± 4.4
$d_{\mathrm{WLLB}}^{(4)}, f_2$	84.7 ± 3.0	64.8 ± 1.8	58.0 ± 10.9	52.0 ± 9.2	69.4 ± 2.5	40.2 ± 3.8	80.4 ± 4.4
WWL	85.1 ± 6.5	64.7 ± 2.8	58.2 ± 8.5	54.3 ± 7.9	65.0 ± 3.3	40.0 ± 3.3	76.1 ± 5.6

A.8.2. EXTRA EXPERIMENTAL RESULTS

In Table 3 and Table 4, we have included the 1-NN and SVM classification accuracies for $k \in \{1,2,3,4\}$, respectively. In general, using a KSVM with kernels $\exp\left(-\gamma d_{\mathrm{WLLB}}^{(k)}\right)$ and $\exp\left(-\gamma d_{\mathrm{WL}}^{(k)}\right)$ yields classification accuracies which are similar to the classification accuracy of a standard SVM with the indefinite kernels. However, note that for the kernel $\exp\left(-\gamma d_{\mathrm{WL}}^{(k)}\right)$, KSVM has a slightly higher classification accuracy on both PTC-FM and IMDB-B than standard SVM. For the kernel $\exp\left(-\gamma d_{\mathrm{WLLB}}^{(k)}\right)$ with the labels generated by f_2, KSVM has a much lower classification accuracy on both PROTEINS and IMDB-B than standard SVM.

A.8.3. TIME COMPARISON

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

B. Proofs

B.1. Proofs from Section 2

B.1.1. PROOF OF PROPOSITION 2.1

The "only if" part is obvious. To prove the "if" part, we assume $(\mathcal{X}_q(G_1), \ell_{G_1})$ is isomorphic to $(\mathcal{X}_q(G_2), \ell_{G_2})$. Then, there exists a bijective map $\psi : V_{G_1} \to V_{G_2}$ such that $\psi_{\#} m_v^{G_1,q} = m_{\psi(v)}^{G_2,q}$, $\psi_{\#}\mu_{G_1} = \mu_{G_2}$ and $\ell_{G_1}(v) = \ell_{G_2}(\psi(v))$ for all $v \in V_{G_1}$. Now, by the definition of $m_\bullet^{G_1,q}$ and $m_\bullet^{G_2,q}$ (see Definition 5), one can easily check that

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

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

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

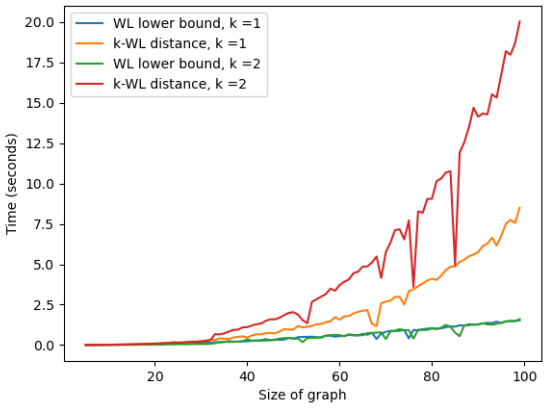

Figure 5. Comparison of runtime of WWL distance against $d_{\mathrm{WL}}^{(k)}$ and its lower bound $d_{\mathrm{WLLB}}^{(k)}$.

Hence, G_1 and G_2 are isomorphic as we required.

B.2. Proofs from Section 3

B.2.1. Proof of the claim in Example 1

By Lemma A.1 we have that

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

B.2.2. Proof of Proposition 3.1

This proposition follows directly from Lemma A.6 and Theorem A.7.

B.2.3. Proof of Proposition 3.2

It is obvious that when $(\mathcal{X},\ell_X)$ is isomorphic to $(\mathcal{Y},\ell_Y)$, $d_{\mathrm{WL}}^{(k)}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = 0$ for all $k\in\mathbb{N}$ and thus $d_{\mathrm{WL}}((\mathcal{X},\ell_X),(\mathcal{Y},\ell_Y)) = 0$. It follows directly from Equation (2) that $d_{\mathrm{WL}}^{(k)}$ satisfies the triangle inequality. Hence, $d_{\mathrm{WL}} := \sup_{k\geq 1} d_{\mathrm{WL}}^{(k)}$ also satisfies the triangle inequality.

B.2.4. Proof of Lemma 3.4

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

$$\ell_{(G_1,\ell_{G_1})}^{(k+1)}(v_1) = \ell_{(G_2,\ell_{G_2})}^{(k+1)}(v_2) \Longrightarrow \ell_{(G_1,\ell_{G_1}^g)}^{(k)}(v_1) = \ell_{(G_2,\ell_{G_2}^g)}^{(k)}(v_2),\ \forall k=0,1,\ldots. \tag{12}$$

We prove Equation (12) by induction on k. When $k=0$, for any $v_1\in V_{G_1}$ and $v_2\in V_{G_2}$, if $\ell_{(G_1,\ell_{G_1})}^{1}(v_1) = \ell_{(G_2,\ell_{G_2})}^{1}(v_2)$,

then

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

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

Now, we assume that Equation (12) holds for some $k \geq 0$. For the case of $k+1$, note that $\ell^{(k+2)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k+2)}_{(G_2,\ell_{G_2})}(v_2)$ implies that

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

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

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

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

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

$$\ell^{(k)}_{(G_1,\ell^g_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell^g_{G_2})}(v_2) \Longrightarrow \ell^{(k)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2,\ell_{G_2})}(v_2),\ \forall k = 0, 1, \ldots. \tag{13}$$

We prove Equation (13) by induction on k. When $k = 0$, for any $v_1 \in V_{G_1}$ and $v_2 \in V_{G_2}$, if $\ell^g_{G_1}(v_1) = \ell^g_{G_2}(v_2)$, then by injectivity of g, we have that $\ell_{G_1}(v_1) = \ell_{G_2}(v_2)$.

Now, we assume that Equation (13) holds for some $k \geq 0$. For the case of $k+1$, note that $\ell^{(k+1)}_{(G_1,\ell^g_{G_1})}(v_1) = \ell^{(k+1)}_{(G_2,\ell^g_{G_2})}(v_2)$ implies that

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

By the induction assumption, it is easy to see that

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

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

B.2.5. Proof of Proposition 3.3

By Lemma 3.4, we only need to prove that the WL test cannot distinguish $(G_1,\ell^g_{G_1})$ and $(G_2,\ell^g_{G_2})$ iff $d_{\mathrm{WL}}\big((\mathcal{X}_q(G_1),\ell^g_G),(\mathcal{X}_q(G_2),\ell^g_{G_2})\big) = 0$. For this purpose, we need to introduce some new notions.

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

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

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

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

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

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

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

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

$$\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G), \ell_G)} = \varphi_q^k \circ \ell^{(k)}_{(G, \ell_G)} : V_G \to \mathcal{P}^{\circ k}(Z). \tag{14}$$

Proof of Lemma B.1. We prove by induction on k.

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

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

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

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

Now, we assume that Equation (14) holds for some $k \geq 1$. Then, for $k+1$ and for any $v \in V$, if $N_G(v) \neq \emptyset$, we have that

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

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

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

This concludes the proof. □

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

$$\ell^{(k)}_{(G_1, \ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2, \ell_{G_2})}(v_2) \;\; \textit{iff} \;\; \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell_{G_1})}(v_1) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell_{G_2})}(v_2). \tag{15}$$

Proof of Lemma B.2. By Lemma B.1, we have that $\ell^{(k)}_{(G_1, \ell_{G_1})}(v_1) = \ell^{(k)}_{(G_2, \ell_{G_2})}(v_2)$ implies that $\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1), \ell_{G_1})}(v_1) = \mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2), \ell_{G_2})}(v_2)$. For the other direction, we prove by induction on k.

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

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

If otherwise $N_{G_1}(v_1) \neq \emptyset$, then $\mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$ again implies that $\frac{1-q}{\deg(v_1)}\sum_{v\in N_{G_1}(v_1)}\delta_{\ell_{G_1}(v)} = \frac{1-q}{\deg(v_2)}\sum_{v\in N_{G_2}(v_2)}\delta_{\ell_{G_2}(v)}$. Hence, $\sum_{v\in N_{G_1}(v_1)}\delta_{\ell_{G_1}(v)} = \sum_{v\in N_{G_2}(v_2)}\delta_{\ell_{G_2}(v)}$ and thus $\{\!\{\ell_{G_1}(v) : v \in N_{G_1}(v_1)\}\!\} = \{\!\{\ell_{G_2}(v) : v \in N_{G_2}(v_2)\}\!\}$. Therefore, $\ell^{(1)}_{(G_1,\ell_{G_1})}(v_1) = \ell^{(1)}_{(G_2,\ell_{G_2})}(v_2)$.

Now, we assume that Equation (15) holds for some $k \geq 1$. Note that

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

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

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

If otherwise $N_{G_1}(v_1) \neq \emptyset$, then $\mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v_1) = \mathfrak{l}^{(k+1)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v_2)$ again implies that $\frac{1-q}{\deg(v_1)}\sum_{v\in N_{G_1}(v_1)}\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v)} = \frac{1-q}{\deg(v_2)}\sum_{v\in N_{G_2}(v_2)}\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v)}$. Hence, $\sum_{v\in N_{G_1}(v_1)}\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell_{G_1})}(v)} = \sum_{v\in N_{G_2}(v_2)}\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell_{G_2})}(v)}$ and thus

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

Then, by the induction assumption again, we have that

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

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

Now, we are ready to prove that

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

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

$$L_k((G_1,\ell^g_1)) = L_k((G_2,\ell^g_2)) \text{ iff } \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell^g_{G_1})}\right)_{\#}\mu_{G_1} = \left(\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell^g_{G_2})}\right)_{\#}\mu_{G_2} \tag{16}$$

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

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

2. By Lemma B.1, for any $v \in V_{G_1}$ we have that

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

Then,

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

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

$$\sum_{v\in V_{G_1}} \frac{\overline{\deg}_{G_1}(v)}{\sum_{v'\in V_{G_1}}\overline{\deg}_{G_1}(v')}\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_1),\ell^g_{G_1})}(v)} = \sum_{v\in V_{G_2}} \frac{\overline{\deg}_{G_2}(v)}{\sum_{v'\in V_{G_2}}\overline{\deg}_{G_2}(v')}\delta_{\mathfrak{l}^{(k)}_{(\mathcal{X}_q(G_2),\ell^g_{G_2})}(v)}. \tag{17}$$

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

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

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

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

Hence, by injectivity of g, we have that $|V_{G_1}| = |V_{G_2}|$, $\deg_{G_1}(v_1) = \deg_{G_2}(v_2)$ and $\overline{\deg}_{G_1}(v_1) = \overline{\deg}_{G_2}(v_2)$. Then, it is easy to see from Equation (17) that

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

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

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

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

B.2.6. Proof of Corollary 3.5

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

Proposition B.3. *For any labeled graphs (G_1, ℓ_{G_1}) and (G_2, ℓ_{G_2}), $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ holds for all $k = 0, \ldots, |V_{G_1}| + |V_{G_2}|$ if and only if $L_k((G_1, \ell_{G_1})) = L_k((G_2, \ell_{G_2}))$ holds for all $k \geq 0$.*

Hence, this corollary is a direct consequence of Proposition 3.3 and Proposition B.3.

B.3. Proofs from Section 4

B.3.1. Proof of Proposition 4.1

We need the following lemma:

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.*

Proof of Lemma B.4. For any $\alpha, \beta \in \mathcal{P}(\mathbb{R}^i)$, pick any $\gamma \in \mathcal{C}(\alpha, \beta)$. Then, we have that

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

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

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

Hence q_φ is C-Lipschitz. □

Now, we start to prove item 1. We introduce some notation. Given a MCNN$_k$ $h := \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1}$ and any $(\mathcal{X}, \ell_X) \in \mathcal{M}^L(Z)$, we let

$$\left(\mathcal{X}, \ell_X^{(\varphi,i)}\right) := F_{\varphi_i} \circ \cdots \circ F_{\varphi_1}((\mathcal{X}, \ell_X)) \tag{18}$$

Now, we assume that for $i = 1, \ldots, k$, the MLP φ_i is C_i-Lipschitz for some $C_i > 0$ (note that MLPs with the Lipschitz activation function σ specified in Section 4 are Lipschitz). Then, by Lemma B.4, we have that q_{φ_i} is a C_i-Lipschitz map for $i = 1, \ldots, k$.

Then, we prove that

$$d_{\mathrm{W}}\left(\left(\ell_X^{(\varphi,k)}\right)_{\#} \mu_X, \left(\ell_Y^{(\varphi,k)}\right)_{\#} \mu_Y\right) \leq \left(\Pi_{i=1}^k C_i\right) \cdot d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)). \tag{19}$$

Given Equation (19), if $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) = 0$, then $d_{\mathrm{W}}\left(\left(\ell_X^{(\varphi,k)}\right)_{\#} \mu_X, \left(\ell_Y^{(\varphi,k)}\right)_{\#} \mu_Y\right) = 0$ and thus $\left(\ell_X^{(\varphi,k)}\right)_{\#} \mu_X = \left(\ell_Y^{(\varphi,k)}\right)_{\#} \mu_Y$. Hence, the MCNN $h = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1}$ satisfies that

$$h((\mathcal{X}, \ell_X)) = \psi\left(q_{\varphi_{k+1}}\left(\left(\ell_X^{(\varphi,k)}\right)_{\#} \mu_X\right)\right) = \psi\left(q_{\varphi_{k+1}}\left(\left(\ell_Y^{(\varphi,k)}\right)_{\#} \mu_Y\right)\right) = h((\mathcal{Y}, \ell_Y)).$$

To prove Equation (19), it suffices to prove that for any $x \in X$ and $y \in Y$ (see Lemma A.1),

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

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

$$\left\|\ell_X^{(\varphi,j)}(x) - \ell_Y^{(\varphi,j)}(y)\right\| \leq \Pi_{i=1}^j C_i \cdot d_{\mathrm{W}}\left(\mathfrak{l}_{(\mathcal{X},\ell_X)}^{(j)}(x), \mathfrak{l}_{(\mathcal{Y},\ell_Y)}^{(j)}(y)\right). \tag{20}$$

When $j = 0$, we have that $\ell_X^{(\varphi,0)} = \ell_X = \mathfrak{l}^{(0)}_{(\mathcal{X},\ell_X)}$ and $\ell_Y^{(\varphi,0)} = \ell_Y = \mathfrak{l}^{(0)}_{(\mathcal{Y},\ell_Y)}$. Therefore, Equation (20) obviously holds (we let $\Pi_{i=1}^0 C_i \coloneqq 1$). We now assume that Equation (20) holds for some $j \geq 0$. For $j+1$, we have that

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

This concludes the induction step and thus the proof of item 1.

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

Lemma B.5. *For any $d \in \mathbb{N}$ and any $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$, if $\alpha \neq \beta$, then there exists a Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that*

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

Proof of Lemma B.5. By Kantorovich duality (see for example Remark 6.5 in (Villani, 2009)),

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

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

Lemma B.6. *For any $d \in \mathbb{N}$ and any $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$, if $\alpha \neq \beta$ and they are both compactly supported, then there exists a single-hidden-layer MLP $\varphi : \mathbb{R}^d \to \mathbb{R}$ (with the activation function σ specified in Section 4) such that*

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

Proof of Lemma B.6. By Lemma B.5, there exists a Lipschitz $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $\int_{\mathbb{R}^d} \psi(x)\alpha(dx) \neq \int_{\mathbb{R}^d} \psi(x)\beta(dx)$. Let $\varepsilon \coloneqq \left|\int_{\mathbb{R}^d} \psi(x)\alpha(dx) - \int_{\mathbb{R}^d} \psi(x)\beta(dx)\right| > 0$. Let $K \coloneqq \operatorname{supp}(\alpha) \cup \operatorname{supp}(\beta)$. Since both probability measures are compactly supported, we have that K is compact. Then, by the classic universal approximation theorem (Pinkus, 1999, Theorem 3.1) and by the assumption that σ is Lipschitz and non-polynomial, there exists a single-hidden-layer MLP $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $\sup_{x\in K} |\varphi(x) - \psi(x)| < \frac{\varepsilon}{4}$. Therefore,

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

and similarly,

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

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

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

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

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

$$\forall x \in X,\ y \in Y,\ \ell_X^{(\varphi,i)}(x) = \ell_Y^{(\varphi,i)}(y) \text{ iff } \mathfrak{l}^{(i)}_{(\mathcal{X},\ell_X)}(x) = \mathfrak{l}^{(i)}_{(\mathcal{Y},\ell_Y)}(y). \tag{21}$$

Given Equation (21), it is obvious that $\left(\mathfrak{l}^{(k)}_{(\mathcal{X},\ell_X)}\right)_{\#}\mu_X \neq \left(\mathfrak{l}^{(k)}_{(\mathcal{Y},\ell_Y)}\right)_{\#}\mu_Y$ implies that $\left(\ell_X^{(\varphi,k)}\right)_{\#}\mu_X \neq \left(\ell_Y^{(\varphi,k)}\right)_{\#}\mu_Y$. Then, by Lemma B.6 and by the fact that X and Y are finite, there exists a MLP $\varphi_{k+1} : \mathbb{R}^{d_k} \to \mathbb{R}$ such that

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

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

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

To conclude the proof, we prove Equation (21) by induction on $i = 1, \ldots, k$. When $i = 1$, let

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

Since X and Y are finite, A_1 is a finite set. We enumerate elements in A_1 and write $A_1 = \{(x_1, y_1), \ldots, (x_{d_1}, y_{d_1})\}$. By Lemma B.6, for each $j = 1, \ldots, d_1$, there exists a MLP $\varphi_1^j : \mathbb{R}^d \to \mathbb{R}$ such that

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

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

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

Equivalent speaking,

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

Now, we assume that Equation (21) holds for some $i \geq 1$. For $i + 1$, we let

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

Since X and Y are finite, A_{i+1} is a finite set. We enumerate elements in A_{i+1} and write $A_{i+1} = \{(x_1, y_1), \ldots, (x_{d_{i+1}}, y_{d_{i+1}})\}$. By Lemma B.6, for each $j = 1, \ldots, d_{i+1}$, there exists a MLP $\varphi_{i+1}^j : \mathbb{R}^{d_i} \to \mathbb{R}$ such that

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

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

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

Equivalent speaking,

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

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

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

This implies that

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

Therefore,

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

and we thus conclude the proof.

B.3.2. Proof of Theorem 4.3

We first prove a relaxation of Theorem 4.3. Let

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

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

Lemma B.7. *For any $k \in \mathbb{N}$, let $\mathcal{K} \subseteq \mathcal{M}_k^L(\mathbb{R}^d)$ be any compact subspace. Then, $\overline{\mathcal{NN}_k^c(\mathbb{R}^d)} = C(\mathcal{K}, \mathbb{R})$.*

The proof of the lemma is based on the following Stone-Weierstrass theorem.

Lemma B.8 (Stone-Weierstrass)**.** *Let X be a compact space. Let $\mathcal{F} \subseteq C(X, \mathbb{R})$ be a subalgebra containing the constant function 1. If moreover $\mathcal{F}$ separates points, then $\mathcal{F}$ is dense in $C(X, \mathbb{R})$.*

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

$\mathcal{NN}_k^c(\mathbb{R}^d)$ **Separates Points.** This follows from item 2 in Proposition 4.1.

$\mathcal{NN}_k^c(\mathbb{R}^d)$ **Is a Subalgebra.** By Equation (19), Lemma B.4 and the continuity of ψ, we have that $\mathcal{NN}_k^c(\mathbb{R}^d) \subseteq C(\mathcal{K}, \mathbb{R})$. Next, we show that $\mathcal{NN}_k^c(\mathbb{R}^d)$ is, in fact, a subalgebra of $C(\mathcal{K}, \mathbb{R})$. Given any constant C and function $h = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots \circ F_{\varphi_1} \in \mathcal{NN}_k^c(\mathbb{R}^d)$, we have that

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

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

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

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

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

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

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

and

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

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

Claim 2. *$h_1 = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} = \psi \circ P \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1}$ and $h_2 = \tilde{\psi} \circ S_{\tilde{\varphi}_{k+1}} \circ F_{\tilde{\varphi}_k} \circ \cdots F_{\tilde{\varphi}_1} = \tilde{\psi} \circ \tilde{P} \circ S_{\Phi_{k+1}} \circ F_{\Phi_k} \circ \cdots F_{\Phi_1}$.*

Proof of Claim 2. Recall notation from Equation (18). Then, we first prove inductively on $i = 1, \ldots, k$ that for any $\mathcal{X} \in \mathcal{K}$

$$\ell_X^{(\Phi,i)}(x) = \left(\ell_X^{(\varphi,i)}(x), \ell_X^{(\tilde{\varphi},i)}(x)\right), \quad \forall x \in X. \tag{22}$$

When $i = 1$,

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

Now, we assume that Equation (22) holds for some $i \geq 1$. Then, for $i+1$, we have that

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

which concludes the proof of Equation (22).

Similarly,

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

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

Given these claims, we then have that

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

and

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

This concludes the proof of Lemma B.7. Now we finish proving Theorem 4.3 by showing that $\mathcal{MN}_k(\mathbb{R}^d)$ is dense in $\mathcal{MN}_k^c(\mathbb{R}^d)$ when restricted to the compact set $\mathcal{K}$. Choose any $h = \psi \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} \in \mathcal{MN}_k^c(\mathbb{R}^d)$. By Equation (19) and Lemma B.4 again, we have that the map $S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1} : \mathcal{M}_k^L(\mathbb{R}^d) \to \mathbb{R}^{d_{k+1}}$ is continuous. Hence, $\mathfrak{K} \coloneqq S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1}(\mathcal{K})$ is a compact subspace of $\mathbb{R}^{d_{k+1}}$. Now, for any $\varepsilon > 0$, by the universal approximation theorem for MLP (Pinkus, 1999, Theorem 3.1), there exists a single-hidden-layer MLP $\tilde{\psi} : \mathbb{R}^{d_{k+1}} \to \mathbb{R}$ such that $\sup_{x \in \mathfrak{K}} \left\|\tilde{\psi}(x) - \psi(x)\right\| < \varepsilon$. If we let $\tilde{h} \coloneqq \tilde{\psi} \circ S_{\varphi_{k+1}} \circ F_{\varphi_k} \circ \cdots F_{\varphi_1}$, then $\tilde{h} \in \mathcal{MN}_k(\mathbb{R}^d)$ and moreover,

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

This implies that $\mathcal{MN}_k(\mathbb{R}^d)$ is dense in $\mathcal{MN}_k^c(\mathbb{R}^d)$ when restricted to the compact set $\mathcal{K}$. This concludes the proof.

B.4. Proofs from Section 5

B.4.1. Proof of Proposition 5.1

The proof is rather lengthy, and we start with some preliminary definitions and lemmas.

Definition 14. *Suppose two finite metric spaces X and Y are given. We say a sequence of measurable maps $\{(\nu_n)_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)\}_{n\in\mathbb{N}}$ weakly converges to $\nu_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)$ if $\{(\nu_n)_{x,y}\}_{n\in\mathbb{N}} \subseteq \mathcal{P}(X \times Y)$ weakly converges to $\nu_{x,y} \in \mathcal{P}(X \times Y)$ for all $(x, y) \in X \times Y$.*

Lemma B.9. *Suppose two finite metric spaces X and Y are given. If a sequence of probability measures $\{\gamma_n\}_{n\in\mathbb{N}} \subseteq \mathcal{P}(X \times Y)$ weakly converges to $\gamma \in \mathcal{P}(X \times Y)$ and a sequence of measurable maps $\{(\nu_n)_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)\}_{n\in\mathbb{N}}$ weakly converges to $\nu_{\bullet,\bullet} : X \times Y \to \mathcal{P}(X \times Y)$, then the sequence $\{(\nu_n)_{\bullet,\bullet} \odot \gamma_n\}_{n\in\mathbb{N}}$ also weakly converges to $\nu_{\bullet,\bullet} \odot \gamma$.*

Proof. Fix an arbitrary continuous bounded map $\phi : X \times Y \longrightarrow \mathbb{R}$. Then,

$$\begin{aligned}
&\left|\int_{X\times Y} \phi(x,y)\,(\nu_n)_{\bullet,\bullet} \odot \gamma_n(dx \times dy) - \int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma(dx \times dy)\right| \\
&\leq \left|\int_{X\times Y} \phi(x,y)\,(\nu_n)_{\bullet,\bullet} \odot \gamma_n(dx \times dy) - \int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma_n(dx \times dy)\right| \\
&\quad + \left|\int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma_n(dx \times dy) - \int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma(dx \times dy)\right|.
\end{aligned}$$

By the weak convergence of $\{(\nu_n)_{\bullet,\bullet}\}_{n\in\mathbb{N}}$ and by applying the bounded convergence theorem, we have that

$$\int_{X\times Y} \phi(x,y)\,(\nu_n)_{\bullet,\bullet} \odot \gamma_n(dx \times dy) = \int_{X\times Y}\int_{X\times Y} \phi(x,y)\,(\nu_n)_{x',y'}(dx \times dy)\gamma_n(dx' \times dy')$$

converges to

$$\int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma_n(dx \times dy) = \int_{X\times Y}\int_{X\times Y} \phi(x,y)\,\nu_{x',y'}(dx \times dy)\gamma_n(dx' \times dy').$$

Also, since $\{\gamma_n\}_{n\in\mathbb{N}}$ weakly converges to γ, by finiteness of X and Y we have that

$$\int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma_n(dx \times dy) = \int_{X\times Y}\int_{X\times Y} \phi(x,y)\,\nu_{x',y'}(dx \times dy)\gamma_n(dx' \times dy')$$

converges to

$$\int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma(dx \times dy) = \int_{X\times Y}\int_{X\times Y} \phi(x,y)\,\nu_{x',y'}(dx \times dy)\gamma(dx' \times dy').$$

Hence, $\left|\int_{X\times Y} \phi(x,y)\,(\nu_n)_{\bullet,\bullet} \odot \gamma_n(dx \times dy) - \int_{X\times Y} \phi(x,y)\,\nu_{\bullet,\bullet} \odot \gamma(dx \times dy)\right|$ converges to zero as we required. This completes the proof. □

Lemma B.10. *Suppose two MCMSs $(\mathcal{X}, d_X)$, $(\mathcal{Y}, d_Y)$, $k \geq 1$, and a sequence of k-step couplings $\{(\nu_n^{(k)})_{\bullet,\bullet}\}_{n\in\mathbb{N}} \subseteq \mathcal{C}^{(k)}\left(m_\bullet^X, m_\bullet^Y\right)$ are given. Then, there is a k-step coupling $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\left(m_\bullet^X, m_\bullet^Y\right)$ to which the sequence $\left\{(\nu_n^{(k)})_{\bullet,\bullet}\right\}_{n\in\mathbb{N}}$ converges.*

Proof. The proof is by induction. $k = 1$ case is obvious since $\mathcal{C}(m_x^X, m_y^Y)$ is compact w.r.t. the weak topology (see Villani (2003, p.49)) for all $(x, y) \in X \times Y$.

Now, suppose the claim holds up to some $k \geq 1$. Consider $k + 1$ case. By the definition, each $(k + 1)$-step coupling $(\nu_n^{(k+1)})_{\bullet,\bullet} \in \mathcal{C}^{(k+1)}\left(m_\bullet^X, m_\bullet^Y\right)$ can be expressed in the following way:

$$(\nu_n^{(k+1)})_{x,y} = \int_{X\times Y} (\nu_n^{(k)})_{x',y'}(\mu_n^{(1)})_{x,y}(dx' \times dy')$$

for all $(x,y) \in X \times Y$ for some $(\nu_n^{(k)})_{\bullet,\bullet} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ and $(\mu_n^{(1)})_{\bullet,\bullet} \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$. Then, by the inductive assumption, there are $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ and $\mu_{\bullet,\bullet}^{(1)} \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$ such that the sequence $\{(\nu_n^{(k)})_{\bullet,\bullet}\}_{n\in\mathbb{N}}$ weakly converges to $\nu_{\bullet,\bullet}^{(k)}$, and the sequence $\{(\mu_n^{(1)})_{\bullet,\bullet}\}_{n\in\mathbb{N}}$ weakly converges to $\mu_{\bullet,\bullet}^{(1)}$. Then, for each $(x,y) \in X \times Y$,

$$(\nu_n^{(k+1)})_{x,y} = \int_{X\times Y} (\nu_n^{(k)})_{x',y'}(\mu_n^{(1)})_{x,y}(dx' \times dy') = (\nu_n^{(k)})_{\bullet,\bullet} \odot (\mu_n^{(1)})_{x,y}$$

weakly converges to

$$\nu_{x,y}^{(k+1)} = \int_{X\times Y} \nu_{x',y'}^{(k)}\mu_{x,y}^{(1)}(dx' \times dy') = \nu_{\bullet,\bullet}^{(k)} \odot \mu_{x,y}^{(1)}$$

by Lemma B.9. This completes the proof. □

Lemma B.11 (Mémoli (2011, Lemma 10.3))**.** *Let (Z, d_Z) be a compact metric space and $\phi : Z \times Z \to \mathbb{R}$ be a Lipschitz map w.r.t. the L^1 metric on $Z \times Z$:*

$$\hat{d}_{Z\times Z}((z_1,z_2),(z_1',z_2')) \coloneqq d_Z(z_1,z_1') + d_Z(z_2,z_2') \text{ for all } (z_1,z_2),(z_1',z_2') \in Z \times Z.$$

Also, for each $\gamma \in \mathcal{P}(Z)$, we define a map $p_{\phi,\gamma}$ in the following way:

$$\begin{aligned} p_{\phi,\gamma} : Z &\longrightarrow \mathbb{R} \\ z &\longmapsto \int_Z \phi(z,z')\,\gamma(dz'). \end{aligned}$$

If a sequence $\{\mu_n\}_{n\in\mathbb{N}} \subseteq \mathcal{P}(Z)$ weakly converges to μ, then p_{ϕ,μ_n} uniformly converges to $p_{\phi,\mu}$.

Corollary B.12. *For any two MCMSs $(\mathcal{X}, d_X)$, $(\mathcal{Y}, d_Y)$, and $k \geq 1$, there exist a coupling measure $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and a k-step coupling $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ such that*

$$d_{\mathrm{GW}}^{(k)}((\mathcal{X},d_X),(\mathcal{Y},d_Y)) = \mathrm{dis}^{(k)}\left(\gamma, \nu_{\bullet,\bullet}^{(k)}\right).$$

Proof. First of all, we define $\phi : (X \times Y) \times (X \times Y) \to \mathbb{R}$ by sending any $((x,y),(x',y')) \in (X \times Y) \times (X \times Y)$ to $|d_X(x,x') - d_Y(y,y')|$.

By Definition 8, there are a sequence of coupling measures $\{\gamma_n\}_{n\in\mathbb{N}} \subseteq \mathcal{C}(\mu_X, \mu_Y)$ and a sequence of k-step couplings $\{(\nu_n^{(k)})_{\bullet,\bullet}\}_{n\in\mathbb{N}} \subseteq \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ such that

$$\begin{aligned} \mathrm{dis}^{(k)}(\gamma, (\nu_n^{(k)})_{\bullet,\bullet}) &= \int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')|\,(\nu_n^{(k)})_{x'',y''}(dx' \times dy')\gamma_n(dx'' \times dy'')\gamma_n(dx \times dy) \\ &= \int_{X\times Y} p_{\phi,(\nu_n^{(k)})_{\bullet,\bullet}\odot\gamma_n}(x,y)\,\gamma_n(dx \times dy) \\ &\leq d_{\mathrm{GW}}^{(k)}((\mathcal{X},d_X),(\mathcal{Y},d_Y)) + \frac{1}{n} \end{aligned}$$

for each $n \geq 1$.

Now, since $\mathcal{C}(\mu_X, \mu_Y)$ is compact w.r.t. the weak topology (see p.49 of (Villani, 2003)), there is a coupling measure $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ such that $\gamma_n \to \gamma$ weakly. Also, by Lemma B.10, there is a k-step coupling $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ such that $(\nu_n^{(k)})_{\bullet,\bullet} \to \nu_{\bullet,\bullet}^{(k)}$ weakly.

Now, let

$$A_n \coloneqq \int_{X\times Y} p_{\phi,\nu^{(k)}_{\bullet,\bullet}\odot\gamma}(x,y)\,\gamma_n(dx\times dy) - \int_{X\times Y} p_{\phi,(\nu^{(k)}_n)_{\bullet,\bullet}\odot\gamma_n}(x,y)\,\gamma_n(dx\times dy),$$
$$B_n \coloneqq \int_{X\times Y} p_{\phi,\nu^{(k)}_{\bullet,\bullet}\odot\gamma}(x,y)\,\gamma(dx\times dy) - \int_{X\times Y} p_{\phi,\nu^{(k)}_{\bullet,\bullet}\odot\gamma}(x,y)\,\gamma_n(dx\times dy),$$
$$C_n \coloneqq \int_{X\times Y} p_{\phi,\nu^{(k)}_{\bullet,\bullet}\odot\gamma}(x,y)\,\gamma(dx\times dy) - \int_{X\times Y} p_{\phi,(\nu^{(k)}_n)_{\bullet,\bullet}\odot\gamma_n}(x,y)\,\gamma_n(dx\times dy).$$

It is easy to see that $C_n = A_n + B_n$ and thus $|C_n| \le |A_n| + |B_n|$. By Lemma B.9 and Lemma B.11, A_n converges to zero. Also, B_n converges to zero by the assumption that X, Y are finite and that γ_n weakly converges to γ. Hence, C_n converges to zero. Therefore,

$$\begin{aligned}\mathrm{dis}^{(k)}\big(\gamma,\nu^{(k)}_{\bullet,\bullet}\big) &= \int_{X\times Y} p_{\phi,\nu^{(k)}_{\bullet,\bullet}\odot\gamma}\gamma(dx\times dy)\\ &= \lim_{n\to\infty}\int_{X\times Y} p_{\phi,(\nu^{(k)}_n)_{\bullet,\bullet}\odot\gamma_n}\gamma_n(dx\times dy)\\ &\le d^{(k)}_{\mathrm{GW}}((\mathcal{X},d_X),(\mathcal{Y},d_Y))\end{aligned}$$

Since we always have that $\mathrm{dis}^{(k)}\big(\gamma,\nu^{(k)}_{\bullet,\bullet}\big) \ge d^{(k)}_{\mathrm{GW}}((\mathcal{X},d_X),(\mathcal{Y},d_Y))$, we conclude that

$$\mathrm{dis}^{(k)}\big(\gamma,\nu^{(k)}_{\bullet,\bullet}\big) = d^{(k)}_{\mathrm{GW}}((\mathcal{X},d_X),(\mathcal{Y},d_Y)).$$

□

Lemma B.13 (Gluing of k-step couplings)**.** *Suppose three MCMSs $(\mathcal{X},d_X),(\mathcal{Y},d_Y),(\mathcal{Z},d_Z)$, $k\ge 1$, and k-step couplings $\mu^{(k)}_{\bullet,\bullet}\in\mathcal{C}^{(k)}(m^X_\bullet,m^Z_\bullet)$, $\eta^{(k)}_{\bullet,\bullet}\in\mathcal{C}^{(k)}(m^Z_\bullet,m^Y_\bullet)$ are given. Then, there are probability measures $\pi^{(k)}_{\bullet,\bullet,\bullet}: X\times Y\times Z\to\mathcal{P}(X\times Y\times Z)$ and k-step coupling $\nu^{(k)}_{\bullet,\bullet}\in\mathcal{C}^{(k)}\big(m^X_\bullet,m^Y_\bullet\big)$ such that $\nu^{(k)}_{x,y}$, $\mu^{(k)}_{x,z}$, and $\eta^{(k)}_{z,y}$ are the marginals of $\pi^{(k)}_{x,y,z}$ for any $(x,y,z)\in X\times Y\times Z$.*

Proof. The proof is by induction on k. First, consider $k=1$ case. Fix an arbitrary $(x,y,z)\in X\times Y\times Z$. Let

$$\pi^{(1)}_{x,y,z}(x',y',z') \coloneqq \begin{cases}\frac{\mu^{(1)}_{x,z}(x',z')\mu^{(1)}_{z,y}(z',y')}{m^Z_z(z')}, & m^Z_z(z')>0\\ 0, & m^Z_z(z')=0\end{cases}$$

for each $(x',y',z')\in X\times Y\times Z$. Observe that

$$\begin{aligned}\sum_{(x',y',z')\in X\times Y\times Z}\pi^{(1)}_{x,y,z}(x',y',z') &= \sum_{z'\in Z,m^Z_z(z')>0}\sum_{y'\in Y}\frac{\mu^{(1)}_{z,y}(z',y')}{m^Z_z(z')}\sum_{x'\in X}\mu^{(1)}_{x,z}(x',z')\\ &= \sum_{z'\in Z,m^Z_z(z')>0}\sum_{y'\in Y}\frac{\mu^{(1)}_{z,y}(z',y')}{m^Z_z(z')}\cdot m^Z_z(z') = \sum_{z'\in Z,m^Z_z(z')>0}\sum_{y'\in Y}\mu^{(1)}_{z,y}(z',y') = 1.\end{aligned}$$

Hence, $\pi^{(1)}_{x,y,z}\in\mathcal{P}(X\times Y\times Z)$. Now, let $\nu^{(1)}_{x,y}(x',y')\coloneqq\sum_{z'\in Z}\pi^{(1)}_{x,y,z}(x',y',z')$ for each $(x',y')\in X\times Y$. Then, for fixed $x'\in X$,

$$\sum_{y'\in Y}\nu^{(1)}_{x,y}(x',y')=\sum_{y'\in Y}\sum_{z'\in Z,m^Z_z(z')>0}\frac{\mu^{(1)}_{x,z}(x',z')\mu^{(1)}_{z,y}(z',y')}{m^Z_z(z')}=\sum_{z'\in Z,m^Z_z(z')>0}\frac{\mu^{(1)}_{x,z}(x',z')}{m^Z_z(z')}\sum_{y'\in Y}\mu^{(1)}_{z,y}(z',y')$$
$$=\sum_{z'\in Z,m^Z_z(z')>0}\frac{\mu^{(1)}_{x,z}(x',z')}{m^Z_z(z')}\cdot m^Z_z(z')=\sum_{z'\in Z,m^Z_z(z')>0}\mu^{(1)}_{x,z}(x',z')=m^X_x(x').$$

Similarly, for each fixed $y'\in Y$, one can prove $\sum_{x'\in X}\nu^{(1)}_{x,y}(x',y')=m^Y_y(y')$. Hence, indeed $\nu^{(1)}_{\bullet,\bullet}\in\mathcal{C}^{(1)}\big(m^X_\bullet,m^Y_\bullet\big)$.

Now, suppose the claim holds up to some $k\geq 1$. We consider $k+1$ case. For a $(k+1)$-step coupling $\mu^{(k+1)}_{\bullet,\bullet}\in\mathcal{C}^{(k+1)}(m^X_\bullet,m^Z_\bullet)$, there are k-step coupling $\mu^{(k)}_{\bullet,\bullet}\in\mathcal{C}^{(k)}(m^X_\bullet,m^Z_\bullet)$ and 1-step coupling $\mu^{(1)}_{\bullet,\bullet}\in\mathcal{C}^{(1)}(m^X_\bullet,m^Z_\bullet)$ such that

$$\mu^{(k+1)}_{x,z}(x',z')=\sum_{(x'',z'')\in X\times Y}\mu^{(k)}_{x'',z''}(x',z')\,\mu^{(1)}_{x,z}(x'',z'')$$

for any $(x,z),(x',z')\in X\times Z$. Similarly, for a $(k+1)$-step coupling $\eta^{(k+1)}_{\bullet,\bullet}\in\mathcal{C}^{(k+1)}(m^Z_\bullet,m^Y_\bullet)$, there are k-step coupling $\eta^{(k)}_{\bullet,\bullet}\in\mathcal{C}^{(k)}(m^Z_\bullet,m^Y_\bullet)$ and 1-step coupling $\eta^{(1)}_{\bullet,\bullet}\in\mathcal{C}^{(1)}(m^Z_\bullet,m^Y_\bullet)$ such that

$$\eta^{(k+1)}_{z,y}(z',y')=\sum_{(z'',y'')\in Z\times Y}\eta^{(k)}_{z'',y''}(z',y')\,\eta^{(1)}_{z,y}(z'',y'')$$

for any $(z,y),(z',y')\in Z\times Y$.

Because of the inductive assumption, we have $\pi^{(k)}_{\bullet,\bullet,\bullet}:X\times Y\times Z\to\mathcal{P}(X\times Y\times Z)$ and $\pi^{(1)}_{\bullet,\bullet,\bullet}:X\times Y\times Z\to\mathcal{P}(X\times Y\times Z)$ satisfying the claim. Then, let

$$\pi^{(k+1)}_{x,y,z}(x',y',z')\coloneqq\sum_{(x'',y'',z'')\in X\times Y\times Z}\pi^{(k)}_{x'',y'',z''}(x',y',z')\,\pi^{(1)}_{x,y,z}(x'',y'',z''),$$

and let $\nu^{(k+1)}_{x,y}(x',y')\coloneqq\sum_{z'\in Z}\pi^{(k+1)}_{x,y,z}(x',y',z')$. Then, we have that

$$\begin{aligned}\nu^{(k+1)}_{x,y}(x',y')&=\sum_{z'\in Z}\sum_{(x'',y'',z'')\in X\times Y\times Z}\pi^{(k)}_{x'',y'',z''}(x',y',z')\,\pi^{(1)}_{x,y,z}(x'',y'',z'')\\&=\sum_{(x'',y'',z'')\in X\times Y\times Z}\pi^{(1)}_{x,y,z}(x'',y'',z'')\sum_{z'\in Z}\pi^{(k)}_{x'',y'',z''}(x',y',z')\\&=\sum_{(x'',y'',z'')\in X\times Y\times Z}\pi^{(1)}_{x,y,z}(x'',y'',z'')\,\nu^{(k)}_{x'',y''}(x',y')\\&=\sum_{(x'',y'')\in X\times Y}\nu^{(k)}_{x'',y''}(x',y')\sum_{z''\in Z}\pi^{(1)}_{x,y,z}(x'',y'',z'')\\&=\sum_{(x'',y'')\in X\times Y}\nu^{(k)}_{x'',y''}(x',y')\,\nu^{(k)}_{x,y}(x'',y'').\end{aligned}$$

Since the choice of $(x,y,z)\in x\times y\times z$ is arbitrary, now we have $\nu^{(k+1)}_{\bullet,\bullet}\in\mathcal{C}^{(k+1)}\big(m^X_\bullet,m^Y_\bullet\big)$ as we required. Hence, this concludes the proof. □

Now we start to prove Proposition 5.1.

First of all, $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ is obviously symmetric.

Next, we prove that $d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X},d_X),(\mathcal{Y},d_Y))=0$ happens if and only if $(\mathcal{X},d_X)$ and $(\mathcal{Y},d_Y)$ are isomorphic. To do this, we first provide a precise definition of MCMS isomorphism.

Definition 15. *Two MCMSs $(\mathcal{X},d_X)$ and $(\mathcal{Y},d_Y)$ are said to be* isomorphic *if there exists an isometry $\psi:X\to Y$ such that $\psi_\#\mu_X=\mu_Y$ and $\psi_\#m^X_x=m^Y_{\psi(x)}$ for all $x\in X$.*

When are $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ are isomorphic, without loss of generality, we simply assume that $(\mathcal{Y}, d_Y) = (\mathcal{X}, d_X)$.

Claim 3. *Let Δ_{μ_X} denote the diagonal coupling between μ_X and itself, i.e.,*

$$\Delta_{\mu_X} = \sum_{x \in X} \mu_X(x) \delta_{(x,x)}.$$

Then, for each $k \in \mathbb{N}$, there exists $\nu^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}(m^X_\bullet, m^X_\bullet)$ such that $\Delta_{\mu_X} = \nu^{(k)}_{\bullet,\bullet} \odot \Delta_{\mu_X}$.

Assume the claim for now. Then, we have that for each $k \in \mathbb{N}$

$$\begin{aligned} d^{(k)}_{\mathrm{GW}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) &\leq \mathrm{dis}\big(\Delta_{\mu_X}, \nu^{(k)}_{\bullet,\bullet}\big) \\ &= \int_{X\times X} \int_{X\times X} |d_X(x,x') - d_X(x_1,x_1')| \nu^{(k)}_{\bullet,\bullet} \odot \Delta_{\mu_X}(dx' \times dx_1') \Delta_{\mu_X}(dx \times dx_1) \\ &= \int_{X\times X} |d_X(x,x') - d_X(x,x')| \mu_X(dx) \mu_X(dx') = 0. \end{aligned}$$

Hence, $d^{\mathrm{MCMS}}_{\mathrm{GW}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = 0$,

Proof of Claim 3. We prove inductively on $k \in \mathbb{N}$ that there exists $\nu^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}(\mathcal{X}, \mathcal{X})$ so that $\nu^{(k)}_{x,x} = \Delta_{m^{X,\otimes k}_x}$ is the diagonal coupling between $m^{X,\otimes k}_x$ and itself for each $x \in X$.

For $k = 1$, we define $\nu^{(1)}_{\bullet,\bullet}$ as follows:

$$\nu^{(1)}_{x,x'} \coloneqq \begin{cases} m^X_x \otimes m^X_{x'} & x \neq x' \\ \Delta_{m^X_x} & x = x' \end{cases}.$$

Since X is finite, obviously we have that $\nu^{(1)}_{\bullet,\bullet} \in \mathcal{C}^{(1)}(\mathcal{X}, \mathcal{X})$.

Assume that the statement holds for some $k \geq 0$. Now, for $k+1$, by the induction assumption, there exists $\nu^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}(\mathcal{X}, \mathcal{X})$ so that $\nu^{(k)}_{x,x} = \Delta_{m^{X,\otimes k}_x}$. We define $\nu^{(k+1)}_{\bullet,\bullet} \in \mathcal{C}^{(k+1)}(\mathcal{X}, \mathcal{X})$ as follows

$$\nu^{(k+1)}_{x,x'} \coloneqq \int_{X\times X} \nu^{(k)}_{x_1,x_1'} \nu^{(1)}_{x,x'}(dx_1 \times dx_1'), \ \forall x, x' \in X.$$

Now, for any $x \in X$, we have that

$$\begin{aligned} \nu^{(k+1)}_{x,x} &= \int_{X\times X} \nu^{(k)}_{x_1,x_1'} \nu^{(1)}_{x,x}(dx_1 \times dx_1') \\ &= \sum_{x' \in X} m^X_x(x') \sum_{x'' \in X} m^{X,\otimes k}_{x'}(x'') \delta_{(x'',x'')} \\ &= \sum_{x'' \in X} \left(\sum_{x' \in X} m^{X,\otimes k}_{x'}(x'') m^X_x(x') \right) \delta_{(x'',x'')} \\ &= \sum_{x'' \in X} m^{X,\otimes (k+1)}_x(x'') \delta_{(x'',x'')} \\ &= \Delta_{m^{X,\otimes(k+1)}_x}. \end{aligned}$$

Now, we turn to prove the claim. For each $k \in \mathbb{N}$, let $\nu^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}(\mathcal{X}, \mathcal{X})$ be such that $\nu^{(k)}_{x,x} = \Delta_{m^{X,\otimes k}_x}$ is the diagonal

coupling between $m_x^{X,\otimes k}$ and itself for each $x \in X$. Then,

$$\begin{aligned}\int_{X\times X} \nu^{(k)}_{x,x'}\, \Delta_{\mu_X}(dx \times dx') &= \sum_{x\in X} \nu^{(k)}_{x,x}\, \mu_X(x) \\ &= \sum_{x,x'\in X} m_x^{X,\otimes k}(x')\delta_{(x',x')}\, \mu_X(x) \\ &= \sum_{x'\in X}\left(\sum_{x\in X} m_x^{X,\otimes k}(x')\mu_X(x)\right)\delta_{(x',x')} \\ &= \sum_{x'\in X} \mu_X(x')\delta_{(x',x')} \\ &= \Delta_{\mu_X}.\end{aligned}$$

□

Now, we assume that $d_{\mathrm{GW}}^{\mathrm{MCMS}}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = 0$ for some MCMSs $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$. Then, $d_{\mathrm{GW}}^{(1)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) = 0$. By Corollary B.12, there exist optimal $\nu_{\bullet,\bullet} \in \mathcal{C}^{(1)}\left(m_\bullet^X, m_\bullet^Y\right)$ and $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ such that

$$\int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')|\nu_{x'',y''}(dx' \times dy')\gamma(dx'' \times dy'')\gamma(dx \times dy) = 0.$$

We let $\gamma' \coloneqq \nu_{\bullet,\bullet} \odot \gamma$. Notice that $\gamma' \in \mathcal{C}(\mu_X, \mu_Y)$. Since X and Y are finite, we rewrite the integral above as finite sums:

$$\sum_{(x,y)\in X\times Y}\ \sum_{(x',y')\in X\times Y} |d_X(x,x') - d_Y(y,y')|\, \gamma'(x',y')\, \gamma(x,y) = 0.$$

By Claim 1, there exists an isometry $\phi : X \to Y$ such that

$$\{(x, \phi(x)) : x \in X\} = \mathrm{supp}(\gamma) = \mathrm{supp}(\gamma').$$

Since $\gamma, \gamma' \in \mathcal{C}(\mu_X, \mu_Y)$, this immediately implies that for any Borel subset $A \subseteq X$, one has

$$\mu_X(A) = \gamma(A \times Y) = \gamma(A \times \phi(A)) = \gamma(X \times \phi(A)) = \mu_Y(\phi(A)).$$

Hence, $\phi_{\#}\mu_X = \mu_Y$. Moreover, we have that

$$\gamma = \gamma' = \sum_{x\in X} \mu_X(x)\delta_{(x,\phi(x))}.$$

Then, by the definition of γ', we have that

$$\begin{aligned}\sum_{x\in X} \mu_X(x)\delta_{(x,\phi(x))} = \gamma' &= \sum_{x_1\in X, y_1\in Y} \gamma(x_1,y_1)\nu_{x_1,y_1} \\ &= \sum_{x\in X,y\in Y}\left(\sum_{x_1\in X,y_1\in Y} \gamma(x_1,y_1)\nu_{x_1,y_1}(x,y)\right)\delta_{(x,y)} \\ &= \sum_{x\in X,y\in Y}\left(\sum_{x_1\in X} \gamma(x_1,\phi(x_1))\nu_{x_1,\phi(x_1)}(x,y)\right)\delta_{(x,y)} \\ &= \sum_{x\in X,y\in Y}\left(\sum_{x_1\in X} \mu_X(x_1)\nu_{x_1,\phi(x_1)}(x,y)\right)\delta_{(x,y)}.\end{aligned}$$

By comparing coefficients for the Dirac delta measures above, one has that

$$\nu_{x_1,\phi(x_1)}(x,y) = \begin{cases} 0, & \text{if } y \neq \phi(x) \\ \nu_{x_1,\phi(x_1)}(x,y), & \text{if } y = \phi(x) \end{cases}$$

for any $x_1 \in X$. This means that $\{(x,\phi(x)) : x \in X\} \supseteq \mathrm{supp}[\nu_{x_1,\phi(x_1)}]$. Since $\nu_{x_1,\phi(x_1)} \in \mathcal{C}\left(m^X_{x_1}, m^Y_{\phi(x_1)}\right)$, for any Borel subset $A \subseteq X$, one has that

$$m^X_{x_1}(A) = \nu_{x_1,\phi(x_1)}(A \times Y) = \nu_{x_1,\phi(x_1)}(A \times \phi(A)) = \nu_{x_1,\phi(x_1)}(X \times \phi(A)) = m^Y_{\phi(x_1)}(\phi(A)).$$

By an argument similar to the one for proving $\phi_\# \mu_X = \mu_Y$, we have that

$$\phi_\# m^X_{x_1} = m^Y_{\phi(x_1)}, \ \forall x_1 \in X.$$

Therefore, $(\mathcal{X}, d_X)$ is isomorphic to $(\mathcal{Y}, d_Y)$.

Finally, we prove that $d_{\mathrm{GW}}^{\mathrm{MCMS}}$ satisfies the triangle inequality. It suffices to prove that for each $k \in \mathbb{N}$, $d_{\mathrm{GW}}^{(k)}$ satisfies the triangle inequality. Fix arbitrary three MCMSs $(\mathcal{X}, d_X)$, $(\mathcal{Y}, d_Y)$, and $(\mathcal{Z}, d_Z)$. Recall the notation $\Gamma_{X,Y}$ which defines a function sending $(x, y, x', y') \in X \times Y \times X \times Y$ to $|d_X(x,x') - d_Y(y,y')|$. Then, for any $x, x' \in X$, $y, y' \in Y$, and $z, z' \in Z$, we obviously have that

$$\Gamma_{X,Y}(x,y,x',y') \leq \Gamma_{X,Z}(x,z,x',z') + \Gamma_{Z,Y}(z,y,z',y').$$

Now, fix arbitrary $\mu^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}(m^X_\bullet, m^Z_\bullet)$, $\gamma_{X,Z} \in \mathcal{C}(\mu_X, \mu_Z)$, $\eta^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}(m^Z_\bullet, m^Y_\bullet)$, and $\gamma_{Z,Y} \in \mathcal{C}(\mu_Z, \mu_Y)$. Then, by the Gluing Lemma (see Villani (2003, Lemma 7.6)), there exists a probability measure $\alpha \in \mathcal{P}(X \times Y \times Z)$ with marginals $\gamma_{X,Z}$ and $\gamma_{Z,Y}$ on $X \times Z$ and $Z \times Y$, respectively. Let $\gamma_{X,Y}$ be the marginal of π on $X \times Y$ which belongs to $\mathcal{C}(\mu_X, \mu_Y)$. By Lemma B.13, there exists $\nu^{(k)}_{\bullet,\bullet} \in \mathcal{C}^{(k)}\left(m^X_\bullet, m^Y_\bullet\right)$ such that $\nu^{(k)}_{x,y}$, $\mu^{(k)}_{x,z}$, $\eta^{(k)}_{z,y}$ are the marginals of some probability measure $\pi^{(k)}_{x,y,z} \in \mathcal{P}(X \times Y \times Z)$ for any $x, y, z \in X \times Y \times Z$. Then, because of the triangle inequality for L^1-norm,

$$\begin{aligned}
&d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \\
&\leq \int_{X\times Y} \int_{X\times Y} \int_{X\times Y} \Gamma_{X,Y}(x,y,x',y') \nu^{(k)}_{x'',y''}(dx' \times dy') \gamma_{X,Y}(dx'' \times dy'') \gamma_{X,Y}(dx \times dy) \\
&= \int_{X\times Y\times Z} \int_{X\times Y\times Z} \int_{X\times Y\times Z} \Gamma_{X,Y}(x,y,x',y') \pi^{(k)}_{x'',y'',z''}(dx' \times dy' \times dz') \alpha(dx'' \times dy'' \times dz'') \alpha(dx \times dy \times dz) \\
&\leq \int_{X\times Y\times Z} \int_{X\times Y\times Z} \int_{X\times Y\times Z} \Gamma_{X,Z}(x,z,x',z') \pi^{(k)}_{x'',y'',z''}(dx' \times dy' \times dz') \alpha(dx'' \times dy'' \times dz'') \alpha(dx \times dy \times dz) \\
&+ \int_{X\times Y\times Z} \int_{X\times Y\times Z} \int_{X\times Y\times Z} \Gamma_{Z,Y}(z,y,z',y') \pi^{(k)}_{x'',y'',z''}(dx' \times dy' \times dz') \alpha(dx'' \times dy'' \times dz'') \alpha(dx \times dy \times dz) \\
&= \int_{X\times Z} \int_{X\times Z} \int_{X\times Z} \Gamma_{X,Z}(x,z,x',z') \mu^{(k)}_{x'',z''}(dx' \times dz') \gamma_{X,Z}(dx'' \times dz'') \gamma_{X,Z}(dx \times dz) \\
&+ \int_{Z\times Y} \int_{Z\times Y} \int_{Z\times Y} \Gamma_{Z,Y}(z,y,z',y') \eta^{(k)}_{z'',y''}(dx' \times dy' \times dz') \gamma_{Z,Y}(dz'' \times dy'') \gamma_{Z,Y}(dz \times dy).
\end{aligned}$$

Since the choice of $\mu^{(k)}_{\bullet,\bullet}, \gamma_{X,Z}, \eta^{(k)}_{\bullet,\bullet}, \gamma_{Z,Y}$ are arbitrary, by taking the infimum one concludes that

$$d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \leq d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Z}, d_Z)) + d_{\mathrm{GW}}^{(k)}((\mathcal{Z}, d_Z), (\mathcal{Y}, d_Y))$$

as we required. Then, we have that $d_{\mathrm{GW}}^{\mathrm{MCMS}} \coloneqq \sup_{k \geq 0} d_{\mathrm{GW}}^{(k)}$ satisfies the triangle inequality.

B.4.2. Proof of the claim in Example 3

The proof is based on the following lemma.

Lemma B.14. *Given two MMSs* $\mathbf{X}$ *and* $\mathbf{Y}$*, for their corresponding MCMSs* $\mathcal{M}(\mathbf{X})$ *and* $\mathcal{M}(\mathbf{Y})$ *we have that*

1. $\mathcal{C}^{(k+1)}\big(m_\bullet^X, m_\bullet^Y\big) \subseteq \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ *for all* $k \geq 1$.

2. *For any coupling measure* $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$*, the constant map* $\nu_{\bullet,\bullet} \equiv \gamma$ *belongs to* $\mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ *for all* $k \geq 1$.

Proof. We first prove item 1. Since $m_x^X = \mu_X$ and $m_y^Y = \mu_Y$ for all $x \in X$ and $y \in Y$, observe that $m_x^{X,\otimes k} = \mu_X$ and $m_y^{Y,\otimes k} = \mu_Y$ for all $x \in X$, $y \in Y$, and $k \geq 1$. Hence, $\mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big) \subseteq \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$ for all $k \geq 1$ by Lemma A.4. Now, for any $k \geq 2$, fix an arbitrary $\nu_{\bullet,\bullet}^{(k+1)} \in \mathcal{C}^{(k+1)}\big(m_\bullet^X, m_\bullet^Y\big)$. Then, by the definition, there are $\nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ and $\nu_{\bullet,\bullet}^{(1)} \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big) = \mathcal{C}(\mu_X, \mu_Y)$ such that

$$\nu_{x,y}^{(k+1)} = \int_{X\times Y} \nu_{x',y'}^{(k)}\, \nu_{x,y}^{(1)}(dx' \times dy')$$

for all $(x, y) \in X \times Y$. Again, by the definition, there are $\nu_{\bullet,\bullet}^{(k-1)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ and $\mu_{\bullet,\bullet}^{(1)} \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big) = \mathcal{C}(\mu_X, \mu_Y)$ such that

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

for all $(x, y) \in X \times Y$. Therefore,

$$\begin{aligned}\nu_{x,y}^{(k+1)} &= \int_{X\times Y}\int_{X\times Y} \nu_{x'',y''}^{(k-1)}\, \mu_{x',y'}^{(1)}(dx'' \times dy'')\, \nu_{x,y}^{(1)}(dx' \times dy') \\ &= \int_{X\times Y} \nu_{x'',y''}^{(k-1)}\, \pi_{x,y}^{(1)}(dx'' \times dy'')\end{aligned}$$

for all $(x, y) \in X \times Y$ where $\pi_{\bullet,\bullet}^{(1)} \coloneqq \int_{X\times Y} \mu_{x',y'}^{(1)}\, \nu_{\bullet,\bullet}^{(1)}(dx' \times dy') \in \mathcal{C}^{(2)}\big(m_\bullet^X, m_\bullet^Y\big) \subseteq \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$. Hence, $\nu_{\bullet,\bullet}^{(k+1)} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ by the definition. The first item is proved.

Next, we prove the second item. The proof is by induction on k. Fix a coupling measure $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and the constant map $\nu_{\bullet,\bullet} \equiv \gamma$. Obviously, $\nu_{\bullet,\bullet} \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$. Then, we also have that the constant map $\mu_X \otimes \mu_Y \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$. Now, suppose the claim holds up to some $k \geq 1$. Consider $k+1$ case. Observe that

$$\gamma = \int_{X\times Y} \gamma\, \mu_X \otimes \mu_Y(dx' \times dy')$$

where $\gamma \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$ by the inductive assumption and the constant map $\mu_X \otimes \mu_Y \in \mathcal{C}^{(1)}\big(m_\bullet^X, m_\bullet^Y\big)$. Hence, $\gamma \in \mathcal{C}^{(k+1)}\big(m_\bullet^X, m_\bullet^Y\big)$ by the definition. This completes the proof. □

Now, fix arbitrary couplings $\gamma, \gamma' \in \mathcal{C}(\mu_X, \mu_Y)$ and consider the constant map $\nu_{\bullet,\bullet} \equiv \gamma'$. Then, by the second item of Lemma B.14, $\nu_{\bullet,\bullet} \in \mathcal{C}^{(k)}\big(m_\bullet^X, m_\bullet^Y\big)$. Hence,

$$
\begin{aligned}
d_{\mathrm{GW}}^{(k)}(\mathcal{M}(\mathbf{X}),\mathcal{M}(\mathbf{Y})) &\leq \int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x')-d_Y(y,y')|\,\nu_{x'',y''}(dx'\times dy')\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&= \int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x')-d_Y(y,y')|\,\gamma'(dx'\times dy')\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&= \int_{X\times Y}\int_{X\times Y} |d_X(x,x')-d_Y(y,y')|\,\gamma'(dx'\times dy')\gamma(dx\times dy).
\end{aligned}
$$

Since the choice of γ,γ' are arbitrary, one concludes that $d_{\mathrm{GW}}^{(k)}((\mathcal{M}(\mathbf{X}),\mathcal{M}(\mathbf{Y}))\leq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y})$.

For the reverse direction, choose arbitrary $\gamma\in\mathcal{C}(\mu_X,\mu_Y)$ and a k-step coupling $\nu_{\bullet,\bullet}^{(k)}\in\mathcal{C}^{(k)}\big(m_\bullet^X,m_\bullet^Y\big)$. Let

$$
\gamma' \coloneqq \int_{X\times Y}\nu_{x'',y''}^{(k)}\,\gamma(dx''\times dy'').
$$

Then,

$$
\begin{aligned}
&\int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x')-d_Y(y,y')|\,\nu_{x'',y''}^{(k)}(dx'\times dy')\gamma(dx''\times dy'')\gamma(dx\times dy)\\
=&\int_{X\times Y}\int_{X\times Y} |d_X(x,x')-d_Y(y,y')|\,\gamma'(dx'\times dy')\gamma(dx\times dy)\\
\geq& d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y}).
\end{aligned}
$$

Since the choice of γ and $\nu_{\bullet,\bullet}^{(k)}$ are arbitrary, one concludes that $d_{\mathrm{GW}}^{(k)}(\mathcal{M}(\mathbf{X}),\mathcal{M}(\mathbf{Y}))\geq d_{\mathrm{GW}}^{\mathrm{bi}}(\mathbf{X},\mathbf{Y})$.

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

B.4.3. Proof of Proposition 5.2

For any $x\in X$, we let $\ell_X^x\coloneqq d_X(x,\bullet)$. For $i=1,\dots,k$, let $\mathfrak{l}_x^{(i)}:X\to\mathcal{P}^{\circ i}(\mathbb{R})$ be the shorthand for the ith WL measure hierarchy $\mathfrak{l}_{(\mathcal{X},\ell_X^x)}^{(k)}$ generated from the label $d_X(x,\bullet)$. We similarly define ℓ_Y^y and $\mathfrak{l}_y^{(i)}:Y\to\mathcal{P}^{\circ i}(\mathbb{R})$ for any $y\in Y$ and each $i=1,\dots,k$.

For any $\nu_{\bullet,\bullet}^{(k)}\in\mathcal{C}^{(k)}\big(m_\bullet^X,m_\bullet^Y\big)$, there exist $(\nu_i)_{\bullet,\bullet}\in\mathcal{C}^{(1)}\big(m_\bullet^X,m_\bullet^Y\big)$ for $i=1,\dots,k$ such that

$$
\nu_{x,y}^{(k)} = \int_{X\times Y}\cdots\int_{X\times Y}(\nu_k)_{x_{k-1},y_{k-1}}\,(\nu_{k-1})_{x_{k-2},y_{k-2}}(dx_{k-1}\times dy_{k-1})\cdots(\nu_1)_{x,y}(dx_1\times dy_1)
$$

for any $x \in X$ and $y \in Y$. Hence, for any $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$, we have that

$$
\begin{aligned}
&\int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')| \nu^{(k)}_{x'',y''}(dx'\times dy')\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&= \int_{X\times Y}\cdots\int_{X\times Y} |d_X(x,x') - d_Y(y,y')| (\nu_k)_{x_{k-1},y_{k-1}}(dx'\times dy')\cdots(\nu_1)_{x'',y''}(dx_1\times dy_1)\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&\geq \int_{X\times Y}\cdots\int_{X\times Y} d_{\mathrm{W}}\Big((\ell^x_X)_{\#}m^X_{x_{k-1}}, (\ell^y_Y)_{\#}m^Y_{y_{k-1}}\Big)\\
&\quad (\nu_{k-1})_{x_{k-2},y_{k-2}}(dx_{k-1}\times dy_{k-1})\cdots(\nu_1)_{x'',y''}(dx_1\times dy_1)\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&= \int_{X\times Y}\cdots\int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(1)}_x(x_{k-1}), \mathfrak{l}^{(1)}_y(y_{k-1})\big)\\
&\quad (\nu_{k-1})_{x_{k-2},y_{k-2}}(dx_{k-1}\times dy_{k-1})\cdots(\nu_1)_{x'',y''}(dx_1\times dy_1)\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&\geq \cdots\\
&\geq \int_{X\times Y}\int_{X\times Y} d_{\mathrm{W}}\big(\mathfrak{l}^{(k)}_x(x''), \mathfrak{l}^{(k)}_y(y'')\big)\,\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&\geq \int_{X\times Y} d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell^x_X),(\mathcal{Y},\ell^y_Y))\,\gamma(dx\times dy)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&d^{(k)}_{\mathrm{GW}}((\mathcal{X},d_X),(\mathcal{Y},d_Y))\\
&= \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y),\nu^{(k)}_{\bullet,\bullet}\in\mathcal{C}^{(k)}(m^X_\bullet,m^Y_\bullet)} \int_{X\times Y}\int_{X\times Y}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')|\nu^{(k)}_{x'',y''}(dx'\times dy')\gamma(dx''\times dy'')\gamma(dx\times dy)\\
&\geq \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell^x_X),(\mathcal{Y},\ell^y_Y))\,\gamma(dx\times dy).
\end{aligned}
$$

B.4.4. Proof of the statement in Remark 5.3

We first recall the third lower bound (TLB) from (Mémoli, 2011):

$$
\mathrm{TLB}(\mathbf{X},\mathbf{Y}) \coloneqq \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y}\left(\inf_{\gamma'\in\mathcal{C}(\mu_X,\mu_Y)}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')|\gamma'(dx'\times dy')\right)\gamma(dx\times dy).
$$

where we omit the $\frac{1}{2}$ factor from (Mémoli, 2011) for simplicity of presentation.

We adopt notation from the previous section. Notice that

$$
\inf_{\gamma'\in\mathcal{C}(\mu_X,\mu_Y)}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')|\gamma'(dx'\times dy') = d^{(1)}_{\mathrm{WL}}((\mathcal{X},\ell^x_X),(\mathcal{Y},\ell^y_Y)).
$$

We hence have that

$$
\mathrm{TLB}(\mathbf{X},\mathbf{Y}) = \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d^{(1)}_{\mathrm{WL}}((\mathcal{X},\ell^x_X),(\mathcal{Y},\ell^y_Y))\gamma(dx\times dy).
$$

For any $k \in \mathbb{N}$, we show that

$$
\mathrm{TLB}(\mathbf{X},\mathbf{Y}) = \inf_{\gamma\in\mathcal{C}(\mu_X,\mu_Y)} \int_{X\times Y} d^{(k)}_{\mathrm{WL}}((\mathcal{X},\ell^x_X),(\mathcal{Y},\ell^y_Y))\gamma(dx\times dy)
$$

by the lemma below.

Lemma B.15. *For any $k \in \mathbb{N}$ we have that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) = d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y))$.*

Proof. By item 2 in Lemma B.14 and Theorem A.7, we have that

$$\begin{aligned}
d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) &= \inf_{\gamma' \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)} \int_{X \times Y} |d_X(x, x') - d_Y(y, y')| \gamma'(dx' \times dy') \\
&= \inf_{\nu_{\bullet,\bullet} \in \mathcal{C}^{(k)}(m_\bullet^X, m_\bullet^Y), \mu \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} \int_{X \times Y} |d_X(x, x') - d_Y(y, y')| \nu_{x_1, y_1}(dx' \times dy') \mu(dx_1 \times dy_1) \\
&\leq \inf_{\mu \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} \int_{X \times Y} |d_X(x, x') - d_Y(y, y')| \mu(dx' \times dy') \mu(dx_1 \times dy_1) \\
&= \inf_{\mu \in \mathcal{C}(\mu_X, \mu_Y)} \int_{X \times Y} |d_X(x, x') - d_Y(y, y')| \mu(dx' \times dy') \\
&= d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)).
\end{aligned}$$

By Proposition 3.1, we conclude that $d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y)) = d_{\mathrm{WL}}^{(1)}((\mathcal{X}, \ell_X^x), (\mathcal{Y}, \ell_Y^y))$. □

B.4.5. Proof of the statement in Example 4

Given $\mathcal{X}$ and $\mathcal{Y}$, we have for any $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ and $\nu_{\bullet,\bullet}^{(k)}$ that

$$\begin{aligned}
&\int_{X \times Y} \int_{X \times Y} |\mathrm{ecc}_X(x') - \mathrm{ecc}_Y(y'))| \, \nu_{x,y}^{(k)}(dx' \times dy') \gamma(dx \times dy) \\
&= \int_{X \times Y} \int_{X \times Y} \left| \int_X d_X(x', x'') \mu_X(dx'') - \int_Y d_Y(y', y'') \mu_Y(y'') \right| \nu_{x,y}^{(k)}(dx' \times dy') \gamma(dx \times dy) \\
&= \int_{X \times Y} \int_{X \times Y} \left| \int_{X \times Y} d_X(x', x'') \gamma(dx'' \times dy'') - \int_{X \times Y} d_Y(y', y'') \gamma(dx'' \times dy'') \right| \nu_{x,y}^{(k)}(dx' \times dy') \gamma(dx \times dy) \\
&\leq \int_{X \times Y} \int_{X \times Y} \int_{X \times Y} |d_X(x', x'') - d_Y(y', y'')| \, \nu_{x,y}^{(k)}(dx' \times dy') \gamma(dx \times dy) \gamma(dx'' \times dy'').
\end{aligned}$$

Hence, we conclude that $\mathrm{ecc}_\bullet$ is stable.

B.4.6. Proof of Proposition 5.4

By Theorem A.7 we have that

$$\begin{aligned}
d_{\mathrm{WL}}^{(k)}((\mathcal{X}, \ell_X), (\mathcal{Y}, \ell_Y)) &= \inf_{\gamma^{(k)} \in \mathcal{C}^{(k)}(\mu_X, \mu_Y)} \int_{X \times Y} d_Z(\ell_X(x), \ell_Y(y)) \gamma^{(k)}(dx \times dy) \\
&= \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y), \nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}(m_\bullet^X, m_\bullet^Y)} \int_{X \times Y} \int_{X \times Y} d_Z(\ell_X(x'), \ell_Y(y')) \, \nu_{x,y}^{(k)}(dx' \times dy') \gamma(dx \times dy) \\
&\leq \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y), \nu_{\bullet,\bullet}^{(k)} \in \mathcal{C}^{(k)}(m_\bullet^X, m_\bullet^Y)} \mathrm{dis}^{(k)}\left(\gamma, \nu_{\bullet,\bullet}^{(k)}\right) \\
&= d_{\mathrm{GW}}^{(k)}((\mathcal{X}, d_X), (\mathcal{Y}, d_Y)) \, .
\end{aligned}$$

The inequality follows from the fact that $\ell_\bullet$ is stable. Hence we conclude the proof.

Table 4. SVM classification accuracy. Let $f_1(G,v) = \deg_G(v)$, $f_2(G,v) = \frac{1}{|V_G|} + \deg_G(v)$. For $d_{\mathrm{WL}}^{(k)}$ and d_{WLLB}, we show SVM classification accuracies using both the indefinite kernel as well as with a KSVM (Loosli et al., 2015).

METHOD	MUTAG	PROTEINS	PTC-FM	PTC-MR	IMDB-B	IMDB-M	COX2
$d_{\mathrm{WL}}^{(1)}$, f_1	87.7 ± 6.2	71.7 ± 3.3	57.6 ± 6.4	55.5 ± 4.5	74.5 ± 4.1	51.3 ± 3.0	78.1 ± 0.8
$d_{\mathrm{WL}}^{(2)}$, f_1	89.9 ± 6.4	71.3 ± 3.3	59.3 ± 3.7	54.9 ± 6.3	75.0 ± 3.0	51.4 ± 3.4	78.1 ± 0.8
$d_{\mathrm{WL}}^{(3)}$, f_1	87.6 ± 8.8	72.6 ± 3.1	59.0 ± 3.7	57.8 ± 7.9	74.9 ± 5.1	51.6 ± 4.0	78.1 ± 0.8
$d_{\mathrm{WL}}^{(4)}$, f_1	87.7 ± 4.1	72.4 ± 4.1	62.1 ± 3.9	56.7 ± 3.7	75.9 ± 2.7	51.4 ± 3.2	78.1 ± 0.8
$d_{\mathrm{WL}}^{(1)}$, f_1, KSVM	88.3 ± 4.6	72.1 ± 2.6	61.0 ± 3.9	52.9 ± 7.6	75.6 ± 5.1	51.4 ± 4.8	77.5 ± 2.8
$d_{\mathrm{WL}}^{(2)}$, f_1, KSVM	87.8 ± 5.7	72.8 ± 2.4	58.1 ± 5.0	56.9 ± 5.8	74.5 ± 3.4	51.4 ± 4.1	77.9 ± 1.0
$d_{\mathrm{WL}}^{(3)}$, f_1, KSVM	88.8 ± 6.8	72.6 ± 3.7	61.3 ± 4.6	55.2 ± 6.1	75.0 ± 3.8	51.8 ± 3.8	78.7 ± 1.7
$d_{\mathrm{WL}}^{(4)}$, f_1, KSVM	87.2 ± 5.5	73.5 ± 2.3	58.1 ± 6.3	53.4 ± 6.3	73.3 ± 3.7	51.2 ± 3.2	77.7 ± 1.3
$d_{\mathrm{WL}}^{(1)}$, f_2	87.3 ± 8.2	71.1 ± 3.2	58.7 ± 7.6	55.2 ± 5.5	74.1 ± 4.1	50.5 ± 4.5	78.1 ± 0.8
$d_{\mathrm{WL}}^{(2)}$, f_2	86.2 ± 7.4	73.5 ± 2.8	60.2 ± 5.3	54.0 ± 6.4	75.0 ± 4.5	51.4 ± 3.9	78.1 ± 0.8
$d_{\mathrm{WL}}^{(3)}$, f_2	88.8 ± 5.4	$\mathbf{74.5 \pm 2.9}$	61.6 ± 5.3	59.2 ± 6.4	75.4 ± 5.4	50.8 ± 4.0	77.0 ± 1.5
$d_{\mathrm{WL}}^{(4)}$, f_2	87.2 ± 5.8	73.9 ± 3.5	60.4 ± 5.1	54.3 ± 7.7	75.7 ± 3.7	50.8 ± 3.2	78.1 ± 0.8
$d_{\mathrm{WL}}^{(1)}$, f_2, KSVM	88.8 ± 7.9	74.3 ± 3.3	63.3 ± 2.3	60.1 ± 6.3	76.2 ± 4.1	51.5 ± 3.3	78.1 ± 0.8
$d_{\mathrm{WL}}^{(2)}$, f_2, KSVM	87.7 ± 9.4	$\mathbf{74.5 \pm 3.8}$	63.2 ± 5.9	54.0 ± 6.6	75.8 ± 4.5	51.0 ± 3.6	78.1 ± 1.7
$d_{\mathrm{WL}}^{(3)}$, f_2, KSVM	86.6 ± 5.9	74.2 ± 2.3	62.7 ± 7.6	58.1 ± 6.1	73.9 ± 3.9	51.7 ± 3.1	$\mathbf{79.2 \pm 1.8}$
$d_{\mathrm{WL}}^{(4)}$, f_2, KSVM	88.2 ± 5.3	74.0 ± 2.8	$\mathbf{63.6 \pm 7.0}$	56.4 ± 6.9	76.1 ± 3.8	51.9 ± 3.5	78.3 ± 2.5
$d_{\mathrm{WLLB}}^{(1)}$, f_1	87.9 ± 5.9	68.0 ± 1.3	59.6 ± 6.4	57.4 ± 8.1	74.7 ± 2.5	$\mathbf{52.0 \pm 1.8}$	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(2)}$, f_1	89.4 ± 5.2	68.9 ± 1.9	58.6 ± 5.7	59.0 ± 8.3	75.1 ± 2.2	50.8 ± 1.6	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(3)}$, f_1	$\mathbf{90.0 \pm 5.6}$	68.6 ± 1.6	57.3 ± 6.2	58.7 ± 8.1	75.2 ± 2.1	51.0 ± 1.6	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(4)}$, f_1	89.4 ± 5.2	66.7 ± 2.0	58.2 ± 6.1	56.6 ± 7.2	74.5 ± 2.0	50.3 ± 1.4	77.5 ± 2.1
$d_{\mathrm{WLLB}}^{(1)}$, f_1, KSVM	88.2 ± 5.7	72.0 ± 1.9	60.1 ± 4.8	54.3 ± 4.8	75.7 ± 3.2	50.2 ± 3.3	78.1 ± 1.9
$d_{\mathrm{WLLB}}^{(2)}$, f_1, KSVM	88.3 ± 10.7	70.6 ± 4.3	59.3 ± 4.5	55.8 ± 8.5	75.2 ± 3.6	50.6 ± 3.6	77.1 ± 2.2
$d_{\mathrm{WLLB}}^{(3)}$, f_1, KSVM	88.8 ± 5.4	71.7 ± 2.7	58.1 ± 6.6	54.6 ± 4.5	75.4 ± 3.8	50.8 ± 4.1	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(4)}$, f_1, KSVM	87.2 ± 5.8	71.6 ± 4.3	60.5 ± 5.8	52.6 ± 7.6	$\mathbf{76.5 \pm 2.9}$	51.5 ± 3.1	78.3 ± 1.1
$d_{\mathrm{WLLB}}^{(1)}$, f_2	88.9 ± 4.5	70.5 ± 1.0	60.5 ± 5.4	56.6 ± 8.0	75.0 ± 2.5	$\mathbf{52.0 \pm 1.8}$	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(2)}$, f_2	$\mathbf{90.0 \pm 4.2}$	70.0 ± 1.4	58.0 ± 5.6	58.5 ± 8.9	75.1 ± 2.2	50.8 ± 1.6	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(3)}$, f_2	$\mathbf{90.0 \pm 4.2}$	70.3 ± 2.4	56.4 ± 6.2	58.8 ± 7.8	75.2 ± 2.1	51.0 ± 1.6	77.9 ± 1.3
$d_{\mathrm{WLLB}}^{(4)}$, f_2	$\mathbf{90.0 \pm 4.2}$	70.2 ± 1.8	58.3 ± 5.6	58.2 ± 6.9	74.5 ± 2.0	50.3 ± 1.4	78.2 ± 0.8
$d_{\mathrm{WLLB}}^{(1)}$, f_2, KSVM	86.7 ± 7.8	32.0 ± 4.2	61.8 ± 2.7	52.0 ± 5.4	27.1 ± 4.5	20.0 ± 2.8	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(2)}$, f_2, KSVM	82.9 ± 8.7	30.9 ± 1.1	63.3 ± 5.6	54.3 ± 7.2	20.1 ± 4.4	19.8 ± 2.7	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(3)}$, f_2, KSVM	86.1 ± 4.9	30.6 ± 2.1	61.6 ± 6.5	55.5 ± 6.3	25.2 ± 1.1	18.0 ± 1.8	78.1 ± 0.8
$d_{\mathrm{WLLB}}^{(4)}$, f_2, KSVM	86.1 ± 7.2	30.7 ± 2.0	58.7 ± 6.0	52.1 ± 7.1	33.7 ± 8.7	17.0 ± 3.2	78.1 ± 0.8
WWL	85.3 ± 7.3	72.9 ± 3.6	62.2 ± 6.1	$\mathbf{63.0 \pm 7.4}$	72.5 ± 3.7	50.0 ± 5.3	78.2 ± 0.8
WL	85.5 ± 1.6	71.6 ± 0.6	56.6 ± 2.1	56.2 ± 2.0	72.4 ± 0.7	50.9 ± 0.4	78.4 ± 1.1
WL-OA	86.3 ± 2.1	72.6 ± 0.7	58.4 ± 2.0	54.2 ± 1.6	73.0 ± 1.1	50.2 ± 1.1	78.8 ± 1.3