Distances for Markov Chains, and Their Differentiation

Tristan Brugère TBRUGERE@UCSD.EDU
Zhengchao Wan ZCWAN@UCSD.EDU
Yusu Wang YUSUWANG@UCSD.EDU
Halıcıoğlu Data Science Institute
University of California, San Diego
9500 Gilman Dr. La Jolla, CA 92093

Editors: Claire Vernade and Daniel Hsu

Abstract

(Directed) graphs with node attributes are a common type of data in various applications and there is a vast literature on developing metrics and efficient algorithms for comparing them. Recently, in the graph learning and optimization communities, a range of new approaches have been developed for comparing graphs with node attributes, leveraging ideas such as the Optimal Transport (OT) and the Weisfeiler-Lehman (WL) graph isomorphism test. Two state-of-the-art representatives are the OTC distance proposed in (O'Connor et al., 2022) and the WL distance in (Chen et al., 2022). Interestingly, while these two distances are developed based on different ideas, we observe that they both view graphs as Markov chains, and are deeply connected. Indeed, in this paper, we propose a unified framework to generate distances for Markov chains (thus including (directed) graphs with node attributes), which we call the *Optimal Transport Markov (OTM)* distances, that encompass both the OTC and the WL distances. We further introduce a special one-parameter family of distances within our OTM framework, called the *discounted WL distance*. We show that the discounted WL distance has nice theoretical properties and can address several limitations of the existing OTC and WL distances. Furthermore, contrary to the OTC and the WL distances, our new discounted WL distance can be differentiated **after** a entropy-regularization similar to the Sinkhorn distance, making it suitable to use in learning frameworks, e.g., as the reconstruction loss in a graph generative model.

Keywords: graphs, Markov chains, Weisfeiler-Lehman, optimal transport

1. Introduction

Graph data is ubiquitous across various application domains, e.g., molecules viewed as node-attribute graphs, citation networks as directed graphs. Developing metrics and efficient algorithms to compare them have been traditionally studied in fields such as graph theory and theoretical computer science. In the last two decades, this problem also received tremendous attention in the graph learning and optimization community, especially for comparing (directed) graphs with node attributes (which we will call **labeled graphs**). In particular, two ideas have become prominent in the modern treatment of labeled graphs. The first idea is to leverage the so-called Weisfeiler-Lehman (WL) graph isomorphism test (Lehman and Weisfeiler, 1968), which is a classic graph isomorphism test that, in linear time, can distinguish a large family of graphs (Babai and Kucera, 1979; Babai and Luks, 1983). It has recently gained renewed interest both in designing WL-inspired graph kernels (Shervashidze et al., 2011; Togninalli et al., 2019) and as a tool for analyzing Message Passing Graph Neural Networks (MP-GNNs) (Xu et al., 2018; Azizian and Lelarge, 2021). The second idea in modern treatment of graphs is to treat labeled graphs (or related structured data) as suitable discrete measure spaces and

then use the idea of Optimal Transport (OT) to compare them. Examples include the Wasserstein WL (WWL) kernel (Togninalli et al., 2019), the Fused Gromov-Wasserstein (FGW) distance (Vayer et al., 2019), and the WL test based tree-mover distance (Chuang and Jegelka, 2022).

Very recently, several studies took a less combinatorial approach and viewed graphs as Markov chains: Chen et al. (2022) introduced the Weisfeiler-Lehman (WL) distance, which generalizes the graph comparing problem to the Markov chain comparison problem through a WL-like process in a natural way. This distance has been found to be more discriminative than the previously popular WWL graph kernel. Around the same time, the optimal transition coupling (OTC) distance was proposed by O'Connor et al. (2022) for comparing stationary Markov chains, i.e., Markov chains with stationary distributions and the study was followed by Yi et al. (2021) with applications in comparing graphs.

The WL distance proposed in (Chen et al., 2022) and the OTC distance in (O'Connor et al., 2022) represent two SOTA approaches in comparing labeled graphs (i.e., graphs with node attributes). In fact, both of them compare more general Markov chains like objects. The Markov chain perspective not only relieves the difficulty in handling combinatorial structures of graphs but also provides a natural and unified way of modelling both directed and undirected graphs. To broaden the use of these distances, especially in graph learning and optimization (e.g., to use such distance as graph reconstruction loss in a generative model), it is crucial that we are able to differentiate such distances w.r.t. changes in input graphs. However, differentiating these distances appears to be challenging.

Our contributions. We propose in Section 3 a unified framework to generate distances between Markov chains (and thus also for labeled graphs), which we call the *Optimal Transport Markov (OTM)* distances. This framework of OTM distances encompasses both the WL distance and the OTC distance and in particular, we prove that the two distances serve as extreme points in the family of OTM distances. We further identify a special one-parameter family of distances within our general framework of OTM distances, and we call our new distance *the δ-discounted WL distance* (for a parameter $\delta \in [0, 1]$) in Section 4. Not only do we unveil succinct connections between our discounted WL distance and both the WL and the OTC distances, but we also show that the discounted WL distance has better theoretical properties than the other two distances:

1. Contrary to the WL and the OTC distances, the discounted WL distance can be used to compare non-stationary Markov chains.
2. The discounted WL distance has the same discriminative power as the OTC distance and possibly stronger discriminative power than the WL distance.
3. All the three types of distances are computed via iterative schemes. We devise an algorithm of the discounted WL distance which converges provably faster than the one for the WL distance introduced in (Chen et al., 2022); whereas to the best of our knowledge, there is no known study on convergence rate of the OTC distance.
4. Furthermore, contrary to both the OTC and the WL distances, a regularized version of the δ-discounted WL distances can be differentiated against its parameters, enabling a range of possible applications as a loss in machine learning or in other optimization tasks. In Section 5, we give a simple formula to compute its gradients.

Note that the effectiveness of the WL distance was already shown in (Chen et al., 2022) where it compared favorably with other graph kernels. Our discounted WL distance is provably more discriminative (e.g, Proposition 5), and thus we expect it will lead to even better practical performance.

Relation to the fused-GW (FGW) distance of Vayer et al. (2019). The fused-GW (FGW) distance also leverages the optimal transport idea, and in fact, uses the Gromov-Wasserstein distance to compute two graphs (equipped with metric structures at nodes). The authors also developed a heuristic algorithm to approximate this algorithm in practice. While the algorithms work well in practice (Vincent-Cuaz et al., 2021; Vayer et al., 2019), there are no theoretical guarantees for them and in fact, the FGW algorithm is only proven to converge to a local minimum (of a provably non-convex function). Current methods on optimizing to minimize FGW as a loss relies on a kind of block coordinate descent, updating alternatively the OT matching (using the FGW algorithm) the parameters by gradient descent with fixed matching (Vincent-Cuaz et al., 2021; Brogat-Motte et al., 2022; Xia et al., 2023). In contrast, we can compute our δ-discounted WL distance –and its gradient in the case of the regularized version– exactly, allowing us to easily optimize it. We further remark that the FGW distance and our OTM distance adopt fundamentally different points of view: FGW stems from the interpretation of graphs as metric spaces, whereas OTM distances rise from random walks on graphs viewed as probabilistic objects (Markov Chains).

2. Preliminaries

We include an appendix within the supplementary material that contains all the detailed proofs, algorithms and experimental details. We provide a Glossary of all notations we use in Appendix F.

2.1. Probability Measures and Markov Chains

In this paper, we use boldface letters, such as $\boldsymbol{X}$, to denote finite sets. We let $\mathcal{P}(X)$ to denote the space of all probability measures on $\boldsymbol{X}$.

A *finite Markov chain* $\mathcal{X} = (\boldsymbol{X}, m_\bullet^{\boldsymbol{X},(\bullet)}, \nu^{\boldsymbol{X}})$ consists of a finite state space $\boldsymbol{X}$, a Markov transition kernel $m_\bullet^{\boldsymbol{X},(\bullet)} : x, t \in \boldsymbol{X} \times \mathbb{N} \to m_x^{\boldsymbol{X},(t)} \in \mathcal{P}(\boldsymbol{X})$, and an initial distribution $\nu^{\boldsymbol{X}}$. A *realization of a Markov chain* $\mathcal{X}$ is a sequence of random variables $(X_t : \Omega \to \boldsymbol{X})_{t\in\mathbb{N}}$ on a common probability space $(\Omega, \mathbb{P})$ such that $\mathrm{law}(X_0) = \nu^{\boldsymbol{X}}$ and $\mathbb{P}(X_{t+1} = x' | X_t = x) = m_x^{\boldsymbol{X},(t)}(x')$ for any $x, x' \in \boldsymbol{X}$. If $m_\bullet^{\boldsymbol{X},(t)}$ is independent of the time t, we call the chain $\mathcal{X}$ *time homogeneous*. In that case, we will omit the t parameter and write $m_\bullet^{\boldsymbol{X}}$. We will also use the notation $m_{xx'}^{\boldsymbol{X}} := m_x^{\boldsymbol{X}}(x')$ for compactness later. If the initial distribution $\nu^{\boldsymbol{X}}$ is stationary, then we call $\mathcal{X}$ a *stationary Markov chain*.

Couplings. Let $\boldsymbol{X}, \boldsymbol{Y}$ be two finite sets and let $\alpha \in \mathcal{P}(\boldsymbol{X}), \beta \in \mathcal{P}(\boldsymbol{Y})$. We call $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$ *a coupling between* α *and* β if for any $A \subseteq \boldsymbol{X}$ and $B \subseteq \boldsymbol{Y}$, one has that $\mu(A \times \boldsymbol{Y}) = \alpha(A)$ and $\mu(\boldsymbol{X} \times B) = \beta(B)$. We let $\mathcal{C}(\alpha, \beta)$ denote the set of all couplings between α and β. Couplings can be also interpreted via random variables. Let $X : \Omega \to \boldsymbol{X}$ and $Y : \Omega \to \boldsymbol{Y}$ be two random variables from some same probability space $(\Omega, \mathbb{P})$ into the spaces $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, such that $\mathrm{law}(X) = \alpha$ (we also write $X \sim \alpha$) and $\mathrm{law}(Y) = \beta$. Then, it is easy to check that $\mathrm{law}((X, Y)) \in \mathcal{C}(\alpha, \beta)$ and that any coupling in $\mathcal{C}(\alpha, \beta)$ can be obtained in this way. We hence also write $(X, Y) \in \mathcal{C}(\alpha, \beta)$.

Markovian couplings. Given two Markov chains $\mathcal{X}$ and $\mathcal{Y}$, a stochastic process $(X_t, Y_t)_{t\in\mathbb{N}}$ on a probability space $(\Omega, \mathbb{P})$ is *a Markovian coupling* (Chen et al., 2023) between them if

- $(X_t, Y_t)_{t\in\mathbb{N}}$ satisfies the (time inhomogeneous) Markov property: for any $t \in \mathbb{N}$,

$$\mathbb{P}((X_{t+1}, Y_{t+1}) | (X_t, Y_t), \dots, (X_0, Y_0)) = \mathbb{P}((X_{t+1}, Y_{t+1}) | (X_t, Y_t)).$$

- For any $t \in \mathbb{N}$, any $x \in \boldsymbol{X}$ and $y \in \boldsymbol{Y}$, $m_{xy}^{\boldsymbol{XY},(t)}$ defined below belongs to $\mathcal{C}(m_x^{\boldsymbol{X}}, m_y^{\boldsymbol{Y}})$:

$$m_{xy}^{\boldsymbol{XY},(t)} := \mathbb{P}((X_{t+1}, Y_{t+1})|(X_t, Y_t) = (x, y)) \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y}).$$

- The initial distribution is a coupling: $\mathrm{law}((X_0, Y_0)) \in \mathcal{C}(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}})$.

We let $\Pi(\mathcal{X}, \mathcal{Y})$ denote the collection of all Markovian couplings. Given a Markovian coupling $(X_t, Y_t)_{t\in\mathbb{N}}$, if for each $t \in \mathbb{N}$ and each $x \in \boldsymbol{X}$ and $y \in \boldsymbol{Y}$, one has that $m_{xy}^{\boldsymbol{XY},(t)} = m_{xy}^{\boldsymbol{XY},(1)}$, then we say $(X_t, Y_t)_{t\in\mathbb{N}}$ is a *time homogeneous* Markovian coupling. We denote by $\Pi_{\mathrm{H}}(\mathcal{X}, \mathcal{Y})$ the collection of all time homogeneous Markovian couplings w.r.t. $\mathcal{X}$ and $\mathcal{Y}$.

2.2. Optimal Transport and Distances between Markov Chains

The Wasserstein distance. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two finite sets. Assume that α, β are probability measures over $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. We call any function $C : \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}_+$ a *cost function* between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Then, the *Optimal Transport (OT) distance* between α and β is defined as follows:

$$d_{\mathrm{W}}(\alpha, \beta; C) := \inf_{(X,Y)\in\mathcal{C}(\alpha,\beta)} \mathbb{E}\, C(X, Y), \tag{1}$$

where $\mathbb{E}$ denotes the expectation. When $\boldsymbol{X} = \boldsymbol{Y}$ and $C := d_{\boldsymbol{X}}$ is a distance function on $\boldsymbol{X}$, the quantity $d_{\mathrm{W}}(\alpha, \beta; d_{\boldsymbol{X}})$ is also called the *Wasserstein distance* between α and β as d_{W} is a metric distance on $\mathcal{P}(\boldsymbol{X})$.

The Weisfeiler-Lehman distance (WL distance). Consider two finite *stationary* Markov chains (i.e., Markov chains with stationary initial distributions) $\mathcal{X}$ and $\mathcal{Y}$, with a cost function $C : \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}_+$. Inspired by (Chen et al., 2022), given any $k \in \mathbb{N}$, the *depth-k Weisfeiler-Lehman (WL) distance* between them is defined as follows:

$$d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y}; C) := \inf_{(X_t,Y_t)_{t\in\mathbb{N}}} \mathbb{E}\, C(X_k, Y_k), \tag{2}$$

where the infimum is taken over all possible Markovian couplings $(X_t, Y_t)_{t\in\mathbb{N}} \in \Pi(\mathcal{X}, \mathcal{Y})$. Then, the Weisfeiler-Lehman distance is defined as

$$d_{\mathrm{WL}}^{(\infty)}(\mathcal{X}, \mathcal{Y}; C) := \sup_{k\in\mathbb{N}} d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y}; C). \tag{3}$$

Remark 1 (Nuance in definition) *The definition of the WL distance above is based on a characterization of the WL distance in Chen et al. (2023). The (depth-k) WL distance was originally inspired by the classical Weisfeiler-Lehman graph isomorphism test. Specifically, the depth-k WL distance was designed to emulate the kth iteration of the WL test*[1]*. Our definition is slightly more general than the one in Chen et al. (2023): the (depth-k) WL distance was originally defined between two Markov chains endowed with label functions $\ell_{\boldsymbol{X}} : \boldsymbol{X} \to \mathbf{Z}$ and $\ell_{\boldsymbol{Y}} : \boldsymbol{Y} \to \mathbf{Z}$ into a common metric space $(\mathbf{Z}, d_{\mathbf{Z}})$. Using our language, this is equivalent to saying that the cost function involved is of the form $C(x, y) = d_{\mathbf{Z}}(l_{\boldsymbol{X}}(x), l_{\boldsymbol{Y}}(y))$.*

The stationary assumption is redundant when $k < \infty$, and hence $d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y}; C)$ is defined for any Markov chains. When $k = \infty$, the situation becomes subtle; see more discussion in Appendix B.

1. In the WL test literature, the index k typically denotes the order of the test. However, in this context, we use it to denote the depth.

The Optimal Transport Coupling distance (OTC distance). Consider two finite stationary Markov chains $\mathcal{X}$ and $\mathcal{Y}$. Let $C : \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}_+$ be any cost function. Then, the *optimal transition coupling (OTC) distance* (O'Connor et al., 2022), which we denote by d_{OTC}, is defined as

$$d_{\mathrm{OTC}}(\mathcal{X}, \mathcal{Y}; C) := \inf_{\substack{(X_t, Y_t)_{t\in\mathbb{N}} \in \Pi_{\mathrm{H}}(\mathcal{X},\mathcal{Y}) \\ \mathrm{law}((X_0,Y_0)) \text{ is stationary}}} \mathbb{E}\, C(X_0, Y_0), \tag{4}$$

where infimum is over *time homogeneous* Markovian couplings with *stationary* initial distributions.

Remark 2 (A note on symbols) *For simplicity of presentation, in what follows, we will sometimes omit the cost matrix C in our notation of distances when its choice is clear. For example, we may write $d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y})$ instead of $d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y}; C)$.*

3. Optimal Transport Markov Distances

Note the similarity between Equation (2) and Equation (4): they are both infimizing certain expected costs between random walk paths. Motivated by this similarity, in this section, we devise a general framework for constructing distances between Markov chains with *arbitrary* initial distributions in contrast to the stationary condition for the two distances mentioned above. We will show how this framework incorporates the (depth-k) WL distance and admits the OTC distance as a limit point, and how these two distances appear as lower and upper bounds for this family of distances. All proofs of results in this section can be found in Appendix E.2.

In what follows, we assume $\mathcal{X}$ and $\mathcal{Y}$ are two (not necessarily stationary) finite Markov chains endowed with any cost function $C : \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}_+$.

Definition 3 (Generalized Optimal Transport Markov Distance (OTM distance)) *Let $p \in \mathcal{P}(\mathbb{N})$ be a distribution over all non-negative integers, $T \sim p$ be a random variable. We define the* Optimal Transport Markov (OTM) distance *associated to p, between two Markov chains $\mathcal{X}$ and $\mathcal{Y}$ as:*

$$d_{\mathrm{OTM}}^{p}(\mathcal{X}, \mathcal{Y}; C) = \inf_{(X_t, Y_t)_{t\in\mathbb{N}}} \mathbb{E}\, C(X_T, Y_T), \tag{5}$$

where the infimum is taken over all Markovian couplings $(X_t, Y_t)_{t\in\mathbb{N}}$ independent *of T.*

Remark 4 (Optimal Markovian couplings exist) *In fact, the infimum in Equation (5) can be replaced by a* minimum*: there exists a Markovian coupling $(X_t, Y_t)_{t\in\mathbb{N}}$ such that $d_{\mathrm{OTM}}^{p}(\mathcal{X}, \mathcal{Y}; C) = \mathbb{E}\, C(X_T, Y_T)$. We refer the reader to Appendix E.2 for a proof.*

We also remark that, just as the Optimal Transport problem gives rise to the Wasserstein distance between probability measures on the same underlying metric space, the OTM distance above becomes a pseudometric on the collection of Markov chains with a pseudometric space $(\boldsymbol{X}, d_{\boldsymbol{X}})$ being their common state space; we refer the reader to Appendix E.2.1 for details.

Example 1 ($d_{\mathrm{WL}}^{(k)}$ is an OTM distance) *Let δ_k denote the Dirac delta measure at $k \in \mathbb{N}$. Then, it is obvious that $d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y}) = d_{\mathrm{OTM}}^{\delta_k}(\mathcal{X}, \mathcal{Y})$. In this way, although $d_{\mathrm{WL}}^{(\infty)}$ is not an instance of OTM distances, it is actually the limit of a sequence of OTM distances.*

Besides the example above, one can establish the following bounds for the OTM distance utilizing the (depth-k) WL distance and the OTC distance.

Proposition 5 (A $d_{\mathrm{WL}}^{(k)}$-based lower bound) *For any distribution p on $\mathbb{N}$, one has that*

$$d_{\mathrm{OTM}}^{p}(\mathcal{X},\mathcal{Y}) \geq \mathbb{E}_{T\sim p}(d_{\mathrm{WL}}^{(T)}(\mathcal{X},\mathcal{Y})) = \sum_{k\in\mathbb{N}} p(k) d_{\mathrm{WL}}^{(k)}(\mathcal{X},\mathcal{Y}).$$

Proposition 6 (d_{OTC} is an upper bound) *For all distributions p on $\mathbb{N}$, and* stationary *Markov chains $\mathcal{X},\mathcal{Y}$, one has that: $d_{\mathrm{OTM}}^{p}(\mathcal{X},\mathcal{Y}) \leq d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y})$.*

The above two propositions suggest that the (depth-k) WL-distance and the OTC distance can be viewed as the two extremes of the family of OTM distances. In fact, in Section 4.3 later, we will show that the OTC distance turns out to be a *tight* upper bound of the family of OTM distances. Furthermore, although the OTC distance serves as an upper bound, the following result states that a large family of OTM distances has the same discriminative power as the OTC distance.

Proposition 7 (Zero-sets) *Suppose that $\mathcal{X},\mathcal{Y}$ are* stationary *and that p is fully supported (i.e., $\forall t\in\mathbb{N}, p(t)>0$). Then, $d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y})=0$ iff $d_{\mathrm{OTM}}^{p}(\mathcal{X},\mathcal{Y})=0$, implying that these two distances "distinguish" the same sets of stationary Markov chains.*

The OTM distances have many interesting theoretical properties. For example, any OTM distance is indeed a "pesudo-distance" satisfying the triangle inequality under suitable conditions (cf. Proposition 34). Furthermore, the OTM distance is continuous with respect to the probability measure p (cf. Lemma 33). Interested readers are referred to Appendix C for more details.

We also refer the reader to the discussion provided in Appendix C.1 for a general way of computing OTM distances with finitely-supported distribution p, and a detailed view of how exactly the geometric distribution – and specifically its memoryless property – is used to obtain the simplified computation. This motivates our study of the *discounted WL distances* (whose distribution p is a geometric law) in Section 4, which will be our focus in the rest of the paper. These distances can be computed more easily than general OTM distances by using a fixed-point algorithm. Furthermore, an entropy-regularized version of them can be differentiated. These nice properties make them a natural choice of OTM distances for applications which we explore in Section 6.

4. The Discounted WL Distance

The OTM distances we introduced above encompass the WL distance and the OTC distance at the two extremes, and are a (significant) generalization of both distances. However, one might wonder why it is useful to consider this general formulation. In this section we first note some limitations of the two distances and then propose the *discounted WL distance*, which is a special instance of our OTM distance, as a remedy of those limitations. Indeed, we will see that the discounted WL distance can compare more general Markov chains, is more efficient to compute, and more importantly, has a relaxed form that can be differentiated (whereas the WL distance and the OTC distance cannot).

All missing proofs from this section are in Appendix E.3.

4.1. Limitations of the WL Distance and the OTC Distance

Stationary initial distributions. Both the WL distance and the OTC distance are only defined for Markov chains with stationary initial distributions. This assumption is quite limited and, in general, does not even accommodate the uniform measure, which assigns equal weight to all states. Note

that while the definition of the (depth-k) WL distance could be extended to Markov chains with any initial distributions, on irreducible and aperiodic Markov chains, depth-∞ WL distance turns out to be independent on the initial distribution. Hence, the extension of the WL distance to non-stationary case is meaningless; see Appendix B for more details.

Rate of convergence. The depth-k WL distance converges as $k \to \infty$ (see our discussion in Appendix B.2) provided that the cost is of the form given in Remark 1. However, this convergence is based on the convergence to a non-unique fixed point of a map. Due to the non-uniqueness feature, this convergence may be vulnerable to numerical errors. We have further established an estimate of the rate of convergence for $d_{\mathrm{WL}}^{(k)} \to d_{\mathrm{WL}}^{(\infty)}$ as $k \to \infty$. Although this limit converges exponentially fast, the rate depends on the input Markov chains. See Appendix B.2 for details.

The algorithm for computing the OTC distance is through the policy iteration on an average-cost Markov Decision Process (O'Connor et al., 2022). Although the policy iteration terminates in a finite number of iterations, as far as we know, this number does not have a reasonable bound (except of the obvious upper bound of $n_{\text{actions}}^{n_{\text{states}}}$ of the number of possible policies)

This all together motivates us to seek for a distance that can be computed via a stable iterative algorithm which can provably converge faster than the one for the WL distance.

Differentiation. To the best of our knowledge, there is no known way to compute the derivative of those two distances with respect to input Markov chains and cost functions. Although the WL distance can be formulated as a fixed point to a certain map, this map lacks the desired contracting property to guarantee uniqueness and smoothness of fixed points. For the OTC distance, differentiating it seems to be even more challenging. Further we are not aware of any way to formulate the OTC distance to a Banach fixed point and thus our strategy for differentiating the discounted WL distance to be introduced (in Section 5) does not apply to differentiate the OTC distance.

4.2. The δ-Discounted WL Distance

We now introduce a one parameter family of instances of OTM distances which has close relationship with the WL distance and the OTC distance. This family of distances addresses those limitations mentioned above in Section 4.1.

The distance that we define next, called the δ-discounted WL distance, is essentially a regularized version of the WL distance (this view will be more evident by considering Proposition 15 later). This regularization enables us to compute the new distance via solving a Banach fixed point problem. This approach is very tractable, addressing the limitations of the original WL distance, and providing the ability to differentiate our distance.

For the purpose of introducing our discounted WL distance, we consider the following two types of distributions on $\mathbb{N}$ given $\delta \in [0, 1]$.
Geometric distribution p_δ^∞: if we let a RV $T_\delta^\infty \sim p_\delta^\infty$, then $\mathbb{P}(T_\delta^\infty = t) = \delta(1-\delta)^t$, $\forall t \in \mathbb{N}$;
Truncated geometric distribution p_δ^k for $k \in \mathbb{N}$: let $T_\delta^k := \min(T_\delta^\infty, k)$, then $T_\delta^k \sim p_\delta^k$:

$$\mathbb{P}(T_\delta^k = t) = \begin{cases} \delta(1-\delta)^t, & t \le k-1 \\ (1-\delta)^k, & t = k \\ 0, & t > k \end{cases}.$$

Definition 8 (δ-discounted WL distance) *For any $k \in \mathbb{N} \cup \{\infty\}$, the* depth-$k$ δ-discounted WL distance *is defined as follows*

$$d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y};C) := d^{p^k_\delta}_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C) \text{ and more explicitly,} \tag{6}$$

$$d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}) := \inf_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi(\mathcal{X},\mathcal{Y})} \begin{cases} \mathbb{E}\Big(\sum_{t=0}^{k-1} \delta(1-\delta)^t C(X_t,Y_t) + (1-\delta)^k C(X_k,Y_k)\Big), & k<\infty \\ \mathbb{E}\big(\sum_{t=0}^{\infty} \delta(1-\delta)^t C(X_t,Y_t)\big), & k=\infty \end{cases}.$$

Remark 9 ($k=\infty$) *$d^{(\infty)}_{\mathrm{WL},\delta}$ is closely related to the* bicausal optimal transport *distance from Moulos (2021): the bicausal optimal transport distance, with a discount factor of $(1-\delta)$ and a* binary *cost matrix C, is the same as our $d^{(\infty)}_{\mathrm{WL},\delta}$ up to a multiplicative constant δ.*

Remark 10 ($\delta=0$) *Note that for any finite k, $d^{(k)}_{\mathrm{WL},0}(\mathcal{X},\mathcal{Y}) = d^{(k)}_{\mathrm{WL}}(\mathcal{X},\mathcal{Y})$.*

The δ-discounted WL distance behaves nicely when k approaches ∞:

Proposition 11 (Convergence w.r.t. k) *For any Markov chains $\mathcal{X}$ and $\mathcal{Y}$, any cost function C, and any $\delta \in [0,1]$, one has that $d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}) = \lim_{k\to\infty} d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y})$.*

We will also establish later in Proposition 15 a convergence rate result.

A second nice property of this distance is, although defined via general Markovian couplings, an optimal Markovian coupling can be chosen to be *time-homogeneous* for $d^{(\infty)}_{\mathrm{WL},\delta}$.

Proposition 12 (Optimal Markovian coupling) *Recall that $\Pi_{\mathrm{H}}(\mathcal{X},\mathcal{Y})$ denotes the collection of all time homogeneous Markovian couplings between $\mathcal{X}$ and $\mathcal{Y}$. Then, for any $\delta > 0$, one has that*

$$d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}) = \min_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi_{\mathrm{H}}(\mathcal{X},\mathcal{Y})} \mathbb{E}\, C(X_{T^\infty_\delta}, Y_{T^\infty_\delta}).$$

4.3. Relationship with the WL Distance and the OTC Distance

Recall from Remark 10 that $d^{(k)}_{\mathrm{WL}}(\mathcal{X},\mathcal{Y}) = d^{(k)}_{\mathrm{WL},0}(\mathcal{X},\mathcal{Y})$ for any finite k. In fact, we have the following stronger result, showing that $d^{(k)}_{\mathrm{WL}}$ is an appropriate limit of $d^{(k)}_{\mathrm{WL},\delta}$:

Theorem 13 *For any Markov chains $\mathcal{X}$ and $\mathcal{Y}$, one has that*

$$d^{(k)}_{\mathrm{WL}}(\mathcal{X},\mathcal{Y}) = \lim_{\delta\to 0} d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}) \text{ and hence } d^{(\infty)}_{\mathrm{WL}}(\mathcal{X},\mathcal{Y}) = \lim_{k\to\infty}\lim_{\delta\to 0} d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}). \tag{7}$$

Interestingly, it turns out that if we fix $k=\infty$, then $d^{(\infty)}_{\mathrm{WL},\delta}$ converges to d_{OTC} as $\delta \to 0$. This closes the loop for our previous claim about the OTM distances, showing that d_{OTC} can also be expressed as a limit of OTM distances.

Theorem 14 *For any* stationary *Markov chains $\mathcal{X}$ and $\mathcal{Y}$, one has that*

$$d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y}) = \lim_{\delta\to 0} d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}) \text{ and hence } d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y}) = \lim_{\delta\to 0}\lim_{k\to\infty} d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}). \tag{8}$$

Besides this convergence result, we note that by Proposition 7 the OTC distance and $d^{(\infty)}_{\mathrm{WL},\delta}$ have the same zero-sets for any $\delta > 0$. This implies that although the OTC distance is an upper bound for $d^{(\infty)}_{\mathrm{WL},\delta}$, our new construction has the same discriminative power as the OTC distance.

From Equation (7) and Equation (8), it is tempting to ask whether the order of the limits can be switched and whether $d^{(\infty)}_{\mathrm{WL}} = d_{\mathrm{OTC}}$. Although we empirically observe that $d^{(\infty)}_{\mathrm{WL}} \neq d_{\mathrm{OTC}}$ in general, due to the approximation nature of the algorithms implemented, we do not know for sure whether $d^{(\infty)}_{\mathrm{WL}} = d_{\mathrm{OTC}}$ or not, and we leave this for future study.

4.4. Algorithm and Convergence

As mentioned earlier, the discounted WL distance is a regularized version of the original WL distance, in this section we will elucidate on this claim and provide a recursive algorithm for computing the (depth-k) discounted WL distance.

In Chen et al. (2022), a recursive algorithm was proposed to compute the depth-k WL distance. We provide a δ-regularized version of that algorithm in the proposition below. The δ-regularization results in a unique fixed point solution to the recursive algorithm (as opposed to the original WL algorithm), which enables differentiation.

Proposition 15 (Recursive computation) *Given any $k \in \mathbb{N}$, we recursively define matrices $C^{\delta,(l)}$ for $l = 0, \dots, k$ as follows:*

$$C^{\delta,(0)}_{ij} = C_{ij}, \quad C^{\delta,(l+1)}_{ij} = \delta C_{ij} + (1-\delta)\, d_{\mathrm{W}}\Big(m^{\boldsymbol{X}}_i, m^{\boldsymbol{Y}}_j; C^{\delta,(l)}\Big). \tag{9}$$

Then, the depth-k δ-discounted WL distance can be computed as follows

$$d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y}; C) = d_{\mathrm{W}}\Big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(k)}\Big). \tag{10}$$

The matrices $C^{\delta,(l)}$ depend on the Markov kernels $m^{\boldsymbol{X}}_\bullet, m^{\boldsymbol{Y}}_\bullet$ and on the cost matrix C. When making this dependency apparent is needed, we will use the notation $C^{\delta,(l)}(m^{\boldsymbol{X}}_\bullet, m^{\boldsymbol{Y}}_\bullet, C)$.

Using Proposition 15, one can devise an algorithm to compute $d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y})$ for any finite k. A similar but more intricate recursive computation exists for general OTM distances. The simplicity of the discounted WL distance's recursive computation within the OTM family stems from the memoryless nature of the (truncated) geometric distribution. See Appendix C.1 for more details.

A natural question is whether $d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y})$ can be a good approximation for $d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y})$. We aim to answer this question below based on the observation that Equation (9) is a fixed point iteration, which enables us to use Banach fixed point theorem to prove convergence, and other properties.

Proposition 16 (Convergence of $C^{\delta,(k)}$) *When $\delta > 0$, $C^{\delta,(k)}$ converges to the* unique fixed point *$C^{\delta,(\infty)}$ of Equation (9) which is not a constant matrix (unless C is a constant matrix itself) such that*

$$d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y}) = d_{\mathrm{W}}\Big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(\infty)}\Big). \tag{11}$$

Moreover, $C^{\delta,(k)}$ converges at rate $\|C^{\delta,(k)} - C^{\delta,(\infty)}\|_\infty \leq \frac{2(1-\delta)^k}{\delta}\|C\|_\infty$. Consequentially,

$$|d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y}) - d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y})| \leq \frac{2(1-\delta)^k}{\delta}\|C\|_\infty.$$

When $\delta = 0$ and $m^{\boldsymbol{X}}_\bullet, m^{\boldsymbol{Y}}_\bullet$ are irreducible *and* aperiodic*, then $C^{\delta,(k)}$ converges to a constant matrix.*

As the WL distance corresponds to the case when $\delta = 0$, this proposition actually implies that the WL distance $d^{(\infty)}_{\mathrm{WL}}(\mathcal{X}, \mathcal{Y})$ is independent of the initial distributions of $\mathcal{X}$ and $\mathcal{Y}$ (see also Proposition 20 in Appendix B). When $\delta > 0$, $d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y})$ behaves completely differently and it of course depends on the initial distributions of $\mathcal{X}$ and $\mathcal{Y}$ since $d_{\mathrm{W}}\big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(\infty)}\big)$ depends on $\nu^{\boldsymbol{X}}$ and $\nu^{\boldsymbol{Y}}$ when $C^{\delta,(\infty)}$ is not constant. Together with the fact that d_{OTC} is only defined for stationary Markov chains, we conclude that $d^{(\infty)}_{\mathrm{WL},\delta}$ distinguishes more Markov chains than both the WL and the OTC distances.

Finally, we remark that when $\delta > 0$, the last step for the computation of $d_{\mathrm{WL},\delta}^{(\infty)}$ (Equation (11)), involves solving for a meaningful (i.e., non-constant) optimal coupling between $\nu^{\boldsymbol{X}}$ and $\nu^{\boldsymbol{Y}}$ that minimizes this cost. That coupling provides a matching between the state spaces of $\mathcal{X}$ and $\mathcal{Y}$ which can be used for some applications. Note that, when $\delta = 0$ (corresponding to the WL distance), as $C^{0,(\infty)}$ is a constant matrix, no meaningful coupling/matching can be obtained.

In Appendix D we provide pseudo-codes for computing $d_{\mathrm{WL},\delta}^{(k)}(\mathcal{X}, \mathcal{Y})$ for both finite and infinite k based on the two propositions above as well as a complexity analysis. We also provide certain acceleration techniques in Appendix D, including a faster (in terms of complexity) algorithm in the case where both transition kernel matrices are sparse in Algorithm 5, and techniques to empirically accelerate the computation.

5. Differentiation of the Discounted WL Distance

Recall from Section 4.4 that $d_{\mathrm{WL}}^{\delta,(k)}$ can be computed recursively for any finite k and $d_{\mathrm{WL}}^{\delta,(\infty)}$ can hence be approximated efficiently. However, in many applications such as graph learning, one requires that the distances involved can be differentiated. This motivates us to devise in this section an algorithm to differentiate $d_{\mathrm{WL},\delta}^{(\infty)}(\mathcal{X}, \mathcal{Y}; C)$ w.r.t. change in parameters $\mathcal{X}, \mathcal{Y}$ or the cost C when $\delta > 0$. All missing proofs and details in this section are in Appendix E.4. This section gives the key results needed to compute the gradient. The detailed steps of the computation of the backwards pass are laid out in Algorithm 3 (with the other algorithms in Appendix D).

5.1. Sinkhorn Approximation

To differentiate our distance, we want the steps to be differentiable. Optimal transport as defined in Equation (1) is not a differentiable problem. In the literature, differentiability is achieved by replacing it with a smooth approximation, called the Sinkhorn distance, originally introduced in Cuturi (2013): Using the same notations as in Equation (1), given $\epsilon \geq 0$, the (ϵ-)regularized OT problem is defined as: $d_{\mathrm{W}}^{\epsilon}(\alpha, \beta; C) := \min_{(X,Y)\in\mathcal{C}(\alpha,\beta)} \mathbb{E}\, C(X,Y) - \epsilon H(X,Y)$. Here H denotes the entropy function, i.e., $H(X,Y) := -\sum_{i\in\boldsymbol{X}, j\in\boldsymbol{Y}} P_{ij} \log(P_{ij})$, where $P_{ij} := \mathbb{P}(X = i, Y = j)$.

We now define the entropy-regularized version of our discounted WL distance, denoted by $d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y}; C)$, via formulas shown in Equation (9) and Equation (10) with the optimal transport distance d_{W} all replaced by the ϵ regularized optimal transport distance $d_{\mathrm{W}}^{\epsilon}$. We then set $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X}, \mathcal{Y}; C) := \lim_{k\to\infty} d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y}; C)$. See Appendix E.4 for the precise definition of $d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y}; C)$ and the well-definedness of $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X}, \mathcal{Y}; C)$. It turns out that the entropy-regularized discounted WL distance is indeed an approximate of our original discounted WL distance.

Theorem 17 (Convergence of the entropy-regularized distance) *For any Markov chains $\mathcal{X}$, $\mathcal{Y}$ over a finite number of states, and cost matrix C between these two Markov chains and any $k \in \mathbb{N} \cup \{\infty\}$, one has that $\lim_{\epsilon\to 0} d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y}; C) = d_{\mathrm{WL},\delta}^{(k)}(\mathcal{X}, \mathcal{Y}; C)$. Moreover, one has the following convergence rate:*

$$|d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y}; C) - d_{\mathrm{WL},\delta}^{(k)}(\mathcal{X}, \mathcal{Y}; C)| \leq \frac{\epsilon}{\delta} \log(|\boldsymbol{X}||\boldsymbol{Y}|).$$

5.2. Differentiation of $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X},\mathcal{Y})$

Fixing the underlying sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, the distance $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X},\mathcal{Y};C)$ can be written down explicitly as a function $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}\left(m_\bullet^{\boldsymbol{X}},m_\bullet^{\boldsymbol{Y}},\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}};C\right)$ which depends on $m_\bullet^{\boldsymbol{X}},m_\bullet^{\boldsymbol{Y}},\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}}$ and C. Furthermore, to compute $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X},\mathcal{Y})$, by Definition 37, one first needs to compute the matrix $C^{\epsilon,\delta,(\infty)}$ which is a function $C^{\epsilon,\delta,(\infty)}(m_\bullet^{\boldsymbol{X}},m_\bullet^{\boldsymbol{Y}},C)$ depending on $m_\bullet^{\boldsymbol{X}},m_\bullet^{\boldsymbol{Y}}$ and C. We now devise an algorithm to compute the gradient of $C^{\epsilon,\delta,(\infty)}$. Based on this, the gradient for $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X},\mathcal{Y})$ can then be computed using the chain rule and the differentiation method for entropy-regularized OT (Peyré et al., 2019, Proposition 4.6 and 9.2).

We use the following tensor notation to represent the target gradient of $C^{\epsilon,\delta,(\infty)}$:

$$\Delta:=\left(\Delta_{ij}^{kl}\right)_{\substack{1\le k\le n,1\le l\le m\\1\le i\le n,1\le j\le m}},\ \Gamma:=\left(\Gamma_{ij}^{kk'}\right)_{\substack{1\le k\le n,1\le k'\le n\\1\le i\le n,1\le j\le m}},\ \Theta:=\left(\Theta_{ij}^{ll'}\right)_{\substack{1\le l\le m,1\le l'\le m\\1\le i\le n,1\le j\le m}},$$

where $\Delta_{ij}^{kl}:=\frac{\partial C_{ij}^{\epsilon,\delta,(\infty)}}{\partial C_{kl}}$, $\Gamma_{ij}^{kk'}:=\frac{\partial C_{ij}^{\epsilon,\delta,(\infty)}}{\partial m_{kk'}^{\boldsymbol{X}}}$, $\Theta_{ij}^{ll'}:=\frac{\partial C_{ij}^{\epsilon,\delta,(\infty)}}{\partial m_{ll'}^{\boldsymbol{Y}}}$ and $m_{kk'}^{\boldsymbol{X}}$ (resp. $m_{ll'}^{\boldsymbol{Y}}$) is the transition probability from state k to state k' (resp. from state l to state l'). For each i,j, given the matrix $C_{ij}^{\epsilon,\delta,(\infty)}$ (approximated by $C_{ij}^{\epsilon,\delta,(k)}$ in practice; see also Proposition 16 for an analysis of convergence rate), we solve the regularized optimal transport problem $d_{\mathrm{W}}^{\epsilon}\left(m_i^{\boldsymbol{X}},m_j^{\boldsymbol{Y}};C_{ij}^{\epsilon,\delta,(\infty)}\right)$ to obtain the following data (defined in Definition 36 in Appendix):

- an optimal transport matrix (also called the primal solution) $P_{ij}=\left(P_{ij}^{kl}\right)_{1\le k\le n,1\le l\le m}$;
- and two dual solutions $f_{ij}=\left(f_{ij}^{k}\right)_{1\le k\le n}$ and $g_{ij}=\left(g_{ij}^{l}\right)_{1\le l\le m}$.

These give rise to the following tensors when considering all i and j:

$$P:=\left(P_{ij}^{kl}\right)_{\substack{1\le k\le n,1\le l\le m\\1\le i\le n,1\le j\le m}},\ F:=\left(f_{ij}^{k'}\mathbb{1}_{i=k}\right)_{\substack{1\le k\le n,1\le k'\le m\\1\le i\le n,1\le j\le m}},\ G:=\left(g_{ij}^{l'}\mathbb{1}_{i=l}\right)_{\substack{1\le l\le m,1\le l'\le m\\1\le i\le n,1\le j\le m}}.$$

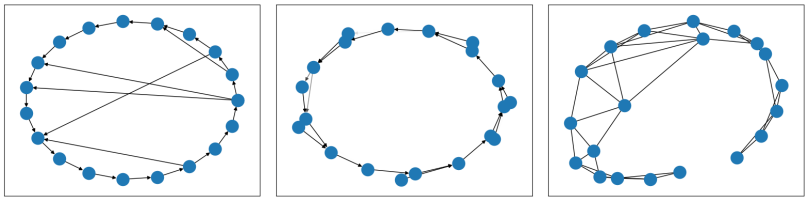
Now that we have computed P, F and G, it turns out that we can use them to directly compute Δ,Γ and Θ, which contain all necessary gradients for $C^{\epsilon,\delta,(\infty)}$.

Theorem 18 (Explicit computation of the gradients) *View the tensors defined above as matrices by flattening their dimensions (and resp. codimensions) together — for example P becomes an $nm\times nm$ square matrix. Let I_{nm} denote the identity matrix of size $nm\times nm$. Then, one has $\Delta=\delta\left(I_{nm}-(1-\delta)P\right)^{-1}$, $\Gamma=(1-\delta)\left(I_{nm}-(1-\delta)P\right)^{-1}F$ and $\Theta=(1-\delta)\left(I_{nm}-(1-\delta)P\right)^{-1}G$.*

Please refer to Appendix D for the pseudo-code that implements gradient computation based on the above theorem as well as analysis on computational complexity.

6. Experiments

In this section, we employ the discounted WL distance for graph classification tasks and the computation of graph barycenters. It is important to highlight that, for computing graph barycenters, we deploy the gradient descent method to minimize the Fréchet functional. This approach necessitates the differentiability of our distance. We also demonstrate how our distance can be used to compute graph coarsening via gradient descent in Appendix A.3.



(a) A noisy circle graph (b) $d^{(\infty)}_{\mathrm{WL},\epsilon,\delta}$ barycenter (c) FGW barycenter

Figure 1: Barycenter computation of 30 noisy circle graphs

Graph Classification. We compared our distance to the fused GW distance (FGW) Vayer et al. (2019) on classification benchmarks on real world datasets from the TUDataset repository (Morris et al., 2020) (see Table 1). The SVM for both FGW and OTM distances are learnt using rbf kernels (with cross-valitated regularization parameter). FGW is run with $\alpha = 0.5$. When attributes are discrete, the cost (distance) used is the Dirac cost (0 if the attributes are the same, 1 otherwise).

1-NN results suggest FGW is superior or similar when labels carry low information, and our distance performs better when the label carries more information. SVM results are mitigated, and suggest similar results as other methods.

dataset	PROTEINS	PTC_MR	PROTEINS_full	ENZYMES
classes	2	2	2	6
attributes	discrete label	discrete label	29	18
FGW 1-NN	$\mathbf{65.1\% \pm 4.6\%}$	$57.6\% \pm 5.0\%$	$69.5\% \pm 4.0\%$	$66.3\% \pm 6.4\%$
$d^{(\infty)}_{WL,\delta}$ ($\delta = 0.2$) 1-NN	$61.4\% \pm 4.0\%$	$\mathbf{61.3\% \pm 7.6\%}$	$\mathbf{70.0\% \pm 4.5\%}$	$\mathbf{74.7\% \pm 6.2\%}$
FGW SVM	$70.5\% \pm 2.9\%$	$57.6\% \pm 4.6\%$	$\mathbf{75.0\% \pm 3.8\%}$	$42.7\% \pm 13.5\%$
$d^{(\infty)}_{WL,\delta}$ ($\delta = 0.2$) SVM	$\mathbf{76.4\% \pm 5.3\%}$	$\mathbf{61.3\% \pm 5.9\%}$	$73.5\% \pm 3.1\%$	$\mathbf{68.3\% \pm 4.1\%}$

Table 1: Results of the classification experiment

Barycenter Computation. In order to show the effectiveness of our proposed OTM distances as optimization targets for learning tasks on directed graphs, we compute a simple graph barycenter. With random noisy cycle graphs $G_1, \ldots, G_n$ as input, we compute the barycenter graph G_{bar} by minimizing the following objective function: $\sum_i d^{(\infty)}_{\mathrm{WL},\delta}(G_{\text{bar}}, G_i; C)$ where the cost C is based on the euclidean distance over the labels. The detailed results (including comparison with the barycenter computed by using fused-GW of Vayer et al. (2019)) and setup can be found in Appendix A.1, and one example is shown in Figure 1 where the discounted WL distance achieves a better barycenter than the fused GW distance.

7. Concluding Remarks

Our paper provides a novel framework of OTM distances comparing Markov chains and hence directed graphs. As our discount-WL distance can be differentiated, the natural next step is to apply our distances to various learning problems, such as to provide effective statistical analysis in the space of graphs (equipped with this metric), or to provide loss for learning models (e.g. graph generative models) over complex networks. In order to make this endeavour easier, we provide the code to

compute it[2], in the form of a packaged python library[3]. We are particularly interested in exploring the use of the discounted WL distance (or variants) to study directed networks, where our current available tool-box has been more limited.

Limitations. On a practical front, the computation of our new distance can be slow on large graphs, although technical optimizations presented in Appendix D mitigate that to some extent. The hyper-parameters (e.g, ϵ, δ) also require careful handling. Though our distance calculation is empirically much slower than the approximate fused-GW distance (See Appendix A.2 for comparisons), it is polynomial-time computable, unlike the NP-hard exact FGW. This difference allows for acceleration techniques, potentially enhancing efficiency. We posit that methods like neural OT (Makkuva et al., 2020; Korotin et al., 2022; Chen and Wang, 2023) could be integrated into our framework for further gains.

Acknowledgments

This work is partially supported by NSF under grants CCF-2112665, CCF-2217058, and CCF-2310411.

References

Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, James Demmel, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, Alan McKenney, and Danny Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999. ISBN 0-89871-447-8 (paperback).

Waïss Azizian and Marc Lelarge. Expressive power of invariant and equivariant graph neural networks. In *International Conference on Learning Representations*, 2021.

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

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

Luc Brogat-Motte, Rémi Flamary, Céline Brouard, Juho Rousu, and Florence d'Alché Buc. Learning to predict graphs with fused Gromov-Wasserstein barycenters. In *International Conference on Machine Learning*, pages 2321–2335. PMLR, 2022.

James R. Bunch and John E. Hopcroft. Triangular factorization and inversion by fast matrix multiplication. *Mathematics of Computation*, 28(125):231–236, 1974. ISSN 00255718, 10886842.

Samantha Chen and Yusu Wang. Neural approximation of wasserstein distance via a universal architecture for symmetric and factorwise group invariant functions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

2. https://github.com/YusuLab/ot_markov_distances
3. https://pypi.org/project/ot-markov-distances/

Samantha Chen, Sunhyuk Lim, Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Weisfeiler-Lehman meets Gromov-Wasserstein. In *International Conference on Machine Learning (ICML)*, pages 3371–3416. PMLR, 2022.

Samantha Chen, Sunhyuk Lim, Facundo Mémoli, Zhengchao Wan, and Yusu Wang. The Weisfeiler-Lehman distance: reinterpretation and connection with GNNs. *ICML workshop: Topology, Algebra, and Geometry in Machine Learning (2023)*, 2023.

Ching-Yao Chuang and Stefanie Jegelka. Tree mover's distance: Bridging graph metrics and stability of graph neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Ran Duan, Hongxun Wu, and Renfei Zhou. Faster matrix multiplication via asymmetric hashing, 2022.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

David Griffeath. A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31(2):95–106, 1975.

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

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2022.

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

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

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6672–6681. PMLR, 2020.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL www.graphlearning.io.

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

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Vittorino Pata et al. *Fixed point theorems and applications*, volume 116. Springer, 2019.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. Wasserstein Weisfeiler-Lehman graph kernels. *Advances in Neural Information Processing Systems*, 32:6439–6449, 2019.

Titouan Vayer, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online graph dictionary learning. In *International Conference on Machine Learning*, pages 10564–10574. PMLR, 2021.

Xinyue Xia, Gal Mishne, and Yusu Wang. Implicit graphon neural representation. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10619–10634. PMLR, 25–27 Apr 2023.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

Bongsoo Yi, Kevin O'Connor, Kevin McGoff, and Andrew B Nobel. Alignment and comparison of directed networks via transition couplings of random walks. *arXiv preprint arXiv:2106.07106*, 2021.

Appendix A. Experiment details

The code to run experiments is available on GitHub[4].

A.1. Barycenter Computation

Target graphs The goal of this experiment is to show that $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}$ produces barycenters that are meaningful with regard to the structure of the input graphs. Here we use simple data: oriented circles with 20 nodes, which we perturb through Erdős-Rényi noise of equal edge addition and deletion probability p. Examples of such data are shown in Figure 2a. The attributes of the nodes of the circles are their (x, y) positions. Our goal is to check that the barycenter approximately recovers the original circle.

Parametric Markov kernels In our experiments, when learning a Markov kernel, it is crucial to ensure that all transition probabilities retain their properties as probability distributions throughout the training process, meaning they remain non-negative and continue to sum to one. We could have used projected gradient descent, but due to better empirical results, we decided to use a *parametric Markov kernel*. An $n \times n$ Markov kernel M is parameterized by an $n \times n$ matrix $\Theta \in \mathbb{R}_+^{n \times n}$ using the parameterization

$$M_i = \mathrm{Softmax}\left(\frac{\Theta_i}{\mathrm{heat}}\right) \tag{12}$$

where heat is a positive floating point parameter, and M_i (resp Θ_i) is the i-th row of M (resp Θ).

This choice of parameterization is both theoretically grounded and practically motivated:

1. **Universality**: This parameterization reaches all dense (without 0 entry) transition matrices and approximates all others.
2. **Standardization**: This aligns with common machine learning practices, where softmax is used to output probability distributions. As a Markov kernel is a collection of probability distributions, this approach is logical. It thus illustrates how our distance could interface with outputs of neural networks.
3. **Convenience**: This method avoids issues like projected gradient descent and degenerated gradients, and is compatible with frameworks like PyTorch (Paszke et al., 2019).
4. **Sparsity Encouragement**: Paradoxically, this parameterization encourages relatively sparse transition matrices via exponentials and certain choice of threshold.

Setup In this experiment, all initial distributions are taken to be uniform.

Let $G^1, \ldots, G^n$ denote the graphs whose barycenter we want to compute. Let $M^1, \ldots, M^n$ denote the transition matrices of the random walks on those graphs, respectively, defined as follows:

$$M^i = (D^i)^{-1} A^i \tag{13}$$

where A^i is the adjacency matrix of G^i and D^i is the diagonal matrix of degrees of G^i. Finally, we let $l^1 \in \mathbb{R}^{s_1 \times d}, \ldots, l^n \in \mathbb{R}^{s_n \times d}$ denote the labels of the graphs, where s_i is the number of nodes of G^i, and d is the label size.

4. https://github.com/YusuLab/ot_markov_distances

Since the size (number of vertices) of the input graphs is not necessarily the same, we need to define the size of the barycenter graph. We leave it as a hyperparameter, and denote it by s.

- the barycenter Markov kernel $M^{\text{bar}} \in \mathbb{R}_+^{s \times s}$
- the labels $l^{\text{bar}} \in \mathbb{R}^{s \times d}$ of the barycenter graph

We encode M^{bar} as a parametric Markov matrix as described in the previous paragraph, and l^{bar} directly as a matrix of parameters.

We then minimize the following objective function:

$$f(M, l) = \sum_i d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}\big(M, M^i; C^i(l)\big) \tag{14}$$

where $C^i(l)$ is the cost matrix defined as

$$C^i(l)_{u,v} = \|l_u - l_v^i\|_2^2 \tag{15}$$

This objective may appear unconventional; however, it is equal to the following:

$$f(M, l) = \sum_i h(M, M^i)^2 \tag{16}$$

where h is the pseudometric defined as in Proposition 35, with $\alpha = 2$ and where the pseudodistance is the L_2 distance between labels. This is the so-called Fréchet variance for the space with pseudometric h, and a minimizer of it as called a Fréchet mean.

And we use the "Adam" optimizer (with the implementation from Pytorch Paszke et al. (2019)) to minimize the objective function.

The parameters for the $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}$ distance we chose are $\delta = 0.5$ and $\epsilon = 0.05$.

The Fused Gromov-Wasserstein barycenters are computed using the official implementation from Vayer et al. (2019). The method is the one described in [Vayer et al, 2019], ie Block Coordinate Descent (BCD). The parameters used are the following

- The tradeoff parameter is $\alpha = 0.95$ (heavily skewed towards the structural loss rather than the attribute loss)
- The weights are not learnt, but fixed to the uniform distribution. This is the same setting as for the delta-discounted WL distance barycenter.

Computational power Each barycenter computation takes about 2.5 to 11 minutes on an Nvidia RTX A6000 GPU depending on the number of target graphs(ranging from 1 to 50). This computation involves 1000 steps with a learning rate of 10^{-2}. Although the time can be reduced by decreasing the number of steps, increasing the learning rate, or increasing either δ or ϵ, these adjustments might degrade the quality of the results.

For a theoretical analysis of the complexity, please refer to Appendix D. Comprehensive performance benchmarking can be found in the Performance.ipynb notebook included in the appended code. The results of this benchmarking are presented in Figure 3.

As a comparison computing one FGW barycenter for this experiment takes between 0.005s and $4.15s$ with an average of 0.67s on CPU (using the code provided by Vayer et al. (2019)). We acknowledge the lack of competitiveness of our method in terms of time complexity, as mentioned in Section 7. We hope that advantages of our distance outweight this problem, and that subsequent work will allow for more compute-efficient approximations.

Results We compare the produced barycenters (in Figure 2b) with the ones produced by the state-of-the-art graph distance, Fused Gromov-Wasserstein distance Vayer et al. (2019) (in Figure 2c). We observe that for higher noise values ($p = 0.01$), our distance recovers the structure significantly better. It is interesting to see that for very high noise ($p = 0.1$), our distance and FGW fail in very similar way: they create one or several "accumulation nodes" that are in the middle (matched with several original nodes) and linked to and from a lot of nodes.

Influence of parameters In this paragraph, we study the influence of the δ and ϵ parameters on the result of this experiment. We run this barycenter expeeriment while varying the values of ϵ and δ, the results are shown in Figure 4 This shows degenerated cases :

- $\delta = 1$ Our distance degenerates to the (regularized) Wasserstein distance between node label sets. Positions are learned through a Wasserstein barycenter problem, but the Markov kernels remain unlearned, with the resulting graph reflecting only random initialization.
- high ϵ Overregularization occurs with high values of , hindering the learning process. The most extreme manifestation of this can be observed in the lower-right part of the grid, where all points are matched equally to each other, resulting in a graph with all nodes clustered at the center.
- low δ A low value of δ introduces instability in learning. This is evident in the upper-left corner of the grid, where the learning process appears erratic.

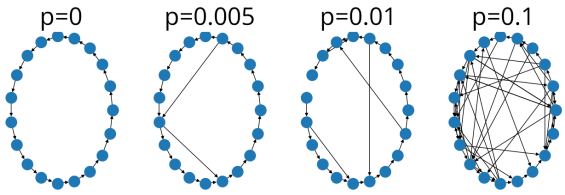
A.2. Graph Classification

The reported accuracy and error margins are the average and standard deviation from a stratified k-fold with 5 splits. In terms of runtime, we measured an average of $0.18s$ to compute one discounted WL distance with parameters $\delta = 0.5$, $\epsilon = 0.1$ (using an Nvidia RTX A6000 graphics card, and an AMD EPYC 7452 32-Core Processor). As a comparison, on the same dataset and hardware, FGW takes an average of $0.0099s$ (approximately 19 times faster). Note that our algorithm runs on a GPU because it is easy to parallelize the many independent optimal transport computations while FGW runs on CPU.

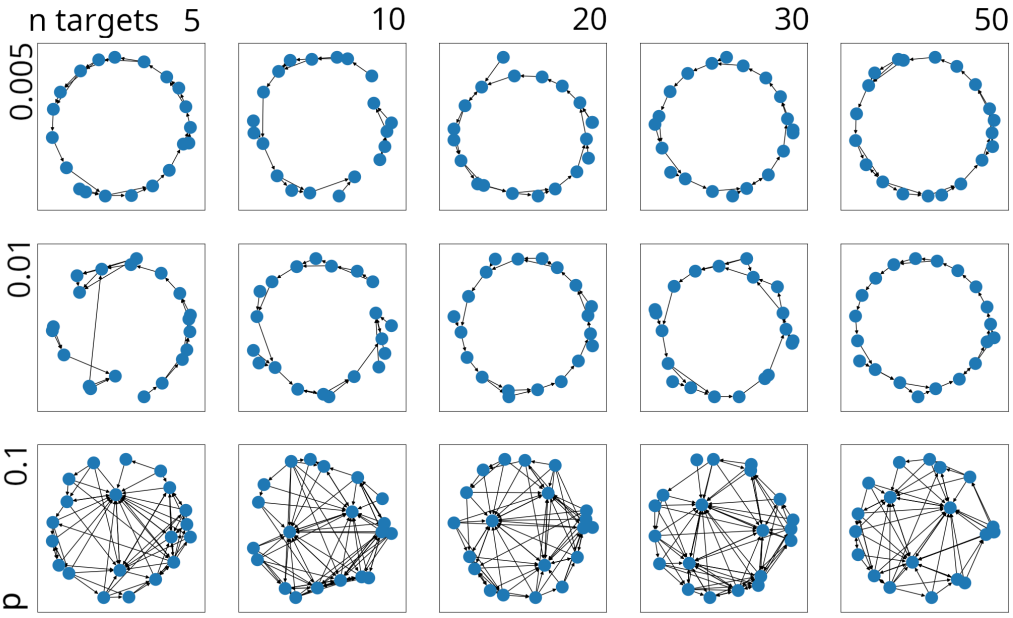
A.3. Graph Coarsening

We also carry out an experiment as a proof of concept on how the discounted WL distance can be used to coarsen graphs. The goal is to coarsen a simple oriented circle of $n = 30$ nodes as in the barycenter experiment into a graph with a given number m of nodes. In order to obtain a coarsening of size m, we minimize an objective on the space of Markov chains of size m, similarly to the barycenter experiment.

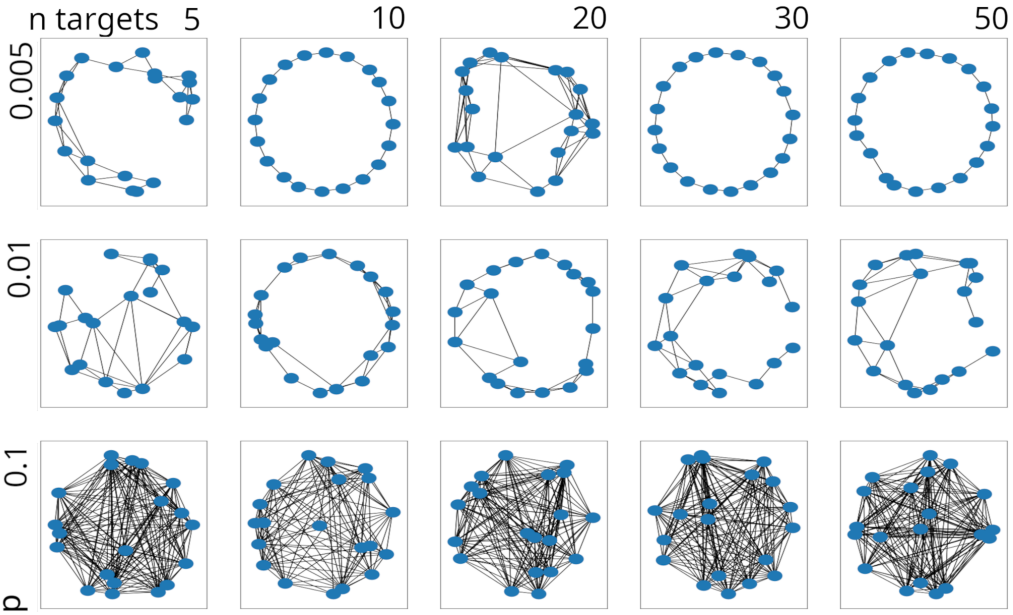
Let M^{target} denote the Markov matrix of the target graph and let $M^{\text{coarsened}}$ denote the Markov matrix of the coarsened graph. A natural objective would be to minimize the δ-discounted WL distance between the original graph and the coarsened graph. This naive approach, however, does not yield good results. An explanation is the following: if the coarsened graph is 4 times smaller (in terms of the number of nodes), then one step of random walk in the coarsened graph should intuitively correspond to 4 steps of random walks in the original graph. In this way, one should think of a coarsened graph as a Markov chain with a larger “time step” than the original graph and hence



(a) Example of perturbed circles depending on the noise level p

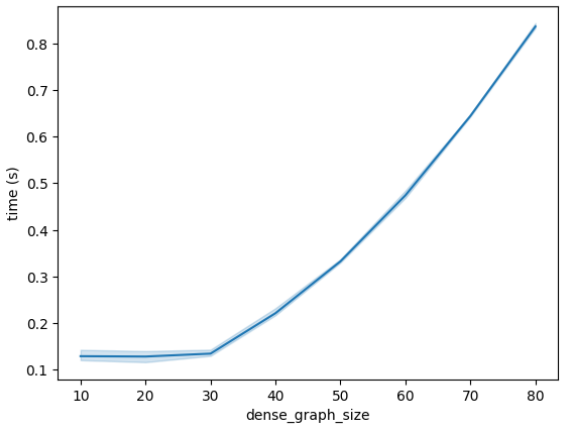


(b) Barycenter computed with different values of p (in ordinate) and for different number of graphs (in abscissa), using our distance

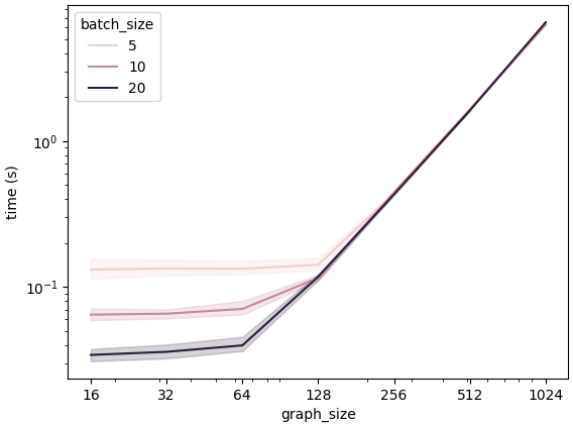


(c) Barycenter computed on the same graphs without orientation, using the FGW barycenters from Vayer et al. (2019)

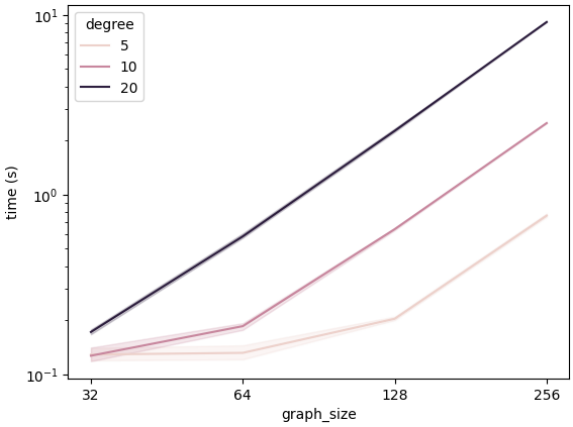
Figure 2: Barycenter experiment



(a) Average time for computing $d^{(\infty)}_{\mathrm{WL},\delta,\epsilon}(G_{\mathrm{dense}}, G_{\mathrm{sparse}})$, where G_{sparse} is of size 64 and degree 5 and the size of G_{dense} is on the abscissa.



(b) Average computing time of $d^{(\infty)}_{\mathrm{WL},\delta,\epsilon}$ with two sparse graphs of varying size with different batch sizes



(c) Average computing time of $d^{(\infty)}_{\mathrm{WL},\delta,\epsilon}$ with two sparse graphs of varying degree and size

Figure 3: Performance analysis results on an Nvidia RTX A6000 GPU

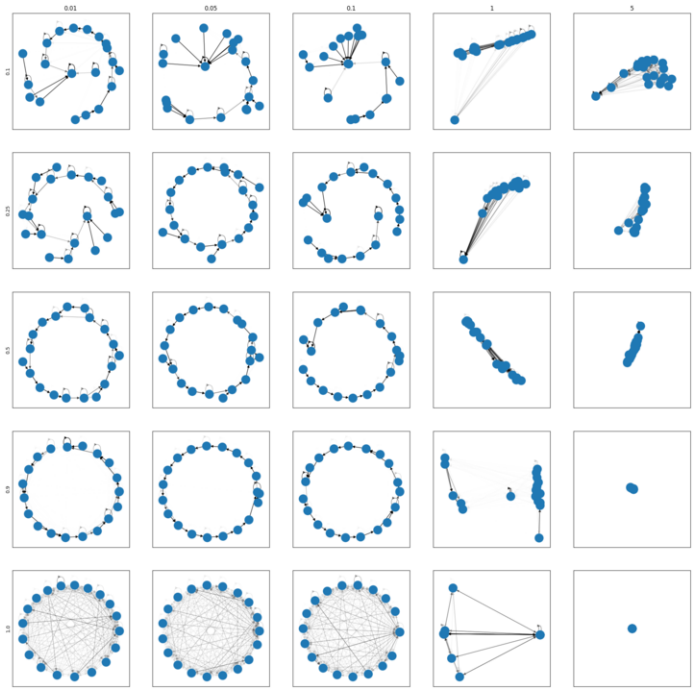


Figure 4: Same barycenter experiment ($n\_targets = 20$, $p = 0.01$), run with different values of ϵ (in abscissa) and δ in ordinate

one should think of the coarsened graph and the original graph induce Markov chains with different time scales.

To adapt to the different time scales, we propose to instead minimize the following objective

$$L = d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}\Big((M^{\text{target}})^k, M^{\text{coarsened}}; C(l)\Big)$$

where $k = \lfloor \frac{n}{m} \rfloor$ is the coarsening factor.

Where l is the set of labels x_i, y_i given to the nodes of the target, and $C(l)$ is the cost matrix so that$C(l)(i,j) = \|(x_i, y_i) - (x_j^{\text{target}}, y_j^{\text{target}})\|$

We use the same parametric markov kernels as in the barycenter experiment (Appendix A.1) and we minimize the objective using the Adam optimizer with a learning rate of 0.005 and 3000 iterations. The results are outlined in Figure 5. We observe that the algorithm gives better results when the coarsened size m is a divisor of the original size n. This hence implies an interesting question of how to coarsen graphs into arbitrary sizes. We leave this as a future work.

Appendix B. New Results on the WL Distance

In this section, we introduce some new results on the WL distance introduced by Chen et al. (2022). Those results justify our motivation for introducing new distances by showing some of the flaws we mentioned in Section 4.1.

Although the original WL distance is defined for Markov chains with stationary initial distributions, Equation (2) can be adapted to define a quantity for Markov chains with arbitrary initial distributions. We can thus define the depth-k WL distance for any Markov chains.

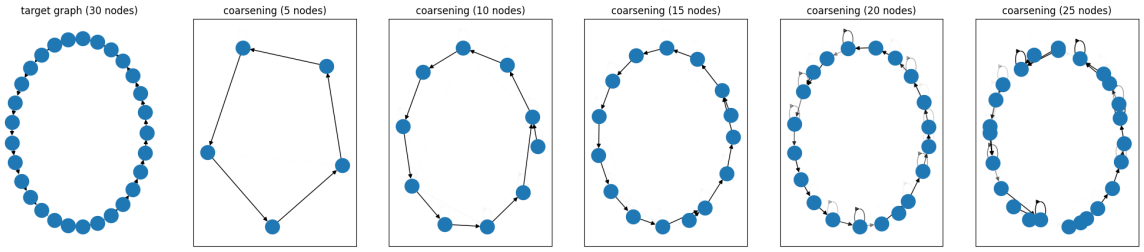


Figure 5: Coarsening results on a circle graph of size 30. The original graph is on the left, the subsequent graphs are coarsenings of different sizes.

It turns out that the depth-k WL distance can be also computed iteratively. We first introduce some constructions.

Definition 19 *Given $k \in \mathbb{N}$, we define $C^{(l)}$ for $l = 0, \dots, k$ recursively as follows.*

$$C_{ij}^{(0)} = C_{ij}, \tag{17}$$

$$C_{ij}^{(l)} = d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{(l-1)}\right). \tag{18}$$

Note that $C^{(l)} = C^{\delta,(l)}$ when $\delta = 0$, where $C^{\delta,(l)}$ is the cost matrix involved in the computation of the discounted WL distance (see Proposition 16).

Those matrices $C^{(l)}$ coincides with the cost matrix computed in the lth iteration in (Chen et al., 2022, Algorithm 1) to compute the depth-k WL distance (for a special type of initial cost function C) in the following way:

$$d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y}) = d_{\mathrm{W}}\left(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{(k)}\right).$$

Notice that, in fact, those matrices above are themselves WL distances in a certain way. More precisely, for any $k \in \mathbb{N}$, one has that

$$d_{\mathrm{WL}}^{(k)}\left((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \delta_i), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \delta_j); C\right) = d_{\mathrm{W}}\left(\delta_i, \delta_j; C^{(k)}\right) = C_{ij}^{(k)}. \tag{19}$$

We analyze $d_{\mathrm{WL}}^{(k)}$ when k approaches ∞ as follows.

Proposition 20 (Convergence of $d_{\mathrm{WL}}^{(k)}$ is independent of initial distributions) *Given a finite set $\boldsymbol{X} = \boldsymbol{Y}$, assume that $m_\bullet^{\boldsymbol{X}}$ and $m_\bullet^{\boldsymbol{Y}}$ are* irreducible *and* aperiodic *Markov transition kernels. Assume also that the cost is defined as a pseudometric in $\boldsymbol{X}$ (This is for example true if the cost is defined as in Chen et al. (2022) or in Remark 1.) Then, for any $\nu^{\boldsymbol{X}} \in \mathcal{P}(\boldsymbol{X})$ and $\nu^{\boldsymbol{Y}} \in \mathcal{P}(\boldsymbol{Y})$, the limit*

$$\lim_{k \to \infty} d_{\mathrm{WL}}^{(k)}\left((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\right) \tag{20}$$

exists and is independent of choices of $\nu^{\boldsymbol{X}} \in \mathcal{P}(\boldsymbol{X})$ and $\nu^{\boldsymbol{Y}} \in \mathcal{P}(\boldsymbol{Y})$.

Note that this property is only true for *irreducible* and *aperiodic* Markov chains. This is a common property in the study of finite Markov chains. See for example (Levin and Peres, 2017) for a good introduction to that theory. A Markov chain is irreducible if all states can be reached from any other state (including itself) in finite positive number of steps (or equivalently if the graph) with positive probability. A Markov chain is aperiodic if for any state s, one has that $\gcd\{k \in \mathbb{N} : \ s \text{ can be reached from } s \text{ in time } k\} = 1$. Irreducibility and aperiodicity ensure the existence of a unique stationary distribution for a finite Markov chain (Levin and Peres, 2017, Corollary 1.17) and its convergence towards that distribution (Levin and Peres, 2017, Theorem 4.9).

In this way, for any irreducible and aperiodic finite Markov chains $\mathcal{X}, \mathcal{Y}$, we redefine $d_{\mathrm{WL}}^{(\infty)}$ (introduced in Equation (3)) as

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

By the proposition above, we know that $d_{\mathrm{WL}}^{(\infty)}$ is independent of choice of initial distributions. In particular, when the initial distributions are stationary, this new definition coincides with the original definition in Chen et al. (2022) since in this case $d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y})$ is increasing w.r.t. k.

B.1. Proof of Proposition 20

In this section, we prove Proposition 20. The proof of Proposition 20 is based on the following observation.

Lemma 21 ($d_{\mathrm{WL}}^{(\infty)}$ does not distinguish initial distributions with the same transitions) *Let $\boldsymbol{X}$ be a finite set and let $m_\bullet^{\boldsymbol{X}}$ denote an irreducible and aperiodic Markov transition kernel on $\boldsymbol{X}$. Assume the assumptions of Proposition 20.*

Then, for any $\nu_1, \nu_2 \in \mathcal{P}(\boldsymbol{X})$, one has that

$$\lim_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_1), (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_2); C\big) = 0,$$

where the cost matrix is defined under the assumptions of Proposition 20 as $C(x, x') = d_{\boldsymbol{X}}(x, x')$

Proof We use symbols $(X_t)_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$ to denote realizations for the two Markov chains $(\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_1)$ and $(\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_2)$, respectively. Then, by definition we have that

$$d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_1), (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_2)\big) = \inf_{\text{Markovian coupling } (X_t, Y_t)_{t\in\mathbb{N}}} \mathbb{E}C(X_k, Y_k). \tag{21}$$

Consider a stochastic process $(X_t, Y_t)_{t\in\mathbb{N}}$ defined as follows:

$$\begin{cases} \text{if } X_t \neq Y_t, & \text{then } \begin{cases} X_{t+1} \sim m_{X_t}^{\boldsymbol{X}} \\ Y_{t+1} \sim m_{Y_t}^{\boldsymbol{X}} \end{cases} \quad \text{independently;} \\ \text{if } X_t = Y_t, & \text{then } X_{t+1} = Y_{t+1} \sim m_{X_t}^{\boldsymbol{X}}. \end{cases}$$

Then, $(X_t, Y_t)_{t\in\mathbb{N}}$ is clearly a time homogeneous Markovian coupling. This coupling has been used for studying convergence of Markov chains, and often called the "classical coupling" (for example, by Griffeath (1975)) since it has the following property:

$$\lim_{k\to\infty} \mathbb{P}(X_k \neq Y_k) = 0. \tag{22}$$

For completeness we provide a proof of the equation above later. In fact, we prove something stronger: there exists $0 \leq \rho \leq 1, t_0 \in N, 0 < \epsilon \leq 1$ depending on the two Markov chains such that

$$\mathbb{P}(X_t \neq Y_t) < (1-\epsilon)^{\lfloor \frac{t}{t_0} \rfloor} \rho. \tag{23}$$

This equation implies that $\lim_{t\to\infty} \mathbb{P}(X_t \neq Y_t)$ converges to 0 exponentially fast. Given this equation, we have that

$$\begin{aligned} &d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_1), (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_2)\big) \leq \mathbb{E}C(X_k, Y_k) && (24)\\ &= \mathbb{P}(X_k = Y_k) \times 0 + \mathbb{P}(X_k \neq Y_k)\mathbb{E}(C(X_k, Y_k) \mid X_k \neq Y_k) && (25)\\ &\leq (1-\epsilon)^{\lfloor \frac{k}{t_0} \rfloor} \rho \cdot \|C\|_\infty. && (26) \end{aligned}$$

Hence,

$$\lim_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_1), (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu_2)\big) = 0.$$

Now, we finish the proof by proving Equation (22).

Proof [Proof of Equation (23)] Let $S = \boldsymbol{X} \times \boldsymbol{X}$ denote the state space of the Markovian coupling $(X_t, Y_t)_{t\in\mathbb{N}}$. Let $E := \{(x, x) : x \in \boldsymbol{X}\} \subseteq S$. For any state $s_0 = (x_0, y_0) \in S$, we define a stopping time as follows

$$T_S := \inf\{t : X_t = Y_t\}.$$

We know that if $(X_t, Y_t) \in E$, then all the subsequent elements also are (by definition of the coupling).

Let $t_0^{s_0} := \inf\{t : \mathbb{P}(T_S \leq t \mid (X_0, Y_0) = s_0) > 0\}$ for any $s_0 \in S$. Note that

- if s^0 in E, then $t_0^{s_0} = 0$ trivially;
- since $m_\bullet^{\boldsymbol{X}}$ is irreducible and aperiodic, one has that $t_0^{s_0} < \infty$ for any $s_0 \in S$: suppose (by contradiction) that $t_0^{s_0} = \infty$ for some $s_0 = (x_0, y_0) \in S \backslash E$. Let $x \in \boldsymbol{X}$ be any state. Given that $(X_0, Y_0) = s_0$, $X_t \neq Y_t$ for all $t \in \mathbb{N}$ almost surely. Then, by (Levin and Peres, 2017, Proposition 1.7), there exists $r_0 \in \mathbb{N}$ so that $\forall t \geq r_0, \mathbb{P}(X_t = x | X_0 = x_0) > 0$ and $\mathbb{P}(Y_t = x | Y_0 = y_0) > 0$. By definition, since $X_t \neq Y_t$ for all $t \geq 1$, X_t is independent of Y_t. Then,

  $$\mathbb{P}(X_t = x, Y_t = x | (X_0, Y_0) = s_0) = \mathbb{P}(X_t = x | (X_0, Y_0) = s_0)\mathbb{P}(Y_t = x | (X_0, Y_0) = s_0) > 0.$$

  This contradicts the fact that $X_t \neq Y_t$ for all $t \in \mathbb{N}$ almost surely. Hence, $t_0^{s_0} < \infty$ for all s_0.

Let $t_0 := \max_{s_0} t_0^{s_0}$, and $\epsilon := \inf_{s_0} \mathbb{P}(T_S \leq t_0 \mid (X_0, Y_0) = s_0)$. Note that $\epsilon > 0$ since $t_0 \geq t_0^{s_0}$. Furthermore, t_0 and ϵ are independent of initial distributions. Then, for any $n \in \mathbb{N}$, we have that

$$
\begin{aligned}
&\mathbb{P}(X_{(n+1)t_0} \neq Y_{(n+1)t_0}) \\
&= \underbrace{\mathbb{P}(X_{(n+1)t_0} \neq Y_{(n+1)t_0} \mid X_{nt_0} = Y_{nt_0})}_{=0} \times \mathbb{P}(X_{nt_0} = Y_{nt_0}) \\
&+ \mathbb{P}(X_{(n+1)t_0} \neq Y_{(n+1)t_0} \mid X_{nt_0} \neq Y_{nt_0}) \times \mathbb{P}(X_{nt_0} \neq Y_{nt_0}) \\
&= \sum_{s_0 \in S \setminus E} \frac{\mathbb{P}(X_{(n+1)t_0} \neq Y_{(n+1)t_0} \mid (X_{nt_0}, Y_{nt_0}) = s_0) \times \mathbb{P}((X_{nt_0}, Y_{nt_0}) = s_0)}{\mathbb{P}(X_{nt_0} \neq Y_{nt_0})} \times \mathbb{P}(X_{nt_0} \neq Y_{nt_0}) \\
&= \sum_{s_0 \in S \setminus E} \mathbb{P}(T_S > t_0 \mid (X_0, Y_0) = s_0) \times \mathbb{P}((X_{nt_0}, Y_{nt_0}) = s_0) \\
&< (1 - \epsilon) \mathbb{P}(X_{nt_0} \neq Y_{nt_0}).
\end{aligned}
$$

We let $\rho := \mathbb{P}(X_0 \neq Y_0)$. Assume that $\rho > 0$ (otherwise $\nu_1 = \nu_2$ and the conclusion holds trivially). Then, by the computation above, one has that

$$\mathbb{P}(X_{nt_0} \neq Y_{nt_0}) < (1 - \epsilon)^n \rho.$$

By our construction of the Markovian coupling $(X_t, Y_t)_{t \in \mathbb{N}}$, we know that $\mathbb{P}(X_t \neq Y_t)$ is decreasing in t since $X_t = Y_t \implies X_{t+1} = Y_{t+1}$. Hence, for any $t \in \mathbb{N}$, we have that

$$\mathbb{P}(X_t \neq Y_t) < (1 - \epsilon)^{\lfloor \frac{t}{t_0} \rfloor} \rho.$$

This concludes the proof. ■

■

Now Proposition 20 follows directly from the lemma above.

Proof [Proof of Proposition 20] Let $\mu^{\boldsymbol{X}}$ and $\mu^{\boldsymbol{Y}}$ be the unique stationary distributions for $m_\bullet^{\boldsymbol{X}}$ and $m_\bullet^{\boldsymbol{Y}}$, respectively. Then, by the triangle inequality we have that

$$
\begin{aligned}
\limsup_{k \to \infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\big) &\leq \underbrace{\limsup_{k \to \infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{Y}})\big)}_{0} \\
&+ \limsup_{k \to \infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big) \\
&+ \underbrace{\limsup_{k \to \infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\big)}_{0} \\
&= d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big).
\end{aligned}
$$

Similarly,

$$\begin{aligned} d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big) &= \liminf_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big) \\ &\leq \underbrace{\liminf_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}})\big)}_{0} \\ &\quad + \liminf_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\big) \\ &\quad + \underbrace{\liminf_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big)}_{0} \\ &= \liminf_{k\to\infty} d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\big). \end{aligned}$$

Therefore, $\lim_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\big)$ exists furthermore, for any $\nu^{\boldsymbol{X}}$ and $\nu^{\boldsymbol{Y}}$, we have that

$$\lim_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\big) = d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big).$$

■

B.2. Convergence of the WL Distance

In this section, we establish that some convergence results on $d_{\mathrm{WL}}^{(k)}$ as k increases.

We assume that $m_\bullet^{\boldsymbol{X}}$ and $m_\bullet^{\boldsymbol{X}}$ are irreducible and aperiodic Markov transition kernels on $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Let $\mu^{\boldsymbol{X}}$ and $\mu^{\boldsymbol{Y}}$ denote their respective unique stationary distributions. We also assume that the cost matrix is defined as in Proposition 20.

Proposition 22 *$(C^{(k)})_{k\in\mathbb{N}}$ converges to a constant matrix.*

Proof By Equation (19) and Proposition 20, one has that for any i, j,

$$\lim_{k\to\infty} C_{ij}^{(k)} = \lim_{k\to\infty} d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \delta_i), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \delta_j)\big) = d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big).$$

■

In fact, we can provide an estimate of the convergence rate of $d_{\mathrm{WL}}^{(k)}$ in the case when C is a pseudometric.

Theorem 23 ($C^{(k)}$ converges exponentially with a pseudometric cost) *Suppose that $\boldsymbol{X} = \boldsymbol{Y}$ is a pseudometric space, and that $C := d_{\boldsymbol{X}}$ is the pseudometric on $\boldsymbol{X}$. If we let*

$$c := d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big),$$

then there exists a rate of convergence $0 \leq \rho < 1$ dependent on $m_\bullet^{\boldsymbol{X}}$ and $m_\bullet^{\boldsymbol{Y}}$ such that

$$\forall (i, j),\ |C_{ij}^{(k)} - c| \leq 2\rho^k \|C\|_\infty. \tag{27}$$

As a direct consequence, we have that

Corollary 24 *For any initial distributions $\nu^{\boldsymbol{X}}$ and $\nu^{\boldsymbol{Y}}$, the quantity $d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}})\big)$ converges to $d_{\mathrm{WL}}^{(\infty)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \mu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \mu^{\boldsymbol{Y}})\big)$* exponentially *fast.*

Proof [Proof of theorem 23] Since C is assumed to be a pseudometric, and $d_{\mathrm{WL}}^{(k)}$ is an OTM distance, using Proposition 34, we can use the triangular inequality on $d_{\mathrm{WL}}^{(k)}$. Let $1 \leq i, u \leq n$, $1 \leq j, l \leq m$, and $k > 0$. Then, one has that

$$\begin{aligned}|C_{ij}^{(k)} - C_{ul}^{(k)}| =& | \, d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \delta_i), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \delta_j)\big) - d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \delta_u), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \delta_l)\big) \, | \\ &\leq d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \delta_i), (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \delta_u)\big) + d_{\mathrm{WL}}^{(k)}\big((\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \delta_j), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, , \delta_l)\big).\end{aligned}$$

Then, from Equation (26), we have that there exists $t_0 \in \mathbb{N}, 0 < \epsilon \leq 1$ such that

$$|C_{ij}^{(k)} - C_{ul}^{(k)}| < 2(1-\epsilon)^{\lfloor \frac{k}{t_0} \rfloor} \cdot \|C\|_\infty.$$

If we let $\rho := (1-\epsilon)^{\frac{1}{t_0}}$, then for any $k \in \mathbb{N}$, one has that

$$|C_{ij}^{(k)} - C_{ul}^{(k)}| < 2\rho^k \|C\|_\infty.$$

Therefore,

$$\max_{ij} C_{ij}^{(k)} - \min_{ij} C_{ij}^{(k)} < 2\rho^k \|C\|_\infty. \tag{28}$$

Using the shorthand $\min C^{(k)} = \min_{ij} C_{ij}^{(k)}$ and $\max C^{(k)} = \max ij C_{ij}^{(k)}$, then we have the following inequalities:

$$\min C^{(0)} \leq \min C^{(1)} \leq \ldots \leq \min C^{(\infty)} = c = \max C^{(\infty)} \leq \ldots \leq \max C^{(1)} \leq \max C^{(0)}. \tag{29}$$

This is a direct consequence of the following inequalities: for any $k > 0$,

$$\min C^{(k)} \leq \min C^{(k+1)} \leq \max C^{(k+1)} \leq \max C^{(k)}. \tag{30}$$

We prove Equation (30) as follows. Let $k > 0$ and let i, j, u, v be such that $\min C^{(k+1)} = C_{ij}^{(k+1)}$ and $\max C^{(k+1)} = C_{uv}^{(k+1)}$. Then,

$$C_{ij}^{(k+1)} = d_{\mathrm{W}}\Big(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{(k)}\Big) \geq \min C^{(k)} \tag{31}$$

and

$$C_{ul}^{(k+1)} = d_{\mathrm{W}}\Big(m_u^{\boldsymbol{X}}, m_l^{\boldsymbol{Y}}; C^{(k)}\Big) \leq \max C^{(k)}. \tag{32}$$

where the inequalities in Equation (31) and Equation (32) follows directly from the definition of optimal transport (the optimal transport cost is smaller than the maximal cost and bigger than the minimal cost).

Then, using Equation (29) and Equation (28), we conclude the proof for Equation (27). ■

Appendix C. Theoretical Details on OTM Distances

In this section, we add conceptual details that can help getting a detailed understanding of OTM distances, and in particular the discounted WL distance.

C.1. Computation of Finite-Time OTM Distances and Usage of the Memoryless Property in the Discounted WL Distance

While we did not include it in the main text for the sake of simplicity, it is possible to compute exactly the value of any finite-time OTM distance, in a way that is similar to the computation of the discounted WL distance.

Theorem 25 *Let p be a finitely supported probability distribution on $\mathbb{N}$. Define n as the maximal value in the support of p (i.e., $n = \max\{t : p(t) > 0\}$). Consider two Markov chains $\mathcal{X}$ and $\mathcal{Y}$ as well as a cost matrix C on $\boldsymbol{X} \times \boldsymbol{Y}$. Then, $d_{\mathrm{OTM}}^{p}(\mathcal{X}, \mathcal{Y}; C)$ can be computed in the following way:*

$$d_{\mathrm{OTM}}^{p}(\mathcal{X}, \mathcal{Y}; C) = d_{\mathrm{W}}\left(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}, C^{p,(n)}\right) \tag{33}$$

where the iterated cost matrices are defined as

$$C_{i,j}^{p,(t)} = \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_T, Y_T) \mid T \geq n - t \text{ and } (X_{n-t}, Y_{n-t}) = (i, j)) \tag{34}$$

where $T \sim p$ is independent of X and Y. The cost matrices can be computed recursively as follows:

$$C_{i,j}^{p,(0)} = C(i, j) \tag{35}$$

$$C_{i,j}^{p,(t)} = \delta_t C(i, j) + (1 - \delta_t) d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{p,(t-1)}\right) \tag{36}$$

where $\delta_t = \mathbb{P}(T = n - t \mid T \geq n - t)$.

A proof of this theorem will be provided at the end of this section.

We remark that Equation (33) has a simpler form in the case of the geometric distribution (ie. in the case of the δ-discounted WL distance).

First note that, in the case of the (truncated) geometric distribution p_δ^k, $\delta_t = \delta$ for all $t = 1, \ldots, n$.

Second, the geometric distribution has the so-called *memoryless* property: Recall the notation $p_\delta^\infty(t) = \delta(1 - \delta)^t$ for the geometric distribution with parameter δ. Assume $T \sim p_\delta^\infty$. Then,

$$\mathbb{P}(T = t + s | T \geq t) = \delta(1 - \delta)^t = \mathbb{P}(T = s) \tag{37}$$

Now, for $k \in \mathbb{N}$, let $T^k = \min(T, k) \sim p_\delta^k$. In this case, a variant of the memoryless property holds as long as $k \geq t + s$ where $t, s \in \mathbb{N}$:

$$\mathbb{P}(T^k = t + s | T^k \geq t) = \mathbb{P}(T^{k-t} = s). \tag{38}$$

The proof is provided later. One immediate consequence of the formula which we will use soon is the following formula:

$$\mathbb{P}(T^k = t + s | T^k \geq t) = \mathbb{P}(T^{k+1} = t + 1 + s | T^{k+1} \geq t + 1). \tag{39}$$

Now, based on the formula above, one has, for $k \geq t$:

$$
\begin{aligned}
&C_{i,j}^{p_\delta^k,(t)} \\
&\quad = \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_{T^k}, Y_{T^k}) \mid T^k \geq k-t \cap (X_{k-t}, Y_{k-t}) = (i,j)) \\
&\quad = \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_{T^{k+1}-1}, Y_{T^{k+1}-1}) \mid T^{k+1} \geq k+1-t \cap (X_{k-t}, Y_{k-t}) = (i,j)) \text{ (Equation (39))} \\
&\quad = \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_{T^{k+1}}, Y_{T^{k+1}}) \mid T^{k+1} \geq k+1-t \cap (X_{k+1-t}, Y_{k+1-t}) = (i,j)) \text{ (Markov property)} \\
&\quad = C_{i,j}^{p_\delta^{k+1},(t)},
\end{aligned}
$$

This shows that the cost matrices $C^{p_\delta^k,(t)}$ are independent of k for $k \geq t$. Thus, we can define a single $C^{\delta,(t)}$ for all truncated geometric distributions p_δ^k, ($C^{\delta,(t)} := C^{p_\delta^k,(t)}$ for any $k \geq t$). Since $\delta_t = \delta$ is independent of k and t as well, we get the simplified formula for computing the discounted WL distance presented in Proposition 15:

$$
\begin{aligned}
C_{ij}^{\delta,(0)} &= C_{ij} \\
C_{ij}^{\delta,(l+1)} &= \delta C_{ij} + (1-\delta)\, d_{\mathrm{W}}\Big(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\delta,(l)}\Big).
\end{aligned}
$$

Moreover, in particular $C^{\delta,(t)} = C^{p_\delta^t,(t)}$ for all t. Thus $d_{\mathrm{WL},\delta}^{(t)}(\mathcal{X},\mathcal{Y})$ can be computed directly from $C^{\delta,(t)}$. This allows us to iterate all the way to convergence to $d_{\mathrm{WL},\delta}^{(\infty)}(\mathcal{X},\mathcal{Y})$. We could not use the same technique to compute the OTM distance with other infinitely-supported distribution q than the geometric distribution p_δ^∞: if we tried to compute $d_{\mathrm{OTM}}^q(\mathcal{X},\mathcal{Y})$ by computing it sequentially for truncated versions of q, we could not reuse the cost matrix of the previous iteration for the subsequent one because the updates in Equation (36) would be different depending on the cutoff.

Additionally, the simplification means the update is always the same, making this computation into a fixed point iteration. This in turn enables the differentiation (that is justified through fixed point properties).

Proof [Proof of Equation (38)] First we assume that $k > t + s$:

$$
\begin{aligned}
\mathbb{P}(T^k = t+s | T^k \geq t) &= \mathbb{P}(\min(T,k) = t+s | \min(T,k) \geq t) \\
&= \mathbb{P}(\min(T,k) = t+s | T \geq t) \\
&= \mathbb{P}(T = t+s | T \geq t) \\
&= \mathbb{P}(T = s) \\
&= \mathbb{P}(\min(T, k-t) = s) \\
&= \mathbb{P}(T^{k-t} = s).
\end{aligned}
$$

Then, we consider the case when $k = t + s$:

$$\begin{aligned}
\mathbb{P}(T^k = t+s | T^k \geq t) &= \mathbb{P}(\min(T,k) = t+s | \min(T,k) \geq t) \\
&= \mathbb{P}(\min(T,k) = t+s | T \geq t) \\
&= \mathbb{P}(T \geq t+s | T \geq t) \\
&= \mathbb{P}(T \geq s) \\
&= \mathbb{P}(\min(T,s) = s) \\
&= \mathbb{P}(T^s = s) \\
&= \mathbb{P}(T^{k-t} = s).
\end{aligned}$$

■

Proof [Proof of Theorem 25] First note that for $t = n$, Equation (34) reduces to

$$C_{i,j}^{p,(n)} = \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_T, Y_T) \mid (X_0, Y_0) = (i,j)). \tag{40}$$

Thus

$$\begin{aligned}
&d_{\mathrm{W}}\left(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{p,(n)}\right) \\
&\quad = \inf_{X_0'\sim\nu^{\boldsymbol{X}}, Y_0'\sim\nu^{\boldsymbol{Y}}} \mathbb{E}(C^{p,(n)}(X_0', Y_0')) \\
&\quad = \inf_{X_0'\sim\nu^{\boldsymbol{X}}, Y_0'\sim\nu^{\boldsymbol{Y}}} \mathbb{E}\left(\inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_T, Y_T) \mid (X_0, Y_0) = (X_0', Y_0'))\right) \\
&\quad = \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_T, Y_T)) \text{ (Markov property)} \\
&\quad = d_{\mathrm{OTM}}^p(\mathcal{X}, \mathcal{Y}; C).
\end{aligned}$$

Now we prove the recursive formula (Equation (36)):

$$\begin{aligned}
C_{i,j}^{p,(t+1)} &= \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_T, Y_T) \mid T \geq n-t-1 \cap (X_{n-t-1}, Y_{n-t-1}) = (i,j)) \\
&= \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{P}(T = n-t-1 | T \geq n-t-1)\mathbb{E}(C(X_{n-t-1}, Y_{n-t-1}) \mid (X_{n-t-1}, Y_{n-t-1}) = (i,j)) \\
&\qquad + \mathbb{P}(T \geq n-t | T \geq n-t-1)\mathbb{E}(C(X_T, Y_T) \mid T \geq n-t \cap (X_{n-t-1}, Y_{n-t-1}) = (i,j)) \\
&= \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \delta_{t+1} C(i,j) + (1-\delta_{t+1})\mathbb{E}(C(X_T, Y_T) \mid T \geq n-t \cap (X_{n-t-1}, Y_{n-t-1}) = (i,j)) \\
&= \delta_{t+1} C(i,j) + (1-\delta_{t+1}) \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_T, Y_T) \mid T \geq n-t \cap (X_{n-t-1}, Y_{n-t-1}) = (i,j)) \\
&= \delta_{t+1} C(i,j) + (1-\delta_{t+1}) \inf_{(X,Y)\sim\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}(C(X_T, Y_T) \mid T \geq n-t \cap (X_{n-t}, Y_{n-t}) \sim (m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}})) \\
&= \delta_{t+1} C(i,j) + (1-\delta_{t+1}) d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{p,(t)}\right).
\end{aligned}$$

■

C.2. A MDP Interpretation of the Discounted WL Distance

In this section, we interpret the algorithm for solving the discounted WL distance (see Proposition 15) as a value iteration process on a certain kind of Markov Decision Process (MDP). This interpretation is not novel, and follows from previous literature (Moulos, 2021; O'Connor et al., 2022), but it gives a different point of view on our iterative algorithm in Proposition 15.

Markov decision processes (MDPs) are a type of model at the center of reinforcement learning theory (Sutton and Barto, 2018). An MDP is defined by a state space S, a collection of action spaces $(A_s)_{s\in S}$, a transition distribution $P : \coprod_{s\in S} A_s \to \mathcal{P}(S)$, and a transition cost $c : S \times A \times S \to \mathbb{R}$.

A policy (or a strategy) is a map $\pi : (s,t) \in S \times \mathbb{N} \mapsto a \in A_s$ which sends a state (and a time) to an action.

Then, given such a policy π and an initial distribution $\nu^S \in \mathcal{P}(S)$, we can define a (time-inhomogeneous) Markov chain $\mathcal{S}_\pi = (S, (m_\pi^{S,(t)})_{t\in\mathbb{N}}, \nu^S)$ whose transition kernels are defined as follows:

$$m_\pi^{S,(t)}(s) = P(\pi(s,t)), \quad \forall t \in \mathbb{N}$$

The goal of MDP theory is to find a policy that minimizes an expected cost $\bar{c} = \mathbb{E}_{S\sim\mathcal{S}_\pi} c(S)$ where $c(S)$ is defined based on the different transitions made in S, and the transition cost $c(s_t, a_t, s_{t+1})$. Examples of the most frequent costs include:

- finite or infinite-time discounted cost (with length $k \in \mathbb{N} \cup \{\infty\}$ and discount factor $\beta \in \mathbb{N}$):

$$c(S) = \sum_{t=0}^{k-1} \beta^t c(S_t, \pi(S_t), S_{t+1});$$

- finite time average cost (with length $k \in \mathbb{N}$):

$$c(S) = \frac{1}{k} \sum_{t=0}^{k-1} c(S_t, \pi(S_t), S_{t+1});$$

- infinite time average cost

$$c(S) = \lim_k \frac{1}{k} \sum_{t=0}^{k-1} c(S_t, \pi(S_t), S_{t+1});$$

- and others such as finite or infinite time total costs.

Such problems can generally be solved by two techniques: value iteration and policy iteration.

In fact, $C^{\delta,(l)}$ defined in Proposition 15 is the lth step cost matrix of a value iteration on the discounted cost Markov Decision Process $(S, (A_s)_{s\in S}, P, c)$ where

- $S = \boldsymbol{X} \times \boldsymbol{Y}$;
- $A_{x,y} = \mathcal{C}(m_x^{\boldsymbol{X}}, m_y^{\boldsymbol{Y}})$;
- $P((x',y')|(x,y),a) = a(x,y)$;
- $c((x,y),a) = C(x,y)$.

We note that this interpretation is not novel, and we list it here only for illustration purpose: As mentioned in Remark 9, $d_{\mathrm{WL},\delta}^{(\infty)}$ coincides with the bicausal optimal transport problem studied in Moulos (2021) when binary cost and discount factor $(1-\delta)$ is considered. Moulos (2021) solves this bicausal optimal transport problem by interpreting it as the same MDP as we described above. Furthermore, this MDP interpretation has also been used in O'Connor et al. (2022) for computing the OTC distance. However, O'Connor et al. (2022) used an average cost (instead of discounted cost) for devising their algorithm.

Remark 26 *Theorem 14 can be interpreted as a convergence result between discounted and average cost MDPs: since $d_{\mathrm{WL},\delta}^{(\infty)}$ is the expected discounted cost of the MDP defined above, and d_{OTC} is defined as the average cost of an MDP, this result can be interpreted as analogous to the classical result that the cost of a discounted-cost finite MDP converges to that of an average-cost MDP if the discount factor approaches zero; see, for example, Puterman (2014) for more details on this result. Note, nevertheless, that this classical result holds only for finite state and finite action spaces, thus it was not applicable here since our action space is infinite; see Appendix C.2 for more details.*

Appendix D. Algorithm and Complexity

In this section, we give algorithmic details on how we compute the discounted WL distance (with Sinkhorn regularization), and its gradient.

We make extensive use of the PyTorch framework (Paszke et al., 2019) to accelerate our code, and we integrate our gradient algorithm into its automatic differentiation engine, so as to make it usable as an optimization target.

The GPU-accelerated code used to compute the distance and its gradient is available as a python library[5], installable with `$ pip install ot_markov_distances`

D.1. Computation and Differentiation of the Discounted WL Distance

Forward pass – computation of the distance The method we use to compute the depth-∞ discounted WL distance is based on Proposition 15. It is described in a simplified way in Algorithms 1 and 2.

Algorithms 1 and 2 are slightly simplified: we note that in both of them, the inner "foreach" loop (line 4 for Algorithm 1 and 7 for Algorithm 2) is embarassingly parallel. In practice, we use GPU acceleration to run the operations in this loop simultaneously.

Sinkhorn distances are computed using the method from Feydy et al. (2019).

Note also that for the depth-∞ version, we take care of saving the primal and dual solutions of every optimal transport computation. These results will be used for the computation of the gradient in Algorithm 3.

The complexity of one pass of computing $d_{\mathrm{W}}^{\epsilon}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C_k\right)$ is $O(n_s nm)$ where n_s is the number of iterations necessary for the Sinkhorn computation to converge.

Thus, the total complexity of this algorithm is $O\left(n_s n_d (nm)^2\right)$, where n_d is the number of iterations needed to converge.

5. https://pypi.org/project/ot-markov-distances/

Algorithm 1: Computation of depth-k discounted WL distance

Input: $m^{\boldsymbol{X}}$: [n, n] float array ;
$m^{\boldsymbol{Y}}$: [m, m] float array ;
$\nu^{\boldsymbol{X}}$: [n] float array ;
$\nu^{\boldsymbol{Y}}$: [m] float array ;
C: [n, m] float array ;
δ: The discount parameter for the WL distance.;
ϵ: Parameter for the Sinkhorn divergence;
k: depth ;
Output: $d^{(k)}_{\mathrm{WL},\delta,\epsilon}\big((\boldsymbol{X}, m^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}}); C\big)$;

$C_{\text{current}} = C$;
foreach $0 \leq l < k$ **do**
  $C_{\text{new}} = (0)_{0 \leq i < n, 0 \leq j < m}$;
  foreach $0 \leq i < n, 0 \leq j < m$ **do**
    $C_{\text{new}}[i, j] = \delta C[i, j] + (1 - \delta) d^{\epsilon}_{\mathrm{W}}\Big(m^{\boldsymbol{X}}_i, m^{\boldsymbol{Y}}_j; C_{\text{current}}\Big)$;
  end
  $C_{\text{current}} = C_{\text{new}}$;
end
Compute $d = d^{\epsilon}_{\mathrm{W}}\big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C_{\text{current}}\big)$;
return d

Backward pass – computation of the gradient: k finite If the distance was computed for a finite (small) depth-k, we can use PyTorch's automatic differentiation engine (Paszke et al., 2019): since all operations done are successive Sinkhorn distances, we only implement the differentiation of a Sinkhorn distance according to (Peyré et al., 2019, Proposition 4.6 and 9.2), and use the automatic differentiation engine to apply the chain rule, to compute the full gradient of the distance.

Note that this approach only works for a small k: the subsequent chain rule applications induce risk of numerical error, as well as the usual problems associated with big differentiation graphs, e.g., high memory footprint, risk of gradient vanishing or explosion, etc.

Backward pass – computation of the gradient: $k = \infty$ If the distance was computed until convergence, we provide two things:

1. A function that compute the backward pass of the Sinkhorn distance, like in the previous paragraph
2. A function that computes the gradient of the $C^{\delta,(\infty)}$ matrix, defined in Algorithm 3, against the parameters $m^{\boldsymbol{X}}, m^{\boldsymbol{Y}}$ and C.

Since the depth-∞ discounted WL distance is computed as

$$d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y}) = d_{\mathrm{W}}\Big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(\infty)}(m^{\boldsymbol{X}}, m^{\boldsymbol{Y}}, C)\Big),$$

provided with those two functions, the automatic differentiation engine is able to apply chain rule to differentiate $d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y})$ against all parameters $\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}, m^{\boldsymbol{X}}, m^{\boldsymbol{Y}}$ and C.

Algorithm 2: Computation of depth-∞ discounted WL distance

Input: $m^{\boldsymbol{X}}$: [n, n] float array ;
$m^{\boldsymbol{Y}}$: [m, m] float array ;
$\nu^{\boldsymbol{X}}$: [n] float array ;
$\nu^{\boldsymbol{Y}}$: [m] float array ;
C: [n, m] float array ;
δ: The discount parameter for the WL distance.;
ϵ: Parameter for the Sinkhorn divergence;
Output: $d^{(\infty)}_{\mathrm{WL},\delta}\big((\boldsymbol{X}, m^{\boldsymbol{X}}_{\bullet}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m^{\boldsymbol{Y}}_{\bullet}, \nu^{\boldsymbol{Y}}); C\big)$;
Stores: f: [n, m, n], g: [n, m, m] float arrays: dual solutions of the last step Sinkhorn computations;
P [n, m, n, m] float array :primal solutions (i.e., optimal matchings) for the last step Sinkhorn computation;

$C_{\text{current}} = C$;
repeat
  $C_{\text{new}} = (0)_{0\leq i<n, 0\leq j<m}$;
  $f = (0)_{0\leq i<n, 0\leq j<m, 0\leq u<n}$;
  $g = (0)_{0\leq i<n, 0\leq j<m, 0\leq l<m}$;
  $P = (0)_{0\leq i<n, 0\leq j<m, 0\leq u<n, 0\leq l<m}$;
  foreach $0 \leq i < n, 0 \leq j < m$ **do**
    $C_{\text{new}}[i,j] = \delta C[i,j] + (1-\delta) d^{\epsilon}_{\mathrm{W}}\Big(m^{\boldsymbol{X}}_i, m^{\boldsymbol{Y}}_j; C_{\text{current}}\Big)$;
    $f[i,j], g[i,j], P[i,j] =$ Dual and primal solutions of the calculation above
  end
  $C_{\text{current}} = C_{\text{new}}$
until $C_{current}$ *converges*;
Compute $d = d^{\epsilon}_{\mathrm{W}}\big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C_{\text{current}}\big)$;
save for backwards pass f, g, P ;
return d

More precisely, denoting by l the value of the loss that we want to differentiate. Denote:

$$\nabla^{C^{\delta,(\infty)}} \mathrm{l}_{ij} := \frac{\partial \mathrm{l}}{\partial C^{\epsilon,\delta,(\infty)}_{ij}} \tag{41}$$

$$\nabla^{C} \mathrm{l}_{kl} := \frac{\partial \mathrm{l}}{\partial C_{kl}}, \tag{42}$$

$$\nabla^{m^{\boldsymbol{X}}} \mathrm{l}_{kk'} := \frac{\partial \mathrm{l}}{\partial m^{\boldsymbol{X}}_{kk'}} \tag{43}$$

$$\nabla^{m^{\boldsymbol{Y}}} \mathrm{l}_{ll'} := \frac{\partial \mathrm{l}}{\partial m^{\boldsymbol{Y}}_{ll'}} \tag{44}$$

The backward pass of the depth-∞ discounted WL distance should input $\nabla^{C^{\delta,(\infty)}} \mathrm{l}$ and output G^C, G^X and G^Y.

Using the notations of the proof of differentiation Section E.4.2 $K := I_{nm} - (1-\delta)P$, then we have:

$$\nabla^C \mathbf{l} = \Delta^T \nabla^{C^{\delta,(\infty)}} \mathbf{l} = \delta (K^T)^{-1} \nabla^{C^{\delta,(\infty)}} \mathbf{l} \tag{45}$$

$$\nabla^{m^{\boldsymbol{X}}} \mathbf{l} = \Gamma^T \nabla^{C^{\delta,(\infty)}} \mathbf{l} = (1-\delta) F^T (K^T)^{-1} \nabla^{C^{\delta,(\infty)}} \mathbf{l} \tag{46}$$

$$\nabla^{m^{\boldsymbol{Y}}} \mathbf{l} = \Theta^T \nabla^{C^{\delta,(\infty)}} \mathbf{l} = (1-\delta) G^T (K^T)^{-1} \nabla^{C^{\delta,(\infty)}} \mathbf{l} \tag{47}$$

Thus, we save some GPU power by applying above formulae, and computing $(K^T)^{-1} \nabla^{C^{\delta,(\infty)}} \mathbf{l}$ only once. Note also that $(K^T)^{-1} \nabla^{C^{\delta,(\infty)}} \mathbf{l}$ can be computed with linear equations solving primitives instead of matrix inversion, for more efficiency and stability.

Algorithm 3: Gradient computation for the depth-∞ discounted WL distance

Input: $\nabla^{C^{\delta,(\infty)}} \mathbf{l}$: [n, m] float array, value of the gradient $(\frac{\partial \mathbf{l}}{\partial C_{ij}^{\delta,(\infty)}})_{0 \leq i < n, 0 \leq j < m}$;

Output: $\nabla^{m^{\boldsymbol{X}}} \mathbf{l}$ [n, n] float array, value of the gradient $(\frac{\partial \mathbf{l}}{\partial m_{ij}^X})_{0 \leq i < n, 0 \leq j < n}$;

$\nabla^{m^{\boldsymbol{Y}}} \mathbf{l}$ [m, m] float array, value of the gradient $(\frac{\partial \mathbf{l}}{\partial m_{ij}^Y})_{0 \leq i < m, 0 \leq j < m}$;

$\nabla^C \mathbf{l}$: [n, m] float array, value of the gradient $(\frac{\partial \mathbf{l}}{\partial C})_{0 \leq i < n, 0 \leq j < m}$;

restore saved variables f, g, P *stored in Algorithm 2 or Algorithm 4*;

Compute $P := \left(P_{ij}^{kl}\right)_{\substack{0 \leq k < n, 0 \leq l < m \\ 0 \leq i < n, 0 \leq j < m}}$;

$F := \left(F_{ij}^{kk'}\right)_{\substack{0 \leq k < n, 0 \leq k' < n \\ 0 \leq i < n, 0 \leq j < m}} = \left(f_{ij}^{k'} \mathbb{1}_{i=k}\right)_{\substack{0 \leq k < n, 0 \leq k' < m \\ 0 \leq i < n, 0 \leq j < m}}$;

$G := \left(G_{ij}^{ll'}\right)_{\substack{0 \leq l < m, 0 \leq l' < m \\ 0 \leq i < n, 0 \leq j < m}} = \left(g_{ij}^{l'} \mathbb{1}_{i=l}\right)_{\substack{0 \leq l < m, 0 \leq l' < m \\ 0 \leq i < n, 0 \leq j < m}}$;

$(\nabla^{C^{\delta,(\infty)}} \mathbf{l})$.reshape(n m) ;

F.reshape(n^2, nm) ;

G.reshape(m^2, nm) ;

P.reshape(nm, nm) ;

$K := I_{nm} - (1-\delta)P$;

$L := \left(K^T\right)^{-1} \nabla^{C^{\delta,(\infty)}} \mathbf{l}$;

$\nabla^C \mathbf{l} = \delta L$;

$\nabla^{m^{\boldsymbol{X}}} \mathbf{l} = (1-\delta) F L$;

$\nabla^{m^{\boldsymbol{Y}}} \mathbf{l} = (1-\delta) G L$;

$\nabla^C \mathbf{l}$.reshape(n, m);

$\nabla^{m^{\boldsymbol{X}}} \mathbf{l}$.reshape(n, n);

$\nabla^{m^{\boldsymbol{Y}}} \mathbf{l}$.reshape(m, m);

return $\nabla^C \mathbf{l}, \nabla^{m^{\boldsymbol{X}}} \mathbf{l}, \nabla^{m^{\boldsymbol{Y}}} \mathbf{l}$,

---

Backward pass complexity The matrix K has size $nm \times nm$. The matrix $\nabla^{C^{\delta,(\infty)}} \mathbf{l}$, is viewed as a vector of size nm. Computing $L := \left(K^T\right)^{-1} \nabla^{C^{\delta,(\infty)}} \mathbf{l}$ thus has complexity $C_{\text{solve}}(nm)$ where $C_{\text{solve}}(k)$ is the complexity of the linear solver for k equations with k unknowns. If the solver used is LAPACKAnderson et al. (1999), $C(k) = O(k^3)$. Theoretically the complexity is lower: from

Bunch and Hopcroft (1974), we know that the complexity of solving this equation is the same as the complexity of a multiplication, which is theoretically $C(k) = O(k^\omega)$, where $\omega < 2.371866$ (Duan et al., 2022). So in theory the complexity for this step is $O((nm)^\omega)$ but in practice solvers will have $O((nm)^3)$. Then the subsequent matrix multiplications have at most the same complexity (the complexity of matrix multiplications since F and G are smaller than $nm \times nm$)

Thus the total complexity of the backward pass is $O((nm)^\omega)$ theoretically but $O((nm)^3)$ in practice.

Acceleration using Sinkhorn scheduling The procedure in Algorithm 2 can in fact be accelerated using a trick related to Sinkhorn distances and fixed point algorithms. Since $C^{\delta,(\infty)}$ can be defined as the unique fixed point of Equation (9) (from Proposition 16), we are guaranteed to reach the right result as long as that result is a fixed point of Equation (9). In particular, it does not matter if some steps are approximative during the algorithm as long as the convergence is reached. Given this observation, we can accept some intermediate steps to be approximate. Thus, to accelerate the algorithm, we can replace the first steps with a good and faster approximation of the right iteration. In practice one such way (mentioned, for example, by Peyré et al. (2019)) is to cap the number of iterations in the Sinkhorn computation, and to use the result even if it has not fully converged. In this way we obtain Algorithm 4, which is empirically faster than Algorithm 2.

Initialization In the same vein, as long as we are computing the depth-∞ distance, the initialization of the matrix C_{current} does not matter. We let C_0 denote the initial value of that variable. Empirically we find that Algorithm 2 and Algorithm 4 converge the fastest when we select $C_0 = \delta C$, (compared to $C_0 = 0$ or $C_0 = C$ which are other sensible choices). This choice of initialization relates to the procedure one obtains if instead of computing $d^{(k)}_{\mathrm{WL},\delta}$, i.e., the OTM distance related to p^k_δ as defined before Definition 8, we compute the distance related to p'^k_δ where $p'^k_\delta(t) = P(T^\infty_\delta = t | T^\infty_\delta < k)$ where $T^\infty_\delta \sim \mathcal{G}(\delta)$.

D.2. Acceleration for Sparse Markov Chains

Sometimes, the Markov chains encountered are sparse, i.e., the transition kernel matrices of Markov chains are sparse. Whenever this happens, we exploit this to develop algorithms for faster computation of the discounted WL distance.

Let $\alpha \in \mathcal{P}(\boldsymbol{X})$. Then, we let $\operatorname{supp}\alpha := \{x \in \boldsymbol{X}, \alpha_x > 0\}$ denote the support of α. Given a Markov chain $\mathcal{X} = (\boldsymbol{X}, m^{\boldsymbol{X}}_\bullet, \nu^{\boldsymbol{X}})$, we define $\operatorname{supp}_{\mathcal{X}} x := \operatorname{supp} m^{\boldsymbol{X}}_x$ for each $x \in \boldsymbol{X}$. We further let $\deg_{\mathcal{X}} x := |\operatorname{supp}_{\mathcal{X}} x|$ and let $d_{\mathcal{X}} := \max_{x \in \boldsymbol{X}} \deg_{\mathcal{X}} x$.

In this section, we propose a modified version of Algorithm 1, that can compute $d^{(k)}_{\mathrm{WL},\delta}(\mathcal{X}, \mathcal{Y})$ in time $O(kl_s nm d_{\mathcal{X}} d_{\mathcal{Y}})$ (which is a performance boost when $d_{\mathcal{X}} << n$ or $d_{\mathcal{Y}} << m$), where l_s is the number of iterations needed for Sinkhorn to converge.

This accelerated version of the algorithm is based on the following observation: if α (resp. β) are probability measures on $\boldsymbol{X}$ (resp. $\boldsymbol{Y}$)

$$d_{\mathrm{W}}(\alpha, \beta; C) = d_{\mathrm{W}}\left(\alpha_{|\operatorname{supp}\alpha}, \beta_{|\operatorname{supp}\beta}; C_{|\operatorname{supp}\alpha \times \operatorname{supp}\beta}\right) \tag{48}$$

where $\alpha_{|\operatorname{supp}\alpha}$ (resp. $\beta_{|\operatorname{supp}\beta}$) denotes the distribution induced by α on its support, and $C_{|\operatorname{supp}\alpha \times \operatorname{supp}\beta}$ denotes the restriction of C to $\operatorname{supp}\alpha \times \operatorname{supp}\beta$. And the same holds true for the Sinkhorn distance:

$$d^\epsilon_{\mathrm{W}}(\alpha, \beta; C) = d^\epsilon_{\mathrm{W}}\left(\alpha_{|\operatorname{supp}\alpha}, \beta_{|\operatorname{supp}\beta}; C_{|\operatorname{supp}\alpha \times \operatorname{supp}\beta}\right) \tag{49}$$

Algorithm 4: Computation of depth-∞ discounted WL distance with Sinkhorn scheduling

Input: $m^{\boldsymbol{X}}$: [n, n] float array ;
$m^{\boldsymbol{Y}}$: [m, m] float array ;
$\nu^{\boldsymbol{X}}$: [n] float array ;
$\nu^{\boldsymbol{Y}}$: [m] float array ;
C: [n, m] float array ;
δ: The discount parameter for the WL distance.;
ϵ: Parameter for the Sinkhorn divergence;
sinkhorn_update_size: integer ;
Output: $d_{\mathrm{WL},\epsilon,\delta}^{(\infty)}\big((\boldsymbol{X}, m_{\bullet}^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m_{\bullet}^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}}); C\big)$;
Stores: f: [n, m, n], g: [n, m, m] float arrays: dual solutions of the last step Sinkhorn computations;
P [n, m, n, m] float array :primal solutions (i.e., optimal matchings) for the last step Sinkhorn computation;

$C_{\text{current}} = C$;
$n_{\text{sinkhorn}} = 1$;
repeat
  $C_{\text{new}} = (0)_{0\le i<n, 0\le j<m}$;
  $f = (0)_{0\le i<n, 0\le j<m, 0\le u<n}$;
  $g = (0)_{0\le i<n, 0\le j<m, 0\le l<m}$;
  $P = (0)_{0\le i<n, 0\le j<m, 0\le u<n, 0\le l<m}$;
  foreach $0 \le i < n, 0 \le j < m$ **do**
    $C_{\text{new}}[i,j] = \delta C[i,j] + (1-\delta) d_{\mathrm{W}}^{\epsilon}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C_{\text{current}}\right)$ limiting the number of iterations to n_{sinkhorn} ;
    $f[i,j], g[i,j], P[i,j] =$ Dual and primal solutions of the calculation above
  end
  $C_{\text{current}} = C_{\text{new}}$;
  if *one of the Sinkhorn computations did not converge* **then**
    $n_{\text{sinkhorn}} += $ sinkhorn_update_size ;
  end
until $C_{current}$ *has converged and all Sinkhorn computations have converged*;
Compute $d = d_{\mathrm{W}}^{\epsilon}\big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C_{\text{current}}\big)$;
save for backwards pass f, g, P ;
return d

Equation (48) and Equation (49) are direct consequences of the definitions of Wasserstein distances and Sinkhorn distances.

We use this simple observation to devise Algorithm 5. Now, the computation of $C_{l+1}[i,j] = d_{\mathrm{W}}^{\epsilon}\left(m^{\boldsymbol{X}}_{i|\operatorname{supp} m_i^{\boldsymbol{X}}}, m^{\boldsymbol{Y}}_{j|\operatorname{supp} m_j^{\boldsymbol{Y}}}; C_{l|\operatorname{supp} m_i^{\boldsymbol{X}} \times \operatorname{supp} m_j^{\boldsymbol{Y}}}\right)$ can be done in time only $O(l_s d_{\mathcal{X}} d_{\mathcal{Y}})$, which is a substantial acceleration compared to the original time complexity $O(l_s nm)$ when the Markov chains are of low degree.

Algorithm 5: Computation of depth-k discounted WL distance with sparse Markov kernels

Input: $m^{\boldsymbol{X}}$: [n, n] float array ;
$m^{\boldsymbol{Y}}$: [m, m] float array ;
$\nu^{\boldsymbol{X}}$: [n] float array ;
$\nu^{\boldsymbol{Y}}$: [m] float array ;
C: [n, m] float array ;
δ: The discount parameter for the WL distance.;
ϵ: Parameter for the Sinkhorn divergence;
k: depth
Output: $d^{(k)}_{\mathrm{WL},\epsilon,\delta}\big((\boldsymbol{X}, m^{\boldsymbol{X}}, \nu^{\boldsymbol{X}}), (\boldsymbol{Y}, m^{\boldsymbol{Y}}, \nu^{\boldsymbol{Y}}); C\big)$;

1 $C_{\text{current}} = C$;
2 **foreach** $0 \leq l < k$ **do**
3 $\quad C_{\text{new}} = (0)_{0\leq i<n, 0\leq j<m}$;
4 $\quad$ **foreach** $0 \leq i < n, 0 \leq j < m$ **do**
5 $\quad\quad C_{\text{new}}[i,j] = \delta C[i,j] + (1-\delta) d^{\epsilon}_{\mathrm{W}}\left(m^{\boldsymbol{X}}_{i|\operatorname{supp} m^{\boldsymbol{X}}_i}, m^{\boldsymbol{Y}}_{j|\operatorname{supp} m^{\boldsymbol{Y}}_j}, ; C_{\text{current}|\operatorname{supp} m^{\boldsymbol{X}}_i \times \operatorname{supp} m^{\boldsymbol{Y}}_j} \right)$;
6 $\quad$ **end**
7 $\quad C_{\text{current}} = C_{\text{new}}$
8 **end**
9 Compute $d = d^{\epsilon}_{\mathrm{W}}\big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C_{\text{current}}\big)$;
10 **return** d

We could also continue until convergence in a similar way as Algorithm 2. But the dual solutions of the computation at line 5 cannot be directly used as the gradient for the whole distributions, thus we need to extend them using the technique from (Feydy et al., 2019, Proposition 2 and comments below). That operation requires recomputing one iteration of Sinkhorn with the full (non-restricted) distributions, which is $O(nm)$. Moreover, this needs to be done nm times, ending in a $O((nm)^2)$ complexity. Additionally, the matrix inversions we do in Algorithm 3 are also $O((nm)^2)$. Thus, in this case, we cannot accelerate the backwards pass.

Appendix E. Proofs and Technical Details

E.1. Preliminary Results

In this section, we will present some lemmas that will be useful in subsequent proofs.

Let us start with a well-known alternative formulation for optimal transport as a linear programming problem, in the finite case, that allows us to justify taking optimal matchings.

Lemma 27 (Equation (2.11) in Peyré et al. (2019)) *Since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two finite sets, without loss of generality, we identify them to $\{1 \dots n\}$ and $\{1 \dots m\}$, respectively. Then, the cost function $C : \{1 \dots n\} \times \{1 \dots m\} \to \mathbb{R}_+$ can be seen as a matrix in $\mathbb{R}_+^{n\times m}$. Then, the Wasserstein distance can be expressed as the following linear program:*

$$d_{\mathrm{W}}(\alpha, \beta; C) = \min_{P} \langle P, C \rangle \tag{50}$$

where the minimum is taken over the (compact) subspace $[0,1]^{n\times m}$ *in which each* P *satisfies that* $\sum_{i=1}^{n} P_{iy} = \beta(y)$ *and* $\sum_{j=1}^{n} P_{xj} = \alpha(x)$ *for any* $x = 1,\dots,n$ *and* $y = 1,\dots,m$*, i.e., the compact space of distributions whose marginals are* α *and* β*. See (Peyré et al., 2019) for more background on these notions.*

In particular, there always exists a coupling (X, Y) *that verifies the infimum in Equation (1).*

Lemma 28 (Optimal transport is 1-lipschitzian in the cost matrix) *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be finite sets such that* $n := |\boldsymbol{X}|$ *and* $m = |\boldsymbol{Y}|$*. Let* $C_1, C_2 \in \mathbb{R}_+^{n\times m}$ *denote two cost matrices between* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$*, and let* $\alpha \in \mathcal{P}(\boldsymbol{X}), \beta \in \mathcal{P}(\boldsymbol{Y})$*. Then, for any* $\epsilon \geq 0$*, one has that*

$$|d_{\mathrm{W}}^{\epsilon}(\alpha,\beta;C_1) - d_{\mathrm{W}}^{\epsilon}(\alpha,\beta;C_2)| \leq \|C_1 - C_2\|_{\infty}. \tag{51}$$

Proof [proof of Lemma 28] Without loss of generality, we assume that $d_{\mathrm{W}}^{\epsilon}(\alpha,\beta;C_1) \geq d_{\mathrm{W}}^{\epsilon}(\alpha,\beta;C_2)$. Let $(X^{\epsilon}, Y^{\epsilon})$ be an optimal coupling for $d_{\mathrm{W}}^{\epsilon}(\alpha,\beta;C_2)$. Then, we have that

$$\begin{aligned}
|d_{\mathrm{W}}^{\epsilon}(\alpha,\beta;C_1) - d_{\mathrm{W}}^{\epsilon}(\alpha,\beta;C_2)| &\leq \mathbb{E}(C_1(X^{\epsilon},Y^{\epsilon}) - \epsilon H(X^{\epsilon},Y^{\epsilon}) - \mathbb{E}C_2(X^{\epsilon},Y^{\epsilon}) + \epsilon H(X^{\epsilon},Y^{\epsilon})) \\
&\leq \mathbb{E}|C_1(X^{\epsilon},Y^{\epsilon}) - C_2(X^{\epsilon},Y^{\epsilon})| \\
&\leq \|C_1 - C_2\|_{\infty}.
\end{aligned}$$

This concludes the proof. ■

Lemma 29 (Continuity of Banach fixed point (Corollary 1.4 in Pata et al. (2019))) *Let* Z *and* Λ *be metric spaces and assume that* Z *is complete. Let* $F : Z \times \Lambda \to Z$ *be a continuous map. Assume that there exists* $\alpha \in [0,1)$ *such that for each* $\lambda \in \Lambda$*,* $F_\lambda : Z \to Z$ *is* α*-Lipschitz. Then, for each* $\lambda \in \Lambda$*,* F_λ *has a unique fixed point* $z(\lambda)$ *and the map* $\lambda \mapsto z(\lambda)$ *is continuous.*

Lemma 30 (Differentiability of Banach fixed point) *Let* Z *and* Λ *be differential manifolds, and assume that* Z *is complete. Let* $F : Z \times \Lambda \to Z$ *be a* $\mathcal{C}^1$ *map. Denote* $\mathrm{d}_{|z}$ *its differential along* Z *and* $\mathrm{d}_{|\lambda}$ *its differential along* Λ *Moreover, suppose that for some* λ*,* $id_Z - \mathrm{d}|zF_{|(z(\lambda),\lambda)}$ *is invertible. Then the fixed point* $z(\lambda)$ *exists for any* $\lambda \in \Lambda$ *and is differentiable in* λ *and more explicitly*

$$dz = (id - \mathrm{d}_{|z}F)^{-1}\,\mathrm{d}_{|\lambda}F \tag{52}$$

Proof Without loss of generality, we prove the result when $Z \subseteq \mathbb{R}^n$ and $\Lambda \subseteq \mathbb{R}^m$ are open subsets in Euclidean spaces and the general result follows from taking charts in manifolds. We define $G : Z \times \Lambda \to \mathbb{R}^n$ by letting $G(z,\lambda) := F(z,\lambda) - z$ for any $z \in Z$ and $\lambda \in \Lambda$. Then, G is also continuously differentiable. The differential (or Jacobian) of G w.r.t. z is computed as $\mathrm{d}_{|z}G = \mathrm{d}_{|z}F - \mathrm{id}$. By assumption, we know that $\mathrm{d}_{|z}G$ is invertible and hence the implicit function theorem applies: there exists a unique differentiable function $z : U \to Z$ defined on a neighborhood of λ such that $G(z(\lambda),\lambda) = 0$ and $dz = -\,\mathrm{d}_{|z}G^{-1}\,\mathrm{d}_{|\lambda}G$. This means that $F(z(\lambda),\lambda) = z(\lambda)$ and

$$dz = (\mathrm{id} - \mathrm{d}_{|z}F)^{-1}\,\mathrm{d}_{|\lambda}F, \tag{53}$$

which concludes the proof. ■

We also provide the following gluing lemma for Markovian couplings, useful for the study of OTM distances:

Lemma 31 (Gluing lemma for Markovian couplings) *Let $(X_t, Z_t^1)_{t\in\mathbb{N}} \in \Pi(\mathcal{X}, \mathcal{Z})$ and $(Z_t^2, Y_t)_{t\in\mathbb{N}} \in \Pi(\mathcal{Z}, \mathcal{Y})$ be Markovian couplings. Then, there exists a (time inhomogeneous) Markov chain on $\boldsymbol{X} \times \boldsymbol{Z} \times \boldsymbol{Y}$*

$$(X'_t, Z'_t, Y'_t)_{t\in\mathbb{N}}$$

so that $(X'_t, Z'_t)_{t\in\mathbb{N}} \sim (X_t, Z_t^1)_{t\in\mathbb{N}}$, $(Z'_t, Y'_t)_{t\in\mathbb{N}} \sim (Z_t^2, Y_t)_{t\in\mathbb{N}}$ and furthermore $(X'_t, Y'_t)_{t\in\mathbb{N}}$ is a Markovian coupling between $\mathcal{X}$ and $\mathcal{Y}$.

Proof We let $\nu^{\boldsymbol{XZ}} := \mathrm{law}((X_0, Z_0^1))$ and $\nu^{\boldsymbol{ZY}} := \mathrm{law}((Z_0^2, Y_0))$. For any $t \in \mathbb{N}$, let $m_{xz}^{\boldsymbol{XZ},(t)} := \mathbb{P}((X_{t+1}, Z_{t+1}^1)|(X_t, Z_t^1) = (x, z)) \in \mathcal{C}(m_x^{\boldsymbol{X}}, m_z^{\boldsymbol{Z}})$ for any $x \in \boldsymbol{X}$ and $z \in \boldsymbol{Z}$. Similarly, let $m_{zy}^{\boldsymbol{ZY},(t)} := \mathbb{P}((Z_{t+1}^2, Y_{t+1})|(Z_t^2, Y_t) = (z, y)) \in \mathcal{C}(m_z^{\boldsymbol{Z}}, m_y^{\boldsymbol{Y}})$ for any $z \in \boldsymbol{Z}$ and $y \in \boldsymbol{Y}$.

By the Gluing Lemma (Villani et al., 2009) for probability measures, one has that

- there exists $\nu^{\boldsymbol{XZY}} \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Z} \times \boldsymbol{Y})$ whose marginals on $\boldsymbol{X} \times \boldsymbol{Z}$ and on $\boldsymbol{Z} \times \boldsymbol{Y}$ coincide with $\nu^{\boldsymbol{XZ}}$ and $\nu^{\boldsymbol{ZY}}$, respectively, and furthermore, the marginal on $\boldsymbol{X} \times \boldsymbol{Y}$, denoted by $\nu^{\boldsymbol{XY}}$, is a coupling between $\nu^{\boldsymbol{X}}$ and $\nu^{\boldsymbol{Y}}$;
- for any $x \in \boldsymbol{X}, y \in \boldsymbol{Y}$ and $z \in \boldsymbol{Z}$, there exists $m_{xzy}^{\boldsymbol{XZY},(t)} \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Z} \times \boldsymbol{Y})$ whose marginals on $\boldsymbol{X} \times \boldsymbol{Z}$ and on $\boldsymbol{Z} \times \boldsymbol{Y}$ coincide with $m_{xz}^{\boldsymbol{XZ},(t)}$ and $m_{zy}^{\boldsymbol{ZY},(t)}$, respectively, and furthermore, the marginal on $\boldsymbol{X} \times \boldsymbol{Y}$, denoted by $\nu^{\boldsymbol{XY}}$, is a coupling between $m_x^{\boldsymbol{X}}$ and $m_y^{\boldsymbol{Y}}$.

By the Kolmogorov extension theorem (Kolmogorov and Bharucha-Reid, 2018), there exists a Markov chain $(X'_t, Z'_t, Y'_t)_{t\in\mathbb{N}}$ with initial distribution $\nu^{\boldsymbol{XZY}}$ and transition kernels at each step $t \in \mathbb{N}$ defined by: $\mathbb{P}((X'_{t+1}, Z'_{t+1}, Y'_{t+1})|(X'_t, Z'_t, Y'_t) = (x, z, y)) := m_{xzy}^{\boldsymbol{XZY},(t)}$ for any $x \in \boldsymbol{X}, y \in \boldsymbol{Y}$ and $z \in \boldsymbol{Z}$. By construction, one obviously has that $(X'_t, Z'_t)_{t\in\mathbb{N}} \sim (X_t, Z_t^1)_{t\in\mathbb{N}}$, $(Z'_t, Y'_t)_{t\in\mathbb{N}} \sim (Z_t^2, Y_t)_{t\in\mathbb{N}}$ and that $(X'_t, Y'_t)_{t\in\mathbb{N}}$ is a Markovian coupling between $\mathcal{X}$ and $\mathcal{Y}$. ■

We end this section with an alternative yet direct description for the expected value involved in the definition of the OTM distances.

Lemma 32 *Given any Markovian coupling $(X_t, Y_t)_{t\in\mathbb{N}}$ between $\mathcal{X}$ and $\mathcal{Y}$, let $\nu^{\boldsymbol{XY}} \in \mathcal{C}(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}})$ denote its initial distribution and let $(m_{\bullet\bullet}^{\boldsymbol{XY},(t)})_{t\in\mathbb{N}} \in \mathcal{C}(m_\bullet^{\boldsymbol{X}}, m_\bullet^{\boldsymbol{Y}})^{\mathbb{N}_+}$ denote its Markov transition kernels at each step $t \in \mathbb{N} := \{0, 1, 2, \dots\}$, where $\mathcal{C}(m_\bullet^{\boldsymbol{X}}, m_\bullet^{\boldsymbol{Y}})$ denotes the space of all Markov transition kernels $m_{\bullet\bullet}^{\boldsymbol{XY}}$ such that $m_{ij}^{\boldsymbol{XY}} \in \mathcal{C}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}})$ for all $i \in \boldsymbol{X}$ and $j \in \boldsymbol{Y}$. Then,*

$$\mathbb{E}\, C(X_T, Y_T) = \sum_{t=0}^{\infty} p(t) \sum_{i_0, j_0, \dots, i_t, j_t} C_{i_t, j_t} m_{i_{t-1}j_{t-1}, i_t j_t}^{\boldsymbol{XY},(t-1)} m_{i_{t-2}j_{t-2}, i_{t-1}j_{t-1}}^{\boldsymbol{XY},(t-2)} \cdots m_{i_0 j_0, i_1 j_1}^{\boldsymbol{XY},(0)} \nu_{i_0 j_0}^{\boldsymbol{XY}} \quad (54)$$

where $m_{ij,kl}^{\boldsymbol{XY},(t)} := m_{ij}^{\boldsymbol{XY},(t)}(k, l)$ is a shorthand for the transition probability.

Proof By properties of expected values, one has that for any give $t \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\, C(X_t, Y_t) &= \sum_{i_t, j_t} C_{i_t j_t} \mathbb{P}(X_t = i_t, Y_t = j_t) \\
&= \sum_{i_{t-1}, j_{t-1}, i_t, j_t} C_{i_t j_t} \mathbb{P}(X_t = i_t, Y_t = j_t | X_{t-1} = i_{t-1}, Y_{t-1} = j_{t-1}) \mathbb{P}(X_{t-1} = i_{t-1}, Y_{t-1} = j_{t-1}) \\
&= \sum_{i_{t-1}, j_{t-1}, i_t, j_t} C_{i_t j_t} m^{\boldsymbol{XY},(t-1)}_{i_{t-1} j_{t-1}, i_t j_t} \mathbb{P}(X_{t-1} = i_{t-1}, Y_{t-1} = j_{t-1}) \\
&= \sum_{i_0, j_0, \ldots, i_t, j_t} C_{i_t, j_t} m^{\boldsymbol{XY},(t-1)}_{i_{t-1} j_{t-1}, i_t j_t} m^{\boldsymbol{XY},(t-2)}_{i_{t-2} j_{t-2}, i_{t-1} j_{t-1}} \cdots m^{\boldsymbol{XY},(0)}_{i_0 j_0, i_1 j_1} \nu^{\boldsymbol{XY}}_{i_0 j_0}.
\end{aligned}$$

The last equality can be proved inductively. Since T is independent of $(X_t, Y_t)_{t \in \mathbb{N}}$, Equation (54) follows directly from the calculation above. ■

E.2. Proofs and Technical Details from Section 3

Proof [Proof of Remark 4] Similarly to Lemma 27 we remark that, under our assumption that the spaces $\boldsymbol{X}$ and $\boldsymbol{Y}$ are finite of sizes n and m respectively, we write them as $\{1 \ldots n\}$ and $\{1 \ldots m\}$. Given any Markovian coupling $(X_t, Y_t)_{t \in \mathbb{N}}$, by Lemma 32, we have that the value $\mathbb{E}\, C(X_T, Y_T)$ is completely determined by the initial distribution

$$\nu^{\boldsymbol{XY}} \in \mathcal{C}(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}})$$

and Markov transition kernels at each step $t \geq 0$ of $(X_t, Y_t)_{t \in \mathbb{N}}$:

$$(m^{\boldsymbol{XY},(t)}_{\bullet\bullet})_{t \in \mathbb{N}} \in \mathcal{C}(m^{\boldsymbol{X}}_{\bullet}, m^{\boldsymbol{Y}}_{\bullet})^{\mathbb{N}}$$

where $\mathcal{C}(m^{\boldsymbol{X}}_{\bullet}, m^{\boldsymbol{Y}}_{\bullet})$ denotes the space of all Markov transition kernels $m^{\boldsymbol{XY}}_{\bullet\bullet}$ such that $m^{\boldsymbol{XY}}_{ij} \in \mathcal{C}(m^{\boldsymbol{X}}_{i}, m^{\boldsymbol{Y}}_{j})$ for all $i \in \boldsymbol{X}$ and $j \in \boldsymbol{Y}$. More precisely,

$$\mathbb{E}\, C(X_T, Y_T) = \sum_{t=0}^{\infty} p(t) \sum_{i_0, j_0, \ldots, i_t, j_t} C_{i_t, j_t} m^{\boldsymbol{XY},(t-1)}_{i_{t-1} j_{t-1}, i_t j_t} m^{\boldsymbol{XY},(t-2)}_{i_{t-2} j_{t-2}, i_{t-1} j_{t-1}} \cdots m^{\boldsymbol{XY},(0)}_{i_0 j_0, i_1 j_1} \nu^{\boldsymbol{XY}}_{i_0 j_0}. \quad (55)$$

Note that for any $\alpha \in \mathcal{P}(\boldsymbol{X})$ and $\beta \in \mathcal{P}(\boldsymbol{Y})$, the set of all couplings $\mathcal{C}(\alpha, \beta)$ can be identified with a compact subset in $[0,1]^{n \times m}$ (see also Lemma 27). Then,

$$\mathcal{C}(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}) \times \mathcal{C}(m^{\boldsymbol{X}}_{\bullet}, m^{\boldsymbol{Y}}_{\bullet})^{\mathbb{N}_+} = \mathcal{C}(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}) \times (\Pi_{x \in \boldsymbol{X}, y \in \boldsymbol{Y}} \mathcal{C}(m^{\boldsymbol{X}}_{x}, m^{\boldsymbol{Y}}_{y}))^{\mathbb{N}_+}$$

is a countable Cartesian product of compact spaces and it is hence compact. Moreover, the right-hand side of Equation (55) is obviously a continuous function defined on $\mathcal{C}(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}) \times \mathcal{C}(m^{\boldsymbol{X}}_{\bullet}, m^{\boldsymbol{Y}}_{\bullet})^{\mathbb{N}_+}$. Therefore, the infimum of the right-hand side of Equation (54) is attainable in $\mathcal{C}(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}) \times \mathcal{C}(m^{\boldsymbol{X}}_{\bullet}, m^{\boldsymbol{Y}}_{\bullet})^{\mathbb{N}_+}$ and hence we conclude the proof. ■

Proof [Proof of Proposition 5] Let p distribution on $\mathbb{N}$,

$$\begin{aligned} d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C) &= \inf_{(X_t,Y_t)_{t\in\mathbb{N}}} \mathbb{E}_{T\sim p, T\perp\!\!\!\perp (X_t,Y_t)}\, C(X_T,Y_T), \\ &= \inf_{(X_t,Y_t)_{t\in\mathbb{N}}} \sum_{t\in\mathbb{N}} p(t) C(X_t,Y_t), \\ &\geq \sum_{k\in\mathbb{N}} p(t) \inf_{(X_t,Y_t)_{t\in\mathbb{N}}} C(X_k,Y_k), \\ &= \sum_{k\in\mathbb{N}} p(k) d^{(k)}_{\mathrm{WL}}(\mathcal{X},\mathcal{Y}). \\ &= \mathbb{E}_{T\sim p}(d^{(T)}_{\mathrm{WL}}(\mathcal{X},\mathcal{Y})) \end{aligned}$$

■

Proof [Proof of Proposition 6] Let p be a distribution on $\mathbb{N}$, and $\mathcal{X}$ and $\mathcal{Y}$ be stationary Markov chains.

Similarly to Remark 4, one can prove that there exists a stationary Markovian coupling that realizes Eq. 4: one can prove that the space of stationary Markovian couplings is compact (as a closed subset of the space of Markovian couplings) in the same sense as in the proof of Remark 4, and the same compactness argument gives the existence of the optimal coupling.

Consequently, let $(\overline{X}_t, \overline{Y}_t)$ be a realization of that coupling.

$$\begin{aligned} d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C) &= \inf_{(X_t,Y_t)_{t\in\mathbb{N}}} \mathbb{E}_{T\sim p, T\perp\!\!\!\perp (X_t,Y_t)}\, C(X_T,Y_T), \\ &\leq \mathbb{E}_{T\sim p, T\perp\!\!\!\perp (\overline{X}_t,\overline{Y}_t)}\, C(\overline{X}_T,\overline{Y}_T) \\ &= \mathbb{E}_{T\sim p, T\perp\!\!\!\perp (\overline{X}_t,\overline{Y}_t)}\, C(\overline{X}_0,\overline{Y}_0) \\ &= C(\overline{X}_0,\overline{Y}_0) \\ &= d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y};C) \end{aligned}$$

■

Proof [Proof of Proposition 7] We, in fact, prove that the following 4 statements are equivalent:

1. $d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y}) = 0$;
2. $d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y}) = 0$;
3. there exists a Markovian coupling $(X_t,Y_t)_{t\in\mathbb{N}} \in \Pi(\mathcal{X},\mathcal{Y})$ so that $\forall t \geq 0, C(X_t,Y_t) = 0$ almost surely;
4. for all distributions q over $\mathbb{N}$, $d^q_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y}) = 0$.

1 $\implies$ 2: This is a direct consequence of Proposition 6: if $d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y}) = 0$, then

$$0 \leq d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y}) \leq d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y}) = 0.$$

Thus, $d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y}) = 0$.

2 $\implies$ 3: Suppose $d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y})=0$. Then, by Remark 4 there exists a Markovian coupling $(X_t,Y_t)_{t\in\mathbb{N}}$ such that:

$$0=\mathbb{E}(C(X_T,Y_T)).$$

Then, suppose (by contradiction) that there is t_0 so that $p(t_0)C(X_{t_0},Y_{t_0})>s>0$ with positive probability $\alpha>0$. Then, $\mathbb{E}(C(X_T,Y_T))\geq\alpha\cdot s>0$. Contradiction.

Thus, $p(t)C(X_t,Y_t)=0$ almost surely for each $t\in\mathbb{N}$ and hence $C(X_t,Y_t)=0$ almost surely (since $p(t)>0$).

3 $\implies$ 4: This holds obviously.

4 $\implies$ 1: Assume 4. We know from Theorem 14 that

$$d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y})=\lim_{\delta\to 0} d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y})=\lim 0=0.$$

■

We finish this section with an interesting stability result of OTM distances with respect to the choice of the distribution $p\in\mathcal{P}(\mathbb{N})$, which will be useful for subsequent proofs.

Lemma 33 *Let $\{p_k\}_{k\in\mathbb{N}}\subseteq\mathcal{P}(\mathbb{N})$ be such that $\lim_{k\to\infty} d_{\mathrm{TV}}(p_k,p)=0$ where d_{TV} denotes the total variation distance. Then for all $\mathcal{X},\mathcal{Y}$, one has that $\lim_{k\to\infty} d^{p_k}_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y})=d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y})$.*

Proof [Proof of Lemma 33] Let $\epsilon>0$, choose N_0 so that there exists a sequence of random variables $T_k\sim p_k$ for any $k\geq N_0$ and $T\sim p$ such that $\mathbb{P}(T_k\neq T)\leq\epsilon$.

Then let $(X_t,Y_t)_{t\in\mathbb{N}}$ (resp. $(X^k_t,Y^k_t)_{t\in\mathbb{N}}$) be an optimal coupling independent of T (resp. independent of T_k) that verifies $d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C)=\mathbb{E}\,C(X_T,Y_T)$, (resp. $d^{p_k}_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C)=\mathbb{E}\,C(X^k_{T_k},Y^k_{T_k})$). Then, one has that

$$\begin{aligned}
d^p_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C)\leq&\mathbb{E}\,C(X^k_T,Y^k_T)\\
=&\mathbb{P}(X=X_k)\mathbb{E}(C(X^k_T,Y^k_T)\mid T=T_k)+\mathbb{P}(X\neq X_k)\mathbb{E}(C(X^k_T,Y^k_T)\mid T\neq T_k)\\
\leq&\mathbb{E}(C(X^k_T,Y^k_T)\mid T=T_k)+\epsilon\|C\|_\infty\\
=&\mathbb{E}(C(X^k_{T_k},Y^k_{T_k})\mid T=T_k)+\epsilon\|C\|_\infty\\
=&\frac{1}{1-\epsilon}\left(\mathbb{E}\,C(X^k_{T_k},Y^k_{T_k})-\mathbb{E}(C(X^k_{T_k},Y^k_{T_k})\mid T\neq T_k)\mathbb{P}(X\neq X_k)\right)+\epsilon\|C\|_\infty\\
\leq&\frac{1}{1-\epsilon}d^{p_k}_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C)+\epsilon\left(1+\frac{1}{1-\epsilon}\right)\|C\|_\infty\\
\leq&d^{p_k}_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C)+2\epsilon d^{p_k}_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C)+\epsilon(1+1+2\epsilon)\|C\|_\infty \text{ if } \epsilon \text{ is small enough}\\
\leq&d^{p_k}_{\mathrm{OTM}}(\mathcal{X},\mathcal{Y};C)+5\epsilon\|C\|_\infty.
\end{aligned}$$

This concludes the proof. ■

E.2.1. THE OTM DISTANCE IS A PSEUDOMETRIC

We first introduce some notation. Let $(\boldsymbol{X}, d_{\boldsymbol{X}})$ be a pseudometric space. We let $\mathcal{M}(\boldsymbol{X})$ denote the collection of all Markov chains $\mathcal{X} = (\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \nu^{\boldsymbol{X}})$ with state space $\boldsymbol{X}$. Then, d_{OTM}^p induces a map as follows

$$d_{\mathrm{OTM}}^p : \mathcal{M}(\boldsymbol{X}) \times \mathcal{M}(\boldsymbol{X}) \to \mathbb{R}_+$$

sending $(\mathcal{X}_1, \mathcal{X}_2)$ to $d_{\mathrm{OTM}}^p(\mathcal{X}_1, \mathcal{X}_2; d_{\boldsymbol{X}})$ with $d_{\boldsymbol{X}}$ being the cost function

Proposition 34 (OTM distances are metrics) *If $\boldsymbol{X} = \boldsymbol{Y}$ is a pseudometric space $(\boldsymbol{X}, d_{\boldsymbol{X}})$ and the cost C is the pseudometric distance on $\boldsymbol{X}$ (i.e., $C(x, y) = d_{\boldsymbol{X}}(x, y)$). For all $p \in \mathcal{P}(\mathbb{N})$, $d_{\mathrm{OTM}}^p : \mathcal{M}(\boldsymbol{X}) \times \mathcal{M}(\boldsymbol{X}) \to \mathbb{R}_+$ defines a pseudometric on $\mathcal{M}(\boldsymbol{X})$.*

When $d_{\boldsymbol{X}}$ is a metric and p is fully supported on $\mathbb{N}$, then $d_{\mathrm{OTM},p}$ is also a metric.

In practice the assumption that C is a pseudometric is respected for example in the framework of Chen et al. (2022), where the states have labels in a common metric space.

One can also derive a slightly more general result, to relate to p-Wassertein distances:

Proposition 35 *Let $\alpha \in [1, \infty)$. If $\boldsymbol{X} = \boldsymbol{Y}$ is a pseudometric space $(\boldsymbol{X}, d_{\boldsymbol{X}})$ and the cost C is defined as $C := d_{\boldsymbol{X}}^\alpha$, i.e., $C(x, y) = d_{\boldsymbol{X}}(x, y)^\alpha$ for any $x, y \in \boldsymbol{X}$.*

Then, for any $p \in \mathcal{P}(\mathbb{N})$, the map $h : \mathcal{M}(\boldsymbol{X}) \times \mathcal{M}(\boldsymbol{X}) \to \mathbb{R}_+$ sending $\mathcal{X}, \mathcal{Y} \in \mathcal{M}(\boldsymbol{X})$ to $\left(d_{\mathrm{OTM}}^p(\mathcal{X}, \mathcal{Y}; d_{\boldsymbol{X}}^\alpha)\right)^{\frac{1}{\alpha}}$ defines a pseudometric on $\mathcal{M}(\boldsymbol{X})$.

When $d_{\boldsymbol{X}}$ is a metric and p is fully supported on $\mathbb{N}$, then h is also a metric.

Proof [Proof of Propositions 34 and 35] Proposition 34 is a direct consequence of Proposition 35 by taking $\alpha = 1$. We thus only need to prove Proposition 35. Let h like in Proposition 35. We prove that h is a pseudometric on $\mathcal{M}(\boldsymbol{X})$ through the following three steps.

- Given any Markov chain $\mathcal{X}$, we let X be any realization of $\mathcal{X}$. Then, (X, X) is a Markovian coupling between $\mathcal{X}$ and itself. Hence,

$$0 \leq (d_{\mathrm{OTM}}^p(\mathcal{X}, \mathcal{X}; d_{\boldsymbol{X}}^\alpha))^{\frac{1}{\alpha}} \leq (\mathbb{E}\, d_{\boldsymbol{X}}^\alpha(X_T, X_T))^{\frac{1}{\alpha}} = 0$$

- **Symmetry**: Given any two Markov chain $\mathcal{X}$ and $\mathcal{Y}$ on $\boldsymbol{X}$, any Markovian coupling $(X_t, Y_t)_{t\in\mathbb{N}}$ between $\mathcal{X}$ and $\mathcal{Y}$ naturally (and bijectively) gives rise to a Markovian coupling $(Y_t, X_t)_{t\in\mathbb{N}}$ between $\mathcal{Y}$ and $\mathcal{X}$. Hence,

$$\begin{aligned} d_{\mathrm{OTM}}^p(\mathcal{X}, \mathcal{Y}; d_{\boldsymbol{X}}^\alpha) &= \inf_{(X_t, Y_t)_{t\in\mathbb{N}}} \mathbb{E}\, d_{\boldsymbol{X}}^\alpha(X_T, Y_T) \\ &= \inf_{(X_t, Y_t)_{t\in\mathbb{N}}} \mathbb{E}\, d_{\boldsymbol{X}}^\alpha(Y_T, X_T) \\ &= d_{\mathrm{OTM}}^p(\mathcal{Y}, \mathcal{X}; d_{\boldsymbol{X}}^\alpha) \\ h(\mathcal{X}, \mathcal{Y}) &= h(\mathcal{Y}, \mathcal{X}) \end{aligned}$$

- **Triangle inequality**: The proof of the triangle inequality is based on Theorem 31. It is similar to the proof of the triangle inequality in Villani et al. (2009, Definition 6.1).

Suppose $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ are three Markov chains on $\boldsymbol{X}$. Then,

$$\begin{aligned}
h(\mathcal{X},\mathcal{Y})^\alpha &= d_{\mathrm{OTM}}^p(\mathcal{X},\mathcal{Y};d_{\boldsymbol{X}}^\alpha)\\
&= \inf_{(X_t,Y_t)_{t\in\mathbb{N}}} \mathbb{E}\, d_{\boldsymbol{X}}^\alpha(X_T,Y_T)\\
&\leq \inf_{(X_t,Z_t,Y_t)_{t\in\mathbb{N}}} \mathbb{E}\, d_{\boldsymbol{X}}^\alpha(X_T,Y_T)\\
&\leq \inf_{(X_t,Z_t,Y_t)_{t\in\mathbb{N}}} \mathbb{E}((d_{\boldsymbol{X}}(X_T,Z_T)+d_{\boldsymbol{X}}(Z_T,Y_T))^\alpha)\\
&\leq \inf_{(X_t,Z_t,Y_t)_{t\in\mathbb{N}}} \left((\mathbb{E}\, d_{\boldsymbol{X}}^\alpha(X_T,Z_T))^{\frac{1}{\alpha}} + (\mathbb{E}\, d_{\boldsymbol{X}}^\alpha(Z_T,Y_T))^{\frac{1}{\alpha}}\right)^\alpha \text{ (Minkowski)}
\end{aligned}$$

where we are infimizing over all Markov chains $(X_t,Z_t,Y_t)_{t\in\mathbb{N}}$ whose marginals $(X_t,Z_t)_{t\in\mathbb{N}}$, $(Z_t,Y_t)_{t\in\mathbb{N}}$ and $(X_t,Y_t)_{t\in\mathbb{N}}$ are Markovian couplings.

Now, by Lemma 31 and Theorem 4 we can take (X'_t,Y'_t,Z'_t) a Markov chain so that

$$\begin{aligned}
h(\mathcal{X},\mathcal{Z}) &= (\mathbb{E}d_{\boldsymbol{X}}^\alpha(X'_T,Z'_T))^{\frac{1}{\alpha}}\\
h(\mathcal{Z},\mathcal{Y}) &= (\mathbb{E}d_{\boldsymbol{X}}^\alpha(Z'_T,Y'_T))^{\frac{1}{\alpha}}.
\end{aligned}$$

Hence

$$\begin{aligned}
h(\mathcal{X},\mathcal{Y})^\alpha &\leq \inf_{(X_t,Z_t,Y_t)_{t\in\mathbb{N}}} \left((\mathbb{E}d_{\boldsymbol{X}}^\alpha(X_T,Z_T))^{\frac{1}{\alpha}} + (\mathbb{E}d_{\boldsymbol{X}}^\alpha(Z_T,Y_T))^{\frac{1}{\alpha}}\right)^\alpha\\
&\leq \left((\mathbb{E}d_{\boldsymbol{X}}^\alpha(X'_T,Z'_T))^{\frac{1}{\alpha}} + (\mathbb{E}d_{\boldsymbol{X}}^\alpha(Z'_T,Y'_T))^{\frac{1}{\alpha}}\right)^\alpha\\
&= (h(\mathcal{X},\mathcal{Z})+h(\mathcal{Z},\mathcal{Y}))^\alpha\\
h(\mathcal{X},\mathcal{Y}) &\leq h(\mathcal{X},\mathcal{Z})+h(\mathcal{Z},\mathcal{Y}).
\end{aligned}$$

Now, for the second part of the statement, assume that p has full support on $\mathbb{N}$ and $d_{\boldsymbol{X}}$ is a metric. Assume further that $d_{\mathrm{OTM}}^p(\mathcal{X},\mathcal{Y};d_{\boldsymbol{X}}^\alpha)=0$. Let $(X_t,Y_t)_{t\in\mathbb{N}}$ be an optimal Markovian coupling ((cf. Remark 4) and let $T\sim p$ be an independent random variable. Then,

$$\mathbb{E}(d_{\boldsymbol{X}}^\alpha(X_T,Y_T)) = \sum_{t=0}^{\infty} p(t)\mathbb{E}(d_{\boldsymbol{X}}^\alpha(X_t,Y_t)) = 0.$$

This implies that $\mathbb{E}(d_{\boldsymbol{X}}^\alpha(X_t,Y_t))=0$ for all $t\in\mathbb{N}$. Since $d_{\boldsymbol{X}}$ is a metric, we have that $X_t=Y_t$ for all $t\in\mathbb{N}$ and this means that $\mathcal{X}$ and $\mathcal{Y}$ are isomorphic to each other. ■

E.3. Proofs and Technical Details from Section 4

Proof [Proof of Proposition 11] Notice that $\lim_{k\to\infty} d_{\mathrm{TV}}(p_\delta^k, p_\delta^\infty)=0$. Then, the proof follows from Lemma 33. ■

Proof [Proof of Proposition 15] We prove by induction the property that,

$$C_{ij}^{\delta,(k)} = \inf_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi((\boldsymbol{X},m_\bullet^{\boldsymbol{X}},\delta_i),(\boldsymbol{Y},m_\bullet^{\boldsymbol{Y}},\delta_j))} \mathbb{E}\left(\sum_{t=0}^{k-1}\delta(1-\delta)^t C(X_t,Y_t) + (1-\delta)^k C(X_k,Y_k)\right) \tag{56}$$

and furthermore, there exists an optimal Markovian coupling for each i, j that shares the same transition kernels $(m_{\bullet\bullet}^{\boldsymbol{XY},(t),k})_{t\geq 0}$, where t denotes the step number.

The case when $k = 0$ is trivial as by definition $C^{\delta,(0)} = C$. Now, suppose that Equation (56) holds true for some $k \in \mathbb{N}$. Then, let us prove that the equation holds for $k + 1$.

By definition, we have that $C_{ij}^{\delta,(k+1)} = \delta C_{ij} + (1-\delta)\, d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\delta,(k)}\right)$. We let $m_{ij}^{\boldsymbol{XY},(0),k+1} \in \mathcal{C}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}})$ be an optimal coupling for each $i \in \boldsymbol{X}, j \in \boldsymbol{Y}$. Then, we expand upon $m_{\bullet\bullet}^{\boldsymbol{XY},(0),k+1}$ to define a new set of transition kernels $(m_{\bullet\bullet}^{\boldsymbol{XY},(t),k+1})_{t\geq 0}$ such that $m_{\bullet\bullet}^{\boldsymbol{XY},(t),k+1} := m_{\bullet\bullet}^{\boldsymbol{XY},(t-1),k}$ for all $t > 0$. Now, for any i, j, let $(X_t, Y_t)_{t\in\mathbb{N}} \in \Pi((\boldsymbol{X}, m_\bullet^{\boldsymbol{X}}, \delta_i), (\boldsymbol{Y}, m_\bullet^{\boldsymbol{Y}}, \delta_j))$ be a Markovian coupling with the prescribed transition kernels $(m_{\bullet\bullet}^{\boldsymbol{XY},(t),k+1})_{t\geq 0}$. Then,

$$\begin{aligned}
&C_{ij}^{\delta,(k+1)} \\
&= \delta C_{ij} + (1-\delta)\, d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\delta,(k)}\right) \\
&= \delta C(X_0,Y_0) + (1-\delta)\mathbb{E}\left(C^{\delta,(k)}(X_1,Y_1)\right) \\
&= \delta C(X_0,Y_0) + (1-\delta)\mathbb{E}\left(\mathbb{E}\left(C_{i'j'}^{\delta,(k)} \mid (X_1,Y_1) = (i',j')\right)\right) \\
&= \delta C(X_0,Y_0) + \\
&\quad (1-\delta)\mathbb{E}\left(\mathbb{E}\left(\mathbb{E}\left(\delta C_{i'j'} + \sum_{t=1}^{k-1}\delta(1-\delta)^t C(X_{t+1},Y_{t+1}) + (1-\delta)^k C(X_{k+1},Y_{k+1})\right) \mid (X_1,Y_1) = (i',j')\right)\right) \\
&= \delta C(X_0,Y_0) + \\
&\quad (1-\delta)\mathbb{E}\left(\delta C(X_1,Y_1) + \sum_{t=1}^{k-1}\delta(1-\delta)^t C(X_{t+1},Y_{t+1}) + (1-\delta)^k C(X_{k+1},Y_{k+1})\right) \\
&= \mathbb{E}\left(\sum_{t=0}^{k}\delta(1-\delta)^t C(X_t,Y_t) + (1-\delta)^{k+1} C(X_{k+1},Y_{k+1})\right).
\end{aligned}$$

This concludes the induction step.

Hence, for all $k \in \mathbb{N}$ we have that

$$C_{ij}^{\delta,(k)} = \inf_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi((\boldsymbol{X},m_\bullet^{\boldsymbol{X}},\delta_i),(\boldsymbol{Y},m_\bullet^{\boldsymbol{Y}},\delta_j))} \mathbb{E}C(X_{T_k^\delta}, Y_{T_k^\delta}). \tag{57}$$

In this way, we have that

$$\begin{aligned} d_{\mathrm{W}}\Big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(k)}\Big) &= \inf_{X_0\sim\nu^{\boldsymbol{X}}, Y_0\sim\nu^{\boldsymbol{Y}}} \mathbb{E}\Big(C^{\delta,(k)}(X_0, Y_0)\Big) \\ &= \inf_{X_0\sim\nu^{\boldsymbol{X}}, Y_0\sim\nu^{\boldsymbol{Y}}} \mathbb{E}\Big(\mathbb{E}\Big(C_{ij}^{\delta,(k)} \mid (X_0, Y_0) = (i,j)\Big)\Big) \\ &= \inf_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi((\boldsymbol{X},m_\bullet^{\boldsymbol{X}},\nu^{\boldsymbol{X}}),(\boldsymbol{Y},m_\bullet^{\boldsymbol{Y}},\nu^{\boldsymbol{Y}}))} \mathbb{E}C(X_{T_k^\delta}, Y_{T_k^\delta}) \\ &= d_{\mathrm{WL},\delta}^{(k)}(\mathcal{X}, \mathcal{Y}). \end{aligned}$$

This concludes the proof. ■

Proof [Proof of Proposition 16] This is actually related to general convergence results on finite discounted MDPs. However, as our action space is infinite (see Section 4.4 for the description of the relevant MDPs), we will provide a complete proof here. We also note that we use arguments similar to the one below for proving other results such as Theorem 17.

When $\delta > 0$, the convergence of $C^{\delta,(k)}$ follows from the Banach fixed point theorem (see for example Lemma 29). Let us define $F : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ by sending $D \in \mathbb{R}^{n\times m}$ to $D' \in \mathbb{R}^{n\times m}$ such that for any i, j:

$$D'_{ij} := \delta C_{ij} + (1-\delta) d_{\mathrm{W}}\big(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{X}}; D\big).$$

By Lemma 28 one has that F is a $(1-\delta)$-Lipschitz function when considering ∞-norm on $\mathbb{R}^{n\times m}$. Since $1-\delta < 1$ whenever $\delta > 0$, the Banach fixed point theorem applies. By definition, $C^{\delta,(k+1)} = F\left(C^{\delta,(k)}\right)$. Hence, using the Banach fixed point theorem, we conclude that $C^{\delta,(k)}$ converges to the unique fixed point $C^{\delta,(\infty)}$ of F.

Then, by Lemma 28 again, one has that

$$\lim_{k\to\infty} d_{\mathrm{W}}\Big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(k)}\Big) = d_{\mathrm{W}}\Big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(\infty)}\Big).$$

Then, by Proposition 11 and Proposition 15 one has that

$$d_{\mathrm{WL},\delta}^{(\infty)}(\mathcal{X}, \mathcal{Y}) = d_{\mathrm{W}}\Big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(\infty)}\Big).$$

Now, suppose that $C^{\delta,(\infty)}$ is a constant matrix with value c. Let $1 \le i \le n$, $1 \le j \le m$. Then, by definition of the fixed point, one has that

$$C_{ij}^{\delta,(\infty)} = \delta C_{ij} + (1-\delta) d_{\mathrm{W}}\Big(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\delta,(\infty)}\Big). \tag{58}$$

This implies that $c = \delta C_{ij} + (1-\delta)c$ and thus $C_{ij} = c$ for any i, j. Hence, this will be a contradiction unless C is also a constant matrix.

Finally, the speed of convergence result is also obtained as a consequence of the fact that the application F is $(1-\delta)$-contracting. Thus, by Pata et al. (2019, Corollary 1.1) we have that

$$\|C^{\delta,(k)} - C^{\delta,(\infty)}\|_\infty \le \frac{(1-\delta)^k}{\delta}\|C^{\delta,(1)} - C\|_\infty \le \frac{2(1-\delta)^k}{\delta}\|C\|_\infty.$$

where the rightmost inequality follows from the fact that $\|C^{\delta,(1)}\|_\infty \le \|C\|_\infty$ which can be proved using an argument similar for proving Equation (29). Finally, $|d_{\mathrm{WL},\delta}^{(k)}(\mathcal{X}, \mathcal{Y}) - d_{\mathrm{WL},\delta}^{(\infty)}(\mathcal{X}, \mathcal{Y})| \le \frac{2(1-\delta)^k}{\delta}\|C\|_\infty$ follows from Lemma 28.

The case when $\delta = 0$ is dealt with in Proposition 22. ■

Proof [Proof of Proposition 12] By Proposition 16, we know that when $\delta > 0$, $C^{\delta,(k)}$ converges to the unique fixed point $C^{\delta,(\infty)}$ of Equation (9). This implies that for any $i \in \boldsymbol{X}$ and $j \in \boldsymbol{Y}$,

$$C^{\delta,(\infty)}_{ij} = \delta C_{ij} + (1-\delta)\, d_{\mathrm{W}}\Big(m^{\boldsymbol{X}}_i, m^{\boldsymbol{Y}}_j; C^{\delta,(\infty)}\Big).$$

Define $m^{\boldsymbol{XY}}_{\bullet\bullet} : \boldsymbol{X} \times \boldsymbol{Y} \to \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$ by sending (i,j) to an optimal coupling between $m^{\boldsymbol{X}}_i$ and $m^{\boldsymbol{Y}}_j$ for the optimal transport problem in the equation above. We finally let $\nu^{\boldsymbol{XY}}$ be an optimal coupling for $d_{\mathrm{W}}\big(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(\infty)}\big)$. For any $i \in \boldsymbol{X}$ and $j \in \boldsymbol{Y}$, we construct a time homogeneous Markovian coupling $(X^{ij}_t, Y^{ij}_t)_{t\in\mathbb{N}}$ with initial distribution $\delta_i \otimes \delta_j$ and transition kernel $m^{\boldsymbol{XY}}_{\bullet\bullet}$. Then, we let $D_{ij} := \mathbb{E}\Big(C(X^{ij}_{T^\infty_\delta}, Y^{ij}_{T^\infty_\delta})\Big)$. We have that

$$\begin{aligned}
D_{ij} =& \mathbb{E}\Big(C(X^{ij}_{T^\infty_\delta}, Y^{ij}_{T^\infty_\delta})\Big)\\
=& \mathbb{E}\left(\sum_{t=0}^{\infty} \delta(1-\delta)^t C(X^{ij}_t, Y^{ij}_t)\right)\\
=& \delta C_{ij} + (1-\delta)\mathbb{E}\left(\sum_{t=1}^{\infty} \delta(1-\delta)^{t-1} C(X^{ij}_t, Y^{ij}_t)\right)\\
=& \delta C_{ij} + (1-\delta)\sum_{t=1}^{\infty} \delta(1-\delta)^{t-1} \sum_{i_1,j_1,\dots,i_t,j_t} C_{i_t,j_t} m^{\boldsymbol{XY}}_{i_{t-1}j_{t-1},i_tj_t} m^{\boldsymbol{XY}}_{i_{t-2}j_{t-2},i_{t-1}j_{t-1}} \cdots m^{\boldsymbol{XY}}_{ij,i_1j_1}\\
=& \delta C_{ij} +\\
& (1-\delta)\sum_{i_1,j_1}\left(\delta C_{i_1,j_1} + \sum_{t=2}^{\infty} \delta(1-\delta)^{t-1} \sum_{i_2,j_2,\dots,i_t,j_t} C_{i_t,j_t} m^{\boldsymbol{XY}}_{i_{t-1}j_{t-1},i_tj_t} \cdots m^{\boldsymbol{XY}}_{i_1j_1,i_2j_2}\right) m^{\boldsymbol{XY}}_{ij,i_1j_1}\\
=& \delta C_{ij} + (1-\delta)\mathbb{E}_{(i_1,j_1)\sim m^{\boldsymbol{XY}}_{ij}}(D_{i_1j_1}).
\end{aligned}$$

Here in the fourth equality we used Equation (54). Then, D is the fixed point (the uniqueness follows from an argument similar to the one for proving Proposition 16) of the equation

$$D_{ij} = \delta C_{ij} + (1-\delta)\mathbb{E}_{m^{\boldsymbol{XY}}_{ij}}(D_{i_1j_1}), \quad \forall i,j.$$

Notice by definition, $C^{\delta,(\infty)}$ is also a fixed point of the equation above. Then, $C^{\delta,(\infty)} = D$. Hence, if one constructs a time homogeneous Markovian coupling $(X_t, Y_t)_{t\in\mathbb{N}}$ with initial distribution $\nu^{\boldsymbol{XY}}$ and transition kernel $m^{\boldsymbol{XY}}_{\bullet\bullet}$. Then,

$$\begin{aligned}
&\mathbb{E}\big(C(X_{T^\infty_\delta}, Y_{T^\infty_\delta})\big) = \mathbb{E}\big(\mathbb{E}\big(C(X_{T^\infty_\delta}, Y_{T^\infty_\delta}) \mid (X_0,Y_0) = (i,j)\big)\big)\\
&= \mathbb{E}_{(i,j)\sim\nu^{\boldsymbol{XY}}}(D_{ij}) = \mathbb{E}_{(i,j)\sim\nu^{\boldsymbol{XY}}}\left(C^{\delta,(\infty)}_{ij}\right) = d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y}).
\end{aligned}$$

This concludes the proof. ■

Proof [Proof of Theorem 14] Choose a sequence $\delta_n \to 0$ such that

$$\lim_{n\to\infty} d^{(\infty)}_{\mathrm{WL},\delta_n}(\mathcal{X},\mathcal{Y};C) = \liminf_{\delta\to 0} d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y};C).$$

By Proposition 12, $d^{(\infty)}_{\mathrm{WL},\delta_n}(\mathcal{X},\mathcal{Y};C)$ can be obtained by an optimal time homogeneous Markovian coupling, which is determined by a transition kernel matrix $P_n \in \mathcal{T}$ and an initial distribution vector $\pi_n \in \mathcal{P}$, where $\mathcal{T} = \left\{P \in [0,1]^{|X||Y|\times|X||Y|}, P\mathbb{1} = \mathbb{1}\right\}$ and $\mathcal{P} = \left\{\pi \in [0,1]^{|X||Y|}, \mathbb{1}^{\mathrm{T}}\pi = 1\right\}$, where $\mathbb{1}$ is the vector containing all ones. $\mathcal{P}$ and $\mathcal{T}$ are both compact spaces. Up to a choice of subsequence, we assume that

- P_n converges to P in ℓ_∞ norm, where P is itself a transition kernel matrix;
- The limit $\lim_{n\to\infty} \delta \sum_{t=0}^{\infty}(1-\delta)^t (P_n^t)^{\mathrm{T}}\pi_n$ exists and is denoted by μ. Obviously, $\mu \in \mathcal{P}$.

Now, we have that

$$\begin{aligned}
P^{\mathrm{T}}\mu &= \lim_{n\to\infty} \delta_n \sum_{t=0}^{\infty}(1-\delta_n)^t P^{\mathrm{T}}(P_n^t)^{\mathrm{T}}\pi_n \\
&= \lim_{n\to\infty} \delta_n \sum_{t=0}^{\infty}(1-\delta_n)^t (P_n^{t+1})^{\mathrm{T}}\pi_n \\
&= \lim_{n\to\infty} \frac{\delta_n}{1-\delta_n} \sum_{i=1}^{\infty}(1-\delta_n)^t (P_n^t)^{\mathrm{T}}\pi_n \\
&= \lim_{n\to\infty} \delta_n \left(\sum_{t=0}^{\infty}(1-\delta_n)^t (P_n^t)^{\mathrm{T}} - P_n^{\mathrm{T}}\right)\pi_n \\
&= \lim_{n\to\infty} \delta_n \sum_{t=0}^{\infty}(1-\delta_n)^t (P_n^t)^{\mathrm{T}}\pi_n - \lim_{n\to\infty} \delta_n P_n^{\mathrm{T}}\pi_n \\
&= \mu - 0 = \mu.
\end{aligned}$$

Note that we have used the fact that $\delta_n \to 0$ several times in the derivation above. This means that μ is stationary w.r.t. the transition kernel P. Moreover, by assumption that ν_X and ν_Y are stationary, it is easy to check that μ, regarded as a probability measure still denote by μ, is a coupling: $\mu \in \mathcal{C}(\nu_X,\nu_Y)$. In this way, there exists a time homogeneous Markovian coupling $(X_t,Y_t)_{t\in\mathbb{N}}$ with transition kernel matrix P and with a stationary initial distribution μ. Hence,

$$\begin{aligned}
\liminf_{\delta\to 0} d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y};C) &= \lim_{n\to\infty} d^{(\infty)}_{\mathrm{WL},\delta_n}(\mathcal{X},\mathcal{Y};C) \\
&= \mathbb{E}\, C(X_0,Y_0) \geq d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y};C).
\end{aligned}$$

On the contrary, we know from Proposition 6 that $\limsup_{\delta\to 0} d^{(\infty)}_{\mathrm{WL},\delta}(\mathcal{X},\mathcal{Y};C) \leq d_{\mathrm{OTC}}(\mathcal{X},\mathcal{Y};C)$ and this concludes the proof. ■

E.4. Proofs and Technical Details from Section 5

E.4.1. THE DEFINITION AND BASIC PROPERTIES

We first provide a precise definition of the entropy-regularized optimal transport, including its primal and dual solution, which will later be useful to compute its gradient.

Definition 36 (Entropy-regularized optimal transport) *Remember that, using the same notations as in Equation (1), given $\epsilon \geq 0$, the (ϵ-)regularized OT problem is defined as:*

$$d_{\mathrm{W}}^{\epsilon}(\alpha, \beta; C) := \min_{(X,Y)\in\mathcal{C}(\alpha,\beta)} \mathbb{E}\, C(X,Y) - \epsilon H(X,Y). \tag{59}$$

Where H denotes the entropy function, i.e., $H(X,Y) := -\sum_{i\in\boldsymbol{X}, j\in\boldsymbol{Y}} P_{ij}\log(P_{ij})$, where $P_{ij} := \mathbb{P}(X=i, Y=j)$.

The distribution $P \in \mathbb{R}_{+}^{|\boldsymbol{X}|\times|\boldsymbol{Y}|}$ of the optimal coupling verifying the minimum is called the primal solution*.*

*Solving this problem is usually done using Sinkhorn's algorithm, an iterative algorithm described by Peyré et al. (2019, Chapter 4.2) — or a variant of described in Peyré et al. (2019, Chapter 4.4). The latter algorithm ends up computing as a byproduct the so-called "*dual solutions*" $f \in \mathbb{R}^{|\boldsymbol{X}|}, g \in \mathbb{R}^{|\boldsymbol{Y}|}$, which are the solutions to the following dual optimization problem:*

$$\operatorname*{arg\,max}_{f\in\mathbb{R}^{|\boldsymbol{X}|}, g\in\mathbb{R}^{|\boldsymbol{Y}|}} \langle f, \alpha\rangle + \langle g, \beta\rangle - \epsilon\langle e^{-f/\epsilon}, K e^{-g/\epsilon}\rangle, \tag{60}$$

where the matrix K is defined by $K_{ij} := e^{-C_{ij}/\epsilon}$.

We then provide a precise definition of the entropy-regularized discounted WL distance.

Definition 37 (Entropy-regularized δ-discounted WL distance) *Analogous to Proposition 15, let C denote the cost matrix, δ be the discount factor, and ϵ be the entropy-regularization parameter. We recursively define matrices $C^{\epsilon,\delta,(l)}$ for $l = 0, \ldots, k$ as follows:*

$$C_{ij}^{\epsilon,\delta,(0)} = C_{ij} \tag{61}$$

$$C_{ij}^{\epsilon,\delta,(l)} = \delta C_{ij} + (1-\delta)\, d_{\mathrm{W}}^{\epsilon}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C_{ij}^{\epsilon,\delta,(l-1)}\right). \tag{62}$$

Then, the δ-discounted entropy-regularized WL distance of depth k is defined as follows

$$d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y}; C) = d_{\mathrm{W}}^{\epsilon}\left(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\epsilon,\delta,(k)}\right). \tag{63}$$

We note that the matrices defined above satisfy some convergence properties similar to those in Proposition 16 for $C^{\delta,(k)}$:

Proposition 38 (Convergence of $C^{\epsilon,\delta,(k)}$) *For any $\delta \in (0,1]$ and any $\epsilon \geq 0$, the matrix $C^{\epsilon,\delta,(k)}$ converges as $k \to \infty$. In particular, $C^{\epsilon,\delta,(k)}$ converges to the unique fixed point $C^{\epsilon,\delta,(\infty)}$ of Equation (62). Moreover, $C^{\epsilon,\delta,(k)}$ converges at rate*

$$\|C^{\epsilon,\delta,(k)} - C^{\epsilon,\delta,(\infty)}\|_{\infty} \leq \frac{(1-\delta)^k}{\delta}(2\|C\|_{\infty} + \epsilon\log(nm)),$$

where $n := |\boldsymbol{X}|$ and $m := |\boldsymbol{Y}|$.

Proposition 39 *The limit $d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X}, \mathcal{Y}) := \lim_{k\to\infty} d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y})$ exists and in fact,*

$$d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X}, \mathcal{Y}) = d_{\mathrm{W}}^{\epsilon}\left(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\epsilon,\delta,(\infty)}\right). \tag{64}$$

Furthermore, one has that

$$|d_{\mathrm{WL},\delta,\epsilon}^{(k)}(\mathcal{X}, \mathcal{Y}) - d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X}, \mathcal{Y})| \leq \frac{(1-\delta)^k}{\delta}(2\|C\|_{\infty} + \epsilon\log(nm)).$$

The proofs of the two propositions above follow essentially the same arguments in the proof of Proposition 16 for the case when $\delta > 0$. The extra term ϵnm is from the fact that the entropy function satisfies that $|H| \leq \log(nm)$. We omit the proofs here.

E.4.2. PROOFS

Proof [Proof of Theorem 17]

Convergence When k is finite, the proof follows from the convergence of Sinkhorn distance to regular optimal transport (Cuturi, 2013, Property 1)

When $k = \infty$, the proof is based on the property of Banach fixed points in Lemma 29. Consider the space $Z = \mathbb{R}_{\geq 0}^{n \times m}$ endowed with ℓ^∞ distance where $n := |\boldsymbol{X}|$ and $m := |\boldsymbol{Y}|$. Let $\Lambda = [0, \infty)$. Let $F : Z \times \Lambda \to Z$ be defined by

$$F(A, \epsilon) := \delta C + (1 - \delta) \left(d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; A) \right)_{1 \leq i \leq n, 1 \leq j \leq m}.$$

Now, consider $A, B \in Z$. Then,

$$\begin{aligned} \|F(A, \epsilon) - F(B, \epsilon)\|_\infty &= (1 - \delta) \left\| \left(d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; A) - d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; B) \right)_{1 \leq i \leq n, 1 \leq j \leq m} \right\|_\infty \\ &= (1 - \delta) \max_{i,j} \left| d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; A) - d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; B) \right|. \end{aligned}$$

Hence, using Lemma 28, we have that for any given i and j,

$$|d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; A) - d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; B)| \leq \|A - B\|_\infty.$$

Therefore,

$$\begin{aligned} \|F(A, \epsilon) - F(B, \epsilon)\|_\infty &= (1 - \delta) \left\| \left(d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; A) - d_{\mathrm{W}}^{\epsilon}(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; B) \right)_{1 \leq i \leq n, 1 \leq j \leq m} \right\|_\infty \\ &\leq (1 - \delta) \|A - B\|_\infty. \end{aligned}$$

Now, by Lemma 29, Proposition 38 and (Peyré et al., 2019, Proposition 4.1), one has that $C^{\epsilon,\delta,(\infty)}$, as the fixed point for F_ϵ, is continuous w.r.t. ϵ. Hence, by Lemma 28 and Proposition 39, one has that

$$d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X}, \mathcal{Y}) = d_{\mathrm{W}}^{\epsilon}\left(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\epsilon,\delta,(\infty)}\right) \xrightarrow{\epsilon \to 0} d_{\mathrm{W}}\left(\nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}; C^{\delta,(\infty)}\right) = d_{\mathrm{WL},\delta}^{(\infty)}(\mathcal{X}, \mathcal{Y}).$$

Convergence rate To prove the bound, denote by $\phi(\epsilon)$ the fixed point of $A \mapsto F(A, \epsilon)$ for any $\epsilon \in [0, \infty)$.

We first note that for any probability measures $\alpha \in P(\boldsymbol{X})$ and $\beta \in P(\boldsymbol{Y})$ and any cost matrix $C : X \times Y \to \mathbb{R}$, we have that

$$|d_{\mathrm{W}}^{\epsilon}(\alpha, \beta; C) - d_{\mathrm{W}}(\alpha, \beta; C)| \leq \epsilon \log nm.$$

To see this, let (X, Y) be an optimal coupling for $d_{\mathrm{W}}^{\epsilon}(\alpha, \beta; C)$, then

$$|d_{\mathrm{W}}^{\epsilon}(\alpha, \beta; C) - d_{\mathrm{W}}(\alpha, \beta; C)| \leq \mathbb{E}C(X, Y) - (\mathbb{E}C(X, Y) - \epsilon H(X, Y)) = \epsilon H(X, Y) \leq \epsilon \log nm.$$

Then

$$\begin{aligned}\|F(A,\epsilon)-F(A,0)\|_\infty &= (1-\delta)\|(d_{\mathrm{W}}^{\epsilon}\big(m_i^{\boldsymbol{X}},m_j^{\boldsymbol{Y}};A\big)-d_{\mathrm{W}}\big(m_i^{\boldsymbol{X}},m_j^{\boldsymbol{Y}};A\big))_{1\le i\le n,1\le j\le m}\|_\infty\\ &\le \epsilon(1-\delta)\log nm.\end{aligned}$$

Thus

$$\begin{aligned}\|\phi(0)-\phi(\epsilon)\|_\infty &= \|F(\phi(0),0)-F(\phi(\epsilon),\epsilon)\|_\infty\\ &\le \|F(\phi(0),0)-F(\phi(\epsilon),0)\|_\infty+\|F(\phi(\epsilon),0)-F(\phi(\epsilon),\epsilon)\|_\infty\\ &\le (1-\delta)\|\phi(0)-\phi(\epsilon)\|_\infty+\epsilon(1-\delta)\log nm.\end{aligned}$$

Therefore, $\|\phi(0)-\phi(\epsilon)\|_\infty \le \frac{\epsilon(1-\delta)}{\delta}\log nm$.

Hence we have the following non-asymptotic bound between the regularized OTM distance and the original OTM distance.

$$\begin{aligned}&|d_{\mathrm{WL},\delta,\epsilon}^{(\infty)}(\mathcal{X},\mathcal{Y})-d_{\mathrm{WL},\delta}^{(\infty)}(\mathcal{X},\mathcal{Y})| = |d_{\mathrm{W}}^{\epsilon}\big(\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}};\phi(\epsilon)\big)-d_{\mathrm{W}}\big(\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}};\phi(0)\big)|\\ &\le |d_{\mathrm{W}}^{\epsilon}\big(\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}};\phi(\epsilon)\big)-d_{\mathrm{W}}^{\epsilon}\big(\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}};\phi(0)\big)|+|d_{\mathrm{W}}^{\epsilon}\big(\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}};\phi(0)\big)-d_{\mathrm{W}}\big(\nu^{\boldsymbol{X}},\nu^{\boldsymbol{Y}};\phi(0)\big)|\\ &\le \|\phi(0)-\phi(\epsilon)\|_\infty+\epsilon\log nm\\ &\le \frac{\epsilon(1-\delta)}{\delta}\log nm+\epsilon\log nm\\ &\quad=\frac{\epsilon(1-\delta+\delta)}{\delta}\log nm\\ &\quad=\frac{\epsilon}{\delta}\log nm.\end{aligned}$$

■

Proof [Proof of Theorem 18] Consider the space $Z=\mathbb{R}_+^{n\times m}$ endowed with ℓ^∞ distance where $n:=|\boldsymbol{X}|$ and $m:=|\boldsymbol{Y}|$. Let $\Lambda = M_n\times M_m\times\mathbb{R}_+^{n\times m}$, where $M_k=\{M\in\mathbb{R}_+^{k\times k}:\forall i,\sum_j M_{ij}=1\}$. Let $F: Z\times\Lambda\to Z$ be defined by

$$F(A,M^1,M^2,C):=\delta C+(1-\delta)(d_{\mathrm{W}}^{\epsilon}\big(M_i^1,M_j^2;A\big))_{1\le i\le n,1\le j\le m}.$$

Recall that $C_{ij}^{\epsilon,\delta,(\infty)}$ satisfies the following equation:

$$C_{ij}^{\epsilon,\delta,(\infty)}=\delta C_{ij}+(1-\delta)d_{\mathrm{W}}^{\epsilon}\Big(m_i^{\boldsymbol{X}},m_j^{\boldsymbol{Y}};C^{\epsilon,\delta,(\infty)}\Big). \tag{65}$$

In other words, $C^{\epsilon,\delta,(\infty)}$ is a fixed point of F. By Lemma 30, one has that $C^{\epsilon,\delta,(\infty)}$ is differentiable on the interior of its definition domain (which is a manifold without boundary). We could also use that lemma directly to compute the gradient, but, for clarity, we will still do the computations in coordinates to show how the gradient is computed in practice.

We differentiate $C_{ij}^{\epsilon,\delta,(\infty)}$ on both sides of Equation (65) below (using Einstein summation convention):

$$\begin{aligned}\Delta_{ij}^{kl} =& \frac{\partial C_{ij}^{\epsilon,\delta,(\infty)}}{\partial C_{kl}} = \delta \mathbb{1}_{(i,j)=(k,l)} + (1-\delta)\frac{\partial}{\partial C_{kl}} d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\epsilon,\delta,(\infty)}\right) \\ =& \delta \mathbb{1}_{(i,j)=(k,l)} + (1-\delta)\frac{\partial d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\epsilon,\delta,(\infty)}\right)}{\partial C_{\alpha\beta}^{\delta,(\infty)}}\frac{\partial C_{\alpha\beta}^{\delta,(\infty)}}{\partial C_{kl}} \\ =& \delta \mathbb{1}_{(i,j)=(k,l)} + (1-\delta) P_{ij}^{\alpha\beta}\Delta_{\alpha\beta}^{kl}.\end{aligned}$$

By identifying tensors with $nm \times nm$-square matrices via flattening together the dimensions (resp codimensions), one has that

$$\Delta = \delta I_{nm} + (1-\delta)P\Delta.$$

Hence,

$$(I_{nm} - (1-\delta)P)\Delta = \delta I_{nm},$$

and therefore, we have that

$$\Delta = \delta(I_{nm} - (1-\delta)P)^{-1}.$$

Here the matrix $K = I_{nm} - (1-\delta)P$ is invertible because it is strictly diagonally dominant: For any $1 \le i \le n$ and $1 \le j \le m$, one has that

$$\begin{aligned}K_{ij}^{ij} =& 1 - (1-\delta)P_{ij}^{ij} = \sum_{k,l} P_{ij}^{kl} - (1-\delta)P_{ij}^{ij} \\ =& \sum_{(k,l)\neq(i,j)} P_{ij}^{kl} + \delta P_{ij}^{ij} > \sum_{(k,l)\neq(i,j)} P_{ij}^{kl} = \sum_{(k,l)\neq(i,j)} |K_{ij}^{kl}|,\end{aligned}$$

where in the second equality we used the fact that $\sum_{k,l} P_{ij}^{kl} = 1$ since P_{ij} represents a coupling.

We apply the same method for calculating Γ: differentiating the fixpoint equation (cf. Equation (65)) on both sides, we have that

$$\begin{aligned}\Gamma_{ij}^{kk'} =& \frac{\partial C_{ij}^{\epsilon,\delta,(\infty)}}{\partial m_{kk'}^{\boldsymbol{X}}} \\ =& (1-\delta)\left(\frac{\partial d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\epsilon,\delta,(\infty)}\right)}{\partial m_{kk'}^{\boldsymbol{X}}} + \frac{\partial d_{\mathrm{W}}\left(m_i^{\boldsymbol{X}}, m_j^{\boldsymbol{Y}}; C^{\epsilon,\delta,(\infty)}\right)}{\partial C_{\alpha\beta}^{\epsilon,\delta,(\infty)}}\frac{\partial C_{\alpha\beta}^{\epsilon,\delta,(\infty)}}{\partial m_{kk'}^{\boldsymbol{X}}}\right) \\ =& (1-\delta)(\mathbb{1}_{k=i} f_{ij}^{k'} + P_{ij}^{\alpha\beta}\Gamma_{\alpha\beta}^{kk'}).\end{aligned}$$

Hence, $\Gamma = (1-\delta)(F + P\Gamma)$ and thus $(I_{nm} - (1-\delta)P)\Gamma = (1-\delta)F$. By invertibility of $K = I_{nm} - (1-\delta)P$ again, one has that

$$\Gamma = (1-\delta)(I_{nm} - (1-\delta)P)^{-1}F.$$

This concludes the proof. ∎

Appendix F. Glossary

$(X_t)_{t\in\mathbb{N}}, (Y_t)_{t\in\mathbb{N}}, (Z_t)_{t\in\mathbb{N}}$ We denote $(X_t)_{t\in\mathbb{N}}$ (resp. $(Y_t)_{t\in\mathbb{N}}$) a realisation of $\mathcal{X}$ (resp. $\mathcal{Y}$) 3

C We denote $C : \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}_+$ a cost function. 4

$\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \nu^{\boldsymbol{X}}, \nu^{\boldsymbol{Y}}, \nu^{\boldsymbol{Z}}, m_\bullet^{\boldsymbol{X}}, m_\bullet^{\boldsymbol{Y}}, m_\bullet^{\boldsymbol{Z}}$ We denote $\mathcal{X}$ (resp. $\mathcal{Y}$) a Markov chain over $\boldsymbol{X}$ (resp. $\boldsymbol{Y}$). It is defined by its initial distribution $\nu^{\boldsymbol{X}}$ (resp. $\nu^{\boldsymbol{Y}}$) and its transition kernel $m_\bullet^{\boldsymbol{X}}$ (resp. $m_\bullet^{\boldsymbol{Y}}$). 3

$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ In this paper, we use boldface letters, such as $\boldsymbol{X}$, to denote finite sets. 3, 53

d_{OTC} The Optimal Transport Coupling distance as defined by O'Connor et al. (2022). It is defined in Equation (4).

$$d_{\mathrm{OTC}}(\mathcal{X}, \mathcal{Y}; C) := \inf_{\substack{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi_{\mathrm{H}}(\mathcal{X},\mathcal{Y}) \\ \mathrm{law}((X_0,Y_0)) \text{ is stationary}}} \mathbb{E}\, C(X_0, Y_0),$$

More details in Section 2.2. 5–9, 31, 41, 42, 48, 53

d_{OTM} Generalized Optimal Transport Markov Distances are a class of Markov distances we define that encompasses or has as limit points WL distances, the OTC distance, and δ-discounted WL distances. More details in Section 3. They are parameterized by a distribution p over the integers, and defined by Equation (5).

$$d_{\mathrm{OTM}}^{p}(\mathcal{X}, \mathcal{Y}) = \inf_{(X_t,Y_t)_{t\in\mathbb{N}}} \mathbb{E}\, C(X_T, Y_T),$$

where $T \sim p$. 5, 6, 8, 27–29, 41–44, 53

$d_{\mathrm{WL},\delta}$ Our δ-discounted WL distance. It is a regularization of the original WL distance. More details in Section 4. They are defined as a parametric class of OTM distance, parameterized by the distributions defined in Section 4.2. For $k \in \mathbb{N}$:

$$d_{\mathrm{WL},\delta}^{(k)}(\mathcal{X}, \mathcal{Y}) := \inf_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}\left(\sum_{t=0}^{k-1} \delta(1-\delta)^t C(X_t, Y_t) + (1-\delta)^k C(X_k, Y_k)\right)$$

and

$$d_{\mathrm{WL},\delta}^{(\infty)}(\mathcal{X}, \mathcal{Y}) := \inf_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}\left(\sum_{t=0}^{\infty} \delta(1-\delta)^t C(X_t, Y_t)\right)$$

2, 7–10, 12, 28, 31, 33–38, 42, 46–48, 50, 51, 53

$d_{\mathrm{WL},\delta,\epsilon}$ The Entropy-regularized δ-discounted WL distance is obtained by replacing all Wasserstein distances by Sinkhorn distances in the computation of δ-discounted WL distance. This operation makes our δ-discounted WL distance into a smooth distance that can be used for learning, using the formulae developed in Section 5. It is defined in Definition 37. 10–12, 16, 17, 20, 21, 32, 49–51, 53

d_{WL} The Weisfeiler-Lehman distance as defined by Chen et al. (2022). It is defined by Equation (2).

$$d_{\mathrm{WL}}^{(k)}(\mathcal{X}, \mathcal{Y}; C) := \inf_{(X_t,Y_t)_{t\in\mathbb{N}}\in\Pi(\mathcal{X},\mathcal{Y})} \mathbb{E}\, C(X_k, Y_k).$$

More details in Section 2.2. 4–9, 22–26, 41, 53

$d_{\mathrm{W}}^{\epsilon}$ The Sinkhorn distance is the entropy-regularized version of the Wasserstein distance. We use it because of its smoothness properties. it is also faster to compute than the Wasserstein distance. It is defined in Theorem 36 of the Appendix.

$$d_{\mathrm{W}}^{\epsilon}(\alpha, \beta; C) := \min_{(X,Y) \in \mathcal{C}(\alpha,\beta)} \mathbb{E}\, C(X, Y) - \epsilon H(X, Y)$$

10, 11, 32, 33, 35–38, 49–51, 53, 54

d_{W} The Wasserstein distance is defined as the solution of the optimal transport between two measures with a cost matrix, in Equation (1).

$$d_{\mathrm{W}}(\alpha, \beta; C) := \inf_{(X,Y)} \mathbb{E}\, C(X, Y)$$

More details in Section 2.2. 4, 5, 9, 22, 27–30, 34, 37, 45–47, 50–54