Orthogonal Gromov Wasserstein Distance with Efficient Lower Bound

Abstract

Comparing structured data from possibly different metric-measure spaces is a fundamental task in machine learning, with applications in, e.g., graph classification. The Gromov-Wasserstein (GW) discrepancy formulates a coupling between the structured data based on optimal transportation, tackling the incomparability between different structures by aligning the intra-relational geometries. Although efficient *local* solvers such as conditional gradient and Sinkhorn are available, the inherent non-convexity still prevents a tractable evaluation. To address this issue, we take inspirations from the connection with the quadratic assignment problem, and propose the orthogonal Gromov-Wasserstein (OGW) discrepancy that admits an efficient and *closed-form* lower bound with the complexity of $\mathcal{O}(n^3)$. It also directly extends to the fused Gromov-Wasserstein (FGW) distance, incorporating node features into the coupling. Extensive experiments on both the synthetic and real-world datasets show the effectiveness and efficiency of our methods.

1 INTRODUCTION

Similarity based learning has been a popular approach in many machine learning applications. Instead of directly modeling each individual object which may pose challenge in some areas, it resorts to models based on pairwise similarity, possibly across different domains. The most common example is the kernels used in support vector machines and Gaussian process, including RBF kernel and covariance matrices that measure similarity in the Euclidean space [Chang et al., 2010], and string or tree kernels that compare discrete objects [Lodhi et al., 2002]. More generally, featured graphs have been a useful tool for capturing similarities and relations in the structured data that are commonly not Euclidean. Examples include social network [Fan et al., 2019], recommendation systems [Wu et al., 2020], fraud detection [Li et al., 2020], quantum chemistry [Coley et al., 2019] and topology-aware IoT applications [Abusnaina et al., 2019].

Despite the inherent challenge in comparing graphs from possibly different metric-measure spaces, there has been a wealth of refined discrepancy measures between graphs, including kernels [Vishwanathan et al., 2010, Shervashidze et al., 2011] and GCNs based approaches [Bronstein et al., 2017, Defferrard et al., 2016]. Recently, the Gromov-Wasserstein discrepancy [GW, Peyre et al., 2016], which extends the Gromov-Wasserstein distance [Memoli, 2011], has emerged as an effective transportation distance between structured data, alleviating the incomparability issue between different structures by aligning the *intra*-relational geometries. Thanks to its favorable properties such as efficiency and isometry-awareness, GW has been extensively applied to domain adaptation [Yan et al., 2018], word embedding [Alvarez-Melis and Jaakkola, 2018], graph classification [Vayer et al., 2019a], metric alignment [Ezuz et al., 2017], generative modeling [Cohen and Sejdinovic, 2019], and graph matching and node embedding [Xu et al., 2019b,a].

In contrast to the standard Wasserstein distance which is a linear program, GW is unfortunately intractable to evaluate. Despite the practical success of non-convex optimization techniques such as conditional gradient method and entropic regularization [Peyre et al., 2016, Gold and Rangarajan, 1996], it remains NP-hard to find the global optimum. Hence, existing practice settles with local solutions, lacking an analyzable guarantee. This significantly challenges the trustworthiness of the GW discrepancy.

The goal of this work, therefore, is to develop a tractable measure between graphs, while preserving important properties such as isometry. Our key contribution is a tight and permutation-invariant lower bound of the OGW that admits efficient and analytical evaluation schemes. The seed inspi-

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

ration, as unrolled in Section 2, stems from the connection of GW to quadratic assignment problems (QAPs), which was tapped into by sliced GW [Vayer et al., 2019b] and Gromov-Monge problems [Memoli and Needham, 2021]. It paved the way for approximating the set of doubly stochastic matrices (used in GW) by orthonormal matrices under marginal constraints [Rendl and Wolkowicz, 1992, Hadley et al., 1992, Anstreicher and Brixius, 2001]. The resulting problem is well known to admit tight approximate solutions, and accommodates fused GW to account for node features [FGW, Vayer et al., 2019a].

Our experiments in Section 4 first validate, through graph classification accuracy, that the relaxations on OGW preserve graph structures. We also demonstrate that the lower bound to our OGW beats the third lower bound of GW in both the tightness and time complexity. In the end, we provide our examples to compute the barycenter based on OGW from both synthetic data and point cloud MNIST dataset.

2 OGW-DISCREPANCY: TRACTABLE LOWER BOUNDS

We represent an undirected graph G with n nodes by an adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$, where $\mathbf{A}_{ii} = 0$, and $\mathbf{A}_{ij} = 1$ if there is an edge between nodes i and j ($i \neq j$), and 0 otherwise. To start with, we consider the discrepancy between two graphs with the same order (*i.e.*, number of nodes). A detailed generalization to any graph orders will be addressed in Section 2.3. Associated with the nodes is a distribution, encoding some prior information about their importance, e.g., the normalized degree of each node [Xu et al., 2019b]. However, many applications [Vayer et al., 2019a,b, Peyre et al., 2016] lack such natural normalization. Therefore, we will stick with a uniform distribution over all nodes and include our extension of non-uniform distribution in Appendix A.4. Letting $\mathbf{1} = (1,\dots,1)^\top$ whose dimensionality can be implicitly induced from the context, the standard GW distance between graph G and H based on ℓ_2 distance can be formulated as (the square root of)

$$\mathsf{GW}(G,H) = \min_{\mathbf{P}\in\mathcal{E}\cap\mathcal{N}} \sum_{i,j,k,l} \left[c_G(i,j) - c_H(k,l)\right]^2 \mathbf{P}_{ik}\mathbf{P}_{jl}, \tag{1}$$

where $\mathcal{E} := \{\mathbf{P} \in \mathbb{R}^{n\times n} : \mathbf{P}\mathbf{1} = \mathbf{P}^\top\mathbf{1} = \mathbf{1}\}$ and $\mathcal{N} := \mathbb{R}_+^{n\times n}$. Here $c_G(i,j)$ represents a distance measure between node i and j on G, and common choices include their shortest-path distance, or simply $1 - \mathbf{A}_{ij}$ (the complement of adjacency). When both c_G and c_H are a metric, Eq. (1) is a squared metric on isomorphism classes of measurable metric spaces. However, as pointed out by Peyre et al. [2016], c_G and c_H do not have to be restricted to metrics and ℓ_2 can be extended to other asymmetric or non-subadditive losses such as f-divergence. They call it GW discrepancy, which broadens its applicability in machine learning. Noting that neither triangle inequality nor symmetry in G and H is so relevant in many applications, we will follow their framework and simply refer to it as GW, without even taking the square root of Eq. (1) just like in [Peyre et al., 2016].

Define two n-by-n symmetric matrices $\mathbf{C}$ and $\mathbf{D}$ whose (i,j)-th elements are $\mathbf{C}_{ij} = c_G(i,j)$ and $\mathbf{D}_{ij} = c_H(i,j)$, respectively. For example, the complement of adjacency can be written as $\mathbf{C} = \mathbf{1}\mathbf{1}^\top - \mathbf{A}$. The above GW can be compactly rewritten in the Koopmans-Beckmann form [Koopmans and Beckmann, 1957]:

$$\mathsf{GW}(G,H) = \frac{1}{n^2}\left(\|\mathbf{C}\|_F^2 + \|\mathbf{D}\|_F^2 - 2\max_{\mathbf{P}\in\mathcal{E}\cap\mathcal{N}} \operatorname{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top)\right). \tag{2}$$

Here $\|\cdot\|_F$ is the Frobenius norm. Obviously, GW is permutation invariant, nonnegative, and equals 0 when G and H are isomorphic. The major drawback is that the maximization over $\mathbf{P}$ is intractable, although efficient *local* algorithms are available such as conditional gradient [Vayer et al., 2019a] and Sinkhorn [Peyre et al., 2016].

2.1 CONNECTING OGW WITH QUADRATIC ASSIGNMENT PROBLEM

Rewriting GW with quadratic optimization over $\mathcal{E}\cap\mathcal{N}$ in Eq. (2) reveals an innate connection to the quadratic assignment problem (QAP). Noting that by the Birkhoff–von Neumann theorem [Birkhoff, 1946], $\mathcal{E}\cap\mathcal{N}$ is the convex hull of the set of $n\times n$ permutation matrices (denoted as Π). Indeed, the connection with QAP has been used to formulate Gromov-Monge distances [Memoli and Needham, 2021], and to accelerate the evaluation of GW via projection (slicing) to 1-D [Vayer et al., 2019b]. Fortunately, a number of tractable relaxations of QAP are available, many of which are based on the following characterization of Π:

$$\Pi = \mathcal{E}\cap\mathcal{N}\cap\mathcal{O}, \tag{3}$$

where $\mathcal{O} := \{\mathbf{P} \in \mathbb{R}^{n\times n} : \mathbf{P}^\top\mathbf{P} = \mathbf{P}\mathbf{P}^\top = \mathbf{I}\}$. Here $\mathbf{I}$ is the identity matrix. Whenever necessary, we will explicitize the dimensionality of $\mathcal{O}$ by writing $\mathcal{O}_n$. Rendl and Wolkowicz [1992] first pointed out that $\operatorname{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top)$ can be maximized exactly by a simple eigen-decomposition if $\mathbf{P}$ is restricted to $\mathcal{O}$. Indeed, assume the eigen-decomposition of $\mathbf{C}$ and $\mathbf{D}$ are $\mathbf{C} = \mathbf{P}_1\mathbf{\Lambda}_1\mathbf{P}_1^\top$ and $\mathbf{D} = \mathbf{P}_2\mathbf{\Lambda}_2\mathbf{P}_2^\top$, respectively, and suppose the eigenvalues in $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are both arranged in a descending order. Then

$$\mathbf{P}_1\mathbf{P}_2^\top \in \arg\max_{\mathbf{P}\in\mathcal{O}} \operatorname{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top), \tag{4}$$

$$\text{and} \quad \max_{\mathbf{P}\in\mathcal{O}} \operatorname{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top) = \mathbf{1}^\top\mathbf{\Lambda}_1\mathbf{\Lambda}_2\mathbf{1}. \tag{5}$$

$$\underset{\mathcal{E}\cap\mathcal{N}}{\text{GW}} \longleftarrow \underset{\mathcal{E}\cap\mathcal{N}\cap\mathcal{O}}{\text{QAP}} \longrightarrow \underset{\mathcal{E}\cap\mathcal{O}}{\text{OGW}}$$

Figure 1: Connection between QAP and GW, OGW

Based on this result, Hadley et al. [1992] proposed tightening the domain approximation from $\mathcal{O}$ to $\mathcal{O}\cap\mathcal{E}$, which, despite the original intention of approximating inhomogeneous QAPs, happens to be useful in our context too. Compared with $\mathcal{N}\cap\mathcal{E}$, $\mathcal{O}\cap\mathcal{E}$ offers more convenience in constructing upper and lower bounds that are tight and efficient. This substitution leads to our proposed new metric, named as *orthogonal Gromov-Wasserstein (OGW)* distance:

$$\mathsf{OGW}(G,H) = \frac{1}{n^2}\left(\|\mathbf{C}\|_F^2 + \|\mathbf{D}\|_F^2 - 2\max_{\mathbf{P}\in\mathcal{O}\cap\mathcal{E}} \operatorname{tr}(\mathbf{CPDP}^\top)\right). \tag{6}$$

Figure 1 provide an overview of connections between QAP, GW and OGW, by relaxing the domain of the permutation matrices with its convex hulls.

2.2 UPPER AND LOWER BOUNDS OF OGW

The evaluation of OGW is hindered by the nonconvex objective *and* nonconvex domain in the optimization of $\mathbf{P}$ in Eq. (6). To address it, Hadley et al. [1992] insightfully characterized $\mathcal{O}\cap\mathcal{E}$ by:

$$\mathcal{O}\cap\mathcal{E} = \left\{\tfrac{1}{n}\mathbf{1}\mathbf{1}^\top + \mathbf{V}\mathbf{Q}\mathbf{V}^\top : \mathbf{Q}\in\mathcal{O}_{n-1}\right\}, \tag{7}$$

where $\mathbf{V}$ is any $n\times(n-1)$ matrix satisfying $\mathbf{V}^\top\mathbf{1}=\mathbf{0}$ and $\mathbf{V}^\top\mathbf{V}=\mathbf{I}_{n-1}$. Plugging $\mathbf{P}=\frac{1}{n}\mathbf{1}\mathbf{1}^\top+\mathbf{V}\mathbf{Q}\mathbf{V}^\top$ into the optimization objective in Eq. (6) yields

$$\begin{aligned}&\max_{\mathbf{P}\in\mathcal{O}\cap\mathcal{E}} \operatorname{tr}(\mathbf{CPDP}^\top)\\ &= \tfrac{1}{n^2}s_{\mathbf{C}}s_{\mathbf{D}} + \underbrace{\max_{\mathbf{Q}\in\mathcal{O}} \operatorname{tr}(\hat{\mathbf{C}}\mathbf{Q}\hat{\mathbf{D}}\mathbf{Q}^\top) + \operatorname{tr}(\hat{\mathbf{E}}^\top\mathbf{Q})}_{=:\ \mathcal{Q}(\hat{\mathbf{C}},\hat{\mathbf{D}},\hat{\mathbf{E}})},\end{aligned} \tag{8}$$

where $\hat{\mathbf{X}} := \mathbf{V}^\top\mathbf{X}\mathbf{V}$ and $s_{\mathbf{X}} := \mathbf{1}^\top\mathbf{X}\mathbf{1}$ for any matrix $\mathbf{X}$, and $\mathbf{E} := \frac{2}{n}\mathbf{C}\mathbf{1}\mathbf{1}^\top\mathbf{D}$. Since $\mathcal{Q}(\hat{\mathbf{C}},\hat{\mathbf{D}},\hat{\mathbf{E}})$ involves both linear and quadratic terms in $\mathbf{Q}$, no closed-form solution remains available. This motivates us to develop lower and upper bounds for $\mathcal{Q}$, which will serve as a tractable solution to the OGW discrepancy under isometry.

Lower bound of $\mathcal{Q}$. Obviously, any locally optimal $\mathbf{Q}$ yields a *lower bound* for $\mathcal{Q}$ (denoted as $\mathcal{Q}_{lb}$), *i.e.*, an upper bound for OGW (denoted as OGW_{ub}). *Locally* optimizing $\mathbf{Q}$ over $\mathcal{O}$ (*a.k.a.* Stiefel manifold) has been very well studied [Absil et al., 2009, Wen and Yin, 2013, Arasu and Mohan, 2018], and we adopt a straightforward approach of projected quasi-Newton, noting that the projection of any matrix $\mathbf{Q}$ on $\mathcal{O}$ is simply $\mathbf{U}_{\mathbf{Q}}\mathbf{V}_{\mathbf{Q}}^\top$, where the singular value decomposition (SVD) of $\mathbf{Q}$ is $\mathbf{U}_{\mathbf{Q}}\Lambda_{\mathbf{Q}}\mathbf{V}_{\mathbf{Q}}^\top$. Incidentally, we can also use it as an approximate evaluation of OGW, akin to the local solutions used for evaluating the vanilla GW.

Upper bound of $\mathcal{Q}$. To develop an *upper bound* for $\mathcal{Q}$ (denoted as $\mathcal{Q}_{ub}$), the most straightforward relaxation is to decouple the $\mathcal{Q}$ into quadratic and linear terms as follows:

$$\mathcal{Q}_{ub}(\hat{\mathbf{C}},\hat{\mathbf{D}},\hat{\mathbf{E}}) = \max_{\mathbf{Q}_1\in\mathcal{O}} \operatorname{tr}(\hat{\mathbf{C}}\mathbf{Q}_1\hat{\mathbf{D}}\mathbf{Q}_1^\top) + \max_{\mathbf{Q}_2\in\mathcal{O}} \operatorname{tr}(\hat{\mathbf{E}}^\top\mathbf{Q}_2). \tag{9}$$

As a result, we obtain a *lower bound* of OGW:

$$\mathsf{OGW}_{lb}(G,H) = \tfrac{1}{n^2}\left(\|\mathbf{C}\|_F^2 + \|\mathbf{D}\|_F^2 - 2\,\mathcal{Q}_{ub}(\hat{\mathbf{C}},\hat{\mathbf{D}},\hat{\mathbf{E}}) - \tfrac{1}{n^2}s_{\mathbf{C}}s_{\mathbf{D}}\right). \tag{10}$$

Theorem 1. *Given two graphs G, H with same order n, OGW is guaranteed to be nonnegative,* i.e.*, OGW$(c_G, c_H) \geq 0$, and if G and H are isomorphic to each other, the equality holds true.*

To achieve closed-form solution of $\mathcal{Q}_{ub}$ in Eq. (9), $\mathbf{Q}_1$ can be solved by Eq. (4). According to von Neumann's trace inequality, $\mathbf{Q}_2$ is optimized at $\mathbf{U}_{\mathbf{E}}\mathbf{V}_{\mathbf{E}}^\top$ where $\mathbf{U}_{\mathbf{E}}\mathbf{\Lambda}_{\mathbf{E}}\mathbf{V}_{\mathbf{E}}^\top$ is the SVD of $\hat{\mathbf{E}}$, and the maximum value of $\operatorname{tr}(\hat{\mathbf{E}}^\top\mathbf{Q}_2)$ is $||\hat{\mathbf{E}}||_*$, the trace norm of $\hat{\mathbf{E}}$, which is the sum of the singular values of $\hat{\mathbf{E}}$. Theorem 1 ensures the non-negativity of OGW by considering its lower bound (proof goes to Appendix A.1). Although the optimal values of $\mathbf{Q}_1$ and $\mathbf{Q}_2$ generally differ, Hadley et al. [1992] showed that such an upper bound in Eq. (9) is often quite tight, which is also observed in our experiments.

Computational complexity. The analytical solution from OGW_{lb} is achieved by singular value decomposition and eigen decomposition, depending on matrix multiplication. Assuming the two graphs have the same graph order, the complexity is $\mathcal{O}(n^3)$ in practice. It is worth mentioning that Memoli [2011] also derived the lower bounds of GW by solving a set of linear assignment problems, named First Lower Bound (FLB), Second Lower Bound (SLB), and Third Lower Bound (TLB). The following inequalities provide the connections between different lower bounds of GW:

$$\mathsf{GW} \geq \begin{cases}\mathsf{GW}_{tlb} \geq \mathsf{GW}_{flb}\\ \mathsf{GW}_{slb}.\end{cases} \tag{11}$$

And the complexities of FLB, SLB and TLB are $\mathcal{O}(n^2), \mathcal{O}(n^4), \mathcal{O}(n^5)$, respectively. In general, TLB provides the tightest lower bound for GW, and SLB has resemblant performance compared with TLB. This is also observed in our experiments.

2.3 GRAPHS WITH DIFFERENT SIZES

So far, we have been restricting the two graphs to have the same number of nodes, primarily because the orthonormal domain $\mathcal{O}$ only contains square matrix. In order to deal with the non-square matrix, *i.e.*, different sizes of graphs, we introduce the semi-orthogonal domain

$$\tilde{\mathcal{O}}_{m,n} := \{\mathbf{T} \in \mathbb{R}^{m\times n} : \mathbf{T}^\top \mathbf{T} = \mathbf{I}_n\}, \tag{12}$$

where $m > n$ without loss of generality. That is a domain of "tall" matrices, whose columns are orthonormal. When $m = n$, $\tilde{\mathcal{O}}$ recovers $\mathcal{O}_n$ because $\mathbf{T}^\top\mathbf{T} = \mathbf{I}_n$ is equivalent to $\mathbf{T}\mathbf{T}^\top = \mathbf{I}_n$ for a square matrix $\mathbf{T}$.

For a non-square matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ and $m > n$, it is no longer feasible to retain the constraint of $\mathcal{E} := \{\mathbf{P} \in \mathbb{R}^{n\times n} : \mathbf{P}\mathbf{1} = \mathbf{P}^\top\mathbf{1} = \mathbf{1}\}$ since the dimensionality does not match. Instead, one can generalize this normalization constraint into

$$\tilde{\mathcal{E}} := \{\mathbf{P}^\top\mathbf{P} = \mathbf{I}_n, \mathbf{P}\mathbf{1}_n = \sqrt{\tfrac{n}{m}}\mathbf{1}_m, \mathbf{P}^\top\mathbf{1}_m = \sqrt{\tfrac{m}{n}}\mathbf{1}_n\}. \tag{13}$$

To summarize, we can extend the OGW in Eq. (6) to graphs with different orders (number of nodes) by replacing the domain of $\mathbf{P}$ into:

$$\mathsf{OGW}(G, H) := \frac{\|\mathbf{C}\|_F^2}{m^2} + \frac{\|\mathbf{D}\|_F^2}{n^2} - \frac{2}{mn}\max_{\mathbf{P}\in\tilde{\mathcal{O}}\cap\tilde{\mathcal{E}}} \operatorname{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top), \tag{14}$$

where G and H have the graph order of m and n, respectively. We reuse the symbol OGW because the constraints $\tilde{\mathcal{O}}$ and $\tilde{\mathcal{E}}$ recover $\mathcal{O}$ and $\mathcal{E}$ respectively when $m = n$.

In the similar spirit to Eq. (7), any $\mathbf{P} \in \tilde{\mathcal{O}} \cap \tilde{\mathcal{E}}$ can be re-parameterized as follows

$$\mathbf{P} \in \tilde{\mathcal{O}} \cap \tilde{\mathcal{E}} \;\Leftrightarrow\; \mathbf{P} = \tfrac{1}{\sqrt{mn}}\mathbf{1}_m\mathbf{1}_n{}^\top + \mathbf{U}\mathbf{Q}\mathbf{V}^\top, \tag{15}$$

where $\mathbf{Q} \in \tilde{\mathcal{O}}_{m-1,n-1}$, and $\mathbf{U} \in \mathbb{R}^{m\times(m-1)}$ and $\mathbf{V} \in \mathbb{R}^{n\times(n-1)}$ are arbitrary projection matrices satisfying

$$\mathbf{U}^\top\mathbf{1}_m = \mathbf{0}_{m-1},\; \mathbf{V}^\top\mathbf{1}_n = \mathbf{0}_{n-1}, \tag{16}$$

$$\mathbf{U}^\top\mathbf{U} = \mathbf{I}_{m-1},\; \mathbf{V}^\top\mathbf{V} = \mathbf{I}_{n-1}. \tag{17}$$

One can then easily check that

$$\mathbf{P}^\top\mathbf{P} = \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top + \mathbf{V}\mathbf{V}^\top = \mathbf{I}_n, \tag{18}$$

where the last equality is by $\mathbf{V}^\top\mathbf{1} = \mathbf{0}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_{n-1}$. The necessity of the parametrization in Eq. (15) is relegated to Appendix A.3.

Letting $\hat{\mathbf{C}} := \mathbf{U}^\top\mathbf{C}\mathbf{U}$ and $\hat{\mathbf{D}} := \mathbf{V}^\top\mathbf{D}\mathbf{V}$, the optimization over $\mathbf{P}$ in Eq. (14) turns into a projected QAP (PQAP) with an additional linear term and some constant terms

$$\max_{\mathbf{Q}\in\tilde{\mathcal{O}}_{m-1,n-1}} \operatorname{tr}(\hat{\mathbf{C}}\mathbf{Q}\hat{\mathbf{D}}\mathbf{Q}^\top) + \operatorname{tr}(\hat{\mathbf{E}}\mathbf{Q}^\top) + \text{const}. \tag{19}$$

Finally, in order to leverage the favorable properties of orthonormal matrices, we right-pad the matrix $\mathbf{Q}$ by an $(m-1)\times(n-m)$ matrix $\mathbf{J}$, such that $[\mathbf{Q},\ \mathbf{J}] \in \mathcal{O}_{m-1}$. Indeed

$$\mathbf{Q} \in \tilde{\mathcal{O}}_{(m-1),(n-1)} \iff \exists\mathbf{J} : \begin{bmatrix}\mathbf{Q} & \mathbf{J}\end{bmatrix} \in \mathcal{O}_{m-1}. \tag{20}$$

Such a $\mathbf{J}$ matrix only needs to be any basis of the kernel space of $\mathbf{Q}$, and the left-most $(n-1)$ columns of any matrix in $\mathcal{O}_{m-1}$ must be in $\tilde{\mathcal{O}}_{m-1,n-1}$. As a result, the quadratic term $\max_{\mathbf{Q}\in\tilde{\mathcal{O}}} \operatorname{tr}(\hat{\mathbf{C}}\mathbf{Q}\hat{\mathbf{D}}\mathbf{Q}^\top)$ is equivalent to

$$\max_{[\mathbf{Q}\ \mathbf{J}]\in\mathcal{O}} \operatorname{tr}\left(\hat{\mathbf{C}}\begin{bmatrix}\mathbf{Q} & \mathbf{J}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{D}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{Q} & \mathbf{J}\end{bmatrix}^\top\right), \tag{21}$$

and we can optimize $[\mathbf{Q}\ \mathbf{J}]$ as a whole, utilizing the closed-form solution to the squared matrix case.

For the linear term, we have

$$\max_{\mathbf{Q}\in\tilde{\mathcal{O}}} \operatorname{tr}(\hat{\mathbf{E}}\mathbf{Q}^\top) = \max_{[\mathbf{Q}\ \mathbf{K}]\in\mathcal{O}} \operatorname{tr}\left(\begin{bmatrix}\hat{\mathbf{E}} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{Q} & \mathbf{K}\end{bmatrix}^\top\right), \tag{22}$$

where $\mathbf{K}$ serves the same role as $\mathbf{J}$.

To conclude, by padding zero on the non-square matrices $\hat{\mathbf{D}}$ and $\hat{\mathbf{E}}$ to square matrices, we can enjoy the analytic solutions to the problems in Eq. (21) and (22) in the same way as the square case.

2.4 BARYCENTER

The Barycenter problem is to minimize the Fréchet mean of the input matrices with respect to the weighted sum over similarities between sample and barycenter. More specific to our designed OGW metric, given a set of cost matrices $\mathbf{D}_i$ for $i \in [1, S]$, the barycenter problem is to minimize the weighted sum as

$$\mathcal{B}(\mathbf{C}) := \min_{\mathbf{C}} \sum_i^S \lambda_i \mathsf{OGW}(\mathbf{C}, \mathbf{D}_i) \tag{23}$$

$$= \min_{\mathbf{C}} \sum_i^S \lambda_i \left[\frac{\|\mathbf{C}\|_F^2}{m^2} - \frac{2}{mn_i}\max_{\mathbf{P}_i} \operatorname{tr}(\mathbf{C}\mathbf{P}_i\mathbf{D}_i\mathbf{P}_i^\top)\right], \tag{24}$$

where λ_i is the weight of the associated sample, having $\sum_i \lambda_i = 1$. And $\mathbf{P}_i \in \tilde{\mathcal{O}}\cap\tilde{\mathcal{E}}$ is the transition matrix between barycenter and i-th sample. For simplicity, we specify the barycenter with a fixed number of vertices m.

We follow the block coordinate update proposed in Peyre et al. [2016], *i.e.*, iteratively minimizing with respect to the couplings $\mathbf{P}_i$ and update the optimal cost matrix from barycenter in closed-form solution. Given a set of coupling matrices $\mathbf{P}_i$, we can retrieve the optimal $\mathbf{C}$ by taking the

partial derivative of Eq. (23)

$$\frac{\partial \mathcal{B}(\mathbf{C})}{\partial \mathbf{C}} = \sum_i \lambda_i \left[\frac{2\mathbf{C}}{m^2} - \frac{2}{mn_i} \mathbf{P}_i \mathbf{D}_i \mathbf{P}_i^\top \right] \tag{25}$$

$$= \frac{2\mathbf{C}}{m^2} - \sum_i \frac{2\lambda_i}{mn_i} \mathbf{P}_i \mathbf{D}_i \mathbf{P}_i^\top = \mathbf{0}. \tag{26}$$

By the first-order optimality, the stationary point can be achieved

$$\mathbf{C}^* \leftarrow \sum_i^S \frac{m\lambda_i}{n_i} \mathbf{P}_i \mathbf{D}_i \mathbf{P}_i^\top. \tag{27}$$

Noting that OGW itself is still intractable, therefore, we resort to the tractable OGW_{lb} and OGW_{ub}, replacing OGW in the definition of $\mathcal{B}(\mathbf{C})$ with its tractable bounds and adjusting the derivative of $\mathcal{B}$ w.r.t. $\mathbf{C}$ accordingly. Appendix A.6 provides more detailed derivatives. Due to the domain of $\mathcal{O} \cap \mathcal{E}$, $\mathbf{C}^*$ is not guaranteed to be elementwise non-negative. Next, we present our method, named as *eigen projection*, to find the proper $\mathbf{C}$ ensuring the the non-negative cost matrix.

2.4.1 Eigen projection

Since the OGW is defined on the domain of $\mathcal{O} \cap \mathcal{E}$ and optimized over $\mathcal{O}$, we want to build a correspondence between $\mathbf{C}^*$ with its domain characters. Start with two graphs with same number of nodes, let us denote $\hat{\mathbf{C}}^* = \mathbf{V}^\top \mathbf{C}^* \mathbf{V}$ and $\hat{\mathbf{D}} = \mathbf{V}^\top \mathbf{D} \mathbf{V}$, where $\mathbf{V}$ are defined in Eq. (7). And from eigen decomposition we have $\hat{\mathbf{C}} = \sum_i^{m-1} \sigma_i \mathbf{r}_i \mathbf{r}_i^\top$ $\hat{\mathbf{D}} = \sum_i^{m-1} \delta_i \mathbf{s}_i \mathbf{s}_i^\top$. Inspired by the observation of Remark 1, we build the $\hat{\mathbf{C}}^* = \sum_i^{m-1} \sigma_i \mathbf{s}_i \mathbf{s}_i^\top$, *i.e.*, assemble $\hat{\mathbf{C}}^*$ from the eigen information of its target matrix. For the graph with different sizes, the trick in Section 2.3, *i.e.*, padding on smaller graph, remedies us to assemble the $\hat{\mathbf{C}}^*$ from the top $m-1$ eigen system. Noting that $\hat{\mathbf{C}}^*$ is still on the projected domain $\mathbb{R}^{m-1 \times m-1}$, *i.e.*, there exist $\mathbf{V}$ such that $\hat{\mathbf{C}} = \mathbf{V}^\top \mathbf{C}^* \mathbf{V}$, where $\mathbf{V}$ is a projection matrix.

Remark 1. *Given two graphs G and H with same order, and their corresponding cost matrices $\mathbf{C}$ and $\mathbf{D}$. The semi-orthogonal projection matrix $\hat{\mathbf{C}}$ and $\hat{\mathbf{D}}$, shares the same set of eigenvalues.*

Next, we want to bring the $\hat{\mathbf{C}}$ back to its original space $\mathbb{R}^{m \times m}$. We can consider a general closed-form solution

$$\mathbf{C}^* = \mathbf{V} \hat{\mathbf{C}} \mathbf{V}^\top + \mathbf{Y} \tag{28}$$

where $\mathbf{Y}$ satisfies $\mathbf{V}^\top \mathbf{Y} \mathbf{V} = \mathbf{0}$ and $\operatorname{diag}(\mathbf{V} \hat{\mathbf{C}} \mathbf{V}^\top + \mathbf{Y}) = \mathbf{0}$. The first condition ensures the projection to the semi-original domain can be reversed, and the latter one ensures self-similarities to be 0 on the barycenter. Thus, we construct the $\mathbf{Y}$ as

$$\mathbf{Y} = -\tfrac{1}{2}(\mathbf{d}\mathbf{1}^\top + \mathbf{1}\mathbf{d}^\top) \tag{29}$$

where $\mathbf{d} := \operatorname{diag}(\mathbf{V} \hat{\mathbf{C}} \mathbf{V}^\top)$. Trivially, it satisfies the $\operatorname{diag}(\mathbf{V} \hat{\mathbf{C}} \mathbf{V}^\top + \mathbf{Y}) = \mathbf{0}$. To show that, we plug in the first condition of $\mathbf{Y}$, we have $\mathbf{V}^\top \left(\mathbf{d}\mathbf{1}^\top + \mathbf{1}\mathbf{d}^\top \right) \mathbf{V} = \mathbf{0}$. This is simply because of $\mathbf{V}^\top \mathbf{1} = \mathbf{1}^\top \mathbf{V} = 0$.

To verify the correctness, consider the barycenter problem with only one sample graph, there exist a set of graphs that satisfy the $\mathsf{OGW}_{lb} = 0$ due to the permutation invariance. Given the optimal solution of $\mathbf{C}^*$ by Eq. (27), represented as cost matrix from one of the permutation invariant graphs, we can retrieve the eigenvalues of its semi-projected matrix. As we can see, $\sigma_i = \delta_i, i \in [1, \cdots, m-1]$ in this scenario. Therefore, we can reconstruct the $\mathbf{C}$ exact to be the cost matrix from sample graph, resulting the $\mathcal{B}(\mathbf{C}) = \mathsf{OGW}_{lb}(\mathbf{C}, \mathbf{D}) = 0$ while finding the sample graph precisely. We provide our experiments on both synthetic and real dataset in Section 4.3.

3 EXTENSION TO FUSED GW-DISCREPANCY

Most applications carry features for each node. To account for this important information, Vayer et al. [2019a] proposed the fused GW (FGW), employing an additional matrix $\mathbf{M}$ whose (i,k)-th entry encodes the ℓ_2 distance between the features of node i in G and of node k in H. $\mathbf{M}$ is asymmetric in general. Then the vanilla FGW-distance was formulated in [Vayer et al., 2019a] as

$$\begin{aligned} \mathsf{FGW}(G, H, \mathbf{M}) := & \tfrac{\alpha}{n^2} \|\mathbf{C}\|_F^2 + \tfrac{\alpha}{n^2} \|\mathbf{D}\|_F^2 \\ & - \tfrac{1}{n^2} \max_{\mathbf{P} \in \mathcal{E} \cap \mathcal{N}} \left[2\alpha \operatorname{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top) - (1-\alpha) \operatorname{tr}(\mathbf{M}^\top \mathbf{P}) \right], \end{aligned} \tag{30}$$

where $\alpha \in [0,1]$ is a trade-off between structure and feature measure. For simplicity, we will only present the treatment for two graphs of the same size. The extension to different sizes can be easily derived in the same way as Section 2.3. Similar to $\mathcal{Q}$, with the additional linear term $\operatorname{tr}(\mathbf{M}^\top \mathbf{P})$, there is no closed-form solution even if we replace the domain of $\mathbf{P}$ by $\mathcal{O}$. However, we can still tighten the domain approximation from $\mathcal{O}$ to $\mathcal{O} \cap \mathcal{E}$, leading to our new approximation

$$\begin{aligned} \mathsf{OFGW}(G, H, \mathbf{M}) := & \tfrac{\alpha}{n^2} \|\mathbf{C}\|_F^2 + \tfrac{\alpha}{n^2} \|\mathbf{D}\|_F^2 \\ & - \tfrac{1}{n^2} \max_{\mathbf{P} \in \mathcal{E} \cap \mathcal{O}} \left[2\alpha \operatorname{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top) - (1-\alpha) \operatorname{tr}(\mathbf{M}^\top \mathbf{P}) \right]. \end{aligned} \tag{31}$$

The pipeline of construction is illustrated in Figure 2. As α tend to be zero, the OFGW distance recovers OGW distance between only structures. A number of favorable properties are enjoyed by OFGW, which are summarized in Theorem 2 below (proof deferred to Appendix A.2). Although OGW must be nonnegative, OFGW is not guaranteed nonnegative for all $\mathbf{M}$. To see a counter-example, set $\mathbf{C} = \mathbf{D} = \mathbf{0}$ and $\mathbf{M} = \mathbf{I}$. Fortunately, the cost matrix $\mathbf{C}$ and $\mathbf{D}$ are nonzero matrices in general. And for a large set of $\mathbf{M}$, it still enjoys nonnegativity.

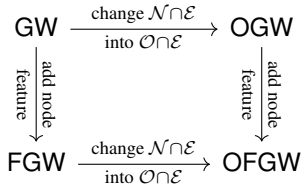


Figure 2: Steps of constructing the graph discrepancy measures

Theorem 2. *Suppose M satisfies $\mathbf{1}^\top\mathbf{M}\mathbf{1} \geq n||\hat{\mathbf{M}}||_*$. Then OFGW$(G, H, \mathbf{M}) \geq 0$ for all G, H, and is invariant to the (different) permutations of G and H. When $\mathbf{M} = \mathbf{0}$, it degenerates to OGW and evaluates to 0 when G and H are isomorphic.*

Since $\mathbf{M}$ encodes the pairwise distance between two sets of node features $\{\mathbf{X}_i\}_{i=1}^n$ and $\{\mathbf{Y}_i\}_{i=1}^n$ with $\mathbf{M}_{ij} = \|\mathbf{X}_i - \mathbf{Y}_j\|^2$, we can easily confirm whether the above assumption holds a priori. Interestingly, this is the case in all the datasets considered in our experiment, and we even conjecture that it always holds for pairwise distance matrices. For datasets without node attributes and labels, we take node degree as their features. In the sequel, we will make this assumption on $\mathbf{M}$. As $\mathbf{M} = \mathbf{0}$ satisfies the assumption, OGW must always be nonnegative, a result that is not so trivial because $\mathbf{P} \in \mathcal{O} \cap \mathcal{E}$ can have negative values, as opposed to $\mathcal{E} \cap \mathcal{N}$ in Eq. (2).

Although OFGW is still intractable in general, local optimization can be performed very efficiently. As will be shown in Section 3.1, it admits a tight *lower bound* as constructed by Hadley et al. [1992]. Now that OFGW is motivated by computational convenience, one may naturally wonder whether it captures as much graph structure as the original FGW does. We verified this in the affirmative by following Vayer et al. [2019a], where FGW is used as a kernel function served in a support vector machine. In Section 4.1, we will show that replacing FGW by OFGW achieves similar or better classification accuracy on a variety of datasets, corroborating the effectiveness of OFGW.

3.1 UPPER AND LOWER BOUNDS OF OFGW

Following Eq. (7), plugging $\mathbf{P} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top + \mathbf{V}\mathbf{Q}\mathbf{V}^\top$ into the optimization objective in Eq. (31) yields

$$\begin{aligned}&\max_{\mathbf{P}\in\mathcal{O}\cap\mathcal{E}}\{2\alpha\,\mathrm{tr}(\mathbf{C}\mathbf{P}\mathbf{D}\mathbf{P}^\top) - (1-\alpha)\,\mathrm{tr}(\mathbf{M}^\top\mathbf{P})\} \qquad (32)\\ &= \tfrac{1}{2n}s_{\mathbf{F}} - \tfrac{\alpha}{2n}s_{\mathbf{M}} + \underbrace{\max_{\mathbf{Q}\in\mathcal{O}} 2\alpha\,\mathrm{tr}(\hat{\mathbf{C}}\mathbf{Q}\hat{\mathbf{D}}\mathbf{Q}^\top) + \mathrm{tr}(\hat{\mathbf{F}}^\top\mathbf{Q})}_{=:\ \mathcal{Q}(\hat{\mathbf{C}},\hat{\mathbf{D}},\hat{\mathbf{F}})},\end{aligned}$$

where $\mathbf{F} := \frac{2\alpha}{n}\mathbf{C}\mathbf{1}\mathbf{1}^\top\mathbf{D} - \frac{1-\alpha}{2\alpha}\mathbf{M}$. It reveals that compared with the expression of GW in Eq. (8), the additional linear term $\mathbf{M}$ in the FGW formulation does not change the structure. As a result, we can again derive the *lower bound* of $\mathcal{Q}$, *i.e.*, *upper bound* of OFGW, by using the projected quasi-Newton as before. And a *lower bound* of OFGW can also be obtained by decoupling the $\mathbf{Q}$ in the two terms of Eq. (32). Specifically, we obtain

$$\begin{aligned}&\mathsf{OFGW}_{lb}(G, H, \mathbf{M}) = \tfrac{\alpha}{n^2}\|\mathbf{C}\|_F^2 + \tfrac{\alpha}{n^2}\|\mathbf{D}\|_F^2 \qquad (33)\\ &- \tfrac{1}{n^2}\left[2\alpha\,\mathcal{Q}_{ub}(\hat{\mathbf{C}}, \hat{\mathbf{D}}, \hat{\mathbf{F}}) + \tfrac{s_{\mathbf{F}}}{n} - \tfrac{1-\alpha}{2\alpha n}s_{\mathbf{M}}\right],\end{aligned}$$

where $\mathcal{Q}_{ub}$ is defined in Eq. (9) by decoupling into $\mathbf{Q}_1$ and $\mathbf{Q}_2$.

4 EXPERIMENTS

4.1 EFFECTIVENESS OF OGW/OFGW

Recall in Figure 2, $\mathcal{N} \cap \mathcal{E}$ was replaced by $\mathcal{O} \cap \mathcal{E}$ because the latter enjoys tight and efficiently computable upper and lower bounds. So it is important to validate the resulting OGW/OFGW as an equally good measure as the vanilla GW/FGW in comparing two graphs.

Datasets. We experimented on six graph classification datasets: BZR, COX2, MUTAG, PTC-MR, IMDB-Binary, and IMDB-Multi [TUDataset]. Their statistics are given in the Appendix A.7. The first four datasets contain a collection of molecules (*e.g.*, chemical compound and ligands), where the vertices represent atoms and edges are chemical bonds. The class label represents a certain property of the molecules, *e.g.*, "mutagenic effect on a specific bacterium" (MUTAG) and carcinogenicity of compounds for male rats (PTC-MR). BZR and COX2 consist collections of ligands for the benzodiazepine receptor and cyclooxygenase-2 inhibitors, respectively. IMDB-Binary and IMDB-Multi are the movie collaboration dataset, where nodes represent actors/actresses who played roles in movies in IMDB, and one edge means two played in the same movie. We group the dataset into three categories according to their feature property: vectorized, discrete, and no features. For the datasets with no features, we take the node degrees as their features.

Settings. In order to evaluate a discrepancy measure d, we follow Vayer et al. [2019a] by studying the graph classification accuracy of an SVM, whose kernel $k(G, H)$ is computed by $\exp(-\gamma d(G, H))$. For both FGW and OFGW, the feature distance matrix $\mathbf{M}$ employed the squared euclidean distance. Since OFGW itself is intractable to evaluate, we resort to the lower bound of OFGW in Eq. (30) which has an analytic form and is nonnegative. For the vanilla GW/FGW, we adopt the implementation from POT [Flamary et al., 2021] package, which initiates the transition matrix by the outer product of marginal distributions. And we initiate the cost matrix c_G by all-pair shortest path for each graph in the datasets, knowing that the structures are all connected. We evaluate the models by cross-validation hyperparameters

Table 1: Graph classification

	Dataset	Graph kernel		GW-based SVM			
		SP	GK ($k = 5$)	GW	FGW	OGW_{lb}	OFGW_{lb}
Vec.	BZR	78.8 ± 3.3	78.8 ± 3.3	**84.9 ± 1.8**	84.8 ± 3.2	83.4 ± 3.4	84.3 ± 3.8
Attr.	COX2	78.2 ± 0.4	78.2 ± 0.4	76.2 ± 2.1	78.5 ± 1.9	78.1 ± 1.8	**80.2 ± 2.4**
Disc.	MUTAG	78.2 ± 4.1	66.5 ± 0.9	85.1 ± 3.4	**85.7 ± 2.4**	82.8 ± 3.0	85.4 ± 1.7
Attr.	PTC-MR	57.3 ± 1.0	55.8 ± 0.7	53.4 ± 4.3	51.8 ± 3.4	**57.9 ± 4.5**	57.1 ± 4.1
No	IMDB-B	57.5 ± 2.6	60.1 ± 2.4	63.4 ± 0.9	65.6 ± 1.8	**68.3 ± 1.7**	67.3 ± 2.1
Attr.	IMDB-M	39.7 ± 1.8	38.2 ± 2.7	47.5 ± 2.3	**49.7 ± 1/7**	48.5 ± 1.9	47.1 ± 2.3

in SVM, setting γ from $\{2^{-10}, 2^{-9}, \cdots, 2^{10}\}$ and C from $\{-4, 4\}$ on evenly log scale with 15 steps. Moreover, for the FGW/OFGW, we cross-validate the value of α from $[0, 1]$ with grid search.

In addition, we provide the graph kernel methods as the baselines. More specifically, we adopt the implementation of shortest path (SP) kernel [Borgwardt and Kriegel, 2005] and graphlet sampling (GK) kernel [Przulj, 2007] from Siglidis et al. [2020]. SP kernel decomposes graphs into shortest paths and compares pairs of shortest paths according to their lengths and the labels of their endpoint, while GK kernel decomposes graphs into graphlets, *i.e.*, small subgraphs with k nodes where $k \in \{3, 4, 5, \cdots\}$, and counts matching graphlets in the input graphs.

Results. The average accuracies achieved after cross-validation are presented in Table 1. To see the effectiveness of OGW/OFGW, we compare the accuracy with the vanilla GW/FGW. As observed from the table, the tractable lower bound of our proposed metrics have comparable performance with GW and its variant. With the additional information from node features (except IMDB datasets), the OFGW provides higher accuracy in general. It corroborates that the chain in Figure 2 preserves important structures in graphs. Furthermore, it is noticeable that the accuracies with GW in the dataset of COX2 and PTC-MR are slightly worse than those in baselines. This is largely due to the intractable non-convex optimization with sub-optimal local minima in GW. Therefore, the tractable lower bound of OGW/OFGW are sound discrepancy measures for graph threat models.

4.2 TIGHTNESS OF THE LOWER BOUND

The tightness of our lower bound OGW_{lb} allows us to evaluate the distance by analytic solutions. To be specific, for the GW distance, we adopt the conditional gradient method to serve as a local optimizer, and use the First Lower Bound (FLB), Second Lower Bound (SLB), and Third Lower Bound (TLB) of GW in Memoli [2011] to serve as different lower bounds. And for the OGW distance, the upper bound comes from the projected quasi-Newton method, and the lower bound comes from the Eq. (10).

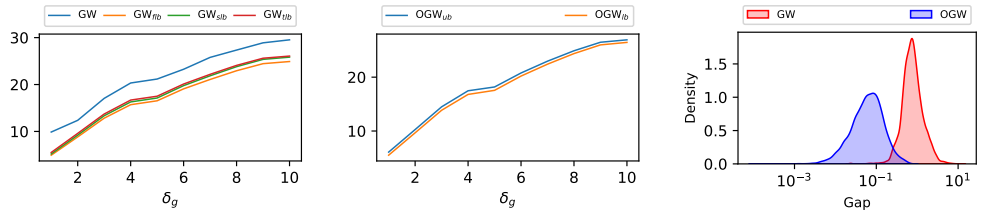
Synthetic data. To demonstrate the tightness in synthetic data, we generate a path graph with 20 nodes and randomly perturb $\delta_g \in [1, 10]$ edges 50 times, so that we can measure the distance (dissimilarity) between the original graph and perturbed graph under different measures. Figure 3a and 3b provide the average distance along with the number of perturbed edges GW and OGW, respectively. The gap (difference between OGW_{ub} and OGW_{lb}) in OGW is much tighter than the best gap (difference between GW and GW_{tlb}) in GW. Noting that even TLB of GW requires $\mathcal{O}(n^5)$ computational time, while the OGW_{lb} only cost $\mathcal{O}(n^3)$ in practice.

Real-world data. We also examine gap between lower and upper bounds on real-world dataset MUTAG by evaluating pairwise distance. Figure 3c provides the density distribution of the gaps from GW and OGW. For GW, we quantify the gap between GW local optimizer and GW_{tlb}, the best known lower bound. Clearly, we can see the gap between OGW_{ub} and OGW_{lb} is around 0.1, while the gap between the GW local optimizer and its third lower bound is around 1. Noting that, our measure is between the upper and lower bounds, a wider range compared with the measure between local optimum and lower bounds of GW. This characterizes the benefits of employing the $\mathcal{O}$ with separate $\mathbf{Q}_1$ and $\mathbf{Q}_2$ in Eq. (9).

4.3 BARYCENTER

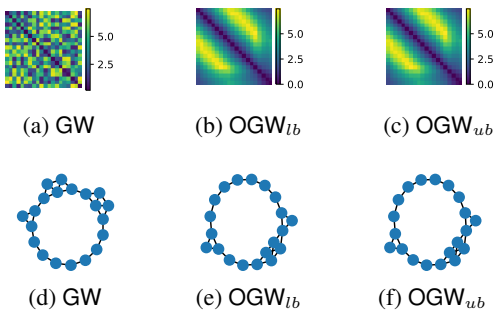
In the last set of experiments, we evaluate the ability of OGW_{lb} and OGW_{ub} to solve the barycenter problem.

Synthetic data. To start with, we generate a set of cycle-like graphs with different sizes ranging from 15 to 25. In addition, we also explicitly add random structural noise to the graphs separately and precompute their shortest path as cost matrices. All synthetic samples are plotted in Appendix A.8. To initialize the barycenter, we fix the number of nodes to be 20 and initiate a random symmetric $\mathbf{C}$ when starting the block coordinate descend to update $\mathbf{P}$. For the GW, we adopt the implementation from Peyre et al. [2016] to find the optimal cost matrix of the center. For the OGW_{lb} and OGW_{ub}, we solve the problem by our proposed eigen



(a) Tightness in GW (b) Tightness in OGW (c) Gap on MUTAG

Figure 3: Tightness of lower and upper bounds for GW and OGW.



(a) GW (b) OGW$_{lb}$ (c) OGW$_{ub}$

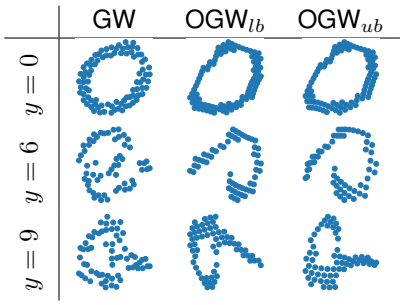
(d) GW (e) OGW$_{lb}$ (f) OGW$_{ub}$

Figure 4: Visualization of the optimized barycenter. Subfigures (a-c) represent the $\mathbf{C}^*$ retrieve by optimizing the barycenter problem. Subfigures (d-f) represent the reconstructed graph from the optimal cost matrix.

projection method in Section 2.4. To demonstrate the correctness of the results, we also reconstruct the adjacency matrix in a heuristic way, minimizing the closest adjacency by line search on optimal cost matrix with a threshold value. From Figure 4, we can see the reconstructed cost matrix for the barycenter from our OGW$_{lb}$ and OGW$_{ub}$ are well aligned to its node ordering. We also recognize that the result from GW is just one of the local minimizers in the context of permutation, Due to the tightness between the lower and upper bounds, it is also hard to differentiate the structures (G) from OGW$_{lb}$ and OGW$_{ub}$.

Point cloud data. We also explore the barycenter on point cloud MNIST-2D dataset [1], where each node represents a non-zero pixel, along with its 2D coordinates. Without constructing the explicit mesh connections between pixels, we take the Euclidean distance as their cost matrix $\mathbf{D}_i$. After retrieving the optimal $\mathbf{C}^*$ of barycenter, we step further to find the optimal coordinates associated with it. To do so, admitting a large set of solutions will satisfy our problem in terms of coordinates rotation and shifting, a BFGS optimizer will find one of the solutions without loss of generality. Table 2 shows the optimal point clouds with labels to be $0, 6, 9$. We sample 5 different point clouds for each digit and set the weight λ_i as uniform distribution. Moreover, we fix the number of nodes on the barycenter as the minimum size from the samples. Refer to Appendix A.8 for plots of corresponding point cloud data. Clearly, OGW finds better point clouds of the barycenter in the cases with digit 6 and 9, acknowledging that they are quite similar originally with their rotations. Moreover, thanks to the tight gap between the bounds, both the OGW$_{lb}$ and OGW$_{ub}$ performance undistinguishably from our perspective.

Table 2: Barycenters from point cloud MNIST-2D samples



5 CONCLUSION

Inspired by quadratic assignment problem, we proposed the orthogonal Gromov-Wasserstein (OGW) distance by optimizing the transition matrix with a relaxed domain from permutation matrices, and developed a tractable and efficient lower bound. In addition, we also provided an extension to OFGW, incorporating the node features on the graphs. From our experiments from graph classification and barycenter problems, we justified our methods serving as a proper alternative to the GW with tighter bounds.

We believe that the tractable bounds will be helpful to provide reliable discrepancies for the structured data, and envision its potential use in the applications such as robust geometric learning and adversarial training.

[1] Point cloud MNIST-2D dataset: `https://www.kaggle.com/cristiangarcia/pointcloudmnist2d`

References

P.-A. Absil, R. Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Ahmed Abusnaina, Aminollah Khormali, Hisham Alasmary, Jeman Park, Afsah Anwar, and Aziz Mohaisen. Adversarial learning attacks on graph-based iot malware detection systems. In *ICDCS*, pages 1296–1305. IEEE, 2019.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Kurt M. Anstreicher and Nathan W. Brixius. A new bound for the quadratic assignment problem based on convex quadratic programming. *Mathematical Programming*, 89: 341–357, 2001.

K. T. Arasu and Manil T. Mohan. Optimization problems with orthogonal matrix constraints. *Numerical Algebra, Control & Optimization*, 8(4):413–440, 2018.

Garrett Birkhoff. Three observations on linear algebra. *Univ. Nac. Tacuman, Rev. Ser. A*, 5:147–151, 1946.

Karsten M. Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Proc. Intl. Conf. Data Mining*, pages 74–81, 2005.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, 11(4), 2010.

Samuel Cohen and Dino Sejdinovic. On the gromov-wasserstein distance and coupled deep generative models. In *NeurIPS 2019 Workshop on Optimal Transport & Machine Learning*, 2019.

Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377, 2019.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Danielle Ezuz, Justin Solomon, Vladimir G. Kim, and Mirela Ben-Chen. Gwcnn: A metric alignment layer for deep shape analysis. *Comput. Graph. Forum*, 36(5): 49–57, 2017.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The World Wide Web Conference*, pages 417–426, 2019.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Steven Gold and Anand Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on pattern analysis and machine intelligence*, 18(4):377–388, 1996.

S. W. Hadley, F. Rendl, and H. Wolkowicz. A new lower bound via projection for the quadratic assignment problem. *Mathematics of Operations Research*, 17(3):727–739, 1992.

T. C. Koopmans and M. Beckmann. Assignment problems and the location of economic activities. In *Econometrica, Journal of the Econometric Society*, 1957.

Xiangfeng Li, Shenghua Liu, Zifeng Li, Xiaotian Han, Chuan Shi, Bryan Hooi, He Huang, and Xueqi Cheng. Flowscope: Spotting money laundering based on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4731–4738, 2020.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, February 2002.

Facundo Memoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

Facundo Memoli and Tom Needham. Distance distributions and inverse problems for metric measure spaces (former title: Gromov-Monge quasi-metrics and distance distributions). *arXiv:1810.09646v3*, 2021.

Gabriel Peyre, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning (ICML)*, 2016.

N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, Jan 2007.

Franz Rendl and Henry Wolkowicz. Applications of parametric programming and eigenvalue maximization to the quadratic assignment problem. *Mathematical Programming: Series A and B*, 53(1–3):63–78, 1992.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011.

Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. Grakel: A graph kernel library in python. *J. Mach. Learn. Res.*, 21(54):1–5, 2020.

TUDataset. Tudataset. `https://chrsmrrs.github.io/datasets/docs/home/`.

Titouan Vayer, Laetitia Chapel, Remi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning (ICML)*, 2019a.

Titouan Vayer, Remi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced Gromov-Wasserstein. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

S. V. N. Vishwanathan, Nicol N. Schraudolph, Imre Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(40):1201–1242, 2010.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142:397–434, 2013.

Shiwen Wu, Fei Sun, Wentao Zhang, and Bin Cui. Graph neural networks in recommender systems: a survey. *arXiv preprint arXiv:2011.02260*, 2020.

Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov-Wasserstein learning for graph matching and node embedding. In *International Conference on Machine Learning (ICML)*, 2019b.

Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.