ROBUST GRAPH DICTIONARY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional Dictionary Learning (DL) aims to approximate data vectors as sparse linear combinations of basis elements (atoms) and is widely used in machine learning, computer vision, and signal processing. To extend DL to graphs, Vincent-Cuaz et al. 2021 propose a method, called GDL, which describes the topology of each graph with a *pairwise relation matrix* (PRM) and compares PRMs via the Gromov-Wasserstein Discrepancy (GWD). However, the lack of robustness often excludes GDL from a variety of real-world applications since GWD is sensitive to the structural noise in graphs. This paper proposes an improved graph dictionary learning algorithm based on a robust Gromov-Wasserstein discrepancy (RGWD) which has theoretically sound properties and an efficient numerical scheme. Based on such a discrepancy, our dictionary learning algorithm can learn atoms from noisy graph data. Experimental results demonstrate that our algorithm achieves good performance on both simulated and real-world datasets.

1 INTRODUCTION

Dictionary learning (DL) seeks to learn a set of basis elements (atoms) from data and approximates data samples by sparse linear combinations of these basis elements (Mallat, 1999; Mairal et al., 2009; Tošić and Frossard, 2011), which has numerous machine learning applications including dimensionality reduction (Feng et al., 2013; Wei et al., 2018), classification (Raina et al., 2007; Mairal et al., 2008), and clustering (Ramirez et al., 2010; Sprechmann and Sapiro, 2010), to name a few.

Although DL has received significant attention, it mostly focuses on vectorized data of the same dimension and is not amenable to graph data (Xu, 2020; Vincent-Cuaz et al., 2021; 2022). Many exciting machine learning tasks use graphs to capture complex structures (Backstrom and Leskovec, 2011; Sadreazami et al., 2017; Naderializadeh et al., 2020; Jin et al., 2017; Agrawal et al., 2018). DL for graphs is more challenging due to the lack of effective means to compare graphs. Specifically, evaluating the similarity between one observed graph and its approximation is difficult, since they are often with different numbers of nodes and the node correspondence across graphs is often unknown (Xu, 2020; Vincent-Cuaz et al., 2021).

The seminal work of Vincent-Cuaz et al. (2021) proposes a DL method for graphs based on the Gromov-Wasserstein Discrepancy (GWD) that is a variant of the Gromov-Wasserstein distance. Gromov-Wasserstein distance compares probability distributions supported on different metric spaces using pairwise distances (Mémoli, 2011). By expressing each graph as a probability measure and capturing the graph topology with a pairwise relation matrix (PRM), comparing graphs can be naturally formulated as computing the GWD, since both the node correspondence and the discrepancy of the compared graphs are calculated (Peyré et al., 2016; Xu et al., 2019b; Chowdhury and Mémoli, 2019). However, observed graphs often contain structural noise including spurious or missing edges, which leads to the differences between the obtained PRMs and the true ones (Donnat et al., 2018; Xu et al., 2019b). Since GWD lacks robustness (Séjourné et al., 2021; Vincent-Cuaz et al., 2022; Tran et al., 2022), the inaccuracies of PRMs may severely affect GWD and the effectiveness of DL in real-world applications.

Contributions. To handle the inaccuracies of PRMs, this paper first proposes a novel robust Gromov-Wasserstein discrepancy (RGWD) which adopts a minimax formulation. We prove that the inner maximization problem has a closed-form solution and derive an efficient numerical scheme to approximate RGWD. Under suitable assumptions, such a numerical scheme is guaranteed to find a δ-stationary solution within $\mathcal{O}(\frac{1}{\delta^2})$ iterations. We further prove that RGWD is lower bounded

and the lower bound is achieved if and only if two graphs are isomorphic. Therefore, RGWD can be employed to compare graphs. RGWD also satisfies the triangle inequality which is of its own interest and allows numerous potential applications. A robust graph dictionary learning (RGDL) algorithm is thereby developed to learn atoms from noisy graph data, which assesses the quality of approximated graphs via RGWD. Numerical experiments on both synthetic and real-world datasets demonstrate that RGDL achieves good performance.

The rest of the paper is organized as follows. In Sec. 2, a comprehensive review of the background is given. Sec. 3 presents RGWD and the numerical approximation scheme for RGWD. RGDL is delineated in Sec. 4. Empirical results are demonstrated in Sec. 5. We finally discuss related work in Sec. 6.

2 PRELIMINARY

2.1 OPTIMAL TRANSPORT

We first present the notation used throughout this paper and then review the definition of the Gromov-Wasserstein distance that originates from the optimal transport theory (Villani, 2008; Peyre and Cuturi, 2018).

Notation. We use bold lowercase symbols (e.g. $\mathbf{x}$), bold uppercase letters (e.g. $\mathbf{A}$), uppercase calligraphic fonts (e.g. $\mathcal{X}$), and Greek letters (e.g. α) to denote vectors, matrices, spaces (sets), and measures, respectively. $\mathbf{1}^d \in \mathbb{R}^d$ is a d-dimensional all-ones vector. Δ^d is the probability simplex with d bins, namely the set of probability vectors $\Delta^d = \left\{\mathbf{a} \in \mathbb{R}_+^d | \sum_{i=1}^d a_i = 1\right\}$. $\mathbf{A}[i,:]$ and $\mathbf{A}[:,j]$ are the ith row and the jth column of matrix $\mathbf{A}$ respectively. Given a matrix $\mathbf{A}$, we denote by $\|\mathbf{A}\|_F$ its Frobenius norm, by $\|\mathbf{A}\|_0$ its number of non-zero elements, and by $\|\mathbf{A}\|_\infty$ its element-wise ℓ_∞ norm (i.e., $\|\mathbf{A}\|_\infty = \max_{ij} |A_{ij}|$). The cardinality of set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. The bracketed notation $[\![n]\!]$ is the shorthand for integer sets $\{1, 2, \ldots, n\}$. A discrete measure α can be denoted by $\alpha = \sum_{i=1}^m p_i \delta_{\mathbf{x}_i}$ where $\delta_{\mathbf{x}}$ is the Dirac at position $\mathbf{x}$, i.e., a unit of mass infinitely concentrated at $\mathbf{x}$.

Gromov-Wasserstein distance. Optimal Transport addresses the problem of transporting one probability measure towards another probability measure with the minimum cost (Villani, 2008; Peyre and Cuturi, 2018). The induced cost defines a distance between the two probability measures. Gromov-Wasserstein (GW) distance extends classic optimal transport to compare probability measures supported on different spaces (Mémoli, 2011). Let $\mathcal{X}$ and $\mathcal{Y}$ be two sample spaces. Endowing the spaces $\mathcal{X}$ and $\mathcal{Y}$ with metrics (distances) $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, the r-GW distance between $\alpha = \sum_{i=1}^m p_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{i'=1}^n q_i \delta_{\mathbf{x}_i}$ is defined as

$$\mathrm{GW}_r(\alpha, \beta) := \left(\min_{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})} \sum_{i,j=1}^m \sum_{i',j'=1}^n D_{ii'jj'}^r T_{ii'} T_{jj'} \right)^{\frac{1}{r}},$$

where the feasible domain of the *transport plan* $\mathbf{T} = [T_{ii'}]$ is given by the set

$$\Pi(\mathbf{p}, \mathbf{q}) = \left\{ \mathbf{T} \in \mathbb{R}_+^{m \times n} \middle| \mathbf{T}\mathbf{1}^n = \mathbf{p}, \mathbf{T}^\top \mathbf{1}^m = \mathbf{q} \right\},$$

and $D_{ii'jj'}$ calculates the difference between pairwise distances, i.e., $D_{ii'jj'} = |d_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) - d_{\mathcal{Y}}(\mathbf{y}_{i'}, \mathbf{y}_{j'})|$ with $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \in \mathcal{Y}$.

2.2 GRAPH REPRESENTATION AND COMPARISON

In this subsection, we formalize the idea of comparing graphs with GWD, which addresses the challenges that graphs are often with different numbers of nodes and the node correspondence is unknown (Xu et al., 2019b; Xu, 2020; Vincent-Cuaz et al., 2021).

Pairwise relation and graph representation. Graphs are often expressed using PRMs. For instance, the atoms making up proteins and other molecules are only known up to translation and rotation, and thus the geometry of a molecule might be better described by the PRM (Peyré et al., 2016). Given a graph $\mathcal{G}$ with n nodes, assigning each node an index $i \in [\![n]\!]$, $\mathcal{G}$ can be expressed

as a tuple $(\mathbf{C}, \mathbf{p})$, where $\mathbf{C} \in \mathbb{R}^{n\times n}$ is a matrix encoding the pairwise relations (e.g. adjacency, shortest-path, Laplacian, or heat kernel) and $\mathbf{p} \in \Delta^n$ is a probability vector modeling the relative importance of nodes within the graph (Peyré et al., 2016; Xu et al., 2019b; Titouan et al., 2019; Vincent-Cuaz et al., 2022).

Gromov-Wasserstein Discrepancy GWD can be derived from the 2-GW distance by replacing the metrics with pairwise relations (Xu et al., 2019b; Vincent-Cuaz et al., 2022). More specifically, given the observed source graph $\mathcal{G}^{\mathrm{s}}$ and the target graph $\mathcal{G}^{\mathrm{t}}$ that can be expressed as $(\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}})$ and $(\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}})$ respectively, GWD is defined as

$$\mathrm{GWD}\left((\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}}), (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}})\right) = \left(\min_{\mathbf{T}\in\Pi(\mathbf{p}^{\mathrm{s}},\mathbf{p}^{\mathrm{t}})} \sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'})^2 T_{ii'} T_{jj'}\right)^{\frac{1}{2}},$$

where n^{s} and n^{t} are the numbers of nodes of $\mathcal{G}^{\mathrm{s}}$ and $\mathcal{G}^{\mathrm{t}}$ respectively. GWD computes both a soft assignment matrix between the nodes of the two graphs and a notion of discrepancy between them. For conciseness, we abbreviate $\mathrm{GWD}\left((\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}}), (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}})\right)$ to $\mathrm{GWD}(\mathbf{C}^{\mathrm{s}}, \mathbf{C}^{\mathrm{t}})$ in the sequel.

2.3 Dictionary Learning

Traditional DL approximates data vectors as sparse linear combinations of basis elements (atoms) (Mallat, 1999; Mairal et al., 2009; Tošić and Frossard, 2011; Jiang et al., 2015), and is usually formulated as

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{W}} \sum_{k=1}^{K} \left\|\mathbf{X}[:,k] - \sum_{m=1}^{M} w^k_m \mathbf{D}[:,m]\right\|_2^2 + \lambda\Omega(\mathbf{w}^k), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{d\times K}$ is the data matrix whose columns represent samples, the matrix $\mathbf{D} \in \mathbb{R}^{d\times M}$ contains M atoms to learn and is constrained to the following set

$$\mathcal{C} = \{\mathbf{D} \in \mathbb{R}^{d\times M} | \forall m \in [\![M]\!], \|\mathbf{D}[:,m]\|_2 \leq 1\},$$

$\mathbf{W} \in \mathbb{R}^{M\times K}$ is the new representation of data whose k^{th}-column $\mathbf{w}^k = [w^k_m]_{m\in[\![M]\!]}$ stores the embedding of the kth sample, and $\lambda\Omega(\mathbf{w}^k)$ promotes the sparsity of $\mathbf{w}^k$. Such a formulation only applies to vectorized data.

Recently, Xu 2020 proposes to approximate graphs via the highly non-linear GW barycenter. Specifically, given a dataset of K graphs which has PRMs $\{\mathbf{C}^k\}_{k\in[\![K]\!]}$ such that $\mathbf{C}^k \in \mathbb{R}^{n^k\times n^k}$, the basis elements $\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}$ are learned by solving

$$\min_{\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]},\{\mathbf{w}^k\}_{k\in[\![K]\!]}} \sum_{k=1}^{K} \mathrm{GWD}^2\left(\mathbf{C}^k, \mathbf{B}\Big(\mathbf{w}^k, \{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}\Big)\right),$$

where $\mathbf{w}^k \in \Delta^M$ is referred to as the embedding of the kth graph $\mathcal{G}^k$, and the GW barycenter $\mathbf{B}\big(\mathbf{w}^k, \{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}\big)$ gives the approximation of $\mathcal{G}^k$ and is defined as

$$\mathbf{B}\Big(\mathbf{w}^k, \{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}\Big) = \operatorname*{argmin}_{\mathbf{B}} \sum_{m=1}^{M} w^k_m \,\mathrm{GWD}^2(\mathbf{B}, \bar{\mathbf{C}}^m).$$

Therefore, a complex bi-level optimization problem is involved, which is computationally inefficient (Vincent-Cuaz et al., 2021).

DL for graphs. To overcome the above computational issues, Vincent-Cuaz et al. 2021 propose GDL which approximates each graph as a weighted sum of PRMs and is formulated as

$$\min_{\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]},\{\mathbf{w}^k\}_{k\in[\![K]\!]}} \sum_{k=1}^{K} \mathrm{GWD}^2\left(\mathbf{C}^k, \sum_{m=1}^{M} w^k_m \bar{\mathbf{C}}^m\right) + \lambda\Omega(\mathbf{w}^k),$$

where each atom $\bar{\mathbf{C}}^m$ is a $n^{\mathrm{a}} \times n^{\mathrm{a}}$ matrix. In contrast to the ℓ_2 loss in Eq. (1), GWD is used to assess the quality of the linear representation $\sum_{m=1}^{M} w^k_m \bar{\mathbf{C}}^m$ for $k \in [\![K]\!]$. However, the observed graphs

often contain noisy edges or miss some edges in real-world applications (Clauset et al., 2008; Xu et al., 2019b; Shi et al., 2019; Piccioli et al., 2022), which leads to the inaccuracies of the PRMs $\mathbf{C}^k$, that is, the deviation between $\mathbf{C}^k$ and the true PRM $\mathbf{C}^{k*}$. Since GWD lacks robustness (Séjourné et al., 2021; Vincent-Cuaz et al., 2022; Tran et al., 2022), the quality of the learned dictionary may be severely affected.

3 Robust Gromov-Wasserstein Discrepancy

To deal with the inaccuracies of PRMs, this section first defines a robust variant of GWD, referred to as RGWD. The properties of RGWD are then rigorously analyzed. We then derive a theoretically guaranteed numerical scheme for calculating RGWD approximately. Due to the limit of space, all proofs can be found in the appendix.

Definition 1 *Given the observed source graph $\mathcal{G}^{\mathrm{s}}$ and the target graph $\mathcal{G}^{\mathrm{t}}$ that can be expressed as $(\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}})$ and $(\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}})$ respectively, RGWD is defined by the solution to the following problem*

$$\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}}), (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}}), \epsilon\right) = \Big(\min_{\mathbf{T}\in\Pi(\mathbf{p}^{\mathrm{s}},\mathbf{p}^{\mathrm{t}})} \max_{\mathbf{E}\in\mathcal{U}_\epsilon} f(\mathbf{T}, \mathbf{E}; \mathbf{C}^{\mathrm{s}}, \mathbf{C}^{\mathrm{t}})\Big)^{\frac{1}{2}},$$

where the objective $f(\cdot)$ is given by

$$f(\mathbf{T}, \mathbf{E}; \mathbf{C}^{\mathrm{s}}, \mathbf{C}^{\mathrm{t}}) = \sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'})^2 T_{ii'} T_{jj'},$$

and the perturbation $\mathbf{E}$ is in the bounded set $\mathcal{U}_\epsilon = \{\mathbf{E} | \mathbf{E} = \mathbf{E}^\top$ and $\|\mathbf{E}\|_\infty \leq \epsilon\}$.

RGWD requires the sought transport plan to have low transportation costs for all perturbation $\mathbf{E}$ in set $\mathcal{U}_\epsilon$. For succinctness, we omit $\mathbf{C}^{\mathrm{s}}$ and $\mathbf{C}^{\mathrm{t}}$ in $f(\mathbf{T}, \mathbf{E}; \mathbf{C}^{\mathrm{s}}, \mathbf{C}^{\mathrm{t}})$ in the following.

3.1 Properties of RGWD

The properties of RGWD are presented as follows. Firstly, although RGWD involves a non-convex non-concave minimax optimization problem, the inner maximization problem has a closed-form solution, which allows an efficient numerical scheme for RGWD. Secondly, RGWD has a lower bound that is achieved if and only if the expressions of compared graphs are identical up to a permutation, which implies RGWD can be employed to evaluate the similarity between one observed graph and its approximation in DL. Finally, RGWD satisfies the triangle inequality, which allows numerous potential applications including clustering (Elkan, 2003; HajKacem et al., 2019), metric learning (Pitis et al., 2019), and Bayesian learning (Moore, 2000; Xiao et al., 2019). More formally, we state the properties in the following theorem.

Theorem 1 *Given two observed graphs $\mathcal{G}^{\mathrm{s}} = (\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}})$ and $\mathcal{G}^{\mathrm{t}} = (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}})$ with n^{s} and n^{t} nodes respectively, RGWD satisfies*

1. *for all $\mathbf{T} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})$, $\mathbf{E}(\mathbf{T}) = [E_{i'j'}(\mathbf{T})]$ where*

$$E_{i'j'}(\mathbf{T}) = \begin{cases} \epsilon, & \text{if } \sum_{i,j=1}^{n^{\mathrm{s}}} T_{ii'} T_{jj'} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'}) \leq 0, \\ -\epsilon, & \text{otherwise.} \end{cases}$$

   solves the inner maximization problem

$$\max_{\mathbf{E}\in\mathcal{U}_\epsilon} f(\mathbf{T}, \mathbf{E}).$$

2. *RGWD is lower bounded, that is,*

$$\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}}), (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}}), \epsilon\right) \geq \epsilon,$$

   where the equality holds if and only if there exists a bijective $\pi^* : [\![n^{\mathrm{s}}]\!] \to [\![n^{\mathrm{t}}]\!]$ such that $p^{\mathrm{s}}_i = p^{\mathrm{t}}_{\pi^*(i)}$ for all $i \in [\![n^{\mathrm{s}}]\!]$ and $C^{\mathrm{s}}_{ij} = C^{\mathrm{t}}_{\pi^*(i)\pi^*(j)}$ for all $i, j \in [\![n^{\mathrm{s}}]\!]$.

3. *The triangle inequality holds for RGWD, i.e.,*

$$\begin{aligned}&\mathrm{RGWD}\left((\mathbf{C}^1,\mathbf{p}^1),(\mathbf{C}^3,\mathbf{p}^3),\epsilon\right)\\ \leq&\,\mathrm{RGWD}\left((\mathbf{C}^1,\mathbf{p}^1),(\mathbf{C}^2,\mathbf{p}^2),\epsilon\right)+\mathrm{RGWD}\left((\mathbf{C}^2,\mathbf{p}^2),(\mathbf{C}^3,\mathbf{p}^3),\epsilon\right).\end{aligned}$$

As is implied by Theorem 1, RGWD does not define a distance between metric-measure spaces. Firstly, the identity axiom is not satisfied. Secondly, the symmetry generally does not hold either, which we exemplify below.

Example 1 (Asymmetry of RGWD) *Consider the case* $\mathbf{p}^{\mathrm{s}}=\mathbf{p}^{\mathrm{t}}=\left[\begin{smallmatrix}0.5\\0.5\end{smallmatrix}\right]$, $\mathbf{C}^{\mathrm{s}}=\left[\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right]$, $\mathbf{C}^{\mathrm{t}}=\left[\begin{smallmatrix}0&4\\4&0\end{smallmatrix}\right]$. *Then* $\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{s}},\mathbf{p}^{\mathrm{s}}),(\mathbf{C}^{\mathrm{t}},\mathbf{p}^{\mathrm{t}}),1\right)=11.5$ *with the solution given by* $\mathbf{T}^*=\left[\begin{smallmatrix}0.25&0.25\\0.25&0.25\end{smallmatrix}\right]$ *and* $\mathbf{E}^*=\left[\begin{smallmatrix}-1&1\\1&-1\end{smallmatrix}\right]$. *In contrast,* $\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{t}},\mathbf{p}^{\mathrm{t}}),(\mathbf{C}^{\mathrm{s}},\mathbf{p}^{\mathrm{s}}),1\right)=10.5$ *with* $\mathbf{T}^*=\left[\begin{smallmatrix}0.25&0.25\\0.25&0.25\end{smallmatrix}\right]$ *and* $\mathbf{E}^*=\left[\begin{smallmatrix}-1&-1\\-1&-1\end{smallmatrix}\right]$. *One has* $\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{s}},\mathbf{p}^{\mathrm{s}}),(\mathbf{C}^{\mathrm{t}},\mathbf{p}^{\mathrm{t}}),1\right)\neq\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{t}},\mathbf{p}^{\mathrm{t}}),(\mathbf{C}^{\mathrm{s}},\mathbf{p}^{\mathrm{s}}),1\right)$

Example 1 showcases that RGWD is asymmetric even if $n^{\mathrm{s}}=n^{\mathrm{t}}$.

3.2 Numerical Scheme of RGWD

We derive a gradient based numerical scheme to solve RGWD by exploiting the property that the inner maximization problem has a closed-form solution, which is summarized in Algorithm 1. In each iteration, $\mathbf{E}_\tau$ that solves the inner problem for current $\mathbf{T}_\tau$ is calculated. Then, the transport plan is updated using the projected gradient descent.

Algorithm 1 Projected Gradient Descent for RGWD

Input: Initialization $\mathbf{T}_0$, step size η, number of iterations N.
Output: Estimated optimal transport plan $\hat{\mathbf{T}}$ and its corresponding perturbation $\hat{\mathbf{E}}$.
for $\tau=0,1,\ldots,N-1$ **do**
  Find $\mathbf{E}_\tau$ that maximizes $f(\mathbf{T}_\tau,\mathbf{E})$.
  Update the transport plan via

$$\mathbf{T}_{\tau+1}=\mathrm{Proj}_{\Pi(\mathbf{p}^{\mathrm{s}},\mathbf{p}^{\mathrm{t}})}\Big(\mathbf{T}_\tau-\eta\nabla_{\mathbf{T}}f(\mathbf{T}_\tau,\mathbf{E}_\tau)\Big),$$

  where the partial gradient takes the form

$$\begin{aligned}\nabla_{\mathbf{T}}f(\mathbf{T}_\tau,\mathbf{E}_\tau)=&2\big(\mathbf{C}^{\mathrm{s}}\odot\mathbf{C}^{\mathrm{s}}\big)\mathbf{T}_\tau\mathbf{1}^{n^{\mathrm{t}}}\mathbf{1}^{n^{\mathrm{t}}\top}2\mathbf{1}^{n^{\mathrm{s}}}\mathbf{1}^{n^{\mathrm{s}}\top}\mathbf{T}_\tau\Big(\big(\mathbf{C}^{\mathrm{t}}+\mathbf{E}_\tau\big)\odot\big(\mathbf{C}^{\mathrm{t}}+\mathbf{E}_\tau\big)\Big)\\&-4\mathbf{C}^{\mathrm{s}}\mathbf{T}_\tau\big(\mathbf{C}^{\mathrm{t}}+\mathbf{E}_\tau\big),\end{aligned}$$

  with $\odot$ denoting the element-wise multiplication.
end for
Pick τ uniformly at random from $\{1,2,\ldots,T\}$.
Set $\hat{\mathbf{T}}\leftarrow\mathbf{T}_\tau$.
Find $\hat{\mathbf{E}}$ that maximizes $f(\hat{\mathbf{T}},\mathbf{E})$.

To present the convergence guarantee of Algorithm 1, we introduce the notion of the Moreau envelope. The stationarity of any function $h(\cdot)$ can be quantified by the norm of the gradient of its Moreau envelope $h_\lambda(\cdot)$ (Rockafellar, 1970; Rafique et al., 2019; Jin et al., 2020),

$$h_\lambda(\mathbf{x})=\min_{\mathbf{x}'}h(\mathbf{x}')+\frac{1}{2\lambda}\|\mathbf{x}-\mathbf{x}'\|^2.$$

The following theorem gives the convergence rate of Algorithm 1 and the proof is deferred to the appendix.

Theorem 2 *Define* $\phi(\cdot)=\max_{\mathbf{E}\in\mathcal{U}_\epsilon}f(\cdot,\mathbf{E})$. *The output* $\hat{\mathbf{T}}$ *of Algorithm 1 with step size* $\eta=\frac{\gamma}{\sqrt{N+1}}$ *satisfies*

$$\mathbb{E}\big[\|\nabla\phi_{1/2l}(\hat{\mathbf{T}})\|^2\big]\leq 2\frac{\phi_{1/2l}(\mathbf{T}_0)-\min_{\mathbf{T}\in\Pi(\mathbf{p}^{\mathrm{s}},\mathbf{p}^{\mathrm{t}})}\phi(\mathbf{T})+lL^2\gamma^2}{\gamma\sqrt{N+1}},$$

where $l = \sqrt{2}\max\{10n^3U_1^2 + 6n^3U_1\epsilon + 4nU_1U_2 + 4n^3\epsilon^2, 6n^2U_1U_2 + 2U_2^2 + 4n^2U_2\epsilon\}$ *and* $L = \sqrt{2}\max\{(4U_1+2\epsilon)U_2^2n^3, 2(2U_1+\epsilon)^2U_2n^3\}$ *with* $n = \max\{n^{\mathrm{s}}, n^{\mathrm{t}}\}$, $U_1 = \max\{\|\mathbf{C}^{\mathrm{s}}\|_\infty, \|\mathbf{C}^{\mathrm{t}}\|_\infty\}$ *and* $U_2 = \max\{\|\mathbf{p}^{\mathrm{t}}\|_2, \max_{\mathbf{T}'\in\Pi(\mathbf{p}^{\mathrm{s}},\mathbf{p}^{\mathrm{t}})}\|\mathbf{T}'\|_F\}$.

When U_1 and ϵ are of the order $\mathcal{O}(\frac{1}{n^3})$, both l and L are of the order $\mathcal{O}(1)$ and Theorem 2 states that an δ-stationary solution can be obtained within $\mathcal{O}(\frac{1}{\delta^2})$ iterations. Note that we can multiply $\mathbf{C}^{\mathrm{s}}$, $\mathbf{C}^{\mathrm{t}}$, and ϵ by the same number without affecting the resulted transport plan.

4 Robust Graph Dictionary Learning

The problem of learning a robust dictionary for graph data is now formulated as follows. Given a dataset of K graphs expressed by $\{(\mathbf{C}^k, \mathbf{p}^k)\}_{k\in[\![K]\!]}$, estimating the optimal dictionary is formalized by

$$\min_{\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]},\{\mathbf{w}^k\}_{k\in[\![K]\!]}} \sum_{k=1}^{K} \mathrm{RGWD}^2\left(\Big(\sum_{m=1}^{M} w_m^k \bar{\mathbf{C}}^m, \bar{\mathbf{p}}\Big), (\mathbf{C}^k, \mathbf{p}^k), \epsilon\right) - \lambda\|\mathbf{w}^k\|^2, \tag{2}$$

where $\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}$ and $\{\mathbf{w}^k\}_{k\in[\![K]\!]}$ are the dictionary and graph embeddings respectively, and $\bar{\mathbf{p}}$ is obtained by sorting and averaging $\{\mathbf{p}^k\}_{k\in[\![K]\!]}$ following Xu et al. (2019a).

To resolve (2), we propose a nested iterative optimization algorithm that is summarized in Algorithm 2. The main idea is that the dictionary and embeddings are updated alternatingly. We discuss some crucial details below.

Algorithm 2 Robust Graph Dictionary Learning (RGDL)

1: **Input:** The dataset $\{\mathbf{C}^k, \mathbf{p}^k\}_{k\in[\![K]\!]}$, the initial dictionary $\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}$, the number of iterations T, mini-batch size b.
2: **Output:** The learned dictionary $\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}$.
3: **for** $t = 0, 1, \ldots, T-1$ **do**
4: Sample a mini-batch of graphs whose indices are denoted by $\mathcal{B}$ such that $|\mathcal{B}| = b$.
5: **for** $k \in \mathcal{B}$ **do**
6: Initialize $\mathbf{w}^k = \frac{1}{M}\mathbf{1}^M$ and $\mathbf{T}^k = \bar{\mathbf{p}}\mathbf{p}^{k\top}$.
7: **repeat**
8: Calculate ($\mathbf{T}^k$, $\mathbf{E}^k$) via Algorithm 1 with fixed $\mathbf{w}^k$.
9: Compute $\mathbf{w}^k$ solving (4) for the fixed $\mathbf{T}^k$ and $\mathbf{E}^k$ with conditional gradient.
10: **until** Convergence
11: **end for**
12: Update the atom $\bar{\mathbf{C}}^m$ for $m \in [\![M]\!]$ with stochastic gradient $\hat{\nabla}_{\bar{\mathbf{C}}^m}$ which has the form

$$\hat{\nabla}_{\bar{\mathbf{C}}^m} = \frac{2}{b}\sum_{k\in\mathcal{B}} w_m^k \Big(\sum_{m'=1}^{M} w_{m'}^k \bar{\mathbf{C}}^{m'} \odot \bar{\mathbf{p}}\bar{\mathbf{p}}^\top - \mathbf{T}^k(\mathbf{C}^k + \mathbf{E}^k)\mathbf{T}^{k\top}\Big). \tag{3}$$

13: **end for**

Solving $\mathbf{w}^k$. We now formulate the problem of obtaining the embedding of the kth graph $\mathcal{G}^k$ when the dictionary is fixed and the PRM is inaccurate. Given dictionary $\{\bar{\mathbf{C}}^m\}_{m\in[\![M]\!]}$ where each $\bar{\mathbf{C}}^m \in \mathbb{R}^{n^{\mathrm{a}}\times n^{\mathrm{a}}}$, the embedding of $\mathcal{G}^k$ expressed by $(\mathbf{C}^k, \mathbf{p}^k)$ is calculated by solving

$$\min_{\mathbf{w}^k\in\Delta^M} \mathrm{RGWD}^2\left(\Big(\sum_{m=1}^{M} w_m^k \bar{\mathbf{C}}^m, \bar{\mathbf{p}}\Big), (\mathbf{C}^k, \mathbf{p}^k), \epsilon\right) - \lambda\|\mathbf{w}^k\|^2, \tag{4}$$

where $\lambda \geq 0$ induces a negative quadratic regularization promoting sparsity on the simplex (Li et al., 2020; Vincent-Cuaz et al., 2021). When $\mathbf{w}^k$ is fixed, updating $\mathbf{T}^k$ and $\mathbf{E}^k$ can be solved by Algorithm 1 whose convergence is guaranteed by Theorem 2. For fixed $\mathbf{T}^k$ and $\mathbf{E}^k$, the problem of updating $\mathbf{w}^k$ is a non-convex problem that can be tackled by a conditional gradient algorithm. Note that for non-convex problems, the conditional gradient algorithm is proved to converge to a local stationary point (Lacoste-Julien, 2016). Such a procedure is described from Line 5 to Line 11 in Algorithm 2, which we observe converges within tens of iterations empirically.

Stochastic updates. To enhance computational efficiency, each atom is updated with stochastic estimates of the gradient. At each stochastic update, b embedding learning problems are solved independently for the current dictionary using the procedure stated above, where b is the size of the sampled mini-batch. Each atom is then updated using the stochastic gradient given in Eq. (3). Note that the symmetry of each atom is preserved as long as the initialized atom is symmetric, since the stochastic gradients are symmetric.

5 EXPERIMENTS

This section provides empirical evidence that RGDL performs well in the unsupervised graph clustering task on both synthetic and real-world datasets. The heat kernel matrix is employed for the PRM since it captures both global and local topology and achieves good performance in many tasks (Donnat et al., 2018; Tsitsulin et al., 2018; Chowdhury and Needham, 2021).

5.1 SIMULATED DATASETS

We first test RGDL in the graph clustering task on datasets simulated according to the well-studied Stochastic Block Model (SBM) (Holland et al., 1983; Wang and Wong, 1987) and demonstrate that RGDL outperforms GDL (Vincent-Cuaz et al., 2021) when PRMs are inaccurate.

Table 1: Average (stdev) ARI scores for the first scenario of synthetic datasets.

	balanced			**unbalanced**		
σ	0.05	0.10	0.15	0.05	0.10	0.15
GDL	0.119(0.017)	0.031(0.012)	0.016(0.006)	0.049(0.019)	0.018(0.004)	0.018(0.001)
RGDL(ϵ=0.01)	0.316(0.005)	0.161(0.005)	0.052(0.002)	0.246(0.013)	0.039(0.009)	0.024 (0.001)
RGDL(ϵ=0.1)	0.853(0.003)	0.756(0.018)	0.439(0.015)	0.765(0.022)	0.694(0.046)	0.499(0.016)
RGDL(ϵ=0.2)	**0.975(0.025)**	0.879(0.023)	0.736(0.020)	0.866(0.023)	0.815(0.028)	0.770(0.016)
RGDL(ϵ=0.3)	**0.975(0.025)**	0.879(0.023)	0.869(0.013)	**0.943(0.001)**	0.916(0.027)	0.848(0.061)
RGDL(ϵ=10)	**0.975(0.025)**	**0.950(0.000)**	**0.950(0.000)**	**0.943(0.001)**	**0.943(0.001)**	**0.943(0.001)**
RGDL(ϵ=30)	0.781(0.046)	0.779(0.070)	0.728(0.085)	0.723(0.067)	0.698(0.057)	0.666(0.040)

Table 2: Average (stdev) ARI scores for the second scenario of synthetic datasets.

	balanced			**unbalanced**		
ρ	0.00	0.05	0.10	0.00	0.05	0.1
GDL	0.260(0.020)	0.187(0.013)	0.024(0.004)	0.152(0.008)	0.018(0.001)	0.005(0.001)
RGDL(ϵ=0.01)	0.451(0.014)	0.449(0.016)	0.449(0.016)	0.401(0.002)	0.401(0.002)	0.399(0.000)
RGDL(ϵ=0.1)	**1.000(0.000)**	**1.000(0.000)**	0.975(0.025)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
RGDL(ϵ=0.2)	**1.000(0.000)**	**1.000(0.000)**	0.975(0.025)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
RGDL(ϵ=0.3)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
RGDL(ϵ=10)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
RGDL(ϵ=30)	0.896(0.070)	0.888(0.080)	0.857(0.044)	0.864(0.057)	0.827(0.043)	0.816(0.074)

Dataset generation. We consider two scenarios of inaccuracies. In the first scenario (S1), Gaussian noise is added into the heat kernel matrix of each graph. More specifically, denoting the heat kernel matrix of the kth graph as $\mathbf{C}^{k*}$ for $k \in [\![K]\!]$, the PRM available to DL methods is $\mathbf{C}^k = \mathbf{C}^{k*} + \mathbf{Z} + \mathbf{Z}^\top$ where each entry Z_{ij} of $\mathbf{Z}$ is sampled from the Gaussian distribution $\mathcal{N}(0, \sigma)$. In the second scenario (S2), we randomly add $\rho|\mathcal{E}|$ edges into the graph and then randomly remove $\rho|\mathcal{E}|$ edges while keeping the graph connected, where $\mathcal{E}$ is the edge set of the graph. The heat kernel matrix is then constructed for the modified graph. Such two scenarios allow us to study the performance of RGDL against different scales of inaccuracies. In both S1 and S2, we generate two datasets, both of which involve three generative structures (also used to label graphs): dense (only one community), two communities, and three communities. We fix $p = 0.1$ as the probability of inter-community connectivity and $1 - p$ as the probability of intra-community connectivity. The first dataset includes 20 graphs for each generative structure and thus is referred to as the balanced dataset. The second dataset consists of 12, 18, and 30 graphs for the three generative structures respectively, and is hence named as the unbalanced dataset. The number of graph nodes is uniformly sampled from $[30, 50]$. The magnitude of the observed PRM $\mathbf{C}^k$ satisfies $\|\mathbf{C}^k\|_\infty \leq 15$.

Evaluating the performance. The learned embeddings of the graphs are used as input for a k-means algorithm to cluster graphs. We use the well-known Adjusted Rand Index (ARI) (Hubert and Arabie, 1985; Steinley, 2004), to evaluate the quality of clustering by comparing it with the graph labels. RGDL with varied ϵ is compared against GDL. Both RGDL and GDL use three atoms which are $\mathbb{R}^{6\times 6}$ matrices. We run each method for 5 times and report the averaged ARI scores and the standard deviations. Experimental results reported in Table 1 and Table 2 demonstrate RGDL outperforms GDL significantly.

Influence of ϵ. RGDL with moderate ϵ values outperforms GDL by a large margin and is more robust to the noise. Even when ϵ is relatively small (ϵ=0.01), RGDL achieves better performance than GDL. Increasing ϵ within a suitable range can boost ARI and RGDL is not sensitive to the choice of ϵ. If ϵ becomes too large, the performance of RGDL slowly decreases. In practice, when a small quantity of data labels are available, ϵ can be chosen according to the performance on this small subset of data.

5.2 Real-world Datasets

We further use RGDL to cluster real-world graphs and compare it with state-of-the-art graph clustering methods. We consider widely utilized benchmark datasets including MUTAG (Debnath et al., 1991), IMDB-B, and IMDB-M (Yanardag and Vishwanathan, 2015), whose statistics are given in Table 3. The labels of the graphs are employed as the ground truth to evaluate the estimated clustering results.

Table 3: Datasets used in our experiments.

Datasets	#graphs	#classes	min #nodes	max #nodes	min #edges	max #edges
IMDB-B	1,000	2	12	136	26	1249
IMDB-M	1,500	2	11	363	12	1467
MUTAG	188	2	10	28	10	33

Table 4: Average (stdev) ARI scores on real-world datasets.

Datasets	**GDL**	**GWF**	**SC**	**RGDL**
IMDB-B	0.002(0.000)	0.001(0.000)	0.001(0.000)	**0.076(0.001)**
IMDB-M	0.005(0.001)	0.004(0.001)	0.014(0.000)	**0.029(0.000)**
MUTAG	0.377(0.006)	0.404(0.015)	0.005(0.000)	**0.458(0.007)**

Experimental setup. RGDL is compared against the following state-of-the-art graph clustering methods: (i) GDL (Vincent-Cuaz et al., 2021) learns graph dictionaries via GWD; (ii) Gromov-Wasserstein Factorization (GWF) (Xu, 2020) that approximates graphs via GW barycenters; (iii) Spectral Clustering (SC) of Shi and Malik (2000); Stella and Shi (2003) applied to the matrix with each entry storing the GWD between two graphs. For each dataset, the size of the atoms is set as the median of the numbers of graph nodes following Vincent-Cuaz et al. (2021). The number of atoms M is set as $M = \beta(\#\,\text{classes})$ where β is chosen from $\{2, 3, 4, 5\}$. For GDL[1] and GWF [2], we use their respective open-source implementations. For SC, POT's implementation [3] is utilized to calculate the GWD between observed graphs.

Results. The experimental results on real-world graphs are reported in Table 4. RGDL outperforms baselines on all datasets, which implies that the observed graphs contain structural noise. Since GDL, GWF, and RGDL can output graph embeddings, we further illustrate the embeddings generated by them respectively based on PCA. As is shown in Figure 1, the embeddings of the two types of the graphs are less likely to be mixed together, which explains why RGDL achieves higher ARI values.

[1] https://github.com/cedricvincentcuaz/GDL

[2] https://github.com/HongtengXu/Relational-Factorization-Model

[3] https://pythonot.github.io/gen_modules/ot.gromov.html

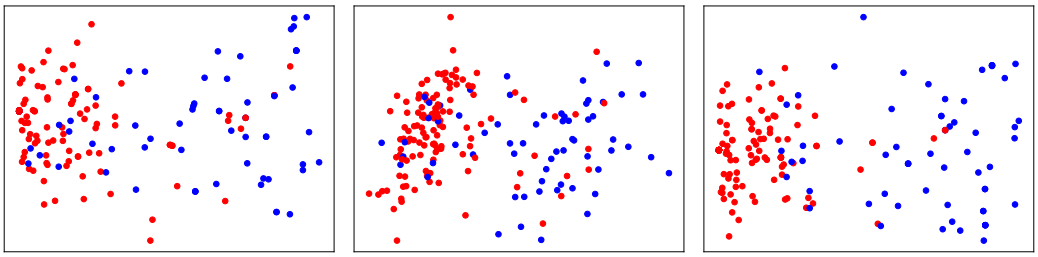

Figure 1: PCA-based visualization of embeddings produced by GDL (left), GWF (middle), and RGDL (right) respectively for the graphs in MUTAG dataset. The color of each point indicate the type of the corresponding graph. RGDL achieves better clustering results.

6 Related Work

Unbalanced OT. Enhancing the robustness of the optimal transport plan has received wide attention recently (Balaji et al., 2020; Mukherjee et al., 2021; Le et al., 2021; Nietert et al., 2022; Chapel et al., 2020; Séjourné et al., 2021; Vincent-Cuaz et al., 2022). Originally, robust variants of classical OT are proposed to compare distributions supported on the *same* metric space (Balaji et al., 2020; Mukherjee et al., 2021; Le et al., 2021; Nietert et al., 2022), which model the noise as outlier supports and reduce the influence of outlier supports by allowing mass destruction and creation. Following the same spirit, variants of the GW distance that also relax the marginal constraints are proposed (Chapel et al., 2020; Séjourné et al., 2021; Vincent-Cuaz et al., 2022). However, these methods do not take the inaccuracies of the pairwise distances/similarities into account. The proposed RGWD aims to handle such cases.

Graph representation learning. Graph neural network-based methods learn the representations of graphs in an end-to-end manner and have shown impressive performance in many tasks (Scarselli et al., 2008; Zhang et al., 2018; Lee et al., 2018; Errica et al., 2019), which however requires a large amount of labeled data. Another family of methods that uses graph representations implicitly is referred to as graph kernels (Shervashidze et al., 2009; Vishwanathan et al., 2010). Dictionary learning can learn representations of graphs in an unsupervised way and obtain graph embeddings which are useful in downstream tasks.

Non-linear combination of atoms. Classic DL methods are linear in the sense that they attempt to approximate each vectorized datum by a linear combination of a few basis elements. Recently, non-linear operations are also considered. In order to exploit the non-linear nature of data, Autoencoder-based methods encode them to low-dimensional vectors using neural networks, and decode data with another neural network (Hinton and Salakhutdinov, 2006; Hu and Tan, 2018). Another family of methods replace the linear combinations by geodesic interpolations (Boissard et al., 2011; Bigot et al., 2013; Seguy and Cuturi, 2015; Schmitz et al., 2018). More closely related to our work, Xu 2020 proposes to approximate graphs via the GW barycenter of graph atoms, which however involves a complicated and computational demanding optimization problem.

7 Conclusion

In this paper, we propose a novel graph dictionary learning algorithm that is robust to the structural noise of graphs. We first propose a robust variant of GWD, referred to as RGWD, which involves a minimax optimization problem. Exploiting the fact that the inner maximization problem has a closed-form solution, an efficient numerical scheme is derived. We further prove that RGWD is always lower bounded and the lower bound is achieved if and only if two compared graphs are isomorphic. Based on RGWD, a robust dictionary learning algorithm for graphs called RGDL is derived to learn atoms from noisy graph data. Numerical results on both simulated and real-world datasets demonstrate that RGDL achieves good performance in the presence of structural noise.

REFERENCES

Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 111–122. World Scientific, 2018.

Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644, 2011.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic pca in the wasserstein space. *arXiv preprint arXiv:1307.7721*, 2013.

Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with wasserstein metrics. *arXiv preprint arXiv:1111.5927*, 2011.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Laetitia Chapel, Gilles Gasso, et al. Partial optimal tranport with applications on positive-unlabeled learning. In *NeurIPS*, 2020.

Samir Chowdhury and Facundo Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

Samir Chowdhury and Tom Needham. Generalized spectral clustering via gromov-wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pages 712–720. PMLR, 2021.

Aaron Clauset, Cristopher Moore, and M E J Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. Learning structural node embeddings via diffusion wavelets. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1320–1329, 2018.

Charles Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th international conference on Machine Learning (ICML-03)*, pages 147–153, 2003.

Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. In *International Conference on Learning Representations*, 2019.

Zhizhao Feng, Meng Yang, Lei Zhang, Yan Liu, and David Zhang. Joint discriminative dimensionality reduction and dictionary learning for face recognition. *Pattern Recognition*, 46(8):2134–2143, 2013.

Mohamed Aymen Ben HajKacem, Chiheb-Eddine Ben N'Cir, and Nadia Essoussi. Overview of scalable partitional methods for big data clustering. In *Clustering Methods for Big Data Analytics*, pages 1–23. Springer, 2019.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Junlin Hu and Yap-Peng Tan. Nonlinear dictionary learning with application to image classification. *Pattern Recognition*, 75:282–291, 2018.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Wenhao Jiang, Feiping Nie, and Heng Huang. Robust dictionary learning with capped l1-norm. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pages 4880–4889. PMLR, 2020.

Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30, 2017.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Khang Le, Huy Nguyen, Quang M Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. *Advances in Neural Information Processing Systems*, 34:21947–21959, 2021.

John Boaz Lee, Ryan Rossi, and Xiangnan Kong. Graph classification using structural attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1666–1674, 2018.

Ping Li, Syama Sundar Rangapuram, and Martin Slawski. Methods for sparse and low-rank recovery under simplex constraints. *Statistica Sinica*, 30(2):557–577, 2020.

Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis Bach. Supervised dictionary learning. *Advances in neural information processing systems*, 21, 2008.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696, 2009.

Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

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

Andrew W Moore. The anchors hierarchy: using the triangle inequality to survive high dimensional data. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 397–405, 2000.

Debarghya Mukherjee, Aritra Guha, Justin M Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. PMLR, 2021.

Navid Naderializadeh, Mark Eisen, and Alejandro Ribeiro. Wireless power control via counterfactual optimization of graph neural networks. In *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5. IEEE, 2020.

Sloan Nietert, Ziv Goldfeld, and Rachel Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 11691–11719. PMLR, 2022.

Gabriel Peyre and Marco Cuturi. Computational optimal transport. *arXiv: Machine Learning*, 2018.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *ICML*, 2016.

Giovanni Piccioli, Guilhem Semerjian, Gabriele Sicuro, and Lenka Zdeborová. Aligning random graphs with a sub-tree similarity message-passing algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(6):063401, 2022.

Silviu Pitis, Harris Chan, Kiarash Jamali, and Jimmy Ba. An inductive bias for distances: Neural nets that respect the triangle inequality. In *International Conference on Learning Representations*, 2019.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. 2019.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766, 2007.

Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3501–3508. IEEE, 2010.

R Tyrrell Rockafellar. Convex analysis princeton university press. *Princeton, NJ*, 1970.

Hamidreza Sadreazami, Arash Mohammadi, Amir Asif, and Konstantinos N Plataniotis. Distributed-graph-based statistical approach for intrusion detection in cyber-physical systems. *IEEE Transactions on Signal and Information Processing over Networks*, 4(1):137–147, 2017.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. *Advances in Neural Information Processing Systems*, 28, 2015.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics*, pages 488–495. PMLR, 2009.

Chuan Shi, Binbin Hu, Wayne Xin Zhao, and Philip S Yu. Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 31 (2):357–370, 2019.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Pablo Sprechmann and Guillermo Sapiro. Dictionary learning and sparse coding for unsupervised clustering. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 2042–2045. IEEE, 2010.

Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3): 386, 2004.

X Yu Stella and Jianbo Shi. Multiclass spectral clustering. In *Computer Vision, IEEE International Conference on*, volume 2, pages 313–313. IEEE Computer Society, 2003.

Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *ICML*, 2019.

Ivana Tošić and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2): 27–38, 2011.

Quang Huy Tran, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, and Ritambhara Singh. Unbalanced co-optimal transport. *arXiv preprint arXiv:2205.14923*, 2022.

Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. Netlsd: hearing the shape of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2347–2356, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

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.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed gromov-wasserstein divergence and applications on graphs. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=RShaMexjc-x`.

S Vishwanathan, N Schraudolph, R Kondor, and KM Borgwardt. Graph kernels. the journal ofmachine learning research. 2010.

Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

Xian Wei, Hao Shen, Yuanxiang Li, Xuan Tang, Fengxiang Wang, Martin Kleinsteuber, and Yi Lu Murphey. Reconstructible nonlinear dimensionality reduction via joint dictionary learning. *IEEE transactions on neural networks and learning systems*, 30(1):175–189, 2018.

Teng Xiao, Jiaxin Ren, Zaiqiao Meng, Huan Sun, and Shangsong Liang. Dynamic bayesian metric learning for personalized product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1693–1702, 2019.

Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. In *NeurIPS*, 2019a.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *ICML*, 2019b.

Hongtengl Xu. Gromov-wasserstein factorization models for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6478–6485, 2020.

Pinar Yanardag and S Vishwanathan. Deep graph kernels in: Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining, 1365–1374. *ACM, New York*, 2015.

Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Appendix

In the appendix, we provide omitted proofs in the main paper.

Theorem 1 *Given two observed graphs $\mathcal{G}^{\mathrm{s}} = (\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}})$ and $\mathcal{G}^{\mathrm{t}} = (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}})$ with n^{s} and n^{t} nodes respectively, RGWD satisfies*

1. *for all $\mathbf{T} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})$, $\mathbf{E}(\mathbf{T}) = [E_{i'j'}(\mathbf{T})]$ where*

$$E_{i'j'}(\mathbf{T}) = \begin{cases} \epsilon, & \text{if } \sum_{i,j=1}^{n^{\mathrm{s}}} T_{ii'} T_{jj'} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'}) \leq 0, \\ -\epsilon, & \text{otherwise.} \end{cases}$$

   solves the inner maximization problem

$$\max_{\mathbf{E} \in \mathcal{U}_\epsilon} f(\mathbf{T}, \mathbf{E}).$$

2. *RGWD is lower bounded, that is,*

$$\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}}), (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}}), \epsilon\right) \geq \epsilon,$$

   where the equality holds if and only if there exists a bijective $\pi^* : [\![n^{\mathrm{s}}]\!] \to [\![n^{\mathrm{t}}]\!]$ such that $p^{\mathrm{s}}_i = p^{\mathrm{t}}_{\pi^*(i)}$ for all $i \in [\![n^{\mathrm{s}}]\!]$ and $C^{\mathrm{s}}_{ij} = C^{\mathrm{t}}_{\pi^*(i)\pi^*(j)}$ for all $i, j \in [\![n^{\mathrm{s}}]\!]$.

3. *The triangle inequality holds for RGWD, i.e.,*

$$\begin{aligned} &\mathrm{RGWD}\left((\mathbf{C}^1, \mathbf{p}^1), (\mathbf{C}^3, \mathbf{p}^3), \epsilon\right) \\ \leq &\mathrm{RGWD}\left((\mathbf{C}^1, \mathbf{p}^1), (\mathbf{C}^2, \mathbf{p}^2), \epsilon\right) + \mathrm{RGWD}\left((\mathbf{C}^2, \mathbf{p}^2), (\mathbf{C}^3, \mathbf{p}^3), \epsilon\right). \end{aligned}$$

Proof: (**i**) The objective can be rewritten as follows,

$$\begin{aligned} &\sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'})^2 T_{ii'} T_{jj'} \\ = &\sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'})^2 T_{ii'} T_{jj'} + \sum_{i',j'=1}^{n^{\mathrm{t}}} \left(\sum_{i,j=1}^{n^{\mathrm{s}}} T_{ii'} T_{jj'} E^2_{i'j'} - 2 \sum_{i,j=1}^{n^{\mathrm{s}}} T_{ii'} T_{jj'} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'}) E_{i'j'} \right) \\ = &\sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'})^2 T_{ii'} T_{jj'} + \sum_{i',j'=1}^{n^{\mathrm{t}}} \left(p^{\mathrm{t}}_{i'} p^{\mathrm{t}}_{j'} E^2_{i'j'} - 2 \sum_{i,j=1}^{n^{\mathrm{s}}} T_{ii'} T_{jj'} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'}) E_{i'j'} \right), \end{aligned} \tag{5}$$

which by the property of quadratic functions yields the closed-form solution $\mathbf{E}(\mathbf{T}) = [E_{i'j'}(\mathbf{T})]$, where

$$E_{i'j'}(\mathbf{T}) = \begin{cases} \epsilon, & \text{if } \sum_{i,j=1}^{n^{\mathrm{s}}} T_{ii'} T_{jj'} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'}) \leq 0, \\ -\epsilon, & \text{otherwise.} \end{cases}$$

It is easy to verify that the such a choice guarantees the symmetry of $\mathbf{E}(\mathbf{T})$.

(**ii**) We now prove the lower boundedness. By Eq. (5), we have

$$\min_{\mathbf{T} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})} \max_{\mathbf{E} \in \mathcal{U}_\epsilon} \sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'})^2 T_{ii'} T_{jj'} \geq \min_{\mathbf{T} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})} \sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} T_{ii'} T_{jj'} \epsilon^2 = \epsilon^2.$$

Note that when there exists a bijective $\pi^* : [\![n^{\mathrm{s}}]\!] \to [\![n^{\mathrm{t}}]\!]$ such that $p^{\mathrm{s}}_i = p^{\mathrm{t}}_{\pi^*(i)}$ for all $i \in [\![n^{\mathrm{s}}]\!]$ and $C^{\mathrm{s}}_{ij} = C^{\mathrm{t}}_{\pi^*(i)\pi^*(j)}$ for all $i, j \in [\![n^{\mathrm{s}}]\!]$, choosing the transport plan $\tilde{\mathbf{T}} = [\tilde{T}_{ii'}]$ where

$$\tilde{T}_{ii'} = \begin{cases} p^{\mathrm{s}}_i, & \text{if } i' = \pi^*(i), \\ 0, & \text{otherwise,} \end{cases}$$

we have for all $i', j' \in [\![n^{\mathrm{t}}]\!]$,

$$\sum_{i,j=1}^{n^{\mathrm{s}}} T_{ii'} T_{jj'} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'}) = T_{\pi^{*-1}(i')i'} T_{\pi^{*-1}(j')j'} \left(C^{\mathrm{s}}_{\pi^{*-1}(i')\pi^{*-1}(j')} - C^{\mathrm{t}}_{i'j'} \right) = 0,$$

which implies that $E_{i'j'}(\mathbf{T}) = \epsilon$ for all $i', j' \in [\![n^{\mathrm{t}}]\!]$. We then have

$$\sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'})^2 T_{ii'} T_{jj'} = \sum_{i,j=1}^{n^{\mathrm{s}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{\pi^*(i)\pi^*(j)} - \epsilon)^2 \tilde{T}_{i\pi^*(i)} T_{j\pi^*(j)} = \epsilon^2.$$

Therefore, in such a case, $\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}}), (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}}), \epsilon\right) = \epsilon$. On the other hand, when such a bijective does not exist,

$$\mathrm{RGWD}\left((\mathbf{C}^{\mathrm{s}}, \mathbf{p}^{\mathrm{s}}), (\mathbf{C}^{\mathrm{t}}, \mathbf{p}^{\mathrm{t}}), \epsilon\right) \geq \sqrt{\epsilon^2 + \min_{\mathbf{T} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})} \sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'})^2 T_{ii'} T_{jj'}} > \epsilon,$$

where the strict inequality is due to the fact that $\sum_{i,j=1}^{n^{\mathrm{s}}} \sum_{i',j'=1}^{n^{\mathrm{t}}} (C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'})^2 T_{ii'} T_{jj'} > 0$. (**iii**) Finally, we prove the triangle inequality. Given tuples $(\mathbf{C}^1, \mathbf{p}^1)$, $(\mathbf{C}^2, \mathbf{p}^2)$, and $(\mathbf{C}^3, \mathbf{p}^3)$ which have the node numbers n^1, n^2, and n^3 respectively, let $(\mathbf{T}^{*12}, \mathbf{E}^{*12})$, $(\mathbf{T}^{*23}, \mathbf{E}^{*23})$, and $(\mathbf{T}^{*13}, \mathbf{E}^{*13})$ be the solutions of $\mathrm{RGWD}\left((\mathbf{C}^1, \mathbf{p}^1), (\mathbf{C}^2, \mathbf{p}^2), \epsilon\right)$, $\mathrm{RGWD}\left((\mathbf{C}^2, \mathbf{p}^2), (\mathbf{C}^3, \mathbf{p}^3), \epsilon\right)$, and $\mathrm{RGWD}\left((\mathbf{C}^1, \mathbf{p}^1), (\mathbf{C}^3, \mathbf{p}^3), \epsilon\right)$. Define $\mathbf{T}^{13} = [T^{13}_{i_1 i_3}]$ where

$$T^{13}_{i_1 i_3} = \sum_{i_2=1}^{n^2} \frac{T^{*12}_{i_1 i_2} T^{*23}_{i_2 i_3}}{p^2_{i_2}}.$$

Then we have

$$\begin{aligned}
&\mathrm{RGWD}\left((\mathbf{C}^1, \mathbf{p}^1), (\mathbf{C}^3, \mathbf{p}^3), \epsilon\right) \\
\leq &\sqrt{\sum_{i_1,j_1=1}^{n^1} \sum_{i_3,j_3=1}^{n^3} \left(C^1_{i_1 j_1} - C^3_{i_3 j_3} - E^{*13}_{i_3 j_3}\right)^2 T^{13}_{i_1 i_3} T^{13}_{j_1 j_3}} \\
= &\sqrt{\sum_{i_1,j_1=1}^{n^1} \sum_{i_3,j_3=1}^{n^3} \left(C^1_{i_1 j_1} - C^3_{i_3 j_3} - E^{*13}_{i_3 j_3}\right)^2 \sum_{i_2=1}^{n^2} \frac{T^{*12}_{i_1 i_2} T^{*23}_{i_2 i_3}}{p^2_{i_2}} \sum_{j_2=1}^{n^2} \frac{T^{*12}_{j_1 j_2} T^{*23}_{j_2 j_3}}{p^2_{j_2}}} \\
= &\sqrt{\sum_{i_1,j_1=1}^{n^1} \sum_{i_2,j_2=1}^{n^2} \sum_{i_3,j_3=1}^{n^3} \left(C^1_{i_1 j_1} - C^2_{i_2 j_2} + C^2_{i_2 j_2} - C^3_{i_3 j_3} - E^{*13}_{i_3 j_3}\right)^2 \frac{T^{*12}_{i_1 i_2} T^{*23}_{i_2 i_3}}{p^2_{i_2}} \frac{T^{*12}_{j_1 j_2} T^{*23}_{j_2 j_3}}{p^2_{j_2}}} \\
\leq &\sqrt{\sum_{i_1,j_1=1}^{n^1} \sum_{i_2,j_2=1}^{n^2} \sum_{i_3,j_3=1}^{n^3} \left(C^1_{i_1 j_1} - C^2_{i_2 j_2}\right)^2 \frac{T^{*12}_{i_1 i_2} T^{*23}_{i_2 i_3}}{p^2_{i_2}} \frac{T^{*12}_{j_1 j_2} T^{*23}_{j_2 j_3}}{p^2_{j_2}}} \\
+ &\sqrt{\sum_{i_1,j_1=1}^{n^1} \sum_{i_2,j_2=1}^{n^2} \sum_{i_3,j_3=1}^{n^3} \left(C^2_{i_2 j_2} - C^3_{i_3 j_3} - E^{*13}_{i_3 j_3}\right)^2 \frac{T^{*12}_{i_1 i_2} T^{*23}_{i_2 i_3}}{p^2_{i_2}} \frac{T^{*12}_{j_1 j_2} T^{*23}_{j_2 j_3}}{p^2_{j_2}}} \\
= &\sqrt{\sum_{i_1,j_1=1}^{n^1} \sum_{i_2,j_2=1}^{n^2} \left(C^1_{i_1 j_1} - C^2_{i_2 j_2}\right)^2 T^{*12}_{i_1 i_2} T^{*12}_{j_1 j_2}} + \sqrt{\sum_{i_2,j_2=1}^{n^2} \sum_{i_3,j_3=1}^{n^3} \left(C^2_{i_2 j_2} - C^3_{i_3 j_3} - E^{*13}_{i_3 j_3}\right)^2 T^{*23}_{i_2 i_3} T^{*23}_{j_2 j_3}} \\
\leq &\,\mathrm{RGWD}\left((\mathbf{C}^1, \mathbf{p}^1), (\mathbf{C}^2, \mathbf{p}^2), \epsilon\right) + \mathrm{RGWD}\left((\mathbf{C}^2, \mathbf{p}^2), (\mathbf{C}^3, \mathbf{p}^3), \epsilon\right).
\end{aligned}$$

■

To prove Theorem 2, we require the following lemma.

Lemma 3 *$f(\cdot)$ is l-smooth and L-Lipschitz, where $l = \sqrt{2}\max\{10n^3U_1^2 + 6n^3U_1\epsilon + 4nU_1U_2 + 4n^3\epsilon^2, 6n^2U_1U_2 + 2U_2^2 + 4n^2U_2\epsilon\}$ and $L = \sqrt{2}\max\{(4U_1 + 2\epsilon)U_2^2 n^3, 2(2U_1 + \epsilon)^2 U_2 n^3\}$ with $n = \max\{n^{\mathrm{s}}, n^{\mathrm{t}}\}$, $U_1 = \max\{\|\mathbf{C}^{\mathrm{s}}\|_\infty, \|\mathbf{C}^{\mathrm{t}}\|_\infty\}$ and $U_2 = \max\{\|\mathbf{p}^{\mathrm{t}}\|_2, \max_{\mathbf{T}' \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})} \|\mathbf{T}'\|_F\}$.*

Proof: (**i**) We first prove that $f(\cdot)$ is L-Lipschitz. For all $\mathbf{T}, \mathbf{T}' \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})$ and $\mathbf{E}, \mathbf{E}' \in \mathcal{U}_\epsilon$,

$$\begin{aligned}
&\big|f(\mathbf{T},\mathbf{E}) - f(\mathbf{T}',\mathbf{E}')\big| \\
\leq&\Big|\sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'})^2 T_{ii'}T_{jj'} - \sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T_{jj'}\Big| \\
+&\Big|\sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T_{jj'} - \sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T'_{jj'}\Big| \\
+&\Big|\sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T'_{jj'} - \sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T'_{ii'}T'_{jj'}\Big|.
\end{aligned}$$

For the first term,

$$\begin{aligned}
&\Big|\sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'})^2 T_{ii'}T_{jj'} - \sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T_{jj'}\Big| \\
=&\Big|\sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'} + C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})(E'_{i'j'} - E_{i'j'})T_{ii'}T_{jj'}\Big| \\
\leq&\sum_{iji'j'}\big|C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E_{i'j'} + C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'}\big|\big|E'_{i'j'} - E_{i'j'}\big|T_{ii'}T_{jj'} \\
\leq&(4U_1 + 2\epsilon)U_2^2 n^2 \sum_{i'j'}\big|E'_{i'j'} - E_{i'j'}\big|.
\end{aligned}$$

For the second term,

$$\begin{aligned}
&\Big|\sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T_{jj'} - \sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T'_{jj'}\Big| \\
\leq&\sum_{iji'j'}|C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'}|^2\big|T_{ii'}\big|\big|T_{jj'} - T'_{jj'}\big| \\
\leq&(2U_1 + \epsilon)^2 U_2 \sum_{iji'j'}\big|T_{jj'} - T'_{jj'}\big| \\
\leq&(2U_1 + \epsilon)^2 U_2 n^2\big|T_{jj'} - T'_{jj'}\big|.
\end{aligned}$$

For the third term,

$$\begin{aligned}
&\Big|\sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T_{ii'}T'_{jj'} - \sum_{iji'j'}(C^{\mathrm{s}}_{ij} - C^{\mathrm{t}}_{i'j'} - E'_{i'j'})^2 T'_{ii'}T'_{jj'}\Big| \\
\leq&(2U_1 + \epsilon)^2 U_2 n^2\big|T_{jj'} - T'_{jj'}\big|.
\end{aligned}$$

Combining the three relations above, we have

$$\begin{aligned}
&\big|f(\mathbf{T},\mathbf{E}) - f(\mathbf{T}',\mathbf{E}')\big| \\
\leq&\max\Big\{(4U_1 + 2\epsilon)U_2^2 n^2, (2U_1 + \epsilon)^2 U_2 n^2\Big\}\Big(\sum_{i'j'}|E'_{i'j'} - E_{i'j'}| + \sum_{jj'}|T_{jj'} - T'_{jj'}|\Big) \\
\leq&L\sqrt{\|\mathbf{E} - \mathbf{E}'\|_F^2 + \|\mathbf{T} - \mathbf{T}'\|_F^2}.
\end{aligned}$$

(**ii**) Now we prove that $f(\cdot)$ is l-smooth, which requires finding a constant l satisfying

$$\left\|\begin{bmatrix}\mathrm{vec}\left(\nabla_{\mathbf{T}} f(\mathbf{T},\mathbf{E})\right)\\ \mathrm{vec}\left(\nabla_{\mathbf{E}} f(\mathbf{T},\mathbf{E})\right)\end{bmatrix} - \begin{bmatrix}\mathrm{vec}\left(\nabla_{\mathbf{T}} f(\mathbf{T}',\mathbf{E}')\right)\\ \mathrm{vec}\left(\nabla_{\mathbf{E}} f(\mathbf{T}',\mathbf{E}')\right)\end{bmatrix}\right\|_2 \leq l\left\|\begin{bmatrix}\mathrm{vec}(\mathbf{T})\\ \mathrm{vec}(\mathbf{E})\end{bmatrix} - \begin{bmatrix}\mathrm{vec}(\mathbf{T}')\\ \mathrm{vec}(\mathbf{E}')\end{bmatrix}\right\|_2,$$

where $\mathrm{vec}(\mathbf{X})$ means the vectorization of matrix $\mathbf{X}$ and $\left[\begin{smallmatrix}\mathbf{a}\\ \mathbf{b}\end{smallmatrix}\right]$ denotes the concatenation of vectors $\mathbf{a}$ and $\mathbf{b}$. Since the left hand side satisfies

$$\begin{aligned}
&\left\|\begin{bmatrix}\mathrm{vec}\left(\nabla_{\mathbf{T}} f(\mathbf{T},\mathbf{E})\right)\\ \mathrm{vec}\left(\nabla_{\mathbf{E}} f(\mathbf{T},\mathbf{E})\right)\end{bmatrix} - \begin{bmatrix}\mathrm{vec}\left(\nabla_{\mathbf{T}} f(\mathbf{T}',\mathbf{E}')\right)\\ \mathrm{vec}\left(\nabla_{\mathbf{E}} f(\mathbf{T}',\mathbf{E}')\right)\end{bmatrix}\right\|_2 \\
=&\sqrt{\|\nabla_{\mathbf{T}} f(\mathbf{T},\mathbf{E}) - \nabla_{\mathbf{T}} f(\mathbf{T}',\mathbf{E}')\|_F^2 + \|\nabla_{\mathbf{E}} f(\mathbf{T},\mathbf{E}) - \nabla_{\mathbf{E}} f(\mathbf{T}',\mathbf{E}')\|_F^2} \\
\leq&\|\nabla_{\mathbf{T}} f(\mathbf{T},\mathbf{E}) - \nabla_{\mathbf{T}} f(\mathbf{T}',\mathbf{E}')\|_F + \|\nabla_{\mathbf{E}} f(\mathbf{T},\mathbf{E}) - \nabla_{\mathbf{E}} f(\mathbf{T}',\mathbf{E}')\|_F,
\end{aligned}$$

and the right hand side satisfies

$$l\left\|\begin{bmatrix}\mathrm{vec}(\mathbf{T})\\ \mathrm{vec}(\mathbf{E})\end{bmatrix}-\begin{bmatrix}\mathrm{vec}(\mathbf{T}')\\ \mathrm{vec}(\mathbf{E}')\end{bmatrix}\right\|_2 = l\sqrt{\|\mathbf{E}-\mathbf{E}'\|_F^2+\|\mathbf{T}-\mathbf{T}'\|_F^2} \geq \frac{l}{\sqrt{2}}\|\mathbf{E}-\mathbf{E}'\|_F+\frac{l}{\sqrt{2}}\|\mathbf{T}-\mathbf{T}'\|_F,$$

it suffices to find a constant l satisfying

$$\|\nabla_{\mathbf{T}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{T}}f(\mathbf{T}',\mathbf{E}')\|_F+\|\nabla_{\mathbf{E}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{E}}f(\mathbf{T}',\mathbf{E}')\|_F \leq \frac{l}{\sqrt{2}}\|\mathbf{E}-\mathbf{E}'\|_F+\frac{l}{\sqrt{2}}\|\mathbf{T}-\mathbf{T}'\|_F.$$

We bound $\|\nabla_{\mathbf{T}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{T}}f(\mathbf{T}',\mathbf{E}')\|_F$ as follows,

$$\begin{aligned}&\|\nabla_{\mathbf{T}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{T}}f(\mathbf{T}',\mathbf{E}')\|_F\\ \leq& 2n\|\mathbf{C}^{\mathrm{s}}\odot\mathbf{C}^{\mathrm{s}}\|_F\|\mathbf{T}-\mathbf{T}'\|_F+4\|\mathbf{C}^{\mathrm{s}}\|_F\|\mathbf{T}(\mathbf{C}^{\mathrm{t}}+\mathbf{E})-\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\|_F\\ &+2n\|\mathbf{T}(\mathbf{C}^{\mathrm{t}}+\mathbf{E})\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E})-\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\|_F.\end{aligned}$$

For the first term,

$$2n\|\mathbf{C}^{\mathrm{s}}\odot\mathbf{C}^{\mathrm{s}}\|_F\|\mathbf{T}-\mathbf{T}'\|_F \leq 2n\|\mathbf{C}^{\mathrm{s}}\|_F^2 \leq 2n^3U_1^2\|\mathbf{T}-\mathbf{T}'\|_F.$$

For the second term,

$$\begin{aligned}&4\|\mathbf{C}^{\mathrm{s}}\|_F\|\mathbf{T}(\mathbf{C}^{\mathrm{t}}+\mathbf{E})-\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\|_F\\ \leq& 4\|\mathbf{C}^{\mathrm{s}}\|_F\|\mathbf{T}-\mathbf{T}'\|_F\Big(\|\mathbf{C}^{\mathrm{t}}\|_F+\|\mathbf{E}\|_F\Big)+4\|\mathbf{C}^{\mathrm{s}}\|_F\|\mathbf{T}'\|_F\|\mathbf{E}-\mathbf{E}'\|_F\\ \leq& (4n^2U_1^2+4n^2U_1\epsilon)\|\mathbf{T}-\mathbf{T}'\|_F+4nU_1U_2\|\mathbf{E}-\mathbf{E}'\|_F.\end{aligned}$$

For the third term,

$$\begin{aligned}&2n\|\mathbf{T}(\mathbf{C}^{\mathrm{t}}+\mathbf{E})\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E})-\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\|_F\\ \leq& 2n\|\mathbf{T}(\mathbf{C}^{\mathrm{t}}+\mathbf{E})\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E})-\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E})\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E})\|_F\\ &+2n\|\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E})\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E})-\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E})\|_F\\ &+2n\|\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E})-\mathbf{T}'(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\odot(\mathbf{C}^{\mathrm{t}}+\mathbf{E}')\|_F\\ \leq& 4n\big(\|\mathbf{C}^{\mathrm{t}}\|_F^2+\|\mathbf{E}\|_F^2\big)\|\mathbf{T}-\mathbf{T}'\|_F+2n\|\mathbf{T}'\|_F\big(\|\mathbf{C}^{\mathrm{t}}\|_F+\|\mathbf{E}\|_F\big)\|\mathbf{E}-\mathbf{E}'\|_F\\ &+2n\|\mathbf{T}'\|_F\big(\|\mathbf{C}^{\mathrm{t}}\|_F+\|\mathbf{E}'\|_F\big)\|\mathbf{E}-\mathbf{E}'\|_F\\ \leq& \big(4n^3U_1^2+4n^3\epsilon^2\big)\|\mathbf{T}-\mathbf{T}'\|_F+(4n^2U_1U_2+4n^2U_2\epsilon)\|\mathbf{E}-\mathbf{E}'\|_F\end{aligned}$$

Since $\nabla_{\mathbf{E}}f(\mathbf{T},\mathbf{E})=2(\mathbf{E}+\mathbf{C}^\top)\mathbf{p}^{\mathrm{t}}\mathbf{p}^{\mathrm{t}\top}-2\mathbf{T}^\top\mathbf{C}^{\mathrm{s}}\mathbf{T}$,$\|\nabla_{\mathbf{E}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{E}}f(\mathbf{T}',\mathbf{E}')\|_F$ can be bounded as follows,

$$\begin{aligned}&\|\nabla_{\mathbf{E}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{E}}f(\mathbf{T}',\mathbf{E}')\|_F\\ \leq& \|2(\mathbf{E}-\mathbf{E}')\|_F\|\mathbf{p}^{\mathrm{t}}\|_F^2+4\|\mathbf{T}\|_F\|\mathbf{C}^{\mathrm{s}}\|_F\|\mathbf{T}-\mathbf{T}'\|_F\\ \leq& 2U_2^2\|(\mathbf{E}-\mathbf{E}')\|_F+4nU_1U_2\|\mathbf{T}-\mathbf{T}'\|_F\\ &.\end{aligned}$$

Combining the above four relations, we have

$$\begin{aligned}&\|\nabla_{\mathbf{T}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{T}}f(\mathbf{T}',\mathbf{E}')\|_F+\|\nabla_{\mathbf{E}}f(\mathbf{T},\mathbf{E})-\nabla_{\mathbf{E}}f(\mathbf{T}',\mathbf{E}')\|_F\\ \leq& \max\{10n^3U_1^2+6n^3U_1\epsilon+4nU_1U_2+4n^3\epsilon^2, 6n^2U_1U_2+2U_2^2+4n^2U_2\epsilon\}\Big(\|\mathbf{E}-\mathbf{E}'\|_F+\|\mathbf{T}-\mathbf{T}'\|_F\Big),\end{aligned}$$

which yields the desired result.

■

Theorem 2 *Define $\phi(\cdot)=\max_{\mathbf{E}\in\mathcal{U}_\epsilon}f(\cdot,\mathbf{E})$. The output $\hat{\mathbf{T}}$ of Algorithm 1 with step size $\eta=\frac{\gamma}{\sqrt{N+1}}$ satisfies*

$$\mathbb{E}\big[\|\nabla\phi_{1/2l}(\hat{\mathbf{T}})\|^2\big] \leq 2\frac{\phi_{1/2l}(\mathbf{T}_0)-\min_{\mathbf{T}\in\Pi(\mathbf{p}^{\mathrm{s}},\mathbf{p}^{\mathrm{t}})}\phi(\mathbf{T})+lL^2\gamma^2}{\gamma\sqrt{N+1}},$$

where $l=\sqrt{2}\max\{10n^3U_1^2+6n^3U_1\epsilon+4nU_1U_2+4n^3\epsilon^2, 6n^2U_1U_2+2U_2^2+4n^2U_2\epsilon\}$ and $L=\sqrt{2}\max\{(4U_1+2\epsilon)U_2^2n^3, 2(2U_1+\epsilon)^2U_2n^3\}$ with $n=\max\{n^{\mathrm{s}},n^{\mathrm{t}}\}$, $U_1=\max\{\|\mathbf{C}^{\mathrm{s}}\|_\infty,\|\mathbf{C}^{\mathrm{t}}\|_\infty\}$ and $U_2=\max\{\|\mathbf{p}^{\mathrm{t}}\|_2,\max_{\mathbf{T}'\in\Pi(\mathbf{p}^{\mathrm{s}},\mathbf{p}^{\mathrm{t}})}\|\mathbf{T}'\|_F\}$.

Proof: By the smoothness of $f(\cdot)$, for any $\tilde{\mathbf{T}} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})$ and $\mathbf{T}_\tau$ from Algorithm 1, we have

$$\begin{aligned}\phi(\tilde{\mathbf{T}}) \geq f(\tilde{\mathbf{T}}, \mathbf{E}_\tau) &\geq f(\mathbf{T}_\tau, \mathbf{E}_\tau) + \langle \nabla_{\mathbf{T}} f(\mathbf{T}_\tau, \mathbf{E}_\tau), \tilde{\mathbf{T}} - \mathbf{T}_\tau \rangle - \frac{l}{2}\|\tilde{\mathbf{T}} - \mathbf{T}_\tau\|_F^2 \\ &= \phi(\mathbf{T}_\tau) + \langle \nabla_{\mathbf{T}} f(\mathbf{T}_\tau, \mathbf{E}_\tau), \tilde{\mathbf{T}} - \mathbf{T}_\tau \rangle - \frac{l}{2}\|\tilde{\mathbf{T}} - \mathbf{T}_\tau\|_F^2.\end{aligned} \tag{6}$$

Let $\hat{\mathbf{T}}_\tau = \mathrm{argmin}_{\mathbf{T} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})} \phi(\mathbf{T}) + l\|\mathbf{T} - \mathbf{T}_\tau\|_F^2$. We have

$$\begin{aligned}\phi_{1/2l}(\mathbf{T}_{\tau+1}) &\leq \phi(\hat{\mathbf{T}}_\tau) + l\|\mathbf{T}_{\tau+1} - \hat{\mathbf{T}}_\tau\|_F^2 \\ &\leq \phi(\hat{\mathbf{T}}_\tau) + l\|\mathbf{T}_\tau - \eta \nabla_{\mathbf{T}} f(\mathbf{T}_\tau, \mathbf{E}_\tau) - \hat{\mathbf{T}}_\tau\|_F^2 \\ &\leq \phi(\hat{\mathbf{T}}_\tau) + l\|\mathbf{T}_\tau - \hat{\mathbf{T}}_\tau\|_F^2 + 2l\eta\langle \nabla_{\mathbf{T}} f(\mathbf{T}_\tau, \mathbf{E}_\tau), \hat{\mathbf{T}}_\tau - \mathbf{T}_\tau \rangle + \eta^2 l \|\nabla_{\mathbf{T}} f(\mathbf{T}_\tau, \mathbf{E}_\tau)\|_F^2 \\ &\leq \phi_{1/2l}(\mathbf{T}_\tau) + 2l\eta\langle \nabla_{\mathbf{T}} f(\mathbf{T}_\tau, \mathbf{E}_\tau), \hat{\mathbf{T}}_\tau - \mathbf{T}_\tau \rangle + \eta^2 l \|\nabla_{\mathbf{T}} f(\mathbf{T}_\tau, \mathbf{E}_\tau)\|_F^2 \\ &\leq \phi_{1/2l}(\mathbf{T}_\tau) + 2\eta l\Big(\phi(\hat{\mathbf{T}}_\tau) - \phi(\mathbf{T}_\tau) + \frac{l}{2}\|\hat{\mathbf{T}}_\tau - \mathbf{T}_\tau\|_F^2\Big) + \eta^2 l L^2,\end{aligned}$$

where the second line uses Lemma 3.1 of Bubeck et al. (2015) and the last line follows from (6). Taking a telescopic sum over τ, we obtain

$$\phi_{1/2l}(\mathbf{T}_N) \leq \phi_{1/2l}(\mathbf{T}_0) + 2\eta l \sum_{\tau=0}^{N} \Big(\phi(\hat{\mathbf{T}}_\tau) - \phi(\mathbf{T}_\tau) + \frac{l}{2}\|\hat{\mathbf{T}}_\tau - \mathbf{T}_\tau\|_F^2\Big) + \eta^2 l L^2.$$

Rearranging this, we obtain

$$\frac{1}{N+1}\sum_{\tau=0}^{N}\Big(-\phi(\hat{\mathbf{T}}_\tau) + \phi(\mathbf{T}_\tau - \frac{l}{2}\|\hat{\mathbf{T}}_\tau - \mathbf{T}_\tau\|_F^2\Big) \leq \frac{\phi_{1/2l}(\mathbf{T}_0) - \min_{\mathbf{T} \in \Pi(\mathbf{p}^{\mathrm{s}}, \mathbf{p}^{\mathrm{t}})} \phi(\mathbf{T})}{2\eta l N} + \frac{\eta L^2}{2}. \tag{7}$$

Since $\phi(\mathbf{T}) + l\|\mathbf{T} - \mathbf{T}_\tau\|_F^2$ is l-strongly convex, we have

$$\begin{aligned}&-\phi(\hat{\mathbf{T}}_\tau) + \phi(\mathbf{T}_\tau) - \frac{l}{2}\|\hat{\mathbf{T}}_\tau - \mathbf{T}_\tau\|_F^2 \\ \geq &\frac{l}{2}\|\mathbf{T}_\tau - \hat{\mathbf{T}}_\tau\|_F^2 + \phi(\mathbf{T}_\tau) + l\|\mathbf{T}_\tau - \mathbf{T}_\tau\|_F^2 - \min_{\mathbf{T}}\Big(\phi(\mathbf{T}) + l\|\mathbf{T} - \mathbf{T}_\tau\|_F^2\Big) \\ \geq &l\|\mathbf{T}_\tau - \hat{\mathbf{T}}_\tau\|_F^2 = \frac{1}{4l}\|\nabla \phi_{1/2l}(\mathbf{T}_\tau)\|_F^2.\end{aligned}$$

Plugging this in (7) and combining Lemma 3 proves the result.

■

.