# Outlier-Robust Gromov Wasserstein for Structured Data

**Anonymous Authors**[1]

## Abstract

Gromov Wasserstein (GW) provides a flexible way to compare and couple probability distributions supported on different metric spaces. Recently, GW acts as the main modeling tool for aligning heterogeneous data for a host of graph learning tasks. However, it is known that the GW distance can be extremely sensitive to outliers when they are weighed similarly as the other samples in its objective function owing to the marginal constraints. To address this issue, we propose a new robust model of the GW distance called RGW with optimistically perturbed marginal constraints in a $\phi$-divergence based ambiguity set. To realize the benefits in practice, we further develop a computationally efficient and yet theoretically provable computational procedure via Bregman proximal alternating linearization minimization algorithm. Empirically, we conduct extensive experiment results to not only corroborate our theoretical findings but also demonstrate the effectiveness of RGW on the subgraph matching task and partial shape correspondence tasks.

## 1. Introduction

Gromov Wasserstein distance (GW) (Mémoli, 2007; 2011) acts as a main model tool in data science to compare data distributions on unaligned metric spaces. Recently, it has received much attention across a host of applications in data analysis, e.g., shape correspondence (Peyré et al., 2016; Mémoli, 2009; Solomon et al., 2016), graph alignment and partition (Chowdhury & Mémoli, 2019; Xu et al., 2019b;a; Chowdhury & Needham, 2021; Gao et al., 2021), graph embedding and classification (Vincent-Cuaz et al., 2021; Xu et al., 2022), unsupervised word embedding and translation (Alvarez-Melis & Jaakkola, 2018; Grave et al., 2019), generative modeling across incomparable spaces (Bunne et al., 2019; Xu et al., 2021).

In practice, the robustness of GW distance suffers heavily from its sensitivity to outliers. Here, outliers mean the samples with large noise, which usually are far away from the clean samples or have different structures from the clean samples. Due to the hard constraints on the marginals, all the mass in the source distribution has to be entirely transported to the target distribution. When the outliers are weighted similarly as other clean samples, even a small fraction of outliers corrupted can largely impact the GW distance value and the optimal coupling, which is unsatisfactory in real-world applications.

To overcome the above issue, some recent works are trying to relax the marginal constraints of GW distance. Rodola et al. (2012) introduces a $L^1$ relaxation of mass conservation of the GW distance. However, this reformulation replaces the strict marginal constraint that the transport plan should be a joint distribution with marginals as specific distributions by the constraint that only requires the transport plan to be a joint distribution, which can easily lead to over-relaxation. On another front, Chapel et al. (2020) propose a so-called partial GW distance (PGW), which only transports a fraction of mass from source distribution to target distribution. The formulation of PGW only allows mass destruction, which hinders PGW from tackling the cases where the outliers are on one side only. A formulation that allows both mass destruction and creation is proposed in (Séjourné et al., 2021) called unbalanced GW (UGW). The UGW relaxes the marginal constraint via adding the quadratic $\varphi$-divergence as the penalty function in the objective function and extends GW distance to compare metric spaces equipped with arbitrary positive measures. Using quadratic -divergence implies that UGW distance exhibits 2-homogeneity, which is critical for establishing a direct link with Conic Gromov Wasserstein (CGW) distance. Additionally, Tran et al. (2023) proved that UGW is robust to outliers and can effectively remove the mass of outliers with high transportation costs. On the computational side, an alternate Sinkhorn minimization method is proposed to calculate the entropy-regularized UGW. Note that the algorithm does not exactly solve UGW but approximates the lower bound of the entropic regularized UGW instead. Additionally, these works do not establish a direct link between the reformulated GW distance and the GW distance between uncontaminated samples.

The classical optimal transport problem is also sensitive to outliers, due to its possession of the same marginal constraints as the GW distance. This has motivated a series of works on outlier-robust optimal transport, which aims to

relax the marginal constraints in various forms. In Staerman et al. (2021), a median-of-means approach was proposed as a robust mean estimator to estimate the Kantorovich-Rubinstein duality of the 1-Wasserstein distance. In (Balaji et al., 2020; Mukherjee et al., 2021; Le et al., 2021), the authors introduce perturbed marginal distributions and use different $\varphi$-divergences to quantify the violation between the perturbed marginal distribution and the original marginal distributions. The purpose of involving perturbed distributions is to find an optimal transport plan between approximate distributions that assign less weight to the outliers, instead of between the original distributions that may be corrupted with outliers.

In this work, we propose the robust Gromov Wasserstein (RGW) that provides a robust estimate of GW distance when outliers are present. The key idea is to relax the strict marginal constraint by finding the optimal marginal distribution in a certain distributionally ambiguity set as an alternative. This idea is closely related to the literature on the optimistic modelings of distribution ambiguity in data-driven optimization, e.g., upper confidence bound in the multi-armed bandit problem and reinforcement learning (Bubeck et al., 2012; Munos et al., 2014; Agarwal et al., 2020), data-driven distributionally robust decision-making with outliers (Jiang & Xie, 2021; Cao & Gao, 2021), etc. For our purpose, the RGW is a relaxation of the original GW distance, which optimizes the transport plan and perturbed marginal distributions concurrently. As a result, the perturbed marginal distributions help to re-weight the samples and lower the weight assigned to the outliers. The RGW relaxes the marginal constraint via adding the $\varphi$-divergence between the marginals of the transport plan and the perturbed distributions as the penalty function in the objective function to lessen the impact of the outliers further. We show that under the Huber $\epsilon$-contamination model, the robust GW is upper bounded by the true GW distance. This indicates that our RGW can effectively ignore the outliers and obtain a robust estimate for the GW distance without outliers.

To realize its modeling benefits, we further propose an algorithm based on the Bregman proximal alternating linearized minimization (BPALM) method to address RGW efficiently. The updates in each iteration of BPALM can be computed in a highly efficient manner. On the theoretical side, we prove that the BPALM algorithm converges to a critical point of the RGW problem.

Empirically, we demonstrate the effectiveness of RGW and the proposed BPALM algorithm through extensive numerical experiments on subgraph alignment and partial shape correspondence tasks. For both tasks, we can view the missing part as the outlier, which motivates us to apply RGW to these kinds of tasks. To the best of our knowledge, it is the first time to apply the GW-based method for solving the partial shape correspondence problem successfully, which remains a challenging problem as pointed out in (Solomon et al., 2016).

The rest of this paper is organized as follows. In Section 2, we introduce the formulation of robust Gromov Wasserstein and present the robustness result. In Section 3, we discuss the details of the proposed algorithm. We then report the numerical results in Section 4 and conclude in Section 5.

*Notation.* Let $(X, d_X)$ be a complete separable metric space and denote the finite, positive Borel measure on $X$ by $\mathcal{M}_+(X)$. Let $\mathcal{P}(X) \subset \mathcal{M}_+(X)$ denotes the space of Borel probability measures on $X$. We use $\Delta^n$ to denote the simplex in $\mathbb{R}^n$. We use $\mathbf{1}_n$ and $\mathbf{1}_{n\times m}$ to denote the $n$-dimensional all-one vector and $n \times m$ all-one matrix. We use $\mathcal{S}^n$ to denote the set of $n \times n$ symmetric matrice. The indicator function of set $C$ is denoted as $\mathbb{I}_C(\cdot)$.

## 2. Problem Formulation

In this section, we review the definition of Gromov Wasserstein distance and formally formulate the robust Gromov Wasserstein. Then, we discuss the statistical properties of the proposed robust Gromov-Wasserstein model under Huber's contamination model.

### 2.1. Robust Gromov Wasserstein

The Gromow Wasserstein (GW) distance aims at matching distributions defined in different metric spaces. It is defined as follows:

**Definition 2.1** (Gromov Wasserstein)**.** Suppose that we are given two unregistered complete separable metric spaces $(X, d_X)$, $(Y, d_Y)$ accompanied with Borel probability measures $\mu, \nu$ respectively. The GW distance between $\mu$ and $\nu$ is defined as

$$\inf_{\pi\in\Pi(\mu,\nu)} \iint |d_X(x, x') - d_Y(y, y')|^2 d\pi(x, y)d\pi(x', y'),$$

where $\Pi(\mu, \nu)$ is the set of all probability measures on $X \times Y$ with $\mu$ and $\nu$ as marginals.

As shown in the definition, the sensitivity to outliers of Gromov Wasserstein distance is due to its hard constraints on marginal distributions. This suggests relaxing the marginal constraints such that the weight assigned to the outliers by the transportation plan can be small. To do it, we invoke the Kullback-Leibler divergence, defined as $d_{\mathbf{KL}}(\alpha, \mu) = \int_X \alpha(x) \log\left(\frac{\alpha(x)}{\mu(x)}\right) dx$, to soften the constraint on marginal distributions. To reduce the weight assigned to the outliers, we perturbed the marginal distributions by an optimistically distributionally robust mechanism we will discuss the details later.

**Definition 2.2** (Robust Gromov Wasserstein)**.** Suppose that we are given two unregistered complete separable metric spaces $(X, d_X)$, $(Y, d_Y)$ accompanied with Borel probability measures $\mu, \nu$ respectively. The Robust GW distance between $\mu$ and $\nu$ is defined as

$$\begin{aligned} \mathrm{GW}^{\mathrm{rob}}_{\rho_1,\rho_2}(\mu,\nu) := \inf_{\alpha\in\mathcal{P}(X),\ \beta\in\mathcal{P}(Y)} \quad & F(\alpha,\beta) \\ \text{s.t.} \quad & d_{\mathbf{KL}}(\mu,\alpha) \le \rho_1, d_{\mathbf{KL}}(\nu,\beta) \le \rho_2, \end{aligned} \tag{1}$$

where $F(\alpha,\beta) =$

$$\inf_{\pi\in\mathcal{M}^+(X\times Y)} \iint |d_X(x,x') - d_Y(y,y'))|^2 d\pi(x,y) d\pi(x',y') + \tau_1 d_{\mathbf{KL}}(\pi_1,\alpha) + \tau_2 d_{\mathbf{KL}}(\pi_2,\beta),$$

and $(\pi_1, \pi_2)$ are two marginals of the joint distribution $\pi$, defined by $\pi_1(A) = \pi(A \times Y)$ for any Borel set $A \subset X$ and $\pi_2(B) = \pi(X \times B)$ for any Borel set $B \subset Y$.

The main idea of our formulation is to optimize the transport plan and perturbed distribution variables in the ambiguity set of the observed marginal distributions jointly. This formulation aims to find the perturbed distributions that can approximate the clean distribution and compute the transport plan based on the perturbed distributions. However, involving the constraints such that the marginals of the transport plan $\pi$ are equal to the perturbed distribution $\alpha$ and $\beta$ directly can bring difficulty in developing an algorithm to tackle the model. Inspired by (Séjourné et al., 2021), we relax these marginal constraints by adding the $\varphi$-divergence $D_\varphi(\pi_1, \alpha)$ and $D_\varphi(\pi_2, \beta)$ as the penalty function in the objective function. Different from (Séjourné et al., 2021), we use $\varphi$-divergence instead of quadratic $\varphi$-divergence since quadratic $\varphi$-divergence is usually non-convex, which is unsatisfactory for algorithm development, while the $\varphi$-divergence is jointly convex. Besides, transforming the hard marginal constraints into penalty functions can further lessen the impact of outliers on the transport plan.

Our new formulation is an extension of the balanced GW distance, and the balanced GW distance can also be recovered when choosing $\rho_1 = \rho_2 = 0$, and $\tau_1$ and $\tau_2$ go to infinity. If $\rho_1$ and $\rho_2$ are chosen properly, the clean distributions may be inside the ambiguity sets. In this case, this relaxed reformulation is close to the original GW distance in a certain way. Inspired by this idea, we prove that the RGW can be a robust approximation of the GW distance without outliers under some mild assumptions on the outliers.

### 2.2. Robustness Guarantees

Robust Gromov Wasserstein is designed to address the problem that the GW distance explodes as the distance between the clean samples and the outliers goes to infinity. In general, one can construct an example where the GW distance changes dramatically when a small number of outliers are added to the marginal distributions. To formalize this, consider the Huber $\epsilon$ contamination model popularized in robust statistics. In that model, a base measure $\mu_c$ is contaminated by an outlier distribution $\mu_a$ to obtain a contaminated measure $\mu$,

$$\mu = (1-\epsilon)\mu_c + \epsilon\mu_a. \tag{2}$$

Under this model, data are drawn from $\mu$ defined in (2).

Under the assumption of the Huber $\epsilon$-contamination model, we can show that the robust Gromov Wasserstein guarantees that outliers cannot arbitrarily increase the transport distance by choosing appropriate $\rho_1$ and $\rho_2$. For robust Gromov Wasserstein, we have the following bound:

**Theorem 2.3.** *Let $\mu$ and $\nu$ be two distributions such that $\mu$ is corrupted with $\epsilon$ fraction of outliers i.e., $\mu = (1-\epsilon)\mu_c + \epsilon\mu_a$, where $\mu_c$ is the clean distribution and $\mu_a$ is the outlier distribution. Then,*

$$\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mu,\nu) \le \mathrm{GW}(\mu_c,\nu) + \max\left(0, \epsilon - \frac{\rho}{d_{\mathbf{KL}}(\mu_a,\mu_c)}\right) \tau_1 d_{\mathbf{KL}}(\mu_c,\mu_a).$$

Our proof in Appendix B constructs a feasible transport plan and feasible relaxed marginal distributions. The relaxation of the strict marginal constraints and relaxed marginal distributions allow us to find a feasible transport plan that can approximate the transport plan between the clean distributions and feasible relaxed marginal distributions that are close to the clean distributions.

This bound implies that robust Gromov Wasserstein obtains a provably robust estimate under the Huber $\epsilon$ contamination model. When the fraction of outliers is known, the robust GW is upper bounded by the true Gromov Wasserstein distance plus a factor of the $\varphi$-divergence between the clean distribution $\mu_c$ and the outlier distribution $\mu_a$. This factor is controlled by the magnitude of the marginal distribution $\rho$. If $\rho$ is chosen appropriately, the robust GW will obtain a value close to the true GW distance (GW distance without outliers). Note that by substituting $\rho = \epsilon D_\varphi(\mu_a, \mu_c)$, we obtain that $\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mu,\nu) \le \mathrm{GW}(\mu_c,\nu)$, which means that the robust GW between the contaminated distribution $\mu$ and $\nu$ is upper bounded by the original GW distance between the clean distribution $\mu_c$ and $\nu$.

## 3. Proposed Algorithm

### 3.1. Problem Setup

To start with our algorithmic developments, we consider the discrete case for simplicity and practicality, where $\mu$ and $\nu$

are two empirical distributions, i.e., $\mu = \sum_{i=1}^{n} \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} \nu_j \delta_{y_j}$. Denote $D \in \mathcal{S}^n$, $D_{ik} = d_X(x_i, x_k)$ and $\bar{D} \in \mathcal{S}^m$ and $\bar{D}_{jl} = d_Y(y_j, y_l)$. The 4-way tensor is given as

$$\mathcal{L}(D, \bar{D}) \coloneqq \left( \left| d_X(x_i, x_k) - d_Y(y_j, y_l) \right|^2 \right)_{i,j,k,l}.$$

We define the tensor-matrix multiplication as

$$(\mathcal{L} \otimes T)_{ij} \coloneqq \left( \sum_{k,\ell} \mathcal{L}_{i,j,k,\ell} T_{k,\ell} \right)_{i,j}.$$

Then, the robust GW admits the following reformulation:

$$\begin{aligned} \min_{\pi,\alpha,\beta} \; & \langle \mathcal{L}(D, \bar{D}) \otimes \pi, \pi \rangle + \tau_1 D_\varphi(\pi_1, \alpha) + \tau_2 D_\varphi(\pi_2, \beta) \\ \text{s.t. } & d_{\mathbf{KL}}(\mu, \alpha) \le \rho_1, d_{\mathbf{KL}}(\nu, \beta) \le \rho_2, \\ & \alpha \in \Delta^n, \beta \in \Delta^m, \pi \ge 0. \end{aligned} \tag{3}$$

Here, $\pi_1 = \pi \mathbf{1}_m$ and $\pi_2 = \pi^T \mathbf{1}_n$.

### 3.2. Bregman Proximal Alternating Linearized Method (BPALM)

Problem (3) is a non-convex optimization problem with three variables. We propose to use Bregman proximal alternating linearized method (Bolte et al., 2014; Ahookhosh et al., 2021). Iterations of this algorithm are given by

$$\begin{aligned} \pi^{k+1} = \arg\min_{\pi \ge 0} \{ & \langle \mathcal{L}(D, \bar{D}) \otimes \pi^k, \pi \rangle + \tau_1 d_{\mathbf{KL}}(\pi_1, \alpha^k) \\ & + \tau_2 d_{\mathbf{KL}}(\pi_2, \beta^k) + \frac{1}{t_k} d_{\mathbf{KL}}(\pi, \pi^k) \}, \end{aligned} \tag{4}$$

$$\alpha^{k+1} = \arg\min_{\substack{\alpha \in \Delta^n \\ D_\varphi(\mu, \alpha) \le \rho_1}} \{ d_{\mathbf{KL}}(\pi_1^{k+1}, \alpha) + \frac{1}{c_k} d_{\mathbf{KL}}(\alpha^k, \alpha) \}, \tag{5}$$

$$\beta^{k+1} = \arg\min_{\substack{\beta \in \Delta^m \\ d_{\mathbf{KL}}(\nu, \beta) \le \rho_2}} \{ d_{\mathbf{KL}}(\pi_2^{k+1}, \beta) + \frac{1}{r_k} d_{\mathbf{KL}}(\beta^k, \beta) \}. \tag{6}$$

Here, $t_k$, $c_k$, and $r_k$ are stepsizes in BPALM.

The $\pi$-subproblem is equivalent to the entropic regularized unbalanced optimal transport problem. Thus, we adopt the Sinkhorn algorithm for unbalanced optimal transport to solve the subproblem for $\pi$ (Chizat et al., 2018; Pham et al., 2020).

As for the $\alpha$-subproblem, we consider the case where $\rho_1$ is strictly larger than 0. Otherwise, when $\rho_1 = 0$, $\alpha$ should equal $\mu$, and we do not need to tackle the subproblem for $\alpha$. To solve the $\alpha$-subprblem, we attempt to find the optimal dual multiplier $w^*$. Specifically, consider the problem

$$\min_{\alpha \in \Delta^n} d_{\mathbf{KL}}(\pi_1^{k+1}, \alpha) + \frac{1}{c_k} d_{\mathbf{KL}}(\alpha^k, \alpha) + w(d_{\mathbf{KL}}(mu, \alpha) - \rho_1). \tag{7}$$

Let $\hat{\alpha}(w)$ be the optimal solution to (7) and define the function $p : \mathbb{R}_+ \to \mathbb{R}$ by $p(w) = d_{\mathbf{KL}}(\mu, \hat{\alpha}(w)) - \rho_1$. We prove the convexity, differentiability, and monotonicity of $p$, which are crucial for developing an efficient algorithm for (5) later.

**Proposition 3.1.** *If $w$ satisfies (i) $w = 0$ and $p(w) \le 0$, or (ii) $w > 0$, $p(w) = 0$, then $\hat{\alpha}(w)$ is the optimal solution to the $\alpha$-subproblem (5). Moreover, $p(\cdot)$ is convex, twice differentiable, and monotonically non-increasing on $\mathbb{R}_+$.*

Problem (7) has a closed-form solution

$$\hat{\alpha}(w) = \frac{\pi^{k+1} \mathbf{1}_m + \frac{1}{c_k} \alpha^k + w\mu}{\sum_{i,j} \pi_{ij}^{k+1} + \frac{1}{c_k} + w}.$$

Given Proposition 3.1, we first check if $p(0) \le 0$. If not, since $p(0) > 0$ and $\lim_{w \to +\infty} p(w) = -\rho_1 < 0$, then $p$ contains at least one root on $\mathbb{R}_+$. The following proposition enables us to search for the root of $p$ using Newton's method initialized at 0. Hence, the $\alpha$-subproblem can be cast to search a root of $p$ in one-dimensional space, in which case it can be solved efficiently.

**Proposition 3.2.** *Let $p(\cdot) : I \to \mathbb{R}$ be a convex, twice differentiable, and monotonically non-increasing on the interval $I \subset \mathbb{R}$. Assume that there exist $\tilde{x}, \bar{x} \in I$ such that $p(\tilde{x}) > 0$ and $p(\bar{x}) < 0$. Then $p$ has a unique root on $I$, and the sequence obtained from Newton's method with initial point $x_0 = \tilde{x}$ will converge to the root of $p$.*

Since the $\beta$-subproblem has the same structure as the $\alpha$-subproblem, we can also use this method to search for the optimal solution to $\beta$-subproblem.

### 3.3. Convergence Analysis

To illustrate the convergence result of BPALM, we consider the compact form for simplicity:

$$\begin{aligned} \min_{\alpha,\beta,\pi} F(\pi, \alpha, \beta) = & f(\pi) + g_1(\pi, \alpha) + g_2(\pi, \beta) + \\ & h_1(\alpha) + h_2(\beta) \end{aligned} \tag{8}$$

where

- $f(\pi) = \langle \mathcal{L}(D, \bar{D}) \otimes \pi, \pi \rangle$,
- $g_1(\pi, \alpha) = \tau_1 d_{\mathbf{KL}}(\pi \mathbf{1}_m, \alpha)$,
- $g_2(\pi, \beta) = \tau_2 d_{\mathbf{KL}}(\pi^T \mathbf{1}_n, \beta)$,
- $h_1(\alpha) = \mathbb{I}_{\{\alpha \in \Delta^n, d_{\mathbf{KL}}(\mu, \alpha) \le \rho_1\}}(\alpha)$,

- and $h_2(\beta) = \mathbb{I}_{\{\beta \in \Delta^m, d_{\mathbf{KL}}(\nu,\beta) \leq \rho_2\}}(\beta)$.

The following theorem states that any limit point of the sequence generated by BPALM belongs to the critical point set of problem (3).

**Theorem 3.3** (Subsequence Convergence)**.** *Suppose that in Problem (1), the step size $t_k$ in (4) satisfies $0 < \underline{t} \leq t_k < \bar{t} \leq \sigma/L_f$ for $k \geq 0$ where $\underline{t}$, $\bar{t}$ are given constants and $L_f$ is the gradient Lipschitz constant of $f$. The step size $c_k$ in (5) and $r_k$ in (6) satisfy $0 < \underline{r} \leq c_k, r_k < \bar{r}$ for $k \geq 0$ where $\underline{r}$, $\bar{r}$ are given constants. Any limit point of the sequence of solutions $\{\pi^k, \alpha^k, \beta^k\}_{k\geq 0}$ belongs to the critical point set $\mathcal{X}$, where $\mathcal{X}$ is defined by*

$$\left\{ (\pi, \alpha, \beta) : \begin{array}{l} \nabla f(\pi) + \nabla_\pi g_1(\pi, \alpha) + \nabla_\pi g_2(\pi, \beta) = 0, \\ 0 \in \nabla_\alpha g_1(\pi, \alpha) + \partial h_1(\alpha), \\ 0 \in \nabla_\beta g_2(\pi, \beta) + \partial h_2(\beta), \\ (\pi, \alpha, \beta) \in \mathbb{R}^{n\times m} \times \mathbb{R}^n \times \mathbb{R}^m \end{array} \right\}. \tag{9}$$

For the sake of brevity, we omit the proof. We refer the reader to Appendix for further details.

## 4. Experiment Results

In this section, we provide extensive experiment results to validate the effectiveness of the proposed RGW model and BPALM algorithm on various graph learning tasks, including subgraph alignment and partial shape correspondence. The balanced GW has been successfully applied to the graph alignment and shape correspondence tasks when the sizes of source and target are similar. Here, the missing part of the target is viewed as outliers and we can apply our RGW. All simulations are implemented using Python 3.9 on a high-performance computing server running Ubuntu 20.10 with an Intel(R) Xeon(R) Silver 4214R CPU.

### 4.1. 2D Matching Example

In this subsection, we study a toy matching problem in 2D to corroborate our theoretical insights and results in Sec 2. Fig. 2 (a) shows an example of mapping a two-dimensional shape without symmetries to a rotated version of the same shape with outliers in the source domain. The original shape is created by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(z) + 0.1\sin(z) \\ 0.5\sin^3(2z) + 0.1\cos(5z) \end{pmatrix}, \; z \in (0, 2\pi).$$

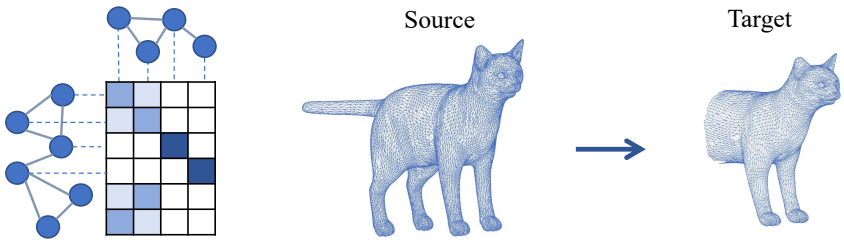
Here, we sample 300 and 400 points from source and target shapes, respectively, and add 50 points of outliers from a discrete uniform distribution on $[-3, -2.5] \times [0, 0.5]$ to the source domain. The distance matrices $D$ and $\bar{D}$ are constructed by the pairwise Euclidean distance.

Figs.2 (b)-(e) provide color maps of coupling matrices and the objective values of all the models to visualize the matching results. As shown in Figure2 (b), even a small fraction of outliers leads to poor coupling and a significant increase in the GW distance estimate. Although unbalanced GW and partial GW can mitigate some impact of outliers, they still fail to obtain an accurate mapping. Our robust GW formulation effectively ignores the outliers and gives us satisfactory solution performance. Besides, the objective value of RGW can also provide an estimate that closely approximates the true GW distance without outliers, which is near 0, as mentioned in Theorem 2.3.

### 4.2. Subgraph Alignment

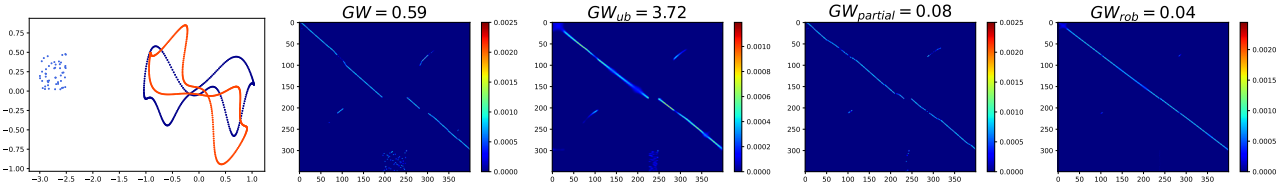
The problem of subgraph alignment is to determine if a query graph is isomorphic to a subgraph of a large target graph (Cordella et al., 2004; Han et al., 2019). For graphs of similar size, instead of solving the restricted quadratic assignment problem (Lawler, 1963; Lacoste-Julien et al., 2006), the GW distance provides the optimal probabilistic correspondence relationship by preserving the isometric property. In the subgraph alignment task, the nodes in the large target graph, except for the nodes in the source graph, can be viewed as outliers in the target graph, which allows us to perform the RGW model on this task. Here, we compare the proposed RGW with unbalanced GW, partial GW, and semi-relaxed GW (Vincent-Cuaz et al., 2022) and also methods for computing the balanced GW: FW (Titouan et al., 2019), BPG (Xu et al., 2019b), SpecGW (Chowdhury & Needham, 2021), eBPG (Solomon et al., 2016) and BAPG (Li et al., 2022).

**Database Statistics** We test all methods on both synthetic and real databases. For the synthetic database, the target graph $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t\}$ is generated by Barabasi-Albert models with different scales, i.e., $|\mathcal{V}_t| \in \{100, 200, 300, 400, 500\}$. Then we generate the source graph $\mathcal{G}_s = \{\mathcal{V}_s, \mathcal{E}_s\}$ by sampling a connected subgraph of the target graph containing $q\%$ of the nodes of the target graph, where $q \in \{20, 30, 40, 50\}$. For each setup, we generate five synthetic graph pairs over different random seeds. In total, the synthetic database contains 200 different graph pairs. We also validate our proposed methods on two other biological graph databases from (Chowdhury & Needham, 2021), *Proteins*, and *Enzymes*, using the same routine to generate subgraphs as source graphs and the original graph as target graph. We match each node in $\mathcal{G}_s$ to the most likely node in $\mathcal{G}_t$ according to the optimized $\pi^\star$. Given the predicted correspondence set $\mathcal{S}_{\text{pred}}$ and the ground truth correspondence set $\mathcal{S}_{\text{gt}}$, we compute the matching accuracy by Acc $= |\mathcal{S}_{\text{gt}} \cap \mathcal{S}_{\text{pred}}|/|\mathcal{S}_{\text{gt}}| \times 100\%$. We also include a social network dataset *Douban Online-Offline*, which contains two social network graphs, the online graph and the




(a) Subgraph Alignment (b) Partial Shape Correspondence

*Figure 1.* Graph learning tasks conducted in this paper.




(a) Shape Geometry (b) Balanced GW (c) Unbalanced GW (d) Partial GW (e) Robust GW

*Figure 2.* (a): 2D shape geometry of the source and target; (b)-(e): visualization of the matching results of balanced GW, unbalanced GW, partial GW, and robust GW.

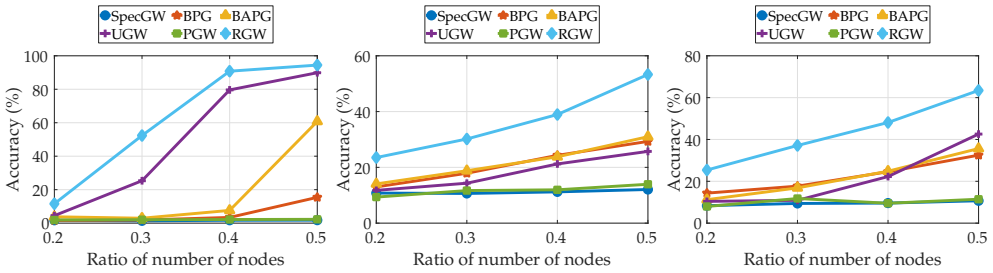
offline graph. In the online graph, nodes represent users and edges represent interaction between users on the site. The offline graph is constructed according to the user's presence in social gatherings. The online graph is larger and includes all the users in the offline graph. In *Douban Online-Offline*, 1,118 users appearing in both graphs are used as ground truth alignments. A user's location is used as a node feature in both graphs. For this dataset, we use Hit@k to evaluate the performance of all graph alignment methods. It calculates the percentage of nodes in $\mathcal{V}_t$ whose ground truth alignment results in $\mathcal{V}_s$ being in the top-k candidates.

**Parameters Setup** We use the unweighted symmetric adjacency matrices as our input distance matrices, i.e. $D$ and $\bar{D}$. Alternatively, SpecGW uses the heat kernel $\exp(-L)$, where $L$ is the normalized graph Laplacian matrix. We set both $\mu$ and $\nu$ as uniform distributions. For SpecGW, BPG, eBPG, and BAPG, we follow the same setup as reported in their papers. For FW, we use the default implementation in the PythonOT package. For semi-relaxed GW, we use the mirror descent method and report the best result obtained among the set of regularization parameter values $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$. For UGW, we report the best results among the set $\{0.5, 0.2, 0.1, 0.01, 0.001\}$ of regularization parameters and the set $\{0.1, 0.01, 0.001\}$ of marginal penalty parameters. For PGW, we report the best results in the range of 0.1 to 0.9 of the transported mass. For RGW, we report the best results in the range $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ of the marginal relaxation parameter $\rho_1, \rho_2$ and the set $\{0.01, 0.1, 0.5\}$ of the marginal penalty parameter $\tau_1, \tau_2$ and the set $\{0.01, 0.05, 0.1, 0.5, 1\}$ for the step size $t_k$, $c_k$ and $r_k$. To initialize the transport plan, we use the $\mathbf{1}_{n\times m}/(nm)$ to initialize the transport plan for all methods on the *Synthetic*, *Proteins*, *Enzymes* datasets. As for the *Douban Online-Offline* dataset, we use two different initializations: the first is $\mathbf{1}_{n\times m}/(nm)$ and the second is to obtain a feature similarity matrix by performing classical optimal transport on the feature space. We use uniform distributions to initialize $\alpha$ and $\beta$ for all datasets.

**Result of All Methods** Table 1 and Figure 3 show the comparison of the matching accuracy and wall-clock time on datasets *Synthetic*, *Proteins*, and *Enzymes* and Hit@1 and Hit@10 on *Douban Online-Offline* dataset. We observe that RGW outperforms other methods in terms of accuracy, and the computation time of RGW is comparable to that of UGW and PGW. In general, the methods for computing the balanced GW perform poorly on all databases, especially on the large graph database *Synthetic*. The reason is that these methods have to satisfy the hard marginal constraint on the source side and fail to eliminate the effect of outliers (the nodes that are not in the target graph). Also, the balanced GW methods can only converge to a stationary point of the balanced GW problem, and the partial structure of the target graph may introduce more local minima into the balanced GW problem. Starting from the initial point, $\mathbf{1}_{n\times m}/(nm)$, the methods for balanced GW can easily be trapped in the local minima far from the ground truth. Besides, it can be observed that UGW works well on when the partiality of the target graph is low, but the performance degrades dramatically when the partiality of the target graph increases. Moreover, the PGW does not perform well in this task, because to reduce the impact of outliers in the PGW, we

*Table 1.* Comparison of the average matching accuracy (%) and wall-clock time (seconds) on subgraph alignment of 50% subgraph on datasets Synthetic, Proteins and Enzymes and Hit@1 and Hit@10 of dataset Douban with two different initialization.

| Method | Synthetic | | Proteins | | Enzymes | | Douban | | Douban (feature) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Time | Acc | Time | Acc | Time | Hit@1 | Hit@10 | Hit@1 | Hit@10 |
| FW | 2.27 | 18.39 | 16.00 | 27.05 | 15.47 | 9.57 | 0.89 | 3.57 | 17.97 | 51.07 |
| SpecGW | 1.78 | 3.72 | 12.06 | 11.07 | 10.69 | 3.96 | 1.79 | 5.37 | 2.68 | 9.83 |
| eBPG | 3.71 | 85.31 | 19.88 | 1975.12 | 21.58 | 1219.81 | 0.09 | 0.54 | 0.08 | 0.53 |
| BPG | 15.41 | 24.67 | 29.30 | 118.26 | 32.49 | 70.42 | 3.31 | 9.30 | 72.72 | 92.39 |
| BAPG | 48.89 | 27.95 | 30.98 | 122.13 | 35.64 | 16.41 | 1.87 | 5.27 | 72.18 | 92.58 |
| srGW | 1.60 | 152.01 | 21.30 | 63.00 | 24.13 | 19.68 | 0.09 | 0.09 | 4.03 | 11.54 |
| UGW | 89.88 | 176.24 | 25.72 | 4026.93 | 42.57 | 3454.45 | 0.09 | 0.72 | 0.09 | 0.72 |
| PGW | 2.28 | 479.99 | 13.94 | 544.79 | 11.43 | 212.09 | 0.09 | 0.36 | 18.24 | 37.03 |
| RGW | **94.44** | 361.44 | **53.30** | 834.76 | **63.43** | 293.84 | **17.17** | **30.05** | **75.58** | **96.24** |

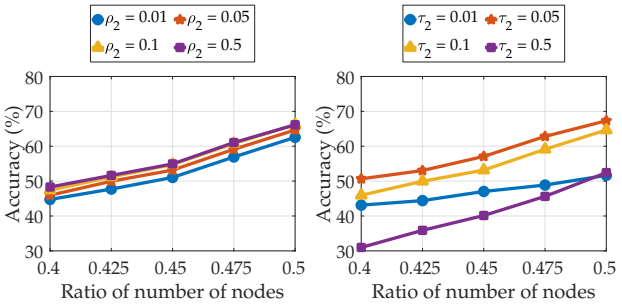



(a) Synthetic (b) Proteins (c) Enzymes

*Figure 3.* Sensitivity of the ratio of the number of nodes of subgraphs and entire graphs for selected methods on Synthetic, Protein, and Enzymes databases.

want to reduce the mass transported from the source domain, which leads to the marginal of the transport plan on the target side, where all samples are clean, will also decrease. As a result, the matching between the clean samples will also be affected. In addition, our proposed RGW achieves the best performance in terms of Hit@1 and Hit@10 on the *Douban Online-Offline* dataset. The Hit@1 and Hit@10 of the initial point created by the features are 4.04% and 14.9%, respectively, showing that RGW improves performance.

**Selection of Hyperparameters $\rho$ and $\tau$** The constants $\rho_1$, $\rho_2$, $\tau_1$, and $\tau_2$ in our formulation are hyperparameters. The values of $\rho_1$ and $\rho_2$ are the amount of marginal relaxation, and $\tau_1$ and $\tau_2$ are the marginal penalty parameters. Although Theorem 2.3 provided a criterion for choosing the values of $\rho_1$ and $\rho_2$ when the fraction of outliers $\epsilon$ is known, this criterion only considers the optimal solution to the RGW. In this case, we use the grid search method to determine $\rho_1$ and $\rho_2$, as well as $\tau_1$ and $\tau_2$. Despite the fact that a larger $\rho_2$ is needed to reduce the impact of outliers as the ratio of the number of nodes of the source graph and the target graph decreases, Figure 4 (a) shows that the alignment accuracy of RGW with a fixed $\rho_2$ does not change dramatically when the ratio of the number of nodes of the target graph and the source graph varies in a small range (0.4 to 0.5). Figure 4 (b) also implies the robustness of $\tau_2$ when the ratio of the number of nodes of the target graph and the source graph varies in a small range. The appendix includes supplementary experiments that demonstrate the robustness of $\rho_2$ and $\tau_2$.




(a) (b)

*Figure 4.* (a): sensitivity of the ratio of the number of nodes of the target graph and the source graph for different $\rho_2$ when $tau_2 = 0.1$ on the Enzymes database. (b): sensitivity of the ratio of the number of nodes of the target graph and source graph for different $\tau_2$ when $\rho_2 = 0.1$ on the Enzymes database.

### 4.3. Partial Shape Correspondence

The problem of partial shape correspondence is to match a subset of the shape to the full version. Analogue to the subgraph, the missing part can be treated as outliers.

**Database Statistics** We evaluate the matching performance of RGW on the TOSCA dataset (Bronstein et al., 2008; Rodolà et al., 2017). In (Rodolà et al., 2017), partial shapes are created by regular cuts, irregular holes, range images, and point clouds. Here, we focus on the matching between the full shape and the partial shapes created by regular cuts. We perform experiments on the 120 shape pairs selected in (Rodolà et al., 2017) and compare the result with the partial functional map method (PFM) introduced in (Rodolà et al., 2017) and nearest-neighbor field method (NNF) (Arbel et al., 2019).

**Parameters Setup** For the large-scale partial shape correspondence problem, we construct $D$ and $\bar{D}$ by computing the Euclidean distance between vertices on the full and partial shapes, respectively. Additionally, we use discrete uniform distributions $\mu$ and $\nu$. It is important to have a good initialization to obtain good performance because the RGW objective is non-convex and partiality introduces undesirable local minima. Therefore, we adopt the partial functional map method using 30 eigenfunctions to obtain a rough initial point for RGW. This method returns a matching relationship between the source and target shapes. It is challenging to construct a feasible $\pi$ for balanced GW methods and PGW from a matching relationship when the marginal distributions $\mu$ and $\nu$ are set to uniform distributions because of the marginal constraint. However, since the RGW formulation relaxes the marginal constraint, its flexibility allows us to use a feasible initial point for $\pi$. The initial point is constructed from the obtained match using the partial functional map as follows: for a given match relation, $\pi_{ik}$ is set to 1 if the pair $(i, k)$ is in the match relation, and 0 otherwise. The resulting transport plan is then scaled with $\|\pi\|_1 = \sum_{ij} \pi_{ij}$. Although this transport plan is not feasible for the balanced GW problem when $\mu$ and $\nu$ are discrete uniform distributions with different dimensions, it is feasible for RGW since it relaxes the marginal constraints. We compare the performance of RGW with partial functional maps using 30 and 50 eigenfunctions and the nearest-neighbor field method.

**Results of All Methods** Given a source shape $\mathcal{N}$ and a target shape $\mathcal{M}$, assume that a correspondence algorithm produces a pair of points $(x, y) \in \mathcal{N} \times \mathcal{M}$, while the ground-truth correspondence is $(x, y^*)$. Then, the geodesic error of the correspondence is calculated by

$$e(x) = \frac{d_{\mathcal{M}}(y, y^*)}{\sqrt{\text{area}(\mathcal{M})}},$$

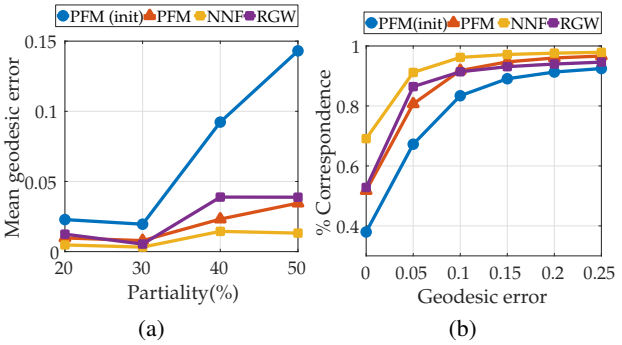



*Figure 5.* Correspondence quality obtain by different methods at increasing levels of partiality.

and the partiality is measured as a percentage of the missing area,

$$\text{partiality} = 1 - \frac{\text{area}(\mathcal{M})}{\text{area}(\mathcal{N})}.$$

The mean geodesic error, which is an indicator of correspondence quality, is shown in Figure 5 (a) for four methods: the partial functional map with 30 eigenfunctions used to generate the initial point for RGW, RGW, partial functional map with 50 eigenfunctions, and nearest-neighbor field method. The level of partiality is increased in each case. As shown in the figure, it is evident that the NNF method outperforms all other methods. Moreover, as the level of partiality increases, the mean geodesic error of the partial functional map with 30 eigenfunctions shows a noticeable increase. The mean geodesic error of RGW remains consistently below 0.05, indicating better performance compared to the partial functional map with 50 eigenfunctions. This demonstrates that using RGW significantly improves the matching accuracy. Additionally, Figure 5 (b) demonstrates through cumulative curves that the percentage of matches drops below the geodesic error threshold. Among the different geodesic error thresholds, the NNF method attains the highest accuracy. When the threshold is greater than 0.05, RGW achieves at least comparable performance to the partial functional map with 50 eigenfunctions. These results demonstrate the potential of utilizing GW-based approaches for partial shape correspondence tasks.

## 5. Conclusion

In this paper, we propose and study RGW, a robust reformulation of Gromov Wasserstein based on distributionally optimistic modeling. Our established theory guarantees that RGW is insensitive to outliers and can serve as a robust estimator for Gromov Wasserstein. We also develop a Bregman proximal alternating linearized method to tackle RGW. Extensive numerical experiments support our theoretical results and demonstrate the effectiveness of the

proposed algorithm. Regarding the robust estimation of Gromov Wasserstein, a natural question is whether we can recover the transport plan from the RGW model. On the computational side, our algorithm suffers from the slow convergence rate of the Sinkhorn algorithm for unbalanced optimal transport, which will further limit its applications in large-scale real-world settings. A future direction is to incorporate a faster algorithm for solving unbalanced optimal transport (Séjourné et al., 2022) and further speed up our algorithm.

## References


Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Ahookhosh, M., Hien, L. T. K., Gillis, N., and Patrinos, P. Multi-block bregman proximal alternating linearized minimization and its application to orthogonal nonnegative matrix factorization. *Computational Optimization and Applications*, 79(3):681–715, 2021.

Alvarez-Melis, D. and Jaakkola, T. S. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

Arbel, N. Y., Tal, A., and Zelnik-Manor, L. Partial correspondence of 3d shapes using properties of the nearest-neighbor field. *Computers & Graphics*, 82:183–192, 2019.

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

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

Bronstein, A. M., Bronstein, M. M., and Kimmel, R. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. Learning generative models across incomparable spaces. In *International conference on machine learning*, pp. 851–861. PMLR, 2019.

Cao, J. and Gao, R. Contextual decision-making under parametric uncertainty and data-driven optimistic optimization. *Available at Optimization Online*, 2021.

Chapel, L., Alaya, M. Z., and Gasso, G. Partial optimal transport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33: 2903–2913, 2020.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

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

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

Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. A (sub) graph isomorphism algorithm for matching large graphs. *IEEE transactions on pattern analysis and machine intelligence*, 26(10):1367–1372, 2004.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

Gao, J., Huang, X., and Li, J. Unsupervised graph alignment with wasserstein distance discriminator. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 426–435, 2021.

Grave, E., Joulin, A., and Berthet, Q. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1880–1890. PMLR, 2019.

Han, M., Kim, H., Gu, G., Park, K., and Han, W.-S. Efficient subgraph matching: Harmonizing dynamic programming, adaptive matching order, and failing set together. In *Proceedings of the 2019 International Conference on Management of Data*, pp. 1429–1446, 2019.

Jiang, N. and Xie, W. Dfo: A framework for data-driven decision-making with endogenous outliers. *Preprint optimization-online.org*, 2021.

Lacoste-Julien, S., Taskar, B., Klein, D., and Jordan, M. Word alignment via quadratic assignment. 2006.

Lawler, E. L. The quadratic assignment problem. *Management science*, 9(4):586–599, 1963.

Le, K., Nguyen, H., Nguyen, Q. M., Pham, T., Bui, H., and Ho, N. On robust optimal transport: Computational complexity and barycenter computation. *Advances in Neural Information Processing Systems*, 34:21947–21959, 2021.

Li, J., Tang, J., Kong, L., Liu, H., Li, J., So, A. M.-C., and Blanchet, J. Fast and provably convergent algorithms for gromov-wasserstein in graph learning. *arXiv preprint arXiv:2205.08115*, 2022.

Mémoli, F. On the use of gromov-hausdorff distances for shape comparison. 2007.

Mémoli, F. Spectral gromov-wasserstein distances for shape matching. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 256–263. IEEE, 2009.

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

Mukherjee, D., Guha, A., Solomon, J. M., Sun, Y., and Yurochkin, M. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pp. 7850–7860. PMLR, 2021.

Munos, R. et al. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7 (1):1–129, 2014.

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

Pham, K., Le, K., Ho, N., Pham, T., and Bui, H. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *International Conference on Machine Learning*, pp. 7673–7682. PMLR, 2020.

Rodola, E., Bronstein, A. M., Albarelli, A., Bergamasco, F., and Torsello, A. A game-theoretic approach to deformable shape matching. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 182–189. IEEE, 2012.

Rodolà, E., Cosmo, L., Bronstein, M. M., Torsello, A., and Cremers, D. Partial functional correspondence. In *Computer graphics forum*, volume 36, pp. 222–236. Wiley Online Library, 2017.

Séjourné, T., Vialard, F.-X., and Peyré, G. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34, 2021.

Séjourné, T., Vialard, F.-X., and Peyré, G. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pp. 4995–5021. PMLR, 2022.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Staerman, G., Laforgue, P., Mozharovskyi, P., and d'Alché Buc, F. When ot meets mom: Robust estimation of wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pp. 136–144. PMLR, 2021.

Titouan, V., Courty, N., Tavenard, R., and Flamary, R. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Tran, Q. H., Janati, H., Courty, N., Flamary, R., Redko, I., Demetci, P., and Singh, R. Unbalanced co-optimal transport. In *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Vincent-Cuaz, C., Vayer, T., Flamary, R., Corneli, M., and Courty, N. Online graph dictionary learning. In *International Conference on Machine Learning*, pp. 10564–10574. PMLR, 2021.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Semi-relaxed gromov wasserstein divergence with applications on graphs. In *International Conference on Learning Representations (ICLR)*, 2022.

Xu, H., Luo, D., and Carin, L. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019a.

Xu, H., Luo, D., Zha, H., and Duke, L. C. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pp. 6932–6941. PMLR, 2019b.

Xu, H., Luo, D., Carin, L., and Zha, H. Learning graphons via structured gromov-wasserstein barycenters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10505–10513, 2021.

Xu, H., Liu, J., Luo, D., and Carin, L. Representing graphs via gromov-wasserstein factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

# A. Organization of the Appendix

We organize the appendix as follows:

- The proof details of Theorem 2.3 is given in Section B.
- The proof details of the algorithm, including the properties of $p$, the convergence analysis of Newton's method, and Bregman proximal alternating linearized minimization method are collected in C.
- Additional experiment results are summarized in Section D.

# B. Proof of Theorem 2.3

*Proof.* $\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mu,\nu)$ is defined as

$$\begin{aligned}\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mathbb{P}_X,\mathbb{P}_Y) = \inf_{\alpha\in\mathcal{P}(X)}\ \inf_{\pi\in\mathcal{M}^+(X\times Y)} &\iint |d_X(x,x')-d_Y(y,y'))|^2 d\pi(x,y)d\pi\left(x',y'\right)+\tau_1 d_{\mathbf{KL}}(\pi_1,\alpha)+\tau_2 d_{\mathbf{KL}}(\pi_2,\nu)\\ &\text{s.t.}\quad d_{\mathbf{KL}}(\mu,\alpha)\le\rho.\end{aligned}$$

Let $\pi_c$ be the optimal transport plans for $\mathrm{GW}(\mu_c,\nu)$. $\pi_c$ is a feasible solution for $\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mu,\nu)$. Then we can have

$$\begin{aligned}\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mu,\nu) &\le \inf_{\alpha\in\mathcal{P}(X)}\iint |d_X(x,x')-d_Y(y,y'))|^2 d\pi_c(x,y)d\pi_c\left(x',y'\right)+\tau_1 d_{\mathbf{KL}}((\pi_c)_1,\alpha)+\tau_2 d_{\mathbf{KL}}((\pi_c)_2,\nu)\\ &\quad\text{s.t.}\quad d_{\mathbf{KL}}(\mu,\alpha)\le\rho\\ &= \inf_{\alpha\in\mathcal{P}(X)}\mathrm{GW}(\mu_c,\nu)+\tau_1 d_{\mathbf{KL}}(\mu_c\|\alpha)\\ &\quad\text{s.t.}\quad d_{\mathbf{KL}}(\mu,\alpha)\le\rho\end{aligned}$$

To give an upper bound of $\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mu,\nu)$, let us focus on the following problem:

$$\begin{aligned}\inf_{\alpha\in\mathcal{P}(X)}\quad & d_{\mathbf{KL}}(\mu_c,\alpha)\\ \text{s.t.}\quad & d_{\mathbf{KL}}(\mu,\alpha)\le\rho\end{aligned}\tag{10}$$

We consider the distribution of the form $(1-\beta)\mu+\beta\mu_c$, for $\beta\in[0,1]$. Then we prove that if $\beta\le\min\left(\dfrac{\rho}{\epsilon d_{\mathbf{KL}}(\mu_a,\mu_c)},1\right)$, then $(1-\beta)\mu+\beta\mu_c$ is a feasible solution for problem (10).

By the joint convexity of f-divergence, we have

$$\begin{aligned}d_{\mathbf{KL}}\left(\mu,(1-\beta)\mu+\beta\mu_c\right) &\le \beta d_{\mathbf{KL}}\left(\mu,\mu_c\right)\\ &= \beta d_{\mathbf{KL}}\left((1-\epsilon)\mu_c+\epsilon\mu_a,\mu_c\right)\\ &\le \beta\epsilon d_{\mathbf{KL}}(\mu_a,\mu_c)\\ &\le \rho.\end{aligned}$$

Therefore,

$$\begin{aligned}d_{\mathbf{KL}}\left(\mu_c,(1-\beta)\mu+\beta\mu_c\right) &\le (1-\beta) d_{\mathbf{KL}}\left(\mu_c,\mu\right)\\ &= (1-\beta) d_{\mathbf{KL}}\left(\mu_c\|(1-\epsilon)\mu_c+\epsilon\mu_a\right)\\ &\le (1-\beta)\epsilon d_{\mathbf{KL}}(\mu_c,\mu_a)\end{aligned}$$

The largest value $\beta$ can take is $\dfrac{\rho}{\epsilon d_{\mathbf{KL}}(\mu_a,\mu_c)}$. This gives

$$\inf_{\alpha\in\mathcal{P}(X),\ d_{\mathbf{KL}}(\mu,\alpha)\le\rho} d_{\mathbf{KL}}(\mu_c,\alpha)\le\max\left(0,1-\frac{\rho}{\epsilon d_{\mathbf{KL}}(\mu_a,\mu_c)}\right)\epsilon d_{\mathbf{KL}}(\mu_c,\mu_a),$$

and

$$\mathrm{GW}^{\mathrm{rob}}_{\rho,0}(\mu,\nu) \leq \mathrm{GW}(\mu_c,\nu) + \max\left(0, \epsilon - \frac{\rho}{d_{\mathbf{KL}}(\mu_a,\mu_c)}\right)\tau_1 d_{\mathbf{KL}}(\mu_c,\mu_a).$$

□

## C. Proof Details of Bregman Proximal Alternating Linearized Method for Robust GW

Given a vector x, we use $\|x\|_2$ to denote its $\ell_2$ norm. We use $\|X\|_F$ to denote the Frobenius norm of martix $X$. For a convex set $C$ and a point $x$, we define the distance between $C$ and $x$ as

$$\mathrm{dist}(x,C) = \min_{y\in C}\|x-y\|_2.$$

### C.1. Proof of Proposition 3.1

*Proof.* Since if $w$ satisfies (i) or (ii), $(\hat{\alpha}(w), w)$ is a solution to KKT conditions of problem (5), therefore, $\hat{\alpha}(w)$ is an optimal solution to problem (5). Next, we prove that $p$ is differentiable when $h$ is relative entropy. Problem (7) admits the closed-form solution

$$\hat{\alpha}(w) = \frac{\pi^{k+1}\mathbf{1}_m + \frac{1}{c_k}\alpha^k + w\mu}{\sum_{i,j}\pi^{k+1}_{ij} + \frac{1}{c_k} + w}. \tag{11}$$

By substituting (11) into $p$, $p(w)$ can be written as

$$p(w) = \sum_{i=1}^n \mu_i \log\left(\frac{\mu_i\left(\sum_{i,j}\pi^{k+1}_{ij} + \frac{1}{c_k} + w\right)}{\left(\pi^{k+1}\mathbf{1}_m\right)_i + \frac{1}{c_k}\alpha^k_i + w\mu_i}\right) - \rho_1. \tag{12}$$

Thus, $p$ is twice differentiable. The first-order derivative and second-order of $p$ are

$$p'(w) = \sum_{i=1}^n \mu_i \frac{\left(\left(\pi^{k+1}\mathbf{1}_m\right)_i + \frac{1}{c_k}\alpha^k_i + \mu_i w\right) - \mu_i\left(\sum_{ij}\pi^{k+1}_{ij} + \frac{1}{c_k} + w\right)}{\left(\sum_{ij}\pi^{k+1}_{ij} + \frac{1}{c_k} + w\right)\left(\left(\pi^{k+1}\mathbf{1}_m\right)_i + \frac{1}{c_k}\alpha^k_i + \mu_i w\right)}, \tag{13}$$

and

$$p''(w) = -\sum_{i=1}^n \mu_i \frac{\left(\left(\pi^{k+1}\mathbf{1}_m\right)_i + \frac{1}{c_k}\alpha^k_i + \mu_i w\right)^2 - \mu_i^2\left(\sum_{ij}\pi^{k+1}_{ij} + \frac{1}{c_k} + w\right)^2}{\left(\sum_{ij}\pi^{k+1}_{ij} + \frac{1}{c_k} + w\right)^2\left(\left(\pi^{k+1}\mathbf{1}_m\right)_i + \frac{1}{c_k}\alpha^k_i + \mu_i w\right)^2}. \tag{14}$$

Then we prove that $p'(w) \leq 0$ and $p''(w) \geq 0$ for $w \geq 0$, so $p$ is monotonically non-increasing and convex on $\mathbb{R}_+$. Let $s_i = \left(\pi^{k+1}\mathbf{1}_m\right)_i + \frac{1}{c_k}\alpha^k_i + \mu_i w$ and $s = \left(\pi^{k+1}\mathbf{1}_m\right)_i + \frac{1}{c_k}\alpha^k_i + \mu_i w$. Note that $s = \sum_{i=1}^n s_i$. Then $p'$ and $p''$ can be written as

$$p'(w) = \sum_{i=1}^n \mu_i \frac{s_i - \mu_i s}{s_i \cdot s} = \frac{1}{s}\left(1 - \sum_{i=1}^n \mu_i \frac{\mu_i s}{s_i}\right),$$

and

$$p''(w) = -\sum_{i=1}^n \mu_i \frac{s_i^2 - \mu_i^2 s^2}{s_i^2 \cdot s^2} = -\frac{1}{s^2}\left(1 - \sum_{i=1}^n \mu_i \frac{\mu_i^2 s^2}{s_i^2}\right).$$

Therefore, it is equivalent to show $\sum_{i=1}^n \mu_i \frac{\mu_i s}{s_i} \geq 1$ and $\sum_{i=1}^n \mu_i \frac{\mu_i^2 s^2}{s_i^2} \geq 1$.

Recall that $\frac{1}{x}$ and $\frac{1}{x^2}$ are convex on $\mathbb{R}_{++}$, then

$$\sum_{i=1}^n \mu_i \frac{\mu_i s}{s_i} = \sum_{i=1}^n \frac{1}{\mu_i \frac{s_i}{\mu_i s}} \geq \frac{1}{\sum_{i=1}^n \mu_i \frac{s_i}{\mu_i s}} = 1,$$

$$\sum_{i=1}^{n}\mu_i\frac{\mu_i^2 s^2}{s_i^2}=\sum_{i=1}^{n}\mu_i\frac{1}{\left(\frac{s_i}{\mu_i s}\right)^2}\geq\frac{1}{\left(\sum_{i=1}^{n}\mu_i\frac{s_i}{\mu_i s}\right)^2}=1.$$

□

### C.2. Proof of Proposition 3.2

*Proof.* First, prove that $p$ only has one root on $I$. Since $p$ is continuous on $I$ and there exists $\tilde{x}$, $\bar{x}\in I$ such that $p(\tilde{x})>0$ and $p(\bar{x})<0$, $p$ contains at least one root on $[\tilde{x},\bar{x}]$. Since $p$ is non-increasing, $p$ cannot have roots outside $[\tilde{x},\bar{x}]$. Suppose that $p$ have two different roots $z_1$ and $z_2$ on $[\tilde{x},\bar{x}]$ and $z_1<z_2$. By convexity of $p$, we have

$$0=p(z_2)=p\left(\frac{\bar{x}-z_2}{\bar{x}-z_1}z_1+\frac{z_2-z_1}{\bar{x}-z_1}\bar{x}\right)\leq\frac{\bar{x}-z_2}{\bar{x}-z_1}p(z_1)+\frac{z_2-z_1}{\bar{x}-z_1}p(\bar{x})=\frac{z_2-z_1}{\bar{x}-z_1}p(\bar{x})<0.$$

This is a contradiction. So $p$ has a unique root on $I$.

$p'(x)\leq 0$ since $p$ is non-increasing on $I$. Denote the root of $p$ as $r$. Claim that $p'(x)<0$ for $x\in[\tilde{x},r]$. Otherwise, there exist $x\in[\tilde{x},r]$ such that $p'(x)=0$, then

$$0>p(\bar{x})\geq p(x)+p'(x)(\bar{x}-x)=p(x)\geq 0,$$

which leads to a contradiction. Especially, $p'(\tilde{x})<0$, and we can set $\tilde{x}$ as the initial point of Newton's method.

The update is Newton's method is

$$x_{k+1}=x_k-\frac{p(x_k)}{p'(x_k)}.$$

Therefore, $x_{k+1}\geq x_k$ and $\{x_k\}_{k\geq 1}$ is an increasing sequence. Since $p$ is convex,

$$p(x_{k+1})\geq p(x_k)+p'(x_k)(x_{k+1}-x_k)=p(x_k)-p(x_k)=0.$$

$x_k\leq r$ because $p$ is a monotonically non-increasing function. $\{x_k\}_{k\geq 1}$ is an increasing sequence with an upper bound, so it has a limit $x^*$ and $\lim_{k\to\infty}(x_k-x_{k+1})=0$. Also, $p'$ is bounded on $[\tilde{x},r]$ since it is continuous. Therefore,

$$p(x^*)=\lim_{k\to+\infty}p(x_k)=\lim_{k\to+\infty}p'(x_k)(x_k-x_{k+1})=0.$$

Hence, the sequence generated by Newton's method converges to a root of $p$. □

### C.3. Proof of Theorem 3.3

**Assumption C.1.** The critical point set $\mathcal{X}$ is non-empty.

**Definition C.2** (Bregman Divergence)**.** We define the proximity measure $D_h:\mathrm{dom}(h)\times\mathrm{int}(\mathrm{dom}(h))\to\mathbb{R}_+$

$$D_h(x,y):=h(x)-h(y)-\langle\nabla h(y),x-y\rangle$$

The proximity measure $D_h$ is the so-called Bregman Distance. It measures the proximity between $x$ and $y$. Indeed, thanks to the gradient inequality, one has $h$ is convex if and only if $D_h(x,y)\geq 0,\forall x\in\mathrm{dom}\,h, y\in\mathrm{int}\,\mathrm{dom}\,h$.

Before the proof of Theorem 3.3, we first prove that sequence $\{\pi^k\}_{k\geq 0}$ generated by BPALM lies in a compact set.

**Proposition C.3.** *Sequence $\{\pi^k\}_{k\geq 0}$ generated by BPALM lies in a compact set.*

*Proof.* We prove that $\{\pi^k\}_{k\geq 0}$ lies in the compact set $\mathcal{A}:=\{\pi\in\mathbb{R}^{n\times m}:0\leq\pi_{ij}\leq 1\}$ by mathematical induction. For $k=0$, we can initialize $\pi^0$ with $0\leq\pi_{ij}^0\leq 1$, which is reasonable for proper initialization. Suppose that $\pi^k\in\mathcal{A}$ and $\pi^{k+1}\notin\mathcal{A}$. Then there exists $i\in[n]$ and $j\in[m]$ such that $\pi_{ij}^{k+1}>1$. Recall that $\pi^{k+1}$ is the optimal solution to the problem

$$\min_{\pi\geq 0}\ \varphi(\pi):=\langle\mathcal{L}(D,\bar{D})\otimes\pi^k,\pi\rangle+\tau_1 d_{\mathbf{KL}}(\pi\mathbf{1}_m,\alpha^k)+\tau_2 d_{\mathbf{KL}}(\pi^T\mathbf{1}_n,\beta^k)+\frac{1}{t_k}d_{\mathbf{KL}}(\pi,\pi^k),\tag{15}$$

Observe that function $\phi(x) \coloneqq x \log \frac{x}{a} - x + a$ is a unimodal function on $\mathbb{R}_+$ and achieve its minimum at $x = a$. Since $\alpha_i$, $\beta_j$, and $\pi_{ij}^k$ are smaller than or equal to 1, $\alpha_i$, $\beta_j$, and $\pi_{ij}^k$ are strictly smaller than $\pi_{ij}^{k+1}$. Let $\tilde{\pi} \in \mathbb{R}^{n \times m}$,

$$\tilde{\pi}_{kl} = \begin{cases} \max\{\alpha_i, \beta_j, \pi_{ij}^k\}, & (k,l) = (i,j), \\ \pi_{kl}^{k+1}, & \text{otherwise.} \end{cases}$$

Then $\varphi(\tilde{\pi}) < \varphi(\pi^k)$, this contradicts to $\pi^{k+1}$ is the optimal solution to problem (15). Thus, $\pi^{k+1} \in \mathcal{A}$. □

For the proof of Theorem 3.3, we first prove the sufficient decrease property of BPALM, i.e., there exist a constant $\kappa_1 > 0$ and an index $k_1 \geq 0$ such that for $k \geq k_1$,

$$F\left(\pi^{k+1}, \alpha^{k+1}, \beta^{k+1}\right) - F\left(\pi^k, \alpha^k, \beta^k\right) \leq -\kappa_1 \left(\left\|\pi^{k+1} - \pi^k\right\|_F^2 + \left\|\alpha^{k+1} - \alpha^k\right\|_2^2 + \left\|\beta^{k+1} - \beta^k\right\|_2^2\right).$$

And then prove the subsequence convergence result.

*Proof.* It is worthwhile noting that $f(\pi)$ is a quadratic function, i.e., $f(\pi) = \left\langle \mathcal{L}(D, \bar{D}) \otimes \pi, \pi \right\rangle$, then $f(\pi)$ is gradient Lipschitz continuous with the constant $\max_{i,j}\left(\sum_{k,l} \mathcal{L}(D,\bar{D})_{i,j,k,l}^2\right)^{1/2}$. To simplify the notation, let $L_f = \max_{i,j}\left(\sum_{k,l} \mathcal{L}(D,\bar{D})_{i,j,k,l}^2\right)^{1/2}$.

$$\begin{aligned}
& F\left(\pi^{k+1}, \alpha^{k+1}, \beta^{k+1}\right) - F\left(\pi^k, \alpha^k, \beta^k\right) \\
\leq\ & \left\langle \nabla f(\pi^k), \pi^{k+1} - \pi^k \right\rangle + \frac{L_f}{2}\left\|\pi^{k+1} - \pi^k\right\|_F^2 + g_1\left(\pi^{k+1}, \alpha^{k+1}\right) + g_2\left(\pi^{k+1}, \beta^{k+1}\right) + h_1\left(\alpha^{k+1}\right) + h_2\left(\beta^{k+1}\right) - \\
& \left(g_1\left(\pi^k, \alpha^k\right) + g_2\left(\pi^k, \beta^k\right) + h_1\left(\alpha^k\right) + h_2\left(\beta^k\right)\right) \\
\overset{(\diamondsuit)}{\leq}\ & \left\langle \nabla f(\pi^k), \pi^{k+1} - \pi^k \right\rangle + \frac{L_f}{\sigma} d_{\mathbf{KL}}\left(\pi^{k+1}, \pi^k\right) + g_1(\pi^{k+1}, \alpha^{k+1}) + g_2\left(\pi^{k+1}, \beta^{k+1}\right) + h_1\left(\alpha^{k+1}\right) + h_2\left(\beta^{k+1}\right) - \\
& \left(g_1\left(\pi^k, \alpha^k\right) + g_2\left(\pi^k, \beta^k\right) + h_1\left(\alpha^k\right) + h_2\left(\beta^k\right)\right) \\
=\ & \left\langle \nabla f(\pi^k), \pi^{k+1} \right\rangle + g_1\left(\pi^{k+1}, \alpha^k\right) + g_2\left(\pi^{k+1}, \beta^k\right) + \frac{1}{t_k} d_{\mathbf{KL}}\left(\pi^{k+1}, \pi^k\right) - \left\langle \nabla f(\pi^k), \pi^k \right\rangle - g_1\left(\pi^k, \alpha^k\right) - g_2\left(\pi^k, \beta^k\right) + \\
& g_1\left(\pi^{k+1}, \alpha^{k+1}\right) + h_1\left(\alpha^{k+1}\right) + \frac{1}{c_k} d_{\mathbf{KL}}\left(\alpha^k, \alpha^{k+1}\right) - g_1\left(\pi^k, \alpha^k\right) - h_1\left(\alpha^k\right) + \\
& g_2\left(\pi^{k+1}, \beta^{k+1}\right) + h_2\left(\beta^{k+1}\right) + \frac{1}{r_k} d_{\mathbf{KL}}\left(\beta^k, \beta^{k+1}\right) - g_2\left(\pi^k, \beta^k\right) - h_2\left(\beta^k\right) - \\
& \left(\frac{1}{t_k} - \frac{L_f}{\sigma}\right) d_{\mathbf{KL}}\left(\pi^{k+1}, \pi^k\right) - \frac{1}{c_k} d_{\mathbf{KL}}\left(\alpha^k, \alpha^{k+1}\right) - \frac{1}{r_k} d_{\mathbf{KL}}\left(\beta^k, \beta^{k+1}\right) \\
\leq\ & -\left(\frac{1}{t_k} - \frac{L_f}{\sigma}\right) d_{\mathbf{KL}}\left(\pi^{k+1}, \pi^k\right) - \frac{1}{c_k} d_{\mathbf{KL}}\left(\alpha^k, \alpha^{k+1}\right) - \frac{1}{r_k} d_{\mathbf{KL}}\left(\beta^k, \beta^{k+1}\right). \\
\overset{(\diamondsuit)}{\leq}\ & -\frac{\sigma}{2}\left(\left(\frac{1}{t_k} - \frac{L_f}{\sigma}\right)\left\|\pi^{k+1} - \pi^k\right\|_F^2 + \frac{1}{c_k}\left\|\alpha^{k+1} - \alpha^k\right\|_2^2 + \frac{1}{r_k}\left\|\beta^{k+1} - \beta^k\right\|_2^2\right).
\end{aligned}$$

$(\diamondsuit)$ is because as $h$ is a $\sigma$-strongly convex, we have

$$d_{\mathbf{KL}}(\pi^{k+1}, \pi^k) \geq \frac{\sigma}{2}\|\pi^{k+1} - \pi^k\|_F^2, \quad d_{\mathbf{KL}}(\alpha^k, \alpha^{k+1}) \geq \frac{\sigma}{2}\|\alpha^{k+1} - \alpha^k\|_2^2, \quad d_{\mathbf{KL}}(\beta^k, \beta^{k+1}) \geq \frac{\sigma}{2}\|\beta^{k+1} - \beta^k\|_2^2.$$

By letting $\kappa_1 = \min\left(\left(\frac{1}{t_k} - \frac{L_f}{\sigma}\right), \frac{1}{\bar{r}}\right) > 0$, we get

$$F\left(\pi^{k+1}, \alpha^{k+1}, \beta^{k+1}\right) - F\left(\pi^k, \alpha^k, \beta^k\right) \leq -\kappa_1 \left(\left\|\pi^{k+1} - \pi^k\right\|_F^2 + \left\|\alpha^{k+1} - \alpha^k\right\|_2^2 + \left\|\beta^{k+1} - \beta^k\right\|_2^2\right). \tag{16}$$

Summing up (16) from $k = 0$ to $+\infty$, we obtain

$$F(\pi^\infty, \alpha^\infty, \beta^\infty) - F(\pi^0, \alpha^0, \beta^0) \leq -\kappa_1 \sum_{k=0}^{\infty} \left(\left\|\pi^{k+1} - \pi^k\right\|_F^2 + \left\|\alpha^{k+1} - \alpha^k\right\|_2^2 + \left\|\beta^{k+1} - \beta^k\right\|_2^2\right).$$

As $F$ is coercive and $\left\{\left(\pi^k, \alpha^k, \beta^k\right)\right\}$ is a bounded sequence, it means the left-hand side, which implies

$$\sum_{k=0}^{\infty}\left(\left\|\pi^{k+1}-\pi^k\right\|_F^2+\left\|\alpha^{k+1}-\alpha^k\right\|_2^2+\left\|\beta^{k+1}-\beta^k\right\|_2^2\right)<+\infty,$$

and

$$\lim_{k\to+\infty}\left(\|\pi^{k+1}-\pi^k\|_F+\|\alpha^{k+1}-\alpha^k\|_2+\|\beta^{k+1}-\beta^k\|_2\right)=0.$$

Let $h(x)=\sum_i x_i\log x_i$. Recall the optimality condition of BPALM, we have

$$0=\nabla f\left(\pi^{k+1}\right)+\nabla_\pi g_1\left(\pi^{k+1},\alpha^k\right)+\nabla_\pi g_2\left(\pi^{k+1},\beta^k\right)+\frac{1}{t_k}\left(\nabla h\left(\pi^{k+1}\right)-\nabla h\left(\pi^k\right)\right), \tag{17}$$

$$0\in\nabla_\alpha g_1\left(\pi^{k+1},\alpha^{k+1}\right)+\partial h_1\left(\alpha^{k+1}\right)+\frac{1}{c_k}\nabla^2 h(\alpha^{k+1})(\alpha^{k+1}-\alpha^k), \tag{18}$$

$$0\in\nabla_\beta g_2\left(\pi^{k+1},\beta^{k+1}\right)+\partial h_2\left(\beta^{k+1}\right)+\frac{1}{r_k}\nabla^2 h(\beta^{k+1})(\beta^{k+1}-\beta^k). \tag{19}$$

Let $(\pi^\infty,\alpha^\infty,\beta^\infty)$ be a limit point of the sequence $\left\{(\pi^k,\alpha^k,\beta^k)\right\}_{k\geq 0}$. Then, there exists a sequence $\{n_k\}_{k\geq 0}$ such that $\{(\pi^{n_k},\alpha^{n_k},\beta^{n_k})\}_{k\geq 0}$ converges to $(\pi^\infty,\alpha^\infty,\beta^\infty)$. Since we assume that $h$ is twice continuous differentiable and $\alpha^k$ and $\beta^k$ are in a compact set, then $\nabla^2 h(\alpha^k)$ and $\nabla^2 h(\beta^k)$ are bounded. Therefore, $\lim_{k\to\infty}\nabla^2 h(\alpha^{k+1})(\alpha^{k+1}-\alpha^k)=0$ and $\lim_{k\to\infty}\nabla^2 h(\beta^{k+1})(\beta^{k+1}-\beta^k)=0$. Replacing the $k$ by $n_k$ in (17), (18), and (19), taking limits on both sides as $k\to\infty$, we obtain that

$$\begin{aligned}
&0=\nabla f\left(\pi^\infty\right)+\nabla_\pi g_1\left(\pi^\infty,\alpha^\infty\right)+\nabla_\pi g_2\left(\pi^\infty,\beta^\infty\right),\\
&0\in\nabla_\alpha g_1\left(\pi^\infty,\alpha^\infty\right)+\partial h_1\left(\alpha^\infty\right),\\
&0\in\nabla_\beta g_2\left(\pi^\infty,\beta^\infty\right)+\partial h_2\left(\beta^\infty\right).
\end{aligned}$$

Thus $(\pi^\infty,\alpha^\infty,\beta^\infty)$ belongs to $\mathcal{X}$.

□

# D. Additional Experiment Results

## D.1. Additional Experiment Results on Subgraph Alignment

*Table 2.* Comparison of the average matching accuracy (%) and wall-clock time (seconds) on subgraph alignment of 30% subgraph and 20% subgraph.

| Method | 30% subgraph | | | | | | 20% subgraph | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Synthetic | | Proteins | | Enzymes | | Synthetic | | Proteins | | Enzymes | |
| | Acc | Time | Acc | Time | Acc | Time | Acc | Time | Acc | Time | Acc | Time |
| FW | 2.22 | 17.06 | 12.96 | 14.83 | 12.08 | 5.37 | 2.24 | 6.65 | 10.83 | 11.34 | 9.53 | 4.92 |
| SpecGW | 1.38 | 2.24 | 10.64 | 11.54 | 9.41 | 3.74 | 1.71 | 2.21 | 10.78 | 10.15 | 8.35 | 3.21 |
| eBPG | 0.65 | 0.49 | 8.12 | 1022.02 | 3.84 | 476.83 | 1.17 | 0.42 | 7.23 | 545.50 | 2.66 | 94.78 |
| BPG | 1.86 | 17.53 | 17.89 | 86.85 | 17.69 | 52.89 | 1.64 | 11.66 | 12.99 | 55.47 | 14.35 | 32.89 |
| BAPG | 2.94 | 35.90 | 18.79 | 36.02 | 16.85 | 10.88 | 3.80 | 23.29 | 14.07 | 23.92 | 11.22 | 8.38 |
| srGW | 3.17 | 86.38 | 22.75 | 89.14 | 27.45 | 41.18 | 5.49 | 88.89 | 18.38 | 17.72 | 23.13 | 17.11 |
| UGW | 25.34 | 2242.55 | 14.32 | 10298 | 10.91 | 5552.27 | 4.48 | 1501.11 | 11.75 | 7813.96 | 10.40 | 4019.62 |
| PGW | 2.06 | 339.23 | 11.68 | 507.11 | 11.77 | 174.26 | 1.90 | 227.87 | 9.34 | 365.88 | 7.97 | 165.27 |
| RGW | **52.35** | 679.00 | **30.17** | 947.48 | **37.12** | 538.04 | **11.58** | 229.05 | **23.51** | 546.15 | **25.39** | 879.93 |

**Source codes of all baselines used in this paper:**

- FW (Flamary et al., 2021): `https://github.com/PythonOT/POT`

*Table 3.* Comparison of the average matching accuracy (%) and wall-clock time (seconds) on subgraph alignment of 40% subgraph.

| Method | 40% subgraph | | | | | |
|---|---|---|---|---|---|---|
| | Synthetic | | Proteins | | Enzymes | |
| | Acc | Time | Acc | Time | Acc | Time |
| FW | 1.84 | 17.96 | 15.34 | 19.64 | 14.22 | 6.36 |
| SpecGW | 1.72 | 3.25 | 11.21 | 12.17 | 9.59 | 3.88 |
| eBPG | 0.38 | 0.51 | 12.16 | 1628.38 | 9.96 | 943.49 |
| BPG | 3.41 | 18.61 | 24.31 | 108.10 | 24.58 | 62.81 |
| BAPG | 7.61 | 22.55 | 23.78 | 36.81 | 24.82 | 11.13 |
| srGW | 2.45 | 120.12 | 22.58 | 74.58 | 27.02 | 32.14 |
| UGW | 79.61 | 960.04 | 21.22 | 11398 | 22.26 | 5589.73 |
| PGW | 2.17 | 483.10 | 11.95 | 491.64 | 9.51 | 182.58 |
| RGW | **90.79** | 662.15 | **38.94** | 769.25 | **48.11** | 291.74 |

- SpecGW (Chowdhury & Needham, 2021): https://github.com/trneedham/Spectral-Gromov-Wasserstein
- eBPG (Flamary et al., 2021): https://github.com/PythonOT/POT
- UGW (Séjourné et al., 2021): https://github.com/thibsej/unbalanced_gromov_wasserstein
- PGW (Chapel et al., 2020; Flamary et al., 2021): https://github.com/PythonOT/POT
- srGW: (Vincent-Cuaz et al., 2022): https://github.com/cedricvincentcuaz/srGW

### D.2. Additional Experiment Results on testing of robustness of $\rho_2$ and $\tau_2$

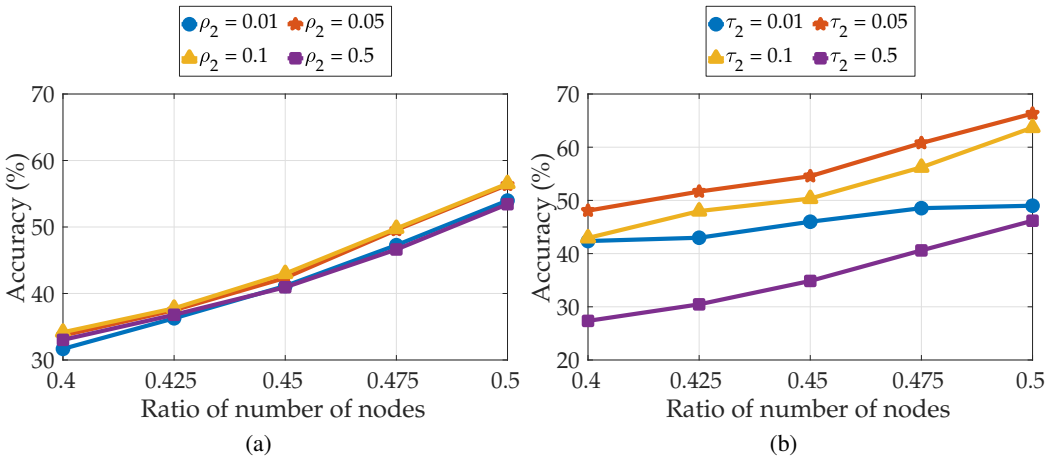



*Figure 6.* (a): sensitivity of the ratio of the number of nodes of the target graph and the source graph for different $\rho_2$ when $\tau_2 = 0.05$ on the Enzymes database. (b): sensitivity of the ratio of the number of nodes of the target graph and source graph for different $\tau_2$ when $\rho_2 = 0.05$ on the Enzymes database.

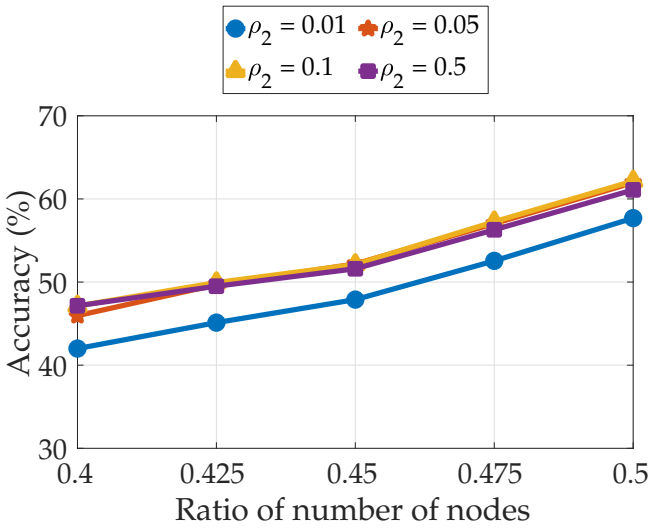



*Figure 7.* sensitivity of the ratio of the number of nodes of the target graph and the source graph for different $\rho_2$ when $\tau_2 = 0.05$ and the marginal distribution is normalized degree on the Enzymes database.