Efficient Solvers for Partial Gromov-Wasserstein

Anonymous Authors[1]

Abstract

The partial Gromov-Wasserstein (PGW) problem facilitates the comparison of measures with unequal masses residing in potentially distinct metric spaces, thereby enabling unbalanced and partial matching across these spaces. In this paper, we demonstrate that the PGW problem can be transformed into a variant of the Gromov-Wasserstein problem, akin to the conversion of the partial optimal transport problem into an optimal transport problem. This transformation leads to two new solvers, mathematically and computationally equivalent, based on the Frank-Wolfe algorithm, that provide efficient solutions to the PGW problem. We further establish that the PGW problem constitutes a metric for metric measure spaces. Finally, we validate the effectiveness of our proposed solvers in terms of computation time and performance on shape-matching and positive-unlabeled learning problems, comparing them against existing baselines.

1. Introduction

The classical optimal transport (OT) problem (Villani, 2009) seeks to match two probability measures while minimizing what is known as the expected transportation cost. At the heart of classical OT theory lies the principle of mass conservation, aiming to optimize the transportation from one probability measure to another under the premise that both measures maintain the same total mass, with strict preservation of mass throughout the transportation process. Statistical distances that arise from OT, such as the Wasserstein distances, have been widely applied across various machine learning domains, ranging from generative modeling (Arjovsky et al., 2017; Gulrajani et al., 2017) to domain adaptation (Courty et al., 2017) and representation learning (Kolouri et al., 2020). Notably, recent advancements have extended the OT problem to address certain limitations within machine learning applications. These advancements include: 1) facilitating the comparison of non-negative measures that possess different total masses via unbalanced (Chizat et al., 2018c) and partial OT (Figalli, 2010), and 2) enabling the comparison of probability measures across distinct metric spaces through Gromov-Wasserstein distances (Mémoli, 2011), with applications spanning from quantum chemistry (Gilmer et al., 2017) to natural language processing (Alvarez-Melis & Jaakkola, 2018).

Regarding the first aspect, many applications in machine learning involve comparing non-negative measures (often empirical measures) with varying total amounts of mass, e.g., domain adaptation (Fatras et al., 2021). Moreover, OT distances (or dissimilarity measures) are often not robust against outliers and noise, resulting in potentially high transportation costs for outliers. Many recent publications have focused on variants of the OT problem that allow for comparing non-negative measures with unequal mass. For instance, the optimal partial transport problem (Caffarelli & McCann, 2010; Figalli, 2010; Figalli & Gigli, 2010), Kantorovich–Rubinstein norm (Guittet, 2002; Heinemann et al., 2023; Lellmann et al., 2014), and the Hellinger–Kantorovich distance (Chizat et al., 2018a; Liero et al., 2018). These methods fall under the broad category of "unbalanced optimal transport" (Chizat et al., 2018c; Liero et al., 2018).

Regarding the second aspect, comparing probability measures across different metric spaces is essential in many machine learning applications, ranging from computer graphics, where shapes and surfaces are compared (Bronstein et al., 2006; Mémoli, 2009), to graph partitioning and matching problems (Xu et al., 2019a). Source and target distributions often arise from varied conditions, such as different times, contexts, or measurement techniques, creating substantial differences in the intrinsic distances among data points. The conventional OT framework necessitates a meaningful distance across diverse domains, a requirement not always achievable. To circumvent this issue, the Gromov-Wasserstein (GW) distances were proposed in (Mémoli, 2009; 2011) as an adaptation of the Gromov-Hausdorff distance, which measures the discrepancy between two metric spaces. The GW distance (Mémoli, 2011; Sturm, 2023) extends OT-based distances to metric measure spaces up to isometries. Its invariance across isomorphic metric measure spaces makes the GW distance particularly valuable

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

for applications like shape comparison and matching, where invariance to rigid motion transformations is crucial.

Given that the Gromov-Wasserstein (GW) distance is limited to the comparison of probability measures, recent works have introduced unbalanced and partial variations of the GW distance (Séjourné et al., 2021; Chapel et al., 2020; De Ponti & Mondino, 2022). These variations, known as unbalanced or partial GW problems, have been applied in diverse contexts, including partial graph matching (Liu et al., 2020) for social network analysis and the alignment of brain images (Thual et al., 2022).

Over the past decade, there has been extensive work on devising fast and efficient solvers for the OT problem and its unbalanced version, involving various techniques such as linear programming, Sinkhorn iterations, dynamic programming, and slicing (see, e.g., (Guittet, 2002; Cuturi, 2013; Papadakis et al., 2014; Benamou et al., 2014; 2015; Peyré et al., 2019; Chizat et al., 2018b; Bonneel & Coeurjolly, 2019; Bai et al., 2023)). In the context of the Gromov-Wasserstein (GW) metric, the main computational challenge is the non-convexity of its formulation (Mémoli, 2011). The conventional computational approach relies on the Frank-Wolfe (FW) Algorithm (Frank et al., 1956; Lacoste-Julien, 2016). In fact, OT computational methods (e.g. Sinkhorn algorithm) can be incorporated into FW iterations. This yields the classical GW solvers (Peyré et al., 2016; Xu et al., 2019b; Titouan et al., 2019).

Motivated by the emerging applications of the partial GW (PGW) problem, this paper focuses on developing efficient solvers for it. We base our formulation of PGW on the general framework by Séjourné et al. (2021), akin to the "Lagrangian form" of Chapel et al. (2020)'s mass-constraint approach, termed primal-PGW. Unlike Séjourné et al. (2021), who introduce a KL-divergence penalty similar to the unbalanced OT problem and a Sinkhorn solver, we employ a total-variation penalty and present novel, efficient solvers for this problem. To the best of our knowledge, this is the first solution to the proposed PGW formulation in this paper.

Contributions. Our specific contributions in this paper are:

- **Proposition** 3.3. Analogous to Caffarelli & McCann (2010)'s technique for turning the POT problem into an OT problem, here, we show that the partial Gromov-Wasserstein (PGW) problem can be turned into a variant of the Gromov-Wasserstein (GW) problem.
- **Propositions** 3.4 & 3.5. We demonstrate that $(PGW_{\lambda,q}(\cdot,\cdot))^{\frac{1}{p}}$ is a metric between metric measure spaces and show that λ provides an upper bound for allowed "transportation" cost.
- **Solvers.** We introduce two distinct solvers for the discrete PGW problem based on the Frank-Wolfe (Frank et al., 1956) algorithm. We show that these solvers are mathematically and computationally equivalent and they differ from previous solvers described by Chapel et al. (2020) and Séjourné et al. (2021).
- **Convergence Analysis.** Inspired by the results of Lacoste-Julien (2016), we prove that, similar to the Frank-Wolfe solver presented in (Chapel et al., 2020), our proposed solvers for the PGW problem converge linearly to a stationary point.
- **Numerical Experiments.** We demonstrate the performance of our proposed algorithms in terms of computation time and efficacy on two problems: shape-matching between 2D and 3D objects, and positive-unlabeled learning, and compare against baselines.

2. Background

Here, we review the basics of the OT and Partial OT theory and their connection due to Caffarelli & McCann (2010). After that, we introduce the Gromov-Wasserstein distance.

2.1. Optimal Transport and Optimal Partial Transport

Let $\Omega \subseteq \mathbb{R}^d$ be a non-empty open (convex) set, $\mathcal{P}(\Omega)$ be the space of probability measures on the Borel σ-algebra on Ω, and $\mathcal{P}_p(\Omega) := \{\sigma \in \mathcal{P}(\Omega) : \int_\Omega |x|^p d\sigma(x) < \infty\}$.

The OT problem for $\mu, \nu \in \mathcal{P}(\Omega)$ is defined as

$$OT(\mu,\nu) := \min_{\gamma\in\Gamma(\mu,\nu)} \int_{\Omega^2} c(x,y)\, d\gamma(x,y), \tag{1}$$

where $c(x,y) : \Omega \times \Omega \to \mathbb{R}_+$ is the lower-semi continuous transportation cost, and $\Gamma(\mu,\nu)$ is the set of all joint probability measures on $\Omega^2 = \Omega \times \Omega$ with marginals μ, ν, i.e., $\Gamma(\mu,\nu) := \{\gamma \in \mathcal{P}(\Omega^2) : \pi_{1\#}\gamma = \mu, \pi_{2\#}\gamma = \nu\}$, where $\pi_1, \pi_2 : \Omega \times \Omega \to \Omega$ are the projections defined by $\pi_1(x,y) := x, \pi_2(x,y) = y$. To simplify our notation, we define $\gamma_i := \pi_{i\#}\gamma$ for $i = 1,2$. The minimizer of (1) exists (Villani, 2021; 2009) and when $c(x,y) = \|x-y\|^p$ for $p \geq 1$, and $\mu, \nu \in \mathcal{P}_p(\Omega)$ it defines a metric in $\mathcal{P}_p(\Omega)$, which is named as the "p-Wasserstein distance:"

$$W_p^p(\mu,\nu) := \min_{\gamma\in\Gamma(\mu,\nu)} \int_{\Omega^2} \|x-y\|^p d\gamma. \tag{2}$$

The Partial OT (POT) problem (Figalli & Gigli, 2010; Piccoli & Rossi, 2014; Chizat et al., 2018c) extends the OT problem to the set of non-negative and finite measures, i.e., Radon measures, denoted as $\mathcal{M}_+(\Omega)$. For $\lambda > 0$ and $\mu, \nu \in \mathcal{M}_+(\Omega)$, the POT problem is defined as:

$$\inf_{\gamma\in\mathcal{M}_+(\Omega^2)} \int_{\Omega^2} c(x,y) d\gamma + \lambda(|\mu-\gamma_1| + |\nu-\gamma_2|), \tag{3}$$

where $|\sigma|$ denotes the total variation norm of a measure σ. The constraint in 3 can be further simplified (Figalli, 2010) into $\gamma \in \Gamma_{\leq}(\mu,\nu)$,

$$\inf_{\gamma\in\Gamma_{\leq}(\mu,\nu)} \int_{\Omega^2} c(x,y)d\gamma + \lambda(|\mu-\gamma_1|+|\nu-\gamma_2|), \quad (4)$$

where $\Gamma_{\leq}(\mu,\nu) := \{\gamma \in \mathcal{M}_+(\Omega^2) : \gamma_1 \leq \mu, \gamma_2 \leq \nu\}$, and the notation $\gamma_1 \leq \mu$ denotes that for any Borel set $B \subseteq \Omega$, $\gamma_1(B) \leq \mu(B)$ (and we say that "γ_1 is dominated by μ"). Interestingly, the constraint can be further restricted to the set of partial transport plans satisfying $\gamma(x,y) = 0$ when $c(x,y) \geq 2\lambda$ for all (x,y) in γ's support (Bai et al., 2023), i.e., the mass will not be transported if the transportation cost is larger than 2λ.

The relationship between POT and OT. By Caffarelli & McCann (2010)'s technique, the POT problem can be transferred into an OT problem, and thus, OT solvers (e.g., Network simplex) can be employed to solve the POT problem. In short, given the POT problem (3), construct measures,

$$\hat{\mu} = \mu + |\nu|\delta_{\hat{\infty}} \quad \text{and} \quad \hat{\nu} = \nu + |\mu|\delta_{\hat{\infty}} \quad (5)$$

on $\hat{\Omega} := \Omega \cup \{\hat{\infty}\}$, for the auxiliary point $\hat{\infty}$ we define the cost as

$$\hat{c}(x,y) := \begin{cases} c(x,y) - 2\lambda & \text{if } x,y \in \Omega, \\ 0 & \text{elsewhere.} \end{cases} \quad (6)$$

Proposition 2.1. *[Caffarelli & McCann (2010); Bai et al. (2023)] Consider the following OT problem*

$$OT(\hat{\mu},\hat{\nu}) = \min_{\gamma\in\Gamma(\hat{\mu},\hat{\nu})} \int_{\Omega^2} \hat{c}(x,y)\, d\hat{\gamma}(x,y). \quad (7)$$

and $F : \Gamma_{\leq}(\mu,\nu) \to \Gamma(\hat{\mu},\hat{\nu})$ that maps $\gamma \mapsto \hat{\gamma}$, where

$$\hat{\gamma} := \gamma + (\mu-\gamma_1)\otimes\delta_{\hat{\infty}} + \delta_{\hat{\infty}}\otimes(\nu-\gamma_2) + |\gamma|\delta_{\hat{\infty},\hat{\infty}} \quad (8)$$

Then, F is a bijection, and γ is optimal for the POT problem (3) *if and only if $\hat{\gamma}$ is optimal for the OT problem* (1).

Finally, it is worth noting that instead of considering the same underlying space Ω for both measures μ and ν, the OT and POT problems can be formulated in the scenario where $\mu \in \mathcal{P}_p(X)$, and $\nu \in \mathcal{P}_p(Y)$, where X and Y are different metric spaces. In this setting, one needs a lower-semi-continuous cost function $c : X \times Y \to \mathbb{R}_+$ to formulate the OT and POT problems. However, such *ground cost*, c, might not exist or might not be known. To address this issue, in the next section, we will review the fundamentals of the *Gromov-Wasserstein* problem (Mémoli, 2011), which relies on intra-domain distances and is invariant under rigid transformations (rotations and translations).

2.2. Gromov-Wasserstein Distances

Given two metric measure spaces (mm-spaces) $\mathbb{X} = (X, d_X, \mu)$, $\mathbb{Y} = (Y, d_Y, \nu)$ with finite moment $q \geq 1$, i.e.,

$$\int_X d_X^q(x-x_0)d\mu(x), \int_X d_Y^q(y-y_0)d\nu(y) < \infty$$

for some $x_0 \in X, y_0 \in Y$ (thus for all $x_0 \in X, y_0 \in Y$), the Gromov-Wasserstein (GW) dissimilarity is defined as

$$GW_q^L(\mathbb{X},\mathbb{Y}) := \quad (9)$$
$$\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{(X\times Y)^2} L(d_X^q(x,x'), d_Y^q(y,y'))\, d\gamma(x,y)d\gamma(x',y'),$$

where $L : \mathbb{R}^2 \to \mathbb{R}_+$ is a non-negative lower semi-continuous function (e.g., Euclidean distance or KL-lost).

Similar to the OT problem, the above GW problem defines a distance for mm-spaces when $L(\cdot,\cdot) = D^p(\cdot,\cdot)$ where $1 \leq p < \infty$ and $D(\cdot,\cdot)$ is a metric in $\mathbb{R}$. In this case, we use the notation $GW_q^p(\mathbb{X},\mathbb{Y})$ for (9). Indeed, a minimizer of the GW problem (9) always exists, and thus, we can replace inf by min (see (Mémoli, 2011)). Moreover, let us define a subset $\mathcal{G}_q$ of mm-spaces

$$\mathcal{G}_q := \{\mathbb{X} = (X, d_X, \mu) : \int_X d_X^q(x,x_0)d\mu(x) < \infty,$$
$$\emptyset \neq X \subseteq \mathbb{R}^d \text{ for some dimension } d\}. \quad (10)$$

Note that dimension d is not fixed. In $\mathcal{G}_q$, we define relation $\sim$ as follows: Given $\mathbb{X} = (X, d_X, \mu), \mathbb{Y} = (Y, d_Y, \nu) \in \mathcal{G}_q$, $\mathbb{X} \sim \mathbb{Y}$ if there exists $\gamma \in \Gamma(\mu,\nu)$ such that

$$\int_{(X\times Y)^2} D^p(d_X^q(x,x'), d_Y^q(y,y'))\, d\gamma^{\otimes 2}((x,y),(x',y')) = 0,$$

where $\gamma^{\otimes 2}$ denotes the product measure of γ with itself. In the remainder of the paper, for brevity, we use $d\gamma^{\otimes 2}$ to denote $d\gamma^{\otimes 2}((x,y),(x',y')) = d\gamma(x,y)d\gamma(x',y')$. Equivalently, if $GW_q^p(\mathbb{X},\mathbb{Y}) = 0$. One can verify that $\sim$ is an equivalence relation. Then, $(GW_q^p(\cdot,\cdot))^{1/p}$ defines a metric in the quotient space $\mathcal{G}_q/\sim$.

3. The Partial Gromov-Wasserstein problem

Similar to the POT problem, the partial GW (PGW) formulation relaxes the assumption that μ,ν are normalized probability measures. In particular, by (Chapel et al., 2020; Séjourné et al., 2021), the partial Gromov-Wasserstein problem for a positive constant $\lambda > 0$ is defined as

$$PGW_{\lambda,q}^L(\mathbb{X},\mathbb{Y}) := \lambda\left(|\mu^{\otimes 2} - \gamma_1^{\otimes 2}| + |\nu^{\otimes 2} - \gamma_2^{\otimes 2}|\right) \quad (11)$$
$$+ \inf_{\gamma\in\mathcal{M}_+(X\times Y)} \int_{(X\times Y)^2} L(d_X^q(x,x'), d_Y^q(y,y'))\, d\gamma^{\otimes 2}.$$

Similar to the POT, the PGW problem can be simplified.

Proposition 3.1. *Given $\mathbb{X} = (X, d_X, \mu), \mathbb{Y} = (Y, d_Y, \nu) \in \mathcal{G}_q$, let $\Gamma_{\leq}(\mu,\nu) = \{\gamma \in \mathcal{P}(X\times Y) : \gamma_1 \leq \mu, \gamma_2 \leq \nu\}$. Then, we can restrict the minimization problem* (11) *from $\mathcal{M}_+(X\times Y)$ to $\Gamma_{\leq}(\mu,\nu)$, that is,*

$$PGW_{\lambda,q}^L(\mathbb{X},\mathbb{Y}) = \lambda\left(|\mu^{\otimes 2} - \gamma_1^{\otimes 2}| + |\nu^{\otimes 2} - \gamma_2^{\otimes 2}|\right) \quad (12)$$
$$+ \inf_{\gamma\in\Gamma_{\leq}(\mu,\nu)} \int_{(X\times Y)^2} L(d_X^q(x,x'), d_Y^q(y,y'))d\gamma^{\otimes 2}.$$

Proposition 3.2. *Similar to the GW problem, if X, Y are compact sets and $L : \mathbb{R}^2 \to \mathbb{R}$ is a C^1-function, then* (11) *and* (12) *admit minimizers and* inf *can be replaced by* min.

The relationship between GW and PGW. We provide the following proposition to relate the GW and PGW problems.

Proposition 3.3. *Define an auxiliary point $\hat{\infty}$ and let $\hat{\mathbb{X}} = (\hat{X}, \hat{d}_X, \hat{\mu})$, where $\hat{X} = X \cup \{\hat{\infty}\}$, $\hat{\mu}$ is constructed by (5), and $\hat{d}_{\hat{X}} : \hat{X}^2 \to \mathbb{R}_+ \cup \{\infty\}$ is the generalized metric*

$$\hat{d}_{\hat{X}}(x, x') = \begin{cases} d_X(x, x') & \text{if } x, x' \in X, \\ \infty & \text{otherwise}, \end{cases} \tag{13}$$

where ∞ is an auxiliary point to $\mathbb{R}_+$ such that for all $x \in \mathbb{R}_+$ we set that $x \leq \infty$ and $\infty + \infty = \infty$. Let $\hat{L} : \mathbb{R} \cup \{\infty\} \to \mathbb{R}$ with

$$\hat{L}(r_1, r_2) := \begin{cases} L(r_1, r_2) - 2\lambda & \text{if } r_1, r_2 \in \mathbb{R}, \\ 0 & \text{elsewhere}. \end{cases} \tag{14}$$

Consider the following GW-variant problem:

$$\widehat{GW}(\hat{\mathbb{X}}, \hat{\mathbb{Y}}) = \inf_{\hat{\gamma} \in \Gamma(\mu,\nu)} \int_{(\hat{X} \times \hat{Y})^2} \hat{L}(\hat{d}^q_{\hat{X}}(x, x'), \hat{d}^q_{\hat{Y}}(y, y')) d\hat{\gamma}^{\otimes 2}. \tag{15}$$

Then, when considering the bijection $\gamma \mapsto \hat{\gamma}$ defined in (8) we have that γ is optimal for partial GW (12) if and only if $\hat{\gamma}$ is optimal for the GW-variant problem (15).

We denote problem (15) by 'GW-variant' since $(\hat{X}, \hat{d}_{\hat{X}})$ is not a metric space ($\hat{d}_{\hat{X}}(\hat{\infty}, \hat{\infty}) = \infty$).

The role of λ in the PGW problem. The following proposition states that λ plays the role of an upper bound for the allowable GW "transportation" cost $L(d^q_X(\cdot,\cdot), d^q_Y(\cdot,\cdot))$.

Proposition 3.4. *Given an optimal transportation plan γ for the partial GW problem (12), then*

$$\gamma^{\otimes 2}(\{((x, y), (x', y')) : L(d^q_X(x, x'), d^q_Y(y, y')) > 2\lambda\}) = 0.$$

Therefore,

$$\begin{aligned} PGW^L_{\lambda,q}(\mathbb{X}, \mathbb{Y}) = \lambda(|\mu^{\otimes 2} - \gamma_1^{\otimes 2}| + |\nu^{\otimes 2} - \gamma_2^{\otimes 2}|) \quad (16) \\ + \inf_{\gamma \in \Gamma_{\leq}(\mu,\nu)} \int_{(X \times Y)^2} L(d^q_X(x, x'), d^q_Y(y, y')) \wedge 2\lambda \, d\gamma^{\otimes 2}, \end{aligned}$$

where $a \wedge b := \min(a, b)$.

Let $L(r_1, r_2) = D^p(r_1, r_2)$ for some metric D on $\mathbb{R}$. For simplicity, consider $D(r_1, r_2) = |r_1 - r_2|^p$. Thus, by the previous results, the partial GW problem can be stated as

$$\begin{aligned} PGW_{\lambda,q}(\mu, \nu) := \lambda \left(|\mu|^2 + |\nu|^2\right) \quad (17) \\ + \min_{\gamma \in \Gamma_{\leq}(\mu,\nu)} \int_{(X \times Y)^2} (|d^q_X(x, x') - d^q_Y(y, y')|^p - 2\lambda) d\gamma^{\otimes 2}. \end{aligned}$$

Proposition 3.5. *Let $1 \leq q, p < \infty$. If $\lambda > 0$, then $(PGW_{\lambda,q}(\cdot, \cdot))^{1/p}$ defines a metric between mm-spaces. Precisely, it defines a distance in $\mathcal{G}_q / \sim$.*

4. Computation of the Partial GW distance

In the discrete setting, $\mathbb{X} = (X, d_X, \sum_{i=1}^n p_i \delta_{x_i})$, $\mathbb{Y} = (Y, d_Y, \sum_{j=1}^m q_j \delta_{y_j})$, where $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$, the weights p_i, q_j are non-negative numbers, and the distances d_X, d_Y are determined by the matrices $C^X \in \mathbb{R}^{n \times n}$, $C^Y \in \mathbb{R}^{m \times m}$ with

$$C^X_{i,i'} := d^2_X(x_i, x_{i'}) \quad \text{and} \quad C^Y_{j,j'} := d^2_Y(y_j, y_{j'}). \tag{18}$$

Let $\mathrm{p} = [p_1, \ldots, p_n]^T$ and $\mathrm{q} = [q_1, \ldots, q_m]^T$ denote the weight vectors corresponding to the discrete measures. We view the set of plans for the GW and PGW problems as the subset of $n \times m$ matrices:

$$\Gamma(\mathrm{p}, \mathrm{q}) := \{\gamma \in \mathbb{R}^{n \times m}_+ : \gamma 1_m = \mathrm{p}, \gamma^T 1_n = \mathrm{q}\} \tag{19}$$

if $|\mathrm{p}| = \sum_{i=1}^n p_i = 1 = \sum_{j=1}^m q_j = |\mathrm{q}|$; and

$$\Gamma_{\leq}(\mathrm{p}, \mathrm{q}) := \{\gamma \in \mathbb{R}^{n \times m}_+ : \gamma 1_m \leq \mathrm{p}, \gamma^T 1_n \leq \mathrm{q}\} \tag{20}$$

for any pair of non-negative vectors $\mathrm{p} \in \mathbb{R}^n_+$, $\mathrm{q} \in \mathbb{R}^m_+$, where 1_n is the vector with all ones in $\mathbb{R}^n$ (resp. 1_m), and $\gamma 1_m \leq \mathrm{p}$ means that component-wise the $\leq$ relation holds.

The transportation cost M, given by a non-negative function $L : \mathbb{R}^{n \times n} \times \mathbb{R}^{m \times m} \to \mathbb{R}_+$, is represented by an $n \times m \times n \times m$ tensor,

$$M_{i,j,i',j'} = L(C^X_{i,i'}, C^Y_{j,j'}). \tag{21}$$

Let

$$\tilde{M} = M - 2\lambda 1_{n,m,n,m}, \tag{22}$$

where $1_{n,m,n,m}$ is the tensor with ones in all its entries. For each $n \times m \times n \times m$ tensor M and each $n \times m$ matrix γ, we define tensor-matrix multiplication $M \circ \gamma \in \mathbb{R}^{n \times m}$ by

$$(M \circ \gamma)_{ij} = \sum_{i',j'} (M_{i,j,i',j'}) \gamma_{i',j'}.$$

Then, the partial GW problem in (12) can be written as

$$PGW^L_\lambda(\mathbb{X}, \mathbb{Y}) = \min_{\gamma \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})} \mathcal{L}_{\tilde{M}}(\gamma) + \lambda(|\mathrm{p}|^2 + |\mathrm{q}|^2), \tag{23}$$

where

$$\begin{aligned} \mathcal{L}_{\tilde{M}}(\gamma) := \tilde{M} \gamma^{\otimes 2} := \sum_{i,j,i',j'} \tilde{M}_{i,j,i',j'} \gamma_{i,j} \gamma_{i',j'} \quad (24) \\ = \sum_{ij} (\tilde{M} \circ \gamma)_{ij} \gamma_{ij} =: \langle \tilde{M} \circ \gamma, \gamma \rangle_F, \end{aligned}$$

here $\langle \cdot, \cdot \rangle_F$ stands for the Frobenius dot product. The constant term $\lambda(|\mathrm{p}|^2 + |\mathrm{q}|^2)$ will be ignored in the rest of the article since it does not depend on γ.

Similarly, consider the discrete version of (5):

$$\hat{\mathrm{p}} = [\mathrm{p}; |\mathrm{q}|] \in \mathbb{R}^{n+1}, \quad \hat{\mathrm{q}} = [\mathrm{q}; |\mathrm{p}|] \in \mathbb{R}^{m+1}, \tag{25}$$

and, in a similar fashion to (14), we define $\hat{M} \in \mathbb{R}^{(n+1)\times(m+1)\times(n+1)\times(m+1)}$ as

$$\hat{M}_{i,j,i',j'} = \begin{cases} \tilde{M}_{i,j,i',j'} & \text{if } i,i' \in [1:n], j,j' \in [1:m], \\ 0 & \text{elsewhere.} \end{cases} \quad (26)$$

Then, the GW-variant problem (15) can be written as

$$\widehat{GW}(\hat{\mathbb{X}}, \hat{\mathbb{Y}}) = \min_{\hat{\gamma}\in\Gamma(\hat{\mathrm{p}},\hat{\mathrm{q}})} \mathcal{L}_{\hat{M}}(\hat{\gamma}). \quad (27)$$

Based on Proposition 3.3 (which relates $PGW_\lambda^L(\cdot,\cdot)$ with $\widehat{GW}(\cdot,\cdot)$) and Proposition 2.1 (which relates POT with OT), we propose two versions of the Frank-Wolfe algorithm (Frank et al., 1956) that can solve the partial GW problem (23). Apart from Algorithm 1 in (Chapel et al., 2020), which solves a different formulation of partial GW, and Algorithm 1 in (Séjourné et al., 2021), which applies the Sinkhorn algorithm to solve an entropic regularized version of (11), to the best of our knowledge, a precise computational method for the above partial GW problem (12) has not been studied.

In our proposed method, we address the discrete partial GW problem (23), highlighting that the *direction-finding subproblem* in the Frank-Wolfe (FW) algorithm is a POT problem for (23) and an OT problem for (27). Specifically, (23) is treated as a discrete POT problem in our Solver 1 (Subsection 4.1), where we apply Proposition 2.1 to solve a discrete OT problem. Conversely, our Solver 2 (Subsection 4.2), based on Proposition 3.3, extends the PGW problem to a discrete GW-variant problem (27), leading to a solution for the original PGW problem by truncating the GW-variant solution. Both algorithms are mathematically and computationally equivalent, owing to the equivalence between the POT problem in Solver 1 and the OT problem in Solver 2. The convergence analysis, detailed in Appendix K, applies the results from (Lacoste-Julien, 2016) to our context, showing that the FW algorithm achieves a stationary point at a rate of $\mathcal{O}(1/\sqrt{k})$ for non-convex objectives with a Lipschitz continuous gradient in a convex and compact domain.

4.1. Frank-Wolfe for the PGW Problem - Solver 1

In this section, we discuss the Frank-Wolfe (FW) algorithm for the PGW problem (23). For each iteration k, the procedure is summarized in three steps detailed below.

Step 1. Computation of gradient and optimal direction.

It is straightforward to verify that the gradient of the objective function in (23) is given by

$$\nabla\mathcal{L}_{\tilde{M}}(\gamma) = 2\tilde{M}\circ\gamma. \quad (28)$$

The classical method to compute $M\circ\gamma$ is the following: First, convert M into a $(n\times m)\times(n\times m)$ matrix, denoted as $v(M)$, and convert γ into a $(n\times m)\times 1$ vector. Then, the computation of $M\circ\gamma$ is equivalent to the matrix multiplication $v(M)v(\gamma)$. The computational cost and the required storage space are $\mathcal{O}(n^2m^2)$. Following Peyré et al. (2016), below we show that the computational and storage costs can be reduced to $\mathcal{O}(n^2m + m^2n)$ and $\mathcal{O}(n^2+m^2)$.

Proposition 4.1 (Proposition 1 (Peyré et al., 2016))**.** *If the cost function can be written as*

$$L(r_1,r_2) = f_1(r_1) + f_2(r_2) - h_1(r_1)h_2(r_2),$$

then

$$M\circ\gamma = u(C^X, C^Y, \gamma) - h_1(C^X)\gamma h_2(C^Y)^T, \quad (29)$$

where $u(C^X,C^Y,\gamma) := f_1(C^X)\gamma_1 1_m^T + 1_n\gamma_2^T f_2(C^Y)$.

Additionally, the following lemma build the connection between $\tilde{M}\circ\gamma$ and $M\circ\gamma$.

Lemma 4.2. *For any $\gamma\in\mathbb{R}^{n\times m}$, we have:*

$$\tilde{M}\circ\gamma = M\circ\gamma - 2\lambda|\gamma|1_{n,m}. \quad (30)$$

Now, we discuss step 1 in the FW algorithm for the k-th iteration. We aim to solve the following problem:

$$\gamma^{(k)'} \leftarrow \arg\min_{\gamma\in\Gamma_\leq(\mathrm{p},\mathrm{q})} \langle\nabla\mathcal{L}_{\tilde{M}}(\gamma^{(k)}),\gamma\rangle_F,$$

which is a discrete POT problem since it is equivalent to

$$\min_{\gamma\in\Gamma_\leq(\mathrm{p},\mathrm{q})} \langle 2M\circ\gamma^{(k)},\gamma\rangle_F + 2\lambda|\gamma^{(k)}|(|\mu|+|\nu|-2|\gamma|),$$

where the transportation cost is given by the tensor M, and the penalization is given by $\lambda > 0$. After calculating $\nabla\mathcal{L}_{\tilde{M}}(\gamma)$, we turn the above POT problem into an OT problem via Proposition 2.1, and solve the following:

$$\hat{\gamma}^{(k)} \leftarrow \arg\min_{\hat{\gamma}\in\Gamma(\hat{\mathrm{p}},\hat{\mathrm{q}})} \langle\hat{\nabla}\mathcal{L}_{\tilde{M}}(\gamma^{(k)}),\hat{\gamma}\rangle_F, \quad (31)$$

where $\hat{\mathrm{p}}$, $\hat{\mathrm{q}}$ are defined in (25), and $\hat{\nabla}\mathcal{L}_{\tilde{M}}(\gamma^{(k)}) \in \mathbb{R}^{(n+1)\times(m+1)}$ is defined by

$$(\hat{\nabla}\mathcal{L}_{\tilde{M}}(\gamma))_{ij} = \begin{cases} (\nabla\mathcal{L}_{\tilde{M}}(\gamma))_{ij} & \text{if } i\in[1:n], j\in[1:m], \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we obtain $\gamma^{(k)'} = \hat{\gamma}^{(k)}[1:n,1:m]$.

Step 2: Line search method.

In this step, at the k-th iteration, we need to determine the optimal step size:

$$\alpha^{(k)} = \arg\min_{\alpha\in[0,1]}\{\mathcal{L}_{\tilde{M}}((1-\alpha)\gamma^{(k)} + \alpha\gamma^{(k)'})\}.$$

The optimal $\alpha^{(k)}$ takes the following values (see Appendix I for details): Let

$$\begin{cases} \delta\gamma^{(k)} = \gamma^{(k)'} - \gamma^{(k)}, \\ a = \langle\tilde{M}\circ\delta\gamma^{(k)},\delta\gamma^{(k)}\rangle_F\ , \\ b = 2\langle\tilde{M}\circ\gamma^{(k)},\delta\gamma^{(k)}\rangle_F. \end{cases} \quad (32)$$

Algorithm 1 Frank-Wolfe Algorithm for partial GW, ver 1

Input: $\mu = \sum_{i=1}^n p_i\delta_{x_i}, \nu = \sum_{j=1}^m q_j\delta_{y_j}, \gamma^{(1)}$
Output: $\gamma^{(final)}$
Compute C^X, C^Y
for $k = 1, 2, \ldots$ **do**
  $G^{(k)} \leftarrow 2\tilde{M} \circ \gamma^{(k)}$ (via Lemma 4.2) // Compute gradient
  $\gamma^{(k)'} \leftarrow \arg\min_{\gamma\in\Gamma_{\leq}(\mathrm{p},\mathrm{q})} \langle G^{(k)}, \gamma\rangle_F$ // Solve the POT problem by (31)
  Compute $\alpha^{(k)} \in [0, 1]$ via (32), (33) // Line search
  $\gamma^{(k+1)} \leftarrow (1-\alpha^{(k)})\gamma^{(k)} + \alpha^{(k)}\gamma^{(k)'}$// Update γ
  if convergence, break
end for
$\gamma^{(final)} \leftarrow \gamma^k$

Then the optimal $\alpha^{(k)}$ is given by

$$\alpha^{(k)} = \begin{cases} 0 & \text{if } a \leq 0, a+b > 0, \\ 1 & \text{if } a \leq 0, a+b \leq 0, \\ \text{clip}(\frac{-b}{2a}, [0,1]) & \text{if } a > 0, \end{cases} \tag{33}$$

where $\text{clip}(\frac{-b}{2a}, [0, 1]) = \min\{\max\{-\frac{b}{2a}, 0\}, 1\}$.

Step 3: Update $\gamma^{(k+1)} \leftarrow (1 - \alpha^{(k)})\gamma^{(k)} + \alpha^{(k)}\gamma^{(k)'}$.

4.2. Frank-Wolfe for the PGW Problem - Solver 2

Here we discuss another version of the FW Algorithm for solving the PGW problem (23). The main idea relies on solving first the GW-variant problem (15), and, at the end of the iterations, by using Proposition 3.3, convert the solution of the GW-variant problem to a solution for the original partial GW problem (23).

First, construct $\hat{\mathrm{p}}, \hat{\mathrm{q}}$ as described in the previous section. Then, for each iteration k, perform the following three steps.
Step 1: Computation of gradient and optimal direction. Solve the OT problem (see appendix H for the computation):

$$\hat{\gamma}^{(k)'} \leftarrow \arg \min_{\hat{\gamma}\in\Gamma(\hat{\mathrm{p}},\hat{\mathrm{q}})} \langle \mathcal{L}_{\hat{M}}(\hat{\gamma}^{(k)}), \hat{\gamma}\rangle_F.$$

Step 2: Line search method. Find optimal step size $\alpha^{(k)}$:

$$\alpha^{(k)} = \arg \min_{\alpha\in[0,1]} \{\mathcal{L}_{\hat{M}}((1-\alpha)\hat{\gamma}^{(k)} + \alpha\hat{\gamma}^{(k)'})\}.$$

Similar to solver 1, let

$$\begin{cases} \delta\hat{\gamma}^{(k)} = \hat{\gamma}^{(k)'} - \hat{\gamma}^{(k)}, \\ a = \langle \hat{M} \circ \delta\hat{\gamma}^{(k)}, \delta\hat{\gamma}^{(k)}\rangle_F, \\ b = 2\langle \hat{M} \circ \delta\hat{\gamma}^{(k)}, \hat{\gamma}^{(k)}\rangle_F. \end{cases} \tag{34}$$

Then the optimal $\alpha^{(k)}$ is given by formula (33). See Appendix J for a detailed discussion.

Step 3. Update $\hat{\gamma}^{(k+1)} \leftarrow (1 - \alpha^{(k)})\hat{\gamma}^{(k)} + \alpha^{(k)}\hat{\gamma}^{(k)'}$.

Algorithm 2 Frank-Wolfe Algorithm for partial GW, ver 2

Input: $\mu = \sum_{i=1}^n p_i\delta_{x_i}, \nu = \sum_{j=1}^m q_j\delta_{y_j}, \gamma^{(1)}$
Output: $\gamma^{(final)}$
Compute $C^X, C^Y, \hat{\mathrm{p}}, \hat{\mathrm{q}}, \hat{\gamma}^{(1)}$
for $k = 1, 2, \ldots$ **do**
  $\hat{G}^{(k)} \leftarrow 2\hat{M} \circ \hat{\gamma}^{(k)}$ // Compute gradient
  $\hat{\gamma}^{(k)'} \leftarrow \arg\min_{\hat{\gamma}\in\Gamma(\hat{\mathrm{p}},\hat{\mathrm{q}})} \langle \hat{G}^{(k)}, \hat{\gamma}\rangle_F$ // Solve the OT problem
  Compute $\alpha^{(k)} \in [0, 1]$ via (34), (33) // Line search
  $\hat{\gamma}^{(k+1)} \leftarrow (1-\alpha^{(k)})\hat{\gamma}^{(k)'} + \alpha\hat{\gamma}^{(k)}$// Update $\hat{\gamma}$
  if convergence, break
end for
$\gamma^{(final)} \leftarrow \hat{\gamma}^{(k)}[1 : n, 1 : m]$

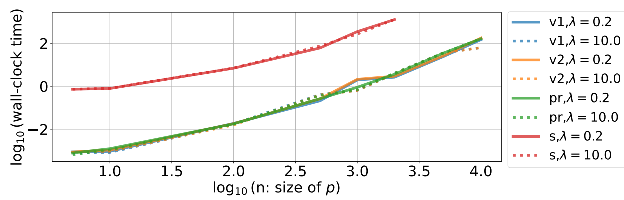

Figure 1. We test the wall-clock time of our Algorithm 1 and Algorithm 2, the partial GW solver (Algorithm 1 in (Chapel et al., 2020)) , and the Sinkhorn algorithm (Peyré et al., 2016). We denote these methods as v1, v2, pr, s respectively. The linear programming solver applied in the first three methods is from POT (Flamary et al., 2021), which is written in C++. The maximum number of iterations for all the methods is set to be 1000. The maximum iteration for OT/OPT solvers is set to be $300n$. The maximum Sinkhorn iteration is set to be 1000. The convergence tolerance for the Frank-Wolfe algorithm and the Sinkhorn algorithm are set to be $1e - 5$. To achieve their best performance, the number of dummy points is set to be 1 for primal-PGW and PGW.

4.3. Numerical Implementation Details

The initial guess, $\gamma^{(1)}$. In the GW problem, the initial guess is simply set to $\gamma^{(1)} = \mathrm{p}\mathrm{q}^T$ if there is no prior knowledge. In PGW, however, as μ, ν may not necessarily be probability measures (i.e. $\sum_i p_i, \sum_j q_j \neq 1$ in general). We set $\gamma^{(1)} = \frac{\mathrm{p}\mathrm{q}^T}{\max(|\mathrm{p}|,|\mathrm{q}|)}$.

It is straightforward to verify that $\gamma^{(1)} \in \Gamma_{\leq}(\mathrm{p}, \mathrm{q})$ as

$$\gamma^{(1)}1_m = \frac{|\mathrm{q}|\mathrm{p}}{\max(|\mathrm{p}|, |\mathrm{q}|)} \leq \mathrm{p}, \;\; \gamma^{(1)T}1_n = \frac{|\mathrm{p}|\mathrm{q}}{\max(|\mathrm{p}|, |\mathrm{q}|)} \leq \mathrm{q}.$$

Column/Row-Reduction According to the interpretation of the penalty weight parameter in partial OT problem (e.g. see Lemma 3.2 in (Bai et al., 2023)), during the POT solving step, for each $i \in [1 : n]$ (and $j \in [1 : m]$), if the i^{th} row (or j^{th} column) of $\tilde{M} \circ \gamma^{(k)}$ contains no negative entry, all the mass of p_i (and q_j) will be destroyed (created). Thus, we can remove the corresponding row (and column) to improve the computation efficiency.

5. Experiments

5.1. Performance comparison

In this section, we present the wall clock time comparison between our method Algorithms 1, 2, the Frank-wolf algorithm proposed in (Chapel et al., 2020), and its Sinkhorn version (Peyré et al., 2016; Chapel et al., 2020). Note that these two baselines solve a mass constraint version of the PGW problem, which we refer to as the "primal-PGW" problem. The proposed PGW formulation in this paper can be regarded as an "Lagragian formulation" of primal-PGW[1] formulation to the PGW problem defined in (12). In this paper, we call these two baselines as "primal-PGW algorithm" and "Sinkhorn PGW algorithm".

The data is generated as follows: let $\mu = \text{Unif}([0,2]^2)$ and $\nu = \text{Unif}([0,2]^3)$, we select i.i.d. samples $\{x_i \sim \mu\}_{i=1}^n, \{y_j \sim \nu\}_{j=1}^m$, where n is selected from $[10, 50, 100, 150, ..., 10000]$ and $m = n + 100$, $\mathrm{p} = 1_n/m, \mathrm{q} = 1_m/m$. For each n, we set $\lambda = 0.2, 1.0, 10.0$. The mass constraint parameter for the algorithm in (Chapel et al., 2020), and Sinkhorn is computed by the mass of the transportation plan obtained by Algorithm 1 or 2. The runtime results are shown in Figure 1.

Regarding the acceleration technique, for the POT problem in step 1, our algorithms and the primal-PGW algorithm apply the linear programming solver provided by Python OT package (Flamary et al., 2021), which is written in C++. The Sinkhorn algorithm from Python OT does not have an acceleration technique. Thus, we only test its wall-clock time for $n \leq 2000$. The data type is 64-bit float number. The experiment is conducted on a computational machine with AMD 64-core Processors clocking 3720 MHZ and 4 NVIDA RTX A6000 49 GB GPUs.

From this Figure 1, we can observe the Algorithms 1, 2 and primal-PGW algorithm have a similar order of time complexity. However, using the column/row-reduction technique for the POT computation discussed in previous sections, and the fact the convergence behaviors of Algorithms 1 and 2 are similar to the primal-PGW algorithm, we observe that the proposed algorithms 1, 2 admits a slightly faster speed than primal-PGW solver.

5.2. Shape matching problem between 2D and 3D spaces

In this section, we consider a simple shape-matching problem. Suppose sets $SQ_2 = \{s_i^2 \in \mathbb{R}^2\}_{i=1}^n, SQ_3 = \{s_i^3 \in \mathbb{R}^3\}_{i=1}^n$ are sampled from a 2D square $[-1,1]^2$ and a 3D cube $[-1,1]^3$, respectively. Similarly, $SP_2 = \{c_i^2 \in \mathbb{R}^2\}_{i=1}^n, SP_3 = \{c_i^3 \in \mathbb{R}^3\}_{i=1}^n$ are sampled from a 2D circle $\{s^2 + [5,0] : \|s^2\|^2 = 1\}$ and a 3D sphere $\{s^3 + [5,0,0] : \|s^3\| = 1\}$. Given $a, b > 0$ with $a + b = 1$, we use $aS_2 + bC_2$ to denote the mixture probability distribution supported on $SQ_2 \cup SP_2$. In particular, the probability mass of each point in SQ_2 is $a\frac{1}{n}$ and the mass of each points in SP_2 is $b\frac{1}{n}$. Similarly, we define $aS_3 + bC_3$.

Given distributions $0.3S_2 + 0.7C_2$ and $0.5S_3 + 0.5C_3$, we visualize the matching between them via balanced Gromov-Wasserstein (GW), Unbalanced Gromov-Wasserstein (UGW), the Primal partial Gromov-Wasserstein approach (primal-PGW), and the partial Gromov-Wasserstein framework proposed in this paper (PGW).

For GW and UGW, the initial guess $\gamma^{(1)}$ is set to be pq^T, for primal-PGW, $\gamma^{(1)}$ is $\frac{m\mathrm{pq}^T}{|\mathrm{p}||\mathrm{q}|}$, and for PGW, $\frac{\mathrm{pq}^T}{\max(|\mathrm{p}|,|\mathrm{q}|)}$. The optimal correspondences for each method are visualized in Figure 2. We observe that UGW, primal-PGW, and PGW induce more natural correspondences in this unbalanced setting. Note that GW matches the same 3D shape (SQ3) to the two different 2D shapes, as it requires mass preservation.

Regarding the wall-clock time, the size of two distributions is set to be 1200, denoted as n. GW requires 4-6 seconds, primal-PGW requires 70-80 seconds, UGW requires 38-46 seconds and PGW requires 26-30 seconds.

The maximum iteration for OT solver in GW, PGW, primal-PGW is set to be $300n$. The maximum iteration in Sinkhorn step in UGW is set to be 1000. The data type for GW, PGW, primal-PGW is 64-bit float number, the data type for UGW is 32-bit float number.

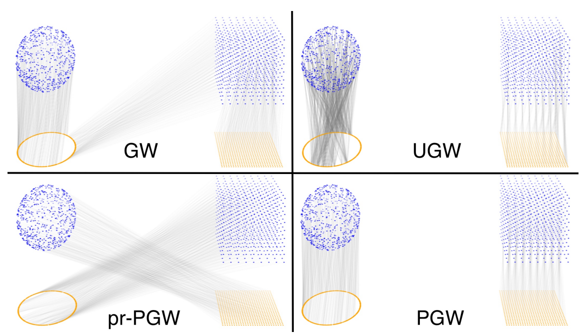

Figure 2. Shape matching results via GW, Primal-PGW, UGW, and PGW. In each subplot, the top two shapes are $SQ3, SP3$, and the bottom two shapes are $SQ2, SP2$. We visualize correspondences if the transported mass is greater than a suitable threshold.

5.3. Positive unlabeled learning problem

Problem Setup. Positive unlabeled (PU) learning (Bekker & Davis, 2020; Elkan & Noto, 2008; Kato et al., 2018) is a semi-supervised binary classification problem for which the training set only contains positive samples. In particular, suppose there exists a fixed unknown overall distribution over triples (x, o, l), where x is data, $l \in \{0, 1\}$ is the label of x, $o \in \{0, 1\}$ where $o = 1$, $o = 0$ denote that l is

[1]Due to the non-convexity of GW, we do not have a strong duality in some of the GW representations. Thus, the Lagrangian form is not a rigorous description.

DATASET	INIT	PR-PGW	UGW	PGW
M → M	POT, 100%	100%	95%	100%
M → M	FLB-U, 75%	96%	95%	96%
M → M	FLB-P, 75%	99%	95%	99%
M → EM	FLB-U, 78%	94%	95%	94%
M → EM	FLB-P, 78%	94%	95%	94%
EM → M	FLB-U, 75%	97%	96%	97%
EM → M	FLB-P, 75%	97%	96%	97%
EM → EM	POT, 100%	100%	95%	100%
EM → EM	FLB-U, 78%	94%	95%	94%
EM → EM	FLB-P, 78%	95%	95%	95%

Table 1. Accuracy comparison of the primal-PGW, UGW, and the proposed PGW method on PU learning. "Init" denotes the initialization method and initial accuracy. 'M' denotes MNIST, and 'EM' denotes EMNIST.

observed or not, respectively. In the PU task, the assumption is that only positive samples' labels can be observed, i.e., $\text{Prob}(o = 1|x, l = 0) = 0$. Consider training labeled data $X^p = \{(x_i^p, l)\}_{i=1}^n \subset \{x : o = 1\}$ and testing data $X^u = \{x_j^u\}_{j=1}^m \subset \{x : o = 0\}$, where $x_i^p \in \mathbb{R}^{d_1}, x_j^u \in \mathbb{R}^{d_2}$. In the classical PU learning setting, $d_2 = d_1$. However, in (Séjourné et al., 2021) this assumption is relaxed. The goal is to leverage X^p to design a classifier $\hat{l} : x^u \to \{0, 1\}$ to predict $l(x^u)$ for all $x^u \in X^u$[2].

Following (Elkan & Noto, 2008; Chapel et al., 2020; Séjourné et al., 2021), in this experiment, we assume that the "select completely at random" (SCAR) assumption holds: $\text{Prob}(o = 1|x, l = 1) = \text{Prob}(o = 1|l = 1)$. In addition, we use $\pi = \text{Prob}(l = 1) \in [0, 1]$ to denote the ratio of positive samples in testing set[3]. Following the PU learning setting in (Kato et al., 2018; Hsieh et al., 2019; Chapel et al., 2020; Séjourné et al., 2021), we assume π is known. In all the PU learning experiments, we fix $\pi = 0.2$.

Our method. Similar to (Chapel et al., 2020) our method is designed as follows: We set $\mathrm{p} \in \mathbb{R}^n, \mathrm{q} \in \mathbb{R}^m$ as $p_i = \frac{\pi}{n}, i \in [1 : n]; \quad q_j = \frac{1}{m}, j \in [1 : m]$. Let $\mathbb{X}^p = (X^p, \|\cdot\|_{d_1}, \sum_{i=1}^n p_i \delta_{x_i}), \mathbb{X}^u = (X^u, \|\cdot\|_{d_2}, \sum_{j=1}^n q_j \delta_{y_j})$. We solve the partial GW problem $PGW_\lambda(\mathbb{X}^p, \mathbb{X}^u)$ and suppose γ is a solution. Let $\gamma_2 = \gamma^T 1_n$. The classifier $\hat{l}$ is defined by the indicator function

$$\hat{l}_\gamma(x^u) = \mathbb{1}_{\{x^u : \gamma_2(x^u) \geq \text{quantile}\}}, \tag{35}$$

where $\text{quantile} := \inf\{\gamma_2(x^u) : \gamma_2(x^u) \geq 1 - \pi\}$ is the quantile value of γ_2 according to $1 - \pi$.

Regarding the initial guess $\gamma^{(1)}$, Chapel et al. (2020) proposed a POT-based approach when X and Y are sampled from the same domain, i.e., $d_1 = d_2$, which we refer to as "POT initialization." The initial guess $\gamma^{(1)}$ is given by the following partial OT variant problem:

$$\gamma^{(1)} = \arg \min_{\gamma \in \Gamma_{PU,\pi}(\mathrm{p},\mathrm{q})} \langle L(X, Y), \gamma \rangle_F, \tag{36}$$

where $L(X, Y) \in \mathbb{R}^{n \times m}$, $(L(X, Y))_{ij} = \|x_i - y_j\|^2$ and

$$\Gamma_{PU,\pi}(\mathrm{p}, \mathrm{q}) := \{\gamma \in \mathbb{R}_+^{n \times m} : (\gamma^T 1_n)_j \in \{q_j, 0\}, \forall j; \\ \gamma 1_m \leq \mathrm{p}, |\gamma| = \pi\}. \tag{37}$$

The above problem can be solved by a Lasso (L^1 norm) regularized OT solver.

When X, Y are sampled from different spaces, that is, $d_1 \neq d_2$, the above technique (36) is not well-defined. Inspired by (Mémoli, 2011; Séjourné et al., 2021), we propose the following "first lower bound-partial OT" (FLB-POT) initialization:

$$\gamma^{(1)} = \arg \min_{\gamma \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})} \int_{X \times Y} |s_{X,2}(x) - s_{Y,2}(y)|^2 d\gamma(x, y) \\ + \lambda(|\mathrm{p} - \gamma_1| + |\mathrm{q} - \gamma_2|), \tag{38}$$

where $s_{X,2}(x) = \int_X |x - x'|^2 d\mu(x)$ and $s_{Y,2}$ is defined similarly. The above formula is analog to Eq. (7) in (Séjourné et al., 2021), which is designed for the unbalanced GW setting. To distinguish them, in this paper we call the Eq. (7) in (Séjourné et al., 2021) as "FLB-UOT initilization".

Dataset. The datasets include MNIST, EMNIST, and the following three domains of Caltech Office: Amazon (A), Webcam (W), and DSLR (D) (Saenko et al., 2010). For each domain, we select the SURF features (Saenko et al., 2010) and DECAF features (Donahue et al., 2014). For MNIST and EMNIST, we train an auto-encoder respectively and the embedding space dimension is 4 and 6 respectively.

Performance. In the Table 1, we present the accuracy comparison between PGW, UGW, and our method on MNIST, EMNIST datasets. The PU learning results on the Caltech Office dataset are included in the appendix. Regarding the wall-clock time, typically Primal-PGW and PGW require 1-35 seconds, UGW requires 100-160 seconds. See the Appendix for the complete accuracy table on all datasets, the visualization of all datasets, and the wall-clock time table.

6. Summary

In this paper, we extend McCann's technique to the Partial Gromov-Wasserstein (PGW) setting and introduce two Frank-Wolfe algorithms for the PGW problem. As a byproduct, we provide pertinent theoretical results, including the relation between PGW and GW, the metric properties of PGW, and the convergence behavior of our proposed solvers. Furthermore, we demonstrate the efficacy of the PGW solver in solving shape-matching problems and PU learning tasks.

[2]In classical setting, the goal is to learn a classifier for all x. In this experiment, we follow the setting in (Séjourné et al., 2021).

[3]In the classical setting, the prior distribution π is the ratio of positive samples of the original dataset. For convenience, we ignore the difference between this ratio in the original dataset and the test dataset.

7. Impact Statement

The work presented in this paper aims to advance the field of machine learning, particularly the supplementary theoretical developments and explorations of computational optimal transport. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alvarez-Melis, D. and Jaakkola, T. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1881–1890, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bai, Y., Schmitzer, B., Thorpe, M., and Kolouri, S. Sliced optimal partial transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13681–13690, 2023.

Bekker, J. and Davis, J. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109:719–760, 2020.

Benamou, J.-D., Froese, B. D., and Oberman, A. M. Numerical solution of the optimal transportation problem using the monge–ampère equation. *Journal of Computational Physics*, 260:107–126, 2014.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Bonneel, N. and Coeurjolly, D. SPOT: sliced partial optimal transport. *ACM Transactions on Graphics*, 38(4): 1–13, 2019. URL `https://dl.acm.org/doi/10.1145/3306346.3323021`.

Bronstein, A. M., Bronstein, M. M., and Kimmel, R. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006.

Caffarelli, L. A. and McCann, R. J. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pp. 673–730, 2010.

Chapel, L., Alaya, M. Z., and Gasso, G. Partial optimal tranport 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. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018a.

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

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018c.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

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

De Ponti, N. and Mondino, A. Entropy-transport distances between unbalanced metric measure spaces. *Probability Theory and Related Fields*, 184(1-2):159–208, 2022.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pp. 647–655. PMLR, 2014.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220, 2008.

Fatras, K., Séjourné, T., Flamary, R., and Courty, N. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pp. 3186–3197. PMLR, 2021.

Figalli, A. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

Figalli, A. and Gigli, N. A new transportation distance between non-negative measures, with applications to gradients flows with dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Frank, M., Wolfe, P., et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

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

Guittet, K. *Extended Kantorovich norms: a tool for optimization*. PhD thesis, INRIA, 2002.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Heinemann, F., Klatt, M., and Munk, A. Kantorovich–rubinstein distance and barycenter for finitely supported measures: Foundations and algorithms. *Applied Mathematics & Optimization*, 87(1):4, 2023.

Hsieh, Y.-G., Niu, G., and Sugiyama, M. Classification from positive, unlabeled and biased negative data. In *International Conference on Machine Learning*, pp. 2820–2829. PMLR, 2019.

Kato, M., Teshima, T., and Honda, J. Learning from positive and unlabeled data with a selection bias. In *International conference on learning representations*, 2018.

Kolouri, S., Naderializadeh, N., Rohde, G. K., and Hoffmann, H. Wasserstein embedding for graph learning. In *International Conference on Learning Representations*, 2020.

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

Lellmann, J., Lorenz, D. A., Schonlieb, C., and Valkonen, T. Imaging with kantorovich–rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

Liero, M., Mielke, A., and Savare, G. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018. URL http://link.springer.com/10.1007/s00222-017-0759-8.

Liu, W., Zhang, C., Xie, J., Shen, Z., Qian, H., and Zheng, N. Partial gromov-wasserstein learning for partial graph matching. *arXiv preprint arXiv:2012.01252*, 2020.

Liu, X., Bai, Y., Tran, H., Zhu, Z., Thorpe, M., and Kolouri, S. Ptlp: Partial transport l^p distances. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023.

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.

Papadakis, N., Peyré, G., and Oudet, E. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 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.

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

Piccoli, B. and Rossi, F. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014. URL http://arxiv.org/abs/1206.3219.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pp. 213–226. Springer, 2010.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Séjourné, T., Vialard, F.-X., and Peyré, G. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

Sturm, K.-T. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*, volume 290. American Mathematical Society, 2023.

Thual, A., TRAN, Q. H., Zemskova, T., Courty, N., Flamary, R., Dehaene, S., and Thirion, B. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in Neural Information Processing Systems*, 35: 21792–21804, 2022.

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.

Villani, C. *Optimal transport: old and new*. Springer, 2009. URL https://www.cedricvillani.org/sites/dev/files/old_images/2012/08/preprint-1.pdf.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

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.

A. Notation and abbreviations

- OT: Optimal Transport.
- POT: Partial Optimal Transport.
- GW: Gromov-Wasserstein.
- PGW: Partial Gromov-Wasserstein minimization problem.
- FW: Frank-Wolfe.
- $S \subset \mathbb{R}^{d_1}, X \subset \mathbb{R}^{d_2}, Y \subset \mathbb{R}^{d_3}$: subsets in Euclidean spaces.
- $\|\cdot\|$: Euclidean norm.
- $X^2 = X \times X$.
- $\mathcal{M}_+(S)$: set of all positive (non-negative) Randon (finite) measures defined on S.
- $\mathcal{P}_2(S)$: set of all probability measures defined on S, whose second moment is finite.
- $\mathbb{R}_+$: set of all non-negative real numbers.
- $\mathbb{R}^{n\times m}$: set of all $n \times m$ matrices with real coefficients.
- $\mathbb{R}_+^{n\times m}$ (resp. $\mathbb{R}_+^n$): set of all $n \times m$ matrices (resp., n-vectors) with non-negative coefficients.
- $\mathbb{R}^{n\times m\times n\times m}$: set of all $n \times m \times n \times m$ tensors with real coefficients.
- $1_n, 1_{n\times m}, 1_{n\times m\times n\times m}$: vector, matrix, and tensor of all ones.
- $\mathbb{1}_E$: characteristic function of a measurable set E

$$\mathbb{1}_E(z) = \begin{cases} 1 & \text{if } z \in E, \\ 0 & \text{otherwise.} \end{cases}$$

- $\mathbb{S}, \mathbb{X}, \mathbb{Y}$: metric measure spaces (mm-spaces): $\mathbb{S} = (S, d_S, \sigma)$, $\mathbb{X} = (X, d_X, \mu)$, $\mathbb{Y} = (Y, d_Y, \nu)$.
- C^X: given a discrete mm-space $\mathbb{X} = (X, d_X, \mu)$, where $X = \{x_1, \ldots, x_n\}$, the symmetric matrix $C^X \in \mathbb{R}^{n\times n}$ is defined as $C^X_{i,i'} = d_X(x_i, x_i')$.
- $\mu^{\otimes 2}$: product measure $\mu \otimes \mu$.
- $T_\#\sigma$: T is a measurable function and σ is a measure on X. $T_\#\sigma$ is the push-forward measure of σ, i.e., its is the measure on Y such that for all Borel set $A \subset Y$, $T_\#\sigma(A) = \sigma(T^{-1}(A))$.
- $\gamma, \gamma_1, \gamma_2$: γ is a joint measure defined for example in $X \times Y$; γ_1, γ_2 are the first and second marginals of γ, respectively. In the discrete setting, they are viewed as matrix and vectors, i.e., $\gamma \in \mathbb{R}_+^{n\times m}$, and $\gamma_1 = \gamma 1_m \in \mathbb{R}_+^n$, $\gamma_2 = \gamma^T 1_n \in \mathbb{R}_+^m$.
- $\Gamma(\mu, \nu)$, where $\mu \in \mathcal{P}_2(X), \nu \in \mathcal{P}_2(Y)$ (where X, Y may not necessarily be the same set): it is the set of all the couplings (transportation plans) between μ and ν, i.e., $\Gamma(\mu, \nu) := \{\gamma \in \mathcal{P}_2(X \times Y) : \gamma_1 = \mu, \gamma_2 = \nu\}$.
- $\Gamma(\mathrm{p}, \mathrm{q})$: it is the set of all the couplings between the discrete probability measures $\mu = \sum_{i=1}^n p_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m y_j \delta_{y_j}$ with weight vectors

$$\mathrm{p} = [p_1, \ldots, p_n]^T \qquad \text{and} \qquad \mathrm{q} = [q_1, \ldots, q_n]^T. \tag{39}$$

  That is, $\Gamma(\mathrm{p}, \mathrm{q})$ coincides with $\Gamma(\mu, \nu)$, but it is viewed as a subset of $n \times m$ matrices defined in (19).
- p, q: real numbers $1 \leq p, q < \infty$.
- p, q: vectors of weights as in (39).

- $\mathrm{p} = [p_1, \ldots, p_n] \leq \mathrm{p}' = [p'_1, \ldots, p'_n]$ if $p_j \leq p'_j$ for all $1 \leq j \leq n$.
- $|\mathrm{p}| = \sum_{i=1}^n p_i$ for $\mathrm{p} = [p_1 \ldots, p_n]$.
- $c(x, y) : X \times Y \to \mathbb{R}_+$ denotes the cost function used for classical and partial optimal transport problems. lower-semi continuous function.
- $OT(\mu, \nu)$: it is the classical optimal transport (OT) problem between the probability measures μ and ν defined in (1).
- $W_p(\mu, \nu)$: it is the p-Wasserstein distance between the probability measures μ and ν defined in (2), for $1 \leq p < \infty$.
- The Partial Optimal Transport (OPT) problem is defined in (3) or, equivalently, in (4).
- $|\mu|$: total variation norm of the positive Randon (finite) measure μ defined on a measurable space X, i.e., $|\mu| = \mu(X)$.
- $\mu \leq \sigma$: denotes that for all Borel set $B \subseteq X$ we have that the measures $\mu, \sigma \in \mathcal{M}_+(X)$ satisfy $\mu(B) \leq \sigma(B)$.
- $\Gamma_{\leq}(\mu, \nu)$, where $\mu \in \mathcal{M}_+(X), \nu \in \mathcal{M}_+(Y)$: it is the set of all "partial transportation plans"
$$\Gamma_{\leq}(\mu, \nu) := \{\gamma \in \mathcal{M}_+(X \times Y) : \; \gamma_1 \leq \mu, \gamma_2 \leq \nu\}.$$
- $\Gamma_{\leq}(\mathrm{p}, \mathrm{q})$: it is the set of all the "partial transportation plans" between the discrete probability measures $\mu = \sum_{i=1}^n p_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m y_j \delta_{y_j}$ with weight vectors $\mathrm{p} = [p_1, \ldots, p_n]$ and $\mathrm{q} = [q_1, \ldots, q_n]$. That is, $\Gamma_{\leq}(\mathrm{p}, \mathrm{q})$ coincides with $\Gamma_{\leq}(\mu, \nu)$, but it is viewed as a subset of $n \times m$ matrices defined in (20).
- $\lambda > 0$: positive real number.
- $\hat{\infty}$: auxiliary point.
- $\hat{X} = X \cup \{\hat{\infty}\}$.
- $\hat{\mu}, \hat{\nu}$: given in (5).
- $\hat{\mathrm{p}}, \hat{\mathrm{q}}$: given in (25).
- $\hat{\gamma}$: given in (8).
- $\hat{c}(\cdot, \cdot) : \hat{X} \times \hat{Y} \to \mathbb{R}_+$: cost as in (6).
- $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$: cost function for the GW problems.
- $D : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$: generic distance on $\mathbb{R}$ used for GW problems.
- $GW_q^L(\cdot, \cdot)$: GW optimization problem given in (9).
- $GW_q^p(\cdot, \cdot)$: GW optimization problem given in (9) when $L(\cdot, \cdot) = D^p(\cdot, \cdot)$.
- $GW_g(\cdot, \cdot)$: generalized GW problem given in (52).
- $\widehat{GW}$: GW-variant problem given in (15) for the general case, and in (27) for the discrete setting.
- $\hat{L}$: cost given in (14) for the GW-variant problem.
- $\hat{d} : \hat{X} \times \hat{X} \to \mathbb{R}_+ \cup \{\infty\}$: "generalized" metric given in (13) for $\hat{X}$.
- $\mathcal{G}_q$: subset of mm-spaces given in (10).
- $\mathbb{X} \sim \mathbb{Y}$: equivalence relation in $\mathcal{G}_q$, where given $\mathbb{X}, \mathbb{Y} \in \mathcal{G}_q$, we define $\mathbb{X} \sim \mathbb{Y}$ if and only if $GW_q^p(\mathbb{X}, \mathbb{Y}) = 0$.
- $PGW_{\lambda,q}^L(\cdot, \cdot)$: partial GW optimization problem given in (11) or, equivalently, in (12).
- $PGW_{\lambda,q}(\cdot, \cdot)$: partial GW problem $PGW_{\lambda,q}^L(\cdot, \cdot)$ for which $L(r_1, r_2) = D(r_1, r_2)^p$ for a metric D in $\mathbb{R}$, and $1 \leq p < \infty$.Typically $D(r_1, r_2) = |r_1 - r_2|$, and $p = 2$. It is explicitly given in (17).

- $PGW_\lambda(\cdot,\cdot)$: discrete partial GW problem given in (23).
- $\mathcal{L}$: functional for the optimization problem $PGW_\lambda(\cdot,\cdot)$.
- M, $\tilde{M}$, and $\hat{M}$: see (21), (22), and (26), respectively.
- $\langle\cdot,\cdot\rangle_F$: Frobenius inner product for matrices, i.e., $\langle A,B\rangle_F=\text{trace}(A^TB)=\sum_{i,j}^{n,m}A_{i,j}B_{i,j}$ for all $A,B\in\mathbb{R}^{n\times m}$.
- $M\circ\gamma$: product between the tensor M and the matrix γ.
- ∇: gradient.
- $[1:n]=\{1,\ldots,n\}$.
- α: step size based on the line search method.
- $\gamma^{(1)}$: initialization of the algorithm.
- $\gamma^{(k)}$, $\gamma^{(k)'}$: previous and new transportation plans before and after step 1 in the $k-$th iteration of version 1 of our proposed FW algorithm.
- $\hat{\gamma}^{(k)}$, $\hat{\gamma}^{(k)'}$: previous and new transportation plans before and after step 1 in the $k-$th iteration of version 2 of our proposed FW algorithm.
- $G=2\tilde{M}\circ\gamma$, $\hat{G}=2\hat{M}\circ\hat{\gamma}$: Gradient of the objective function in version 1 and version 2, respectively, of our proposed FW algorithm for solving the discrete version of partial GW problem.
- $(\delta\gamma,a,b)$ and $(\delta\hat{\gamma},a,b)$: given in (32) and (34) for versions 1 and 2 of the algorithm, respectively.
- C^1-function: continuous and with continuous derivatives.
- $\Gamma_{PU,\pi}(\mathrm{p},\mathrm{q})$: defined in (37)

B. Proof of Proposition 3.1

For the first statement, the idea of the proof is inspired by the proof of Proposition 1 in (Piccoli & Rossi, 2014).

The goal is to verify that

$$\begin{aligned}PGW_{\lambda,q}^L(\mathbb{X},\mathbb{Y}) &:= \underbrace{\inf_{\gamma\in\mathcal{M}_+(X,Y)}\int_{(X\times Y)^2}L(d_X^q(x,x'),d_Y^q(y,y'))d\gamma(x,y)d\gamma(x',y')}_{\text{transport GW cost}}+\underbrace{\lambda\left(|\mu^{\otimes 2}-\gamma_1^{\otimes 2}|+|\nu^{\otimes 2}-\gamma_2^{\otimes 2}|\right)}_{\text{mass penalty}}\\ &=\inf_{\gamma\in\Gamma_\leq(\mu,\nu)}\int_{(X\times Y)^2}L(d_X^q(x,x'),d_Y^q(y,y'))d\gamma(x,y)d\gamma(x',y')+\lambda\left(|\mu^{\otimes 2}-\gamma_1^{\otimes 2}|+|\nu^{\otimes 2}-\gamma_2^{\otimes 2}|\right).\end{aligned}\tag{40}$$

Given one $\gamma\in\mathcal{M}_+(X\times Y)$ such that $\gamma_1\leq\mu$ does not hold. Then we can write the Lebesgue decomposition of γ_1 with respect to μ:

$$\gamma_1=f\mu+\mu^\perp,$$

where $f\geq 0$ is the Radon-Nikodym derivative of γ_1 with respect to μ, and $\mu^\perp,\mu$ are mutually singular, that is, there exist measurable sets A,B such that $A\cap B=\emptyset$, $X=A\cup B$ and $\mu^\perp(A)=0,\mu(B)=0$. Without loss of generality, we can assume that the support of f lies on A, since

$$\gamma_1(E)=\int_{E\cap A}f(x)\,d\mu(x)+\mu^\perp(E\cap B)\qquad\forall E\subseteq X\text{ measurable.}$$

Define $A_1=\{x\in A:\ f(x)>1\}, A_2=\{x\in A:\ f(x)\leq 1\}$ (both are measurable, since f is measurable), and define $\bar{\mu}=\min\{f,1\}\mu$. Then,

$$\bar{\mu}\leq\mu\qquad\text{and}\qquad\bar{\mu}\leq f\mu\leq f\mu+\mu^\perp=\gamma_1.$$

There exists a $\bar\gamma \in \mathcal{M}_+(X \times Y)$ such that $\bar\gamma_1 = \bar\mu, \bar\gamma \leq \gamma$, and $\bar\gamma_2 \leq \gamma_2$. Indeed, we can construct $\bar\gamma$ in the following way: First, let $\{\gamma^x\}_{x\in X}$ be the set of conditional measures (disintegration) such that for every measurable (test) function $\psi : X \times Y \to \mathbb{R}$ we have

$$\int \psi(x,y)\, d\gamma(x,y) = \int_X \int_Y \psi(x,y)\, d\gamma^x(y)\, d\gamma_1(x).$$

Then, define $\bar\gamma$ as

$$\bar\gamma(U) := \int_X \int_Y \mathbb{1}_U(x,y)\, d\gamma^x(y)\, d\bar\mu(x) \qquad \forall U \subseteq X \times Y \text{ Borel.}$$

Then, $\bar\gamma$ verifies that $\bar\gamma_1 = \bar\mu$, and since $\bar\mu \leq \gamma_1$, we also have that $\bar\gamma \leq \gamma$, which implies $\bar\gamma_2 \leq \gamma_2$.

Since $|\gamma_1| = |\gamma_2|$ and $|\bar\gamma_1| = |\bar\gamma_2|$, then we have $|\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}| = |\gamma_2^{\otimes 2} - \bar\gamma_2^{\otimes 2}|$.

We claim that

$$|\mu^{\otimes 2} - \gamma_1^{\otimes 2}| \geq |\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}| + |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}|. \tag{41}$$

- *Left-hand side of* (41): Since $\{A, B\}$ is a partition of X, we first spit the left-hand side of (41) as

$$|\mu^{\otimes 2} - \gamma_1^{\otimes 2}| = \underbrace{|\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A\times A)}_{(I)} + \underbrace{|\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A\times B) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(B\times A)}_{(II)} + \underbrace{|\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(B\times B)}_{(III)}.$$

  Then we have

$$(III) = |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(B\times B) = \mu^\perp \otimes \mu^\perp (B\times B) = |\mu^\perp|^2,$$

$$(II) = |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A\times B) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(B\times A) = 2|\mu^\perp| \cdot (|\mu - \gamma_1|(A)).$$

  Since $\gamma_1 = f\mu$ in A, then $\bar\gamma_1 = \gamma_1$ in A_2 and $\bar\gamma_1 = \mu$ in A_1, so we have

$$|\mu - \gamma_1|(A) = |\mu - \gamma_1|(A_1) + |\mu - \gamma_1|(A_2) = (\gamma_1 - \bar\gamma_1)(A_1) + (\mu - \bar\gamma_1)(A_2) = |\gamma_1 - \bar\gamma_1|(A) + |\mu - \bar\gamma_1|(A).$$

  Thus,

$$(II) = 2|\mu^\perp| \cdot (|\gamma_1 - \bar\gamma_1|(A) + |\mu - \bar\gamma_1|(A)),$$

  and we also get that

$$\begin{aligned}
(I) &= |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A\times A)\\
&= |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_1\times A_1) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_2\times A_2) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_1\times A_2) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_2\times A_1)\\
&= (\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2})(A_1\times A_1) + (\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2})(A_2\times A_2) +\\
&\qquad + |\bar\gamma_1 \otimes \mu - \gamma_1 \otimes \bar\gamma_1|(A_1\times A_2) + |\mu \otimes \bar\gamma_1 - \bar\gamma_1 \otimes \gamma_1|(A_2\times A_1)\\
&= (\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2})(A_1\times A_1) + (\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2})(A_2\times A_2) + 2|\bar\gamma_1 - \gamma_1|(A_1) \cdot |\mu - \bar\gamma_1|(A_2)\\
&= |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times A) + |\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times A) + \underbrace{2|\bar\gamma_1 - \gamma_1|(A_1) \cdot |\mu - \bar\gamma_1|(A_2)}_{\geq 0}.
\end{aligned}$$

- *Right-hand side of* (41)*:* First notice that

$$|\gamma_1 - \bar\gamma_1|(B) = (\gamma_1 - \bar\gamma_1)(B) \leq \gamma_1(B) = |\mu^\perp|, \tag{42}$$

  and since $\bar\gamma_1 \leq \mu$ and $\mu(B) = 0$, we have

$$|\mu - \bar\gamma_1|(B) = 0. \tag{43}$$

  Then,

$$\begin{aligned}
&|\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}| + |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}| =\\
&= |\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times A) + |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times A) + |\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(B\times B) + |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(B\times B) +\\
&\quad + \cancel{|\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times B)} + |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times B) + \cancel{|\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(B\times A)} + |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(B\times A)\\
&\leq \underbrace{|\mu^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times A) + |\gamma_1^{\otimes 2} - \bar\gamma_1^{\otimes 2}|(A\times A)}_{\leq (I)} + \underbrace{|\mu^\perp|^2}_{=(III)} + \underbrace{2|\mu^\perp| \cdot (|\gamma_1 - \bar\gamma_1|(A))}_{=(II)}.
\end{aligned}$$

Thus, (41) holds.

We finish the proof of the proposition by noting that

$$\begin{aligned}|\mu^{\otimes 2}-\bar{\gamma}_1^{\otimes 2}|+|\nu^{\otimes 2}-\bar{\gamma}_2^{\otimes 2}| &\leq |\mu^{\otimes 2}-\gamma_1^{\otimes 2}|-|\gamma_1^{\otimes 2}-\bar{\gamma}_1^{\otimes 2}|+|\nu^{\otimes 2}-\bar{\gamma}_2^{\otimes 2}| \\ &= |\mu^{\otimes 2}-\gamma_1^{\otimes 2}|-|\gamma_2^{\otimes 2}-\bar{\gamma}_2^{\otimes 2}|+|\nu^{\otimes 2}-\bar{\gamma}_2^{\otimes 2}| \\ &\leq |\mu^{\otimes 2}-\gamma_1^{\otimes 2}|+|\nu^{\otimes 2}-\gamma_2^{\otimes 2}|\end{aligned}$$

where the first inequality follows from (41), and the second inequality holds from the fact the total variation norm $|\cdot|$ satisfies triangular inequality. Therefore $\bar{\gamma}$ induces a smaller transport GW cost than γ (since $\bar{\gamma}\leq\gamma$), and also $\bar{\gamma}$ decreases the mass penalty in comparison that corresponding to γ. Thus, $\bar{\gamma}$ is a better GW transportation plan, which satisfies $\bar{\gamma}_1\leq\mu$. Similarly, we can further construct $\bar{\gamma}'$ based on $\bar{\gamma}$ such that $\bar{\gamma}_1'\leq\mu, \bar{\gamma}_2'\leq\nu$. Therefore, we can restrict the minimization in (11) from $\mathcal{M}_+(X\times Y)$ to $\Gamma_\leq(\mu,\nu)$. Thus, the equality (40) is satisfied.

Remark B.1. Given $\gamma\in\Gamma\leq(\mu,\nu)$, since $\gamma_1\leq\mu$, $\gamma_2\leq\nu$, and $\gamma_1(X)=|\gamma_1|=|\gamma|=|\gamma_2|=\gamma_2(Y)$, we have

$$|\mu^{\otimes 2}-\gamma_1^{\otimes 2}|+|\nu^{\otimes 2}-\gamma_2^{\otimes 2}|=\mu^{\otimes 2}(X^2)-\gamma_1^{\otimes 2}(X^2)+\nu^{\otimes 2}(Y^2)-\gamma_2^{\otimes 2}(Y^2)=|\mu|^2+|\nu|^2-2|\gamma|^2,$$

and so the transportation cost in partial GW problem (12) becomes

$$\begin{aligned}C(\gamma;\lambda) &:= \int_{(X\times Y)^2} L(d_X^q(x,x'),d_Y^q(y,y'))\,d\gamma(x,y)d\gamma(x',y')+\lambda\left(|\mu^{\otimes 2}-\gamma_1^{\otimes 2}|+|\nu^{\otimes 2}-\gamma_2^{\otimes 2}|\right) \\ &= \int_{(X\times Y)^2} L(d_X^q(x,x'),d_Y^q(y,y'))\,d\gamma(x,y)d\gamma(x',y')+\lambda\left(|\mu|^2+|\nu|^2-2|\gamma|^2\right) \\ &= \int_{(X\times Y)^2} \left(L(d_X^q(x,x'),d_Y^q(y,y'))-2\lambda\right)\,d\gamma(x,y)d\gamma(x',y')+\underbrace{\lambda\left(|\mu|^2+|\nu|^2\right)}_{\text{does not depend on }\gamma}. \end{aligned} \tag{44}$$

C. Proof of Proposition 3.2

In this section, we discuss the minimizer of the partial GW problem. Trivially, $\Gamma_\leq(\mu,\nu)\subseteq\mathcal{M}_+(X\times Y)$, and by using Proposition 3.1 it is enough to show that a minimizer for problem (12) exists.

and that this minimizer can also solve (11).

From Proposition B.1 in (Liu et al., 2023) , we have that $\Gamma_\leq(\mu,\nu)$ is a compact set with respect to weak-convergence topology.

Consider a sequence $(\gamma^n)\in\Gamma_\leq(\mu,\nu)$, such that

$$C(\gamma^n;\lambda)\to\inf_{\gamma\in\Gamma_\leq(\mu,\nu)}C(\gamma^n;\lambda).$$

Then, there exists a sub-sequence that converges weakly: $\gamma^{n_k}\xrightarrow{w}\gamma^*$ for some $\gamma^*\in\Gamma_\leq(\mu,\nu)$.

We claim that

$$\Gamma_\leq(\mu,\nu)\ni\gamma\mapsto C(\gamma;\lambda)=\int_{(X\times Y)^2}L(d_X^q(x,x'),d_Y^q(y,y'))d\gamma(x,y)d\gamma(x',y')\in\mathbb{R}$$

is a lower-semi continuous function.

Since L is a C^1-function, and X,Y are compact, we have that the following mapping

$$X^2\times Y^2\ni((x,x'),(y,y'))\mapsto L(d_X^q(x,y),d_Y^q(x',y'))\in\mathbb{R}$$

is Lipschitz with respect to the metric $d_X(x,x')+d_Y(y,y')$. By [(Mémoli, 2011) Lemma 10.1], we have

$$\Gamma_\leq(\mu,\nu)\ni\gamma\mapsto C(\gamma;\lambda)$$

is a continuous mapping, thus, it is lower-semi-continuous.

By Weierstrass Theorem, the facts $\gamma^{n_k} \xrightarrow{w} \gamma^*$ and $\gamma \mapsto C(\gamma;\lambda)$ lower-semi-continuous, imply that

$$PGW^L_{\lambda,q}(\mathbb{X},\mathbb{Y}) = \lim_{k\to\infty} C(\gamma^{n_k};\lambda) = C(\gamma^*;\lambda).$$

Thus, we prove γ^* is a minimizer for the problem $PGW^L_{\lambda,q}(\mathbb{X},\mathbb{Y})$ defined in (12).

D. Proof of Proposition 3.3

Without loss of generality, we can assume $X, Y \subset \Omega \subset \mathbb{R}^d$ for some d large enough. Moreover, we can assume $X = Y$. (Notice that the measures μ and ν might have very different supports, even be singular measures in $\mathbb{R}^d$).

For convenience, we denote the mapping defined (8) as F. By (Bai et al., 2023; Caffarelli & McCann, 2010), F is a well-defined bijection.

Given $\gamma \in \Gamma_\leq(\mu,\nu)$, we have $\hat\gamma = F(\gamma) \in \Gamma(\hat\mu,\hat\nu)$. Let $\hat C(\hat\gamma)$ denote the transformation cost in the GW-variant problem (15), i.e.,

$$\hat C(\hat\gamma) = \int_{(\hat X\times\hat Y)^2} \hat L(\hat d^q_{\hat X}(x,x'), \hat d^q_{\hat Y}(y,y'))\, d\hat\gamma(x,y) d\hat\gamma(x',y')$$

Then, we have

$$\begin{aligned}
C(\gamma;\lambda) &= \int_{(X\times Y)^2} (L(d^q_X(x,x'), d^q_Y(y,y')) - 2\lambda)\, d\gamma^{\otimes 2} + \underbrace{\lambda(|\mu|+|\nu|)}_{\text{does not depend on } \gamma} \\
&= \int_{(X\times Y)^2} (L(d^q_X(x,x'), d^q_Y(y,y')) - 2\lambda)\, d\hat\gamma^{\otimes 2} + \lambda(|\mu|+|\nu|) \qquad (\text{since } \hat\gamma|_{X\times Y} = \gamma) \\
&= \int_{(X\times Y)^2} (L(\hat d^q_{\hat X}(x,x'), \hat d^q_{\hat Y}(y,y')) - 2\lambda)\, d\hat\gamma^{\otimes 2} + \lambda(|\mu|+|\nu|) \quad (\text{since } \hat d_{\hat X}|_{X\times X} = d_X,\ \hat d_{\hat Y}|_{Y\times Y} = d_Y) \\
&= \int_{(X\times Y)^2} \hat L(\hat d^q_{\hat X}(x,x'), \hat d^q_{\hat Y}(y,y'))\, d\hat\gamma^{\otimes 2} + \lambda(|\mu|+|\nu|) \quad (\text{since } \hat L|_{\mathbb{R}\times\mathbb{R}}(\cdot,\cdot) = L(\cdot,\cdot) - 2\lambda) \\
&= \int_{(\hat X\times\hat Y)^2} \hat L(\hat d^q_{\hat X}(x,x'), \hat d^q_{\hat Y}(y,y'))\, d\hat\gamma^{\otimes 2} + \underbrace{\lambda(|\mu|+|\nu|)}_{\text{does not depend on } \hat\gamma}. \quad (\text{since } \hat L \text{ assigns 0 to } \hat\infty)
\end{aligned}$$

Combined with the fact $F: \gamma \mapsto \hat\gamma$ is a bijection, we have that γ is optimal for (12) if and only if $\hat\gamma$ is optimal for (15). Since an optimal $\hat\gamma$ for (15) always exists because it is a GW problem, we have:

$$\arg\min_{\hat\gamma\in\Gamma(\hat\mu,\hat\nu)} \hat C(\hat\gamma) = \arg\min_{\gamma\in\Gamma_\leq(\mu,\nu)} C(\gamma;\lambda). \tag{45}$$

E. Proof of Proposition 3.4

We will prove it by contradiction. Let

$$A = \{(x,y,x',y') :\ L(d^q_X(x,x'), d^q_Y(y,y')) > 2\lambda\} \subseteq (X\times Y)^2. \tag{46}$$

Assume that there exists a transportation plan $\gamma \in \Gamma_\leq(\mu,\nu)$ such that $\gamma^{\otimes 2}(A) > 0$. We claim A is a product set.

Let $\mathcal{F}$ be the family of sets in $X \times Y$ such that for each $B' \in \mathcal{F}$ we have

$$L(d^q_X(x,x'), d^q_Y(y,y')) \geq 2\lambda, \qquad \forall (x,y),(x',y') \in B'.$$

Let B be the largest set in $\mathcal{F}$, i.e., $B = \bigcup_{B'\in\mathcal{F}} B'$.

It is that clear $B^2 := B \times B \subset A$. For the other direction, let $(x,y,x',y') \in A$ and by definition of A, we have $L(d^q_X(x,x'), d^q_Y(y,y')) \geq 2\lambda$, then $(x,y),(x',y') \in B'$ for some $B' \in \mathcal{F}$, and so $(x,y),(x',y') \in B' \subset B$ by the definition of B. Thus $A = B \times B$.

Next, by definition of measure, there exists $\epsilon > 0$ such that

$$\gamma(A_\epsilon) > 0, \qquad \text{where } A_\epsilon = \{(x,y,x',y') : \ L(d_X^q(x,x'), d_Y^q(y,y')) > 2\lambda + \epsilon\} \subseteq A.$$

We restrict the measure $\gamma^{\otimes 2}$ to the complement set of A, and we denote such measure as $\gamma^{\otimes 2}\mid_A$:

$$\gamma^{\otimes 2}\mid_A (D) = \gamma^{\otimes 2}(D \cap A), \qquad \forall \text{ Borel set } D \subset (X \times Y)^2.$$

Since $\gamma^{\otimes 2}$ is a product measure and $A = B^2$ is a product set, it is straightforward to verify that $\gamma^{\otimes 2}\mid_A$ is a product measure, in fact, $\gamma^{\otimes 2}\mid_A = (\gamma\mid_B)^{\otimes 2}$. Since, $0 \le \gamma\mid_B \le \gamma$, we have $\gamma\mid_B \in \Gamma_\le(\mu,\nu)$. We define $\gamma' = \gamma\mid_{B^c}$, where $B^c := X \times Y \setminus B$ is the complement of B on $X \times Y$. Notice that $(\gamma')^{\otimes 2}|_A = ((\gamma'|_B)|_{B^c})^{\otimes 2} = 0$. Also, $\gamma^{\otimes 2} - (\gamma')^{\otimes 2} \ge \gamma^{\otimes 2}|_A$. Also,

Thus, when considering the partial GW transportation cost as in (44) we obtain,

$$\begin{aligned} C(\gamma;\lambda) - C(\gamma';\lambda) &= \int_{(X\times Y)^2} (L(d_X^q(x,x'), d_Y^q(y,y')) - 2\lambda)(d\gamma^{\otimes 2} - d(\gamma')^{\otimes 2}) \\ &\ge \int_A (L(d_X^q(x,x'), d_Y^q(y,y')) - 2\lambda)\, d\gamma^{\otimes 2} \ge \int_{A_\epsilon} \epsilon\, d\gamma^{\otimes 2} = \epsilon\, \gamma^{\otimes 2}(A_\epsilon) > 0. \end{aligned}$$

That is, for any $\gamma \in \Gamma_\le(\mu,\nu)$, we can find a better transportation plan γ' such that $(\gamma')^{\otimes 2}(A) = 0$.

Notice that the same result holds if we redefine A as $A := \{(x,y,x',y') : L(d_X^q(x,x'), d_Y^q(y,y'))^2 \ge 2\lambda\}$. By a similar process, we can prove the existence of an optimal γ for partial GW problem (12) such that $\gamma^{\otimes 2}(A) = 0$.

F. Proof of Proposition 3.5: Metric property of partial GW

For simplicity in the notation, consider $q = 1$. Let $L(r_1, r_2) = D^2(r_1, r_2)$ for D a metric on $\mathbb{R}$. That is, for simplicity we assume $p = 2$. Since all the metrics in $\mathbb{R}$ are equivalent, for simplicity, consider $D(r_1, r_2) = |r_1 - r_2|^2$ (notice that this satisfies the hypothesis of Proposition 4.1 used in the experiments). Consider the GW problem

$$GW(\mathbb{X}, \mathbb{Y}) = \inf_{\gamma' \in \Gamma(\mu,\nu)} \int_{(X\times Y)^2} |d_X(x,x') - d_Y(y,y')|^2 d(\gamma')^{\otimes 2},$$

in the space $\mathcal{G}_1$ with the equivalence relation $\mathbb{X} \sim \mathbb{Y}$ if and only if $GW(\mathbb{X}, \mathbb{Y}) = 0$. By Remark B.1, the PGW problem can be formulated as in (17), and we denote it here by $PGW_\lambda(\cdot,\cdot)$.

Proof of Proposition 3.5. Consider $\mathbb{X} = (X, d_X, \mu), \mathbb{Y} = (Y, d_Y, \nu) \in \mathcal{G}_1$. It is straightforward to verify $PGW_\lambda(\mathbb{X}, \mathbb{Y}) \ge 0$, and that $PGW_\lambda(\mathbb{X}, \mathbb{Y}) = PGW_\lambda(\mathbb{Y}, \mathbb{X})$.

If $PGW_\lambda(\mathbb{X}, \mathbb{Y}) = 0$, we claim that $|\nu| = |\mu|$ and that there exist an optimal γ for PGW_λ such that $|\mu| = |\gamma| = |\nu|$.

Assume $|\nu| \ne |\mu|$. For convenience, suppose $|\nu|^2 \le |\mu|^2 - \epsilon$ for some $\epsilon > 0$. Then, for each $\gamma \in \Gamma_\le(\mu,\nu)$, we have $|\gamma^{\otimes 2}| \le |\nu|^2 \le |\mu|^2 - \epsilon$, and so

$$PGW_\lambda(\mathbb{X}, \mathbb{Y}) \ge \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma|) \ge \lambda(|\mu^2| - |\gamma|) \ge \lambda\epsilon > 0. \tag{47}$$

Thus, $PGW_\lambda(\mathbb{X}, \mathbb{Y}) > 0$, which is a contradiction. Then we have $|\nu| = |\mu|$.

Similarly, assume for all optimal γ for PGW_λ, it holds that $|\gamma| < |\nu| = |\mu|$. Thus, for any of such γ, we have

$$PGW_\lambda(\mathbb{X}, \mathbb{Y}) \ge \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma^{\otimes 2}|) > 0, \tag{48}$$

which is a contradiction since $PGW_\lambda(\mathbb{X}, \mathbb{Y}) = 0$. Thus there exist an optimal γ for PGW_λ with $|\gamma| = |\nu| = |\mu|$.

This, combined with the fact that for $\gamma \in \Gamma_\le(\mu,\nu)$ (i.e., $\pi_{1\#}\gamma = \mu, \pi_{2\#}\gamma = \nu$) results in having $\gamma \in \Gamma(\nu,\mu)$. Therefore, for an optimal γ,

$$\begin{aligned} 0 = PGW_\lambda(\mathbb{X}, \mathbb{Y}) &= \int_{(X\times Y)^2} |d_X(x,x') - d_Y(y,y')|^2 d\gamma^{\otimes 2} + \underbrace{\lambda(|\mu|^2 + |\nu|^2 - 2|\gamma^{\otimes 2}|)}_{=0} \\ &= \int_{(X\times Y)^2} |d_X(x,x') - d_Y(y,y')|^2 d\gamma^{\otimes 2} = \inf_{\gamma' \in \Gamma(\mu,\nu)} \int_{(X\times Y)^2} |d_X(x,x') - d_Y(y,y')|^2 d(\gamma')^{\otimes 2} = GW(\mathbb{X}, \mathbb{Y}) \end{aligned}$$

Thus, we have $\mathbb{X} \sim \mathbb{Y}$.

It remains to show the triangular inequality. Let $\mathbb{S} = (S, d_S, \sigma)$, $\mathbb{X} = (X, d_X, \mu)$, $\mathbb{Y} = (Y, d_Y, \nu)$ in $\mathcal{G}_1$, and define $\hat{\mathbb{S}} = (\hat{S}, \hat{d}_S, \hat{\sigma})$, $\hat{\mathbb{X}} = (\hat{X}, \hat{d}_{\hat{X}}, \hat{\mu})$, $\hat{\mathbb{Y}} = (\hat{Y}, \hat{d}_{\hat{Y}}, \hat{\nu})$ in a similar way to that of Proposition 3.3: We slightly modify the definition of $\hat{\sigma}, \hat{\mu}, \hat{\nu}$ as follows:

$$\begin{cases} \hat{\sigma} & = \sigma + (B - |\sigma|)\delta_{\hat{\infty}}, \\ \hat{\mu} & = \mu + (B - |\mu|)\delta_{\hat{\infty}}, \\ \hat{\nu} & = \nu + (B - |\nu|)\delta_{\hat{\infty}}, \end{cases} \tag{49}$$

where $B = |\sigma| + |\mu| + |\nu|$. Thus, $|\hat{\nu}| = |\hat{\mu}| = |\hat{\nu}| = B$. (For a similar idea in unbalanced optimal transport see, for example, (Heinemann et al., 2023).) The mapping (8) is modified as

$$\Gamma_{\leq}(\mu, \nu) \ni \gamma \mapsto \hat{\gamma} := \gamma + (\mu - \gamma_1) \otimes \delta_{\hat{\infty}} + \delta_{\hat{\infty}} \otimes (\nu - \gamma_2) + (|\gamma| + B - |\mu| - |\nu|)\delta_{\hat{\infty}, \hat{\infty}} \in \Gamma(\hat{\mu}, \hat{\nu}) \tag{50}$$

which is still a well-defined bijection by [Proposition B.1. (Bai et al., 2023)].

We define the following mapping $\tilde{D} : (\mathbb{R} \cup \{\infty\}) \times (\mathbb{R} \cup \{\infty\}) \to \mathbb{R}_+$:

$$\tilde{D}^2(r_1, r_2) = \begin{cases} |r_1 - r_2|^2 \wedge 2\lambda & \text{if } r_1, r_2 < \infty, \\ \lambda & \text{if } r_1 = \infty, r_2 < \infty \text{ or vice versa}, \\ 0 & \text{if } r_1 = r_2 = \infty. \end{cases} \tag{51}$$

It is straightforward to verify that $\tilde{D}$ defines a metric in $\mathbb{R} \cup \{\infty\}$. Then the following defines a *generalized GW problem*:

$$GW_g(\mathbb{X}, \mathbb{Y}) = \inf_{\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})} \underbrace{\int_{(\hat{X} \times \hat{Y})^2} \tilde{D}^2(\hat{d}_{\hat{X}}(x, x'), \hat{d}_{\hat{Y}}(y, y'))\, d\hat{\gamma}^{\otimes 2}}_{\tilde{C}_G}. \tag{52}$$

Similarly, we define $GW_g(\mathbb{X}, \mathbb{S})$, and $GW_g(\mathbb{S}, \mathbb{Y})$. For each $\gamma \in \Gamma_{\leq}(\mu, \nu)$, define $\hat{\gamma}$ by (50). Then,

$$\begin{aligned}
&\tilde{C}_G(\hat{\gamma}) = \int_{(\hat{X} \times \hat{Y})^2} \tilde{D}^2(\hat{d}_{\hat{X}}(x, x'), \hat{d}_{\hat{y}}(x, x'))d\hat{\gamma}^{\otimes 2} \\
&= \left[\int_{(X \times Y)^2} |d_X(x, x') - d_Y(y, y')|^2 \wedge 2\lambda d\hat{\gamma}^{\otimes 2}\right] \\
&+ \left[2\int_{(\{\hat{\infty}\} \times Y) \times (X \times Y)} \lambda d\hat{\gamma}^{\otimes 2} + \int_{(\{\hat{\infty}\} \times Y)^2} \lambda d\hat{\gamma}^{\otimes 2}\right] + \left[2\int_{(X \times \{\hat{\infty}\}) \times (X \times Y)} \lambda d\hat{\gamma}^{\otimes 2} + \int_{(X \times \{\hat{\infty}\})^2} \lambda \hat{\gamma}^{\otimes 2}\right] \\
&+ \left[2\int_{(\{\hat{\infty}\} \times Y) \times (X \times \{\hat{\infty}\})} \tilde{D}^2(\infty, \infty)d\tilde{\gamma}^{\otimes 2} + 2\int_{(\{\hat{\infty}\} \times \{\hat{\infty}\}) \times (X \times Y)} \tilde{D}(\infty, \infty)d\hat{\gamma}^{\otimes 2} + \int_{(\{\hat{\infty}\} \times \{\hat{\infty}\})^2} \tilde{D}(\infty, \infty)d\hat{\gamma}^{\otimes 2}\right] \\
&+ 2\left[\int_{(\{\hat{\infty}\} \times \{Y\}) \times \{\hat{\infty}\}^2} \tilde{D}(\infty, \infty)d\hat{\gamma}^{\otimes 2}\right] + 2\left[\int_{(X \times \{\hat{\infty}\}) \times \{\hat{\infty}\}^2} \tilde{D}(\infty, \infty)d\hat{\gamma}^{\otimes 2}\right] \\
&= \left[\int_{(X \times Y)^2} |d_X(x, x') - d_Y(y, y')|^2 \wedge 2\lambda d\gamma^{\otimes 2})\right] \\
&+ \left[2\lambda(|\nu| - |\gamma|)(|\gamma|) + \lambda(|\nu| - |\gamma|)^2\right] + \left[2\lambda(|\mu| - |\gamma|)(|\gamma|) + \lambda(|\mu| - |\gamma|)^2\right] + 0 \\
&= \left[\int_{(X \times Y)^2} |d_X(x, y') - d_Y(y, y')|^2 \wedge 2\lambda\, d\gamma^{\otimes 2})\right] + \lambda(|\nu^2| + |\mu|^2 - 2|\gamma|^2) = C(\gamma; \lambda)
\end{aligned}$$

since the mapping $\gamma \mapsto \hat{\gamma}$ defined in (50) is a bijection. Then, $\gamma \in \Gamma_{\leq}(\mu, \nu)$ is optimal for partial GW problem (17) if and only if $\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})$ is optimal for generalized GW problem (52). Thus, we have

$$GW_g(\hat{\mathbb{X}}, \hat{\mathbb{Y}}) = PGW_\lambda(\mathbb{X}, \mathbb{Y}).$$

Similarly, $GW_g(\hat{\mathbb{X}},\hat{\mathbb{S}}) = PGW_\lambda(\mathbb{X},\mathbb{S})$, and $GW_g(\hat{\mathbb{S}},\hat{\mathbb{Y}}) = PGW_\lambda(\mathbb{S},\mathbb{Y})$.

In addition, $(GW_g(\cdot,\cdot))^{1/p}$ satisfies the triangle inequality (see Lemma F.1 below for its proof). Therefore, we have

$$(PGW_\lambda(\mathbb{X},\mathbb{Y}))^{1/p} \leq (PGW_\lambda(\mathbb{X},\mathbb{S}))^{1/p} + (PGW_\lambda(\mathbb{S},\mathbb{Y}))^{1/p}.$$

$\square$

Lemma F.1. *Consider the generalized GW problem (52) for three give mm-spaces $\mathbb{S},\mathbb{X},\mathbb{Y}$. Then, for any $p\in[1,\infty)$, we have $(GW_g(\cdot,\cdot))^{1/p}$ satisfies the triangle inequality*

$$(GW_g(\mathbb{X},\mathbb{Y}))^{1/p} \leq (GW_g(\mathbb{X},\mathbb{S}))^{1/p} + (GW_g(\mathbb{S},\mathbb{Y}))^{1/p}.$$

Proof. We provide the proof for case $p=2$. For general case $p\geq 1$, it can be derived similarly.

First, we notice that as a by-product of proof of Proposition 3.5 and Proposition 3.2, we have that there exists a minimizer for (52).

Now, we proceed as in the classical proof for checking the triangle inequality for the Wasserstein distance (see Lemmas 5.4 in (Santambrogio, 2015)). Indeed, we will use the approach based on *disintegration of measures*.

The spaces $(\hat{S},\hat{\sigma})$, $(\hat{X},\hat{\mu})$, and $(\hat{Y},\hat{\nu})$ are measure spaces. Consider $\hat{\gamma}_{X,S}\in\Gamma(\hat{\mu},\hat{\sigma})$ and $\hat{\gamma}_{S,Y}\in\Gamma(\hat{\sigma},\hat{\nu})$ to be optimal for $GW_g(\mathbb{X},\mathbb{S})$ and $GW_g(\mathbb{S},\mathbb{Y})$, respectively. By disintegration of measures, there exists a measure $\sigma\in\mathcal{P}(\hat{X}\times\hat{S}\times\hat{Y})$ such that $(\pi_{X,S})_\#\sigma = \hat{\gamma}_{X,S}$ and $(\pi_{S,Y})_\#\sigma=\hat{\gamma}_{S,Y}$, where $\pi_{X,S}:\hat{X}\times\hat{S}\times\hat{Y}\to\hat{X}\times\hat{S}$, $\pi_{X,S}(x,s,y)=(x,s)$ denotes the projection on the first two variables, and $\pi_{S,Y}:\hat{X}\times\hat{S}\times\hat{Y}\to\hat{S}\times\hat{Y}$, $\pi_{S,Y}(x,s,y)=(s,y)$ denotes the projection on the last two variables (see Lemma 5.5 (Santambrogio, 2015) -called as the *gluing lemma*-). Now, let us define $\hat{\gamma}:=(\pi_{X,Y})_\#\sigma$ ($\pi_{X,Y}:\hat{X}\times\hat{S}\times\hat{Y}\to\hat{X}\times\hat{Y}$, $\pi_{S,Y}(x,s,y)=(x,y)$ denotes the projection on the first and last variables). By composition of projections, it holds that $\hat{\gamma}$ has first and second marginals $\hat{\mu}$ and $\hat{\nu}$, respectively, and so $\hat{\gamma}\in\Gamma(\hat{\mu},\hat{\nu})$. Moreover, since $\tilde{D}(\cdot,\cdot)$ given by (51) defines a metric in $\mathbb{R}\cup\{\infty\}$, we have

$$\begin{aligned}
(GW_g(\mathbb{X},\mathbb{Y}))^{1/2} &\leq \left(\int_{(\hat{X}\times\hat{Y})^2}\tilde{D}^2(\hat{d}_{\hat{X}}(x,x'),\hat{d}_{\hat{Y}}(y,y'))\,d\hat{\gamma}^{\otimes 2}\right)^{1/2}\\
&= \left(\int_{(\hat{X}\times\hat{S}\times\hat{Y})^2}\tilde{D}^2(\hat{d}_{\hat{X}}(x,x'),\hat{d}_{\hat{Y}}(y,y'))\,d\sigma^{\otimes 2}\right)^{1/2}\\
&\leq \left(\int_{(\hat{X}\times\hat{S}\times\hat{Y})^2}|\tilde{D}(\hat{d}_{\hat{X}}(x,x'),\hat{d}_{\hat{S}}(s,s'))-\tilde{D}(\hat{d}_{\hat{S}}(s,s'),\hat{d}_{\hat{Y}}(y,y'))|^2\,d\sigma^{\otimes 2}\right)^{1/2}\\
&\leq \left(\int_{(\hat{X}\times\hat{S}\times\hat{Y})^2}\tilde{D}^2(\hat{d}_{\hat{X}}(x,x'),\hat{d}_{\hat{S}}(s,s'))\,d\sigma^{\otimes 2}\right)^{1/2} + \left(\int_{(\hat{X}\times\hat{S}\times\hat{Y})^2}\tilde{D}^2(\hat{d}_{\hat{S}}(s,s'),\hat{d}_{\hat{Y}}(y,y'))\,d\sigma^{\otimes 2}\right)^{1/2}\\
&= \left(\int_{(\hat{X}\times\hat{S}\times)^2}\tilde{D}^2(\hat{d}_{\hat{X}}(x,x'),\hat{d}_{\hat{S}}(s,s'))\,d\hat{\gamma}_{X,S}^{\otimes 2}\right)^{1/2} + \left(\int_{(\hat{S}\times\hat{Y})^2}\tilde{D}^2(\hat{d}_{\hat{S}}(s,s'),\hat{d}_{\hat{Y}}(y,y'))\,d\hat{\gamma}_{S,Y}^{\otimes 2}\right)^{1/2}\\
&= (GW_g(\mathbb{X},\mathbb{S}))^{1/2} + (GW_g(\mathbb{S},\mathbb{Y}))^{1/2}
\end{aligned}$$

where in the third inequality we used Minkowski inequality in $L^2((\hat{X}\times\hat{S}\times\hat{Y})^2,\sigma^{\otimes 2})$. $\square$

G. Tensor product computation

Lemma G.1. *Given a tensor $M\in\mathbb{R}^{n\times m\times n\times n}$ and $\gamma,\gamma'\in\mathbb{R}^{n\times m}$, the tensor product operator $M\circ\gamma$ satisfies the following:*

(i) The mapping $\gamma\mapsto M\circ\gamma$ is linear with respect to γ.

(ii) If M is symmetric, in particular, $M_{i,j,i',j'} = M_{i',j',i,j}, \forall i, i' \in [1:n], j, j' \in [1:m]$, then

$$\langle M \circ \gamma, \gamma' \rangle_F = \langle M \circ \gamma', \gamma \rangle_F.$$

Proof. (i) For the first part, consider $\gamma, \gamma' \in \mathbb{R}^{n\times m}$ and $k \in \mathbb{R}$. For each $i, j \in [1:n] \times [1:m]$, we have we have

$$\begin{aligned}
(M \circ (\gamma + \gamma'))_{ij} &= \sum_{i',j'} M_{i,j,i',j'}(\gamma + \gamma')_{i'j'} \\
&= \sum_{i',j'} M_{i,j,i',j'}\gamma_{i'j'} + \sum_{i',j'} M_{i,j,i',j'}\gamma'_{i'j'} \\
&= (M \circ \gamma)_{ij} + (M \circ \gamma)_{i'j'}, \\
(M \circ (k\gamma))_{ij} &= \sum_{i',j'} M_{i,j,i',j'}(k\gamma)_{ij} \\
&= k \sum_{i',j'} M_{i,j,i',j'}\gamma_{ij} \\
&= k(M \circ \gamma)_{ij}.
\end{aligned} \tag{53}$$

Thus, $M \circ (\gamma + \gamma') = M \circ \gamma + M \circ \gamma'$ and $M \circ (k\gamma) = kM \circ \gamma$. Therefore, $\gamma \mapsto M \circ \gamma$ is linear.

(ii) For the second part, we have

$$\begin{aligned}
\langle M \circ \gamma, \gamma' \rangle_F &= \sum_{iji'j'} M_{i,j,i',j'},\gamma_{ij}\gamma'_{i'j'} \\
&= \sum_{i,j,i',j'} M_{i',j',i,j}\gamma_{i',j'}\gamma_{i,j} \qquad (54) \\
&= \langle M\gamma', \gamma \rangle
\end{aligned}$$

where (54) follows from the fact that M is symmetric.

□

H. Gradient computation in Algorithm 1 and 2

In this section, we discuss the computation of Gradient $\nabla \mathcal{L}_{\mathcal{M}}(\gamma)$ in Algorithm 1 and $\nabla \mathcal{L}_{\hat{M}}(\hat{\gamma})$ in Algorithm 2.

First, in the setting of algorithm 1, for each $\gamma \in \mathbb{R}^{n\times m}$, we have

$$\nabla \mathcal{L}_{\tilde{M}}(\gamma) = 2\tilde{M} \circ \gamma \tag{55}$$

In lemma 4.2, $\tilde{M} \circ \gamma$ is given by

$$\tilde{M} \circ \gamma = M\tilde{\gamma} - 2\lambda|\gamma|1_{n,m}.$$

We provide the proof as the following.

Proof. For any $\gamma \in \mathbb{R}^{n\times m}$, we have

$$\begin{aligned}
\tilde{M} \circ \gamma &= (M - 2\lambda 1_{n,m}) \circ \gamma \\
&= M \circ \gamma - 2\lambda 1_{n,m,n,m} \circ \gamma \\
&= M \circ \gamma - 2(\langle 1_{n,m}, \gamma \rangle_F)1_{n,m} \\
&= M \circ \gamma - 2\lambda|\gamma|1_{n,m}
\end{aligned}$$

where the second equality follows from lemma G.1. □

Next, in setting of algorithm 2, for any $\hat{\gamma} \in \mathbb{R}^{(n+1)\times(m+1)}$, we have

$$\nabla \mathcal{L}_{\hat{M}}(\hat{\gamma}) = 2\hat{M} \circ \hat{\gamma} \tag{56}$$

$\hat{M} \circ \hat{\gamma}$ can be computed by the following lemma.

Lemma H.1. *For each $\hat{\gamma} \in \mathbb{R}^{(n+1)\times(m+1)}$, we have $\hat{M} \circ \hat{\gamma} \in \mathbb{R}^{(n+1)\times(m+1)}$ with the following:*

$$(\hat{M} \circ \hat{\gamma})_{ij} = \begin{cases} (\tilde{M} \circ \hat{\gamma}[1:n, 1:m])_{ij} & \text{if } i \in [1:n], j \in [1:m] \\ 0 & \text{elsewhere} \end{cases}. \tag{57}$$

Proof. Recall the definition of $\hat{M}$ is given by (26), choose $i \in [1:n], j \in [1:m]$, we have

$$\begin{aligned}
(\hat{M} \circ \hat{\gamma})_{ij} &= \sum_{i'=1}^{n}\sum_{j'=1}^{m} \hat{M}_{i,j,i',j'}\hat{\gamma}_{i',j'} + \sum_{j'=1}^{m} \hat{M}_{i,j,n+1,j}\hat{\gamma}_{n+1,j'} + \sum_{i'=1}^{n} \hat{M}_{i,j,i',m+1}\hat{\gamma}_{i,m+1} + \hat{M}_{i,j,n+1,m+1}\hat{\gamma}_{n+1,m+1} \\
&= \hat{M}_{i,j,i',j'}\hat{\gamma}_{i',j'} + 0 + 0 + 0 \\
&= \tilde{M}_{i,j,i',j'}\hat{\gamma}_{i',j'} \\
&= \tilde{M} \circ (\hat{\gamma}[1:n, 1:m])
\end{aligned}$$

If $i = n+1$, we have

$$(\hat{M} \circ \hat{\gamma})_{n+1,j} = \sum_{i'=1}^{n+1}\sum_{j'=1}^{m+1} \hat{M}_{n+1,j,i',j'}\hat{\gamma}_{i',j'} = 0$$

Similarly, $(\hat{M} \circ \hat{\gamma})_{i,m+1} = 0$. Thus we complete the proof. □

I. Line search in Algorithm 1

In this section, we discuss the derivation of the line search algorithm.

We observe that in the partial GW setting, for each $\gamma \in \Gamma_{\leq}(\mu, \nu)$, the marginals of γ are not fixed. Thus, we can not directly apply the classical algorithm (e.g. (Titouan et al., 2019)).

In iteration k, let $\gamma^{(k)}, \gamma^{(k)'}$ be the previous and new transportation plans from step 1 of the algorithm. For convenience, we denote them as γ, γ', respectively.

The goal is to solve the following problem:

$$\min_{\alpha \in [0,1]} \mathcal{L}(\tilde{M}, (1-\alpha)\gamma + \alpha\gamma') \tag{58}$$

where $\mathcal{L}(\tilde{M}, \gamma) = \langle \tilde{M} \circ \gamma, \gamma \rangle_F$. By denoting $\delta\gamma = \gamma' - \gamma$, we have

$$\mathcal{L}(\tilde{M}, (1-\alpha)\gamma + \alpha\gamma') = \mathcal{L}(\tilde{M}, \gamma + \alpha\delta\gamma).$$

Then,

$$\langle \tilde{M} \circ (\gamma + \alpha\delta\gamma), (\gamma + \alpha\delta\gamma) \rangle_F = \langle \tilde{M} \circ \gamma, \gamma \rangle_F + \alpha \left(\langle \tilde{M} \circ \gamma, \delta\gamma \rangle_F + \langle \tilde{M} \circ \delta\gamma, \gamma \rangle_F \right) + \alpha^2 \langle \tilde{M} \circ \delta\gamma, \delta\gamma \rangle_F$$

Let

$$\begin{aligned}
a &= \langle \tilde{M} \circ \delta\gamma, \delta\gamma \rangle_F, \\
b &= \langle \tilde{M} \circ \gamma, \delta\gamma \rangle_F + \langle \tilde{M} \circ \delta\gamma, \gamma \rangle_F = 2\langle \tilde{M} \circ \gamma, \delta\gamma \rangle_F, \\
c &= \langle \tilde{M} \circ \gamma, \gamma \rangle_F,
\end{aligned} \tag{59}$$

where the second identity in (59) follows from Lemma G.1 and the fact that $\tilde{M} = M - 2\lambda 1_{n,m,n',m'}$ is symmetric.

Therefore, the above problem (58) becomes

$$\min_{\alpha\in[0,1]} a\alpha^2 + b\alpha + c.$$

The solution is the following:

$$\alpha^* = \begin{cases} 1 & \text{if } a \leq 0, a+b \leq 0, \\ 0 & \text{if } a \leq 0, a+b > 0, \\ \text{clip}(\frac{-b}{2a}, [0,1]) & \text{if } a > 0, \end{cases} \tag{60}$$

where

$$\text{clip}(\frac{-b}{2a}, [0,1]) = \min\left\{1, \max\{0, \frac{-b}{2a}\}\right\} = \begin{cases} \frac{-b}{2a} & \text{if } \frac{-b}{2a} \in [0,1], \\ 0 & \text{if } \frac{-b}{2a} < 0, \\ 1 & \text{if } \frac{-b}{2a} > 1. \end{cases}$$

It remains to discuss the computation of a and b. If the assumption in Proposition 4.1 holds, by (29) and (30), we have

$$\begin{aligned} a &= \langle \tilde{M} \circ \delta\gamma, \delta\gamma \rangle_F \\ &= \langle (M \circ \delta\gamma - 2\lambda|\delta\gamma| I_{n,m}), \delta\gamma \rangle_F \\ &= \langle M \circ \delta\gamma, \delta\gamma \rangle_F - 2\lambda|\delta\gamma|^2 \\ &= \left\langle u(C^X, C^Y, \delta\gamma) - h_1(C^X)\delta\gamma h_2(C^Y)^T, \delta\gamma \right\rangle_F - 2\lambda|\delta\gamma|^2, \end{aligned} \tag{61}$$

$$\begin{aligned} b &= 2\langle \tilde{M} \circ \gamma, \delta\gamma \rangle_F \\ &= 2\langle M \circ \gamma - 2\lambda|\gamma| I_{n,m}, \delta\gamma \rangle \\ &= 2(\langle M \circ \gamma, \delta\gamma \rangle_F - 2\lambda|\delta\gamma||\gamma|) \end{aligned} \tag{62}$$

Note that in the classical GW setting (Titouan et al., 2019), the term $u(C^X, C^Y, \delta\gamma) = 0_{n\times m}$ and $|\delta\gamma| = 0$. Therefore, in such line search algorithm (Algorithm 2 in (Titouan et al., 2019)), the terms $u(C^X, C^Y, \delta\gamma), 2\lambda|\delta\gamma|1_{n\times m}$ are not required. (61),(62) are applied in our numerical implementation. In addition, in equation (62), $M \circ \gamma, 2\lambda|\gamma|$ have been computed in the gradient computation step, thus these two terms can be directly applied in this step.

J. Line search in Algorithm 2

Similar to the previous section, in iteration k, let $\hat{\gamma}^{(k)}, \hat{\gamma}^{(k)'}$ denote the previous transportation plan and the updated transportation plan. For convenience, we denote them as $\hat{\gamma}, \hat{\gamma}'$, respectively.

Let $\delta\hat{\gamma} = \hat{\gamma} - \hat{\gamma}'$.

The goal is to find the following optimal α:

$$\alpha = \arg\min_{\alpha\in[0,1]} \mathcal{L}(\hat{M}, (1-\alpha)\hat{\gamma}, \alpha\hat{\gamma}') = \arg\min_{\alpha\in[0,1]} \mathcal{L}(\hat{M}, \alpha\delta\hat{\gamma} + \hat{\gamma}), \tag{63}$$

where $\hat{M} \in \mathbb{R}^{(n+1)\times(m+1)\times(n+1)\times(m+1)}$, with $\hat{M}[1:n, 1:m, 1:n, 1:m] = \tilde{M} = M - 2\lambda 1_{n\times m\times n\times m}$.

Similar to the previous section, let

$$\begin{aligned} a &= \langle \hat{M} \circ \delta\hat{\gamma}, \delta\hat{\gamma} \rangle_F, \\ b &= \langle \hat{M} \circ \delta\hat{\gamma}, \hat{\gamma} \rangle_F + \langle \hat{M} \circ \hat{\gamma}, \delta\hat{\gamma} \rangle_F = 2\langle \hat{M} \circ \delta\hat{\gamma}, \hat{\gamma} \rangle_F, \\ c &= \langle \hat{M} \circ \hat{\gamma}, \hat{\gamma} \rangle_F, \end{aligned} \tag{64}$$

where (64) holds since $\hat{M}$ is symmetric. Then, the optimal α is given by (60).

It remains to discuss the computation. By Lemma G.1, we set $\gamma = \hat{\gamma}[1:n, 1:m], \delta\gamma = \delta\hat{\gamma}[1:n, 1:m]$. Then,

$$a = \langle (\hat{M} \circ \delta\hat{\gamma})[1:n, 1:m], \delta\gamma \rangle_F = \langle (\tilde{M} \circ \delta\gamma, \delta\gamma \rangle_F,$$
$$b = \langle (\hat{M} \circ \delta\hat{\gamma})[1:n, 1:m], \gamma \rangle_F = \langle (\tilde{M} \circ \delta\gamma, \gamma \rangle_F.$$

Thus, we can apply (61), (62) to compute a, b in this setting by plugging in $\gamma = \hat{\gamma}[1:n, 1:m]$ and $\delta\gamma = \delta\hat{\gamma}[1:n, 1:m]$.

K. Convergence

As in (Chapel et al., 2020) we will use the results from (Lacoste-Julien, 2016) on the convergence of the Frank-Wolfe algorithm for non-convex objective functions.

Consider the minimization problems

$$\min_{\gamma \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})} \mathcal{L}_{\tilde{M}}(\gamma) \qquad \text{and} \qquad \min_{\hat{\gamma} \in \Gamma(\hat{\mathrm{p}},\hat{\mathrm{q}})} \mathcal{L}_{\hat{M}}(\hat{\gamma}) \tag{65}$$

that corresponds to the discrete partial GW problem, and the discrete GW-variant problem (used in version 2), respectively. The objective functions $\gamma \mapsto \mathcal{L}_{\tilde{M}}(\gamma) = \tilde{M}\gamma^{\otimes 2}$ (where $\tilde{M} = M - 2\lambda 1_{n,m}$ for a fixed matrix $M \in \mathbb{R}^{n\times m}$ and $\lambda > 0$), and $\hat{\gamma} \mapsto \mathcal{L}_{\hat{M}}(\hat{\gamma}) = \hat{M}\hat{\gamma}^{\otimes 2}$ (where $\hat{M}$ is given by (26)) are non-convex in general (for $\lambda > 0$, the matrices $\tilde{M}$ and $\hat{M}$ symmetric but not positive semi-definite), but the constraint sets $\Gamma_{\leq}(\mathrm{p},\mathrm{q})$ and $\Gamma(\hat{\mathrm{p}},\hat{\mathrm{q}})$ are convex and compact on $\mathbb{R}^{n\times m}$ (see Proposition B.2 (Liu et al., 2023)) and on $\mathbb{R}^{(n+1)\times(m+1)}$, respectively.

From now on we will concentrate on the first minimization problem in (65) and the convergence analysis for the second one will be analogous.

Consider the *Frank-Wolfe gap* of $\mathcal{L}_{\tilde{M}}$ at the approximation $\gamma^{(k)}$ of the optimal plan γ:

$$g_k = \min_{\gamma \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})} \langle \nabla \mathcal{L}_{\tilde{M}}(\gamma^{(k)}), \gamma^{(k)} - \gamma \rangle_F. \tag{66}$$

It provided a good criterion to measure the distance to a stationary point at iteration k. Indeed, a plan $\gamma^{(k)}$ is a stationary transportation plan for the corresponding constrained optimization problem in (65) if and only if $g_k = 0$. Moreover, g_k is always non-negative ($g_k \geq 0$).

From Theorem 1 in (Lacoste-Julien, 2016), after K iterations we have the following upper bound for the minimal Frank-Wolf gap:

$$\tilde{g}_K := \min_{1\leq k\leq K} g_k \leq \frac{\max\{2L_1, D_L\}}{\sqrt{K}}, \tag{67}$$

where

$$L_1 := \mathcal{L}_{\tilde{M}}(\gamma^{(1)}) - \min_{\gamma \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})} \mathcal{L}_{\tilde{M}}(\gamma)$$

is the initial global suboptimal bound for the initialization $\gamma^{(1)}$ of the algorithm, and $D_L := \mathrm{Lip} \cdot (\mathrm{diam}(\Gamma_{\leq}(\mathrm{p},\mathrm{q})))^2$, where Lip is the Lipschitz constant of $\nabla \mathcal{L}_{\tilde{M}}$ and $\mathrm{diam}(\Gamma_{\leq}(\mathrm{p},\mathrm{q}))$ is the $\|\cdot\|_F$ diameter of $\Gamma_{\leq}(\mathrm{p},\mathrm{q})$ in $\mathbb{R}^{n\times m}$.

The important thing to notice is that the constant $\max\{2L_1, D_L\}$ does not depend on the iteration step k. Thus, according to Theorem 1 in (Lacoste-Julien, 2016), the rate on $\tilde{g}_K$ is $\mathcal{O}(1/\sqrt{K})$. That is, the algorithm takes at most $\mathcal{O}(1/\varepsilon^2)$ iterations to find an approximate stationary point with a gap smaller than ε.

Finally, we adapt Lemma 1 in Appendix B.2 in (Chapel et al., 2020) to our case characterizing the convergence guarantee, precisely, determining such a constant $\max\{2L_1, D_L\}$ in (67). Essentially, we will estimate upper bounds for the Lipschitz constant Lip and for the diameter $\mathrm{diam}(\Gamma_{\leq}(\mathrm{p},\mathrm{q}))$.

- Let us start by considering the diameter of the couplings of $\Gamma_{\leq}(\mathrm{p},\mathrm{q})$ with respect to the Frobenious norm $\|\cdot\|_F$. By definition,

$$\mathrm{diam}(\Gamma_{\leq}(\mathrm{p},\mathrm{q})) := \sup_{\gamma,\gamma' \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})} \|\gamma - \gamma'\|_F.$$

For any $\gamma \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})$, since $\gamma_1 \leq \mathrm{p}$ and $\gamma_2 \leq \mathrm{q}$, we obtain that, in particular, $|\gamma_1| \leq |\mathrm{p}|$ and $|\gamma_2| \leq |\mathrm{q}|$. Thus, since $|\gamma_1| = |\gamma| = |\gamma_2|$ (recall that $\gamma_1 = \pi_{1\#}\gamma$ and $\gamma_2 = \pi_{2\#}\gamma$) we have

$$|\gamma| \leq \min\{|\mathrm{p}|, |\mathrm{q}|\} =: \sqrt{s} \qquad \forall \gamma \in \Gamma_{\leq}(\mathrm{p},\mathrm{q}).$$

Thus, given $\gamma, \gamma' \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})$, we obtain

$$\begin{aligned}\|\gamma - \gamma'\|_F^2 &\leq 2\|\gamma\|_F^2 + 2\|\gamma'\|_F^2 = 2\sum_{i,j}(\gamma_{i,j})^2 + 2\sum_{i,j}(\gamma'_{i,j})^2 \\ &\leq 2\left(\sum_{i,j}|\gamma_{i,j}|\right)^2 + 2\left(\sum_{i,j}|\gamma'_{i,j}|\right)^2 = 2|\gamma|^2 + 2|\gamma'|^2 \leq 4s\end{aligned}$$

(essentially, we used that $\|\cdot\|_F$ is the 2-norm for matrices viewed as vectors, that $|\cdot|$ is the 1-norm for matrices viewed as vectors, and the fact that $\|\cdot\|_2 \leq \|\cdot\|_1$). As a result,

$$\mathrm{diam}(\Gamma_{\leq}(\mathrm{p},\mathrm{q})) \leq 2\sqrt{s}, \tag{68}$$

where s only depends on p and q that are fixed weight vectors in $\mathbb{R}_+^n$ and $\mathbb{R}_+^m$, respectively.

- Now, let us analyze the Lipschitz constant of $\nabla\mathcal{L}_{\tilde{M}}$ with respect to $\|\cdot\|_F$. For any $\gamma, \gamma' \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})$ we have,

$$\begin{aligned}\|\nabla\mathcal{L}_{\tilde{M}}(\gamma) - \nabla\mathcal{L}_{\tilde{M}}(\gamma')\|_F^2 &= \|\tilde{M} \circ \gamma - \tilde{M} \circ \gamma'\|_F^2 \\ &= \|[M - 2\lambda 1_{n,m}] \circ (\gamma - \gamma')\|_F^2 = \langle [M - 2\lambda 1_{n,m}] \circ (\gamma - \gamma'), [M - 2\lambda 1_{n,m}] \circ (\gamma - \gamma')\rangle_F \\ &= \sum_{i,j}\left([M - 2\lambda 1_{n,m}] \circ (\gamma - \gamma')_{i,j}\right)^2 \\ &= \sum_{i,j}\left(\sum_{i',j'}(M_{i,j,i',j'} - 2\lambda)(\gamma_{i',j'} - \gamma'_{i',j'})\right)^2 \\ &\leq \left(\max_{i,j,i',j'}\{M_{i,j,i',j'} - 2\lambda\}\right)^2 \left(\sum_{i,j}^{n,m}\left(\sum_{i',j'}^{n,m}(\gamma_{i',j'} - \gamma'_{i',j'})\right)^2\right) \\ &= \left(\max_{i,j,i',j'}\{M_{i,j,i',j'}\} - 2\lambda\right)^2 \left(\sum_{i,j}^{n,m}\|\gamma - \gamma'\|_F^2\right) \\ &\leq nm\left(\max_{i,j,i',j'}\{M_{i,j,i',j'}\} - 2\lambda\right)^2 \|\gamma - \gamma'\|_F^2.\end{aligned}$$

Hence, the Lipschitz constant of the gradient of $\mathcal{L}_{\tilde{M}}$ is bounded by

$$\mathrm{Lip} \leq \sqrt{nm}\left|\max_{i,j,i',j'}\{M_{i,j,i',j'}\} - 2\lambda\right|.$$

In the particular case where $L(r_1, r_2) = |r_1 - r_2|^2$ we have $M_{i,j,i',j'} = |C^X_{i,i'} - C^Y_{j,j'}|^2$ (as in (21)) where C^X, C^Y are given $n \times n$ and $m \times m$ non-negative symmetric matrices defined in (18), that depend on the given discrete mm-spaces $\mathbb{X}$ and $\mathbb{Y}$. Here, we obtain

$$\max_{i,j,i',j'}\{M_{i,j,i',j'}\} = \max_{i,j,i',j'}\{|C^X_{i,i'} - C^Y_{j,j'}|^2\} \leq \frac{1}{2}\left((\max_{i,i'}\{C^X_{i,i'}\})^2 + (\max_{j,j'}\{C^Y_{j,j'}\})^2\right)$$

and so the Lipschitz constant verifies

$$\mathrm{Lip} \leq \sqrt{nm}\left|\frac{(\max_{i,i'}\{C^X_{i,i'}\})^2 + (\max_{j,j'}\{C^Y_{j,j'}\})^2}{2} - 2\lambda\right|$$

Combining all together, we obtain that after K iterations, the minimal Frank-Wolf gap verifies

$$\tilde{g}_K = \min_{1\leq k\leq K} g_k \leq \frac{\max\{2L_1, 4s\sqrt{nm}\,|\max_{i,j,i',j'}\{M_{i,j,i',j'}\} - 2\lambda|\}}{\sqrt{K}}$$

$$\leq 2\frac{\max\{L_1, s\sqrt{nm}\,\left|(\max_{i,i'}\{C^X_{i,i'}\})^2 + (\max_{j,j'}\{C^Y_{j,j'}\})^2 - 4\lambda\right|\}}{\sqrt{K}} \qquad \text{(if } M \text{ is as in (21))}$$

where L_1 dependents on the initialization of the algorithm.

Finally, we mention that there is a dependence in the constant $\max\{2L_1, D_L\}$ on the number of points (n and m) of our discrete spaces $X = \{x_1, \ldots x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ which was not pointed out in (Chapel et al., 2020).

L. Numerical Details of PU learning experiment

Wall-clock time comparison. In this experiment, to prevent unexpected convergence to local minima in the Frank-Wolf algorithms, we manually set $\alpha = 1$ during the line search step for both primal-PGW and PGW methods.

For the convergence criteria, we set the tolerance term for Frank-Wolfe convergence and the main loop in the UGW algorithm to be $1e-5$. Additionally, the tolerance for Sinkhorn convergence in UGW was set to $1e-6$. The maximum number of iterations for the POT solver in PGW and primal-PGW was set to $500n$.

Regarding data types, we used 64-bit floating-point numbers for primal-PGW and PGW, and 32-bit floating-point numbers for UGW.

For the MNIST and EMNIST datasets, we set $n = 1000$ and $m = 5000$. In the Surf(A) and Decaf(A) datasets, each class contained an average of 100 samples. To ensure the SCAR assumption, we set $n = 1/2 * 100 = 50$ and $m = 250$. Similarly, for the Surf(D) and Decaf(D) datasets, we set $n = 15$ and $m = 75$. Finally, for Surf(W) and Decaf(W), we used $n = 20$ and $m = 100$.

In Table 2, we provide a comparison of wall-clock times for the MNIST and EMNIST datasets.

Source	Target	Init	Pr-PGW	UGW	PGW
M(1000)	M(5000)	POT, 6.08	0.64	154.87	0.71
M(1000)	M(5000)	FLB-U, 0.02	13.73	157.31	14.79
M(1000)	M(5000)	FLB-P, 0.60	23.86	171.16	31.17
M(1000)	EM(5000)	FLB-U, 0.03	20.43	167.07	24.08
M(1000)	EM(5000)	FLB-P, 0.70	26.98	169.87	32.46
EM(1000)	M(5000)	FLB-U, 0.03	23.14	152.43	22.44
EM(1000)	M(5000)	FLB-P, 0.61	26.04	160.33	29.14
EM(1000)	EM(5000)	POT, 5.67	0.54	156.40	0.68
EM(1000)	EM(5000)	FLB-U, 0.04	14.90	179.55	15.03
EM(1000)	EM(5000)	FLB-P, 0.57	12.20	173.56	15.20

Table 2. In this table, we present the wall-clock time for three initialization method: POT, FLB-UOT, FLB-POT and three GW-methods, Primal-PGW, UGW, and PGW. In the "Source" (or "Target") columm, M (or EM) denotes the MNIST (or EMNIST) dataset, the value 1000 (or 5000) denote the sample size of X (or Y). In the "Init" Colume, the first entry is initilization method, the second entry is its corresponding wall-clock time. The unit of wall-clock time is second.

DATASET	INIT	PR-PGW	UGW	PGW
SURF(A) → SURF(A)	POT, 100%	100%	65%	100%
SURF(A) → SURF(A)	FLB-U, 69%	83%	65%	83%
SURF(A) → SURF(A)	FLB-P, 67%	81%	65%	81%
DECAF(A) → DECAF(A)	POT, 100%	100%	100%	100%
DECAF(A) → DECAF(A)	FLB-U, 65%	63%	60%	63%
DECAF(A) → DECAF(A)	FLB-P, 65%	62%	61%	62%
SURF(D) → SURF(D)	POT, 100%	100%	89%	100%
SURF(D) → SURF(D)	FLB-U, 63%	73%	84%	73%
SURF(D) → SURF(D)	FLB-P, 60%	60%	79%	60%
DECAF(D) → DECAF(D)	POT, 100%	100%	100%	100%
DECAF(D) → DECAF(D)	FLB-U, 76%	68%	71%	68%
DECAF(D) → DECAF(D)	FLB-P, 73%	73%	87%	73%
SURF(W) → SURF(W)	POT, 100%	100%	80%	100%
SURF(W) → SURF(W)	FLB-U, 77%	66%	80%	66%
SURF(W) → SURF(W)	FLB-P, 71%	71%	77%	71%
DECAF(W) → DECAF(W)	POT, 100%	100%	100%	100%
DECAF(W) → DECAF(W)	FLB-U, 71%	74%	71%	74%
DECAF(W) → DECAF(W)	FLB-P, 71%	71%	77%	71%
SURF(A) → DECAF(A)	POT, 92%	90%	69%	90%
SURF(A) → DECAF(A)	FLB-U, 64%	81%	69%	81%
SURF(A) → DECAF(A)	FLB-P, 71%	65%	69%	65%
DECAF(A) → SURF(A)	POT, 97%	95%	97%	95%
DECAF(A) → SURF(A)	FLB-U, 63%	60%	60%	60%
DECAF(A) → SURF(A)	FLB-P, 63%	60%	62%	60%
SURF(D) → DECAF(D)	POT, 89%	73%	81%	73%
SURF(D) → DECAF(D)	FLB-U, 60%	68%	79%	68%
SURF(D) → DECAF(D)	FLB-P, 63%	71%	63%	71%
DECAF(D) → SURF(D)	POT, 95%	95%	71%	95%
DECAF(D) → SURF(D)	FLB-U, 73%	65%	63%	65%
DECAF(D) → SURF(D)	FLB-P, 73%	73%	60%	73%
SURF(W) → DECAF(W)	POT, 86%	69%	77%	69%
SURF(W) → DECAF(W)	FLB-U, 77%	63%	66%	63%
SURF(W) → DECAF(W)	FLB-P, 69%	74%	77%	74%
DECAF(W) → SURF(W)	POT, 94%	94%	69%	94%
DECAF(W) → SURF(W)	FLB-U, 69%	69%	69%	69%
DECAF(W) → SURF(W)	FLB-P, 69%	69%	71%	69%

Table 3. In this table, we present the accuracy comparison of the primal-PGW, UGW, and the proposed PGW method in this paper. In "Init" column, the first entry is the name of initialization method. The second entry is its accuracy. The prior distribution $\pi = p(l = 1)$ is set to be 0.2 in all the experiment. To guarantee the SCAR assumption, for Surf(A) and Decaf(A), we set $n = 50$, which is the half of the total number of data in one single class. m is set to be 250. Similarly, we set suitable n, m for Surf(D), Decaf(D), Surf(W), Decaf(W).

Accuracy Comparison.

In Table 1 and 3, we present the accuracy results for the primal-PGW, UGW, and the proposed PGW methods when using three different initialization methods: POT, FLB-UOT, and FLB-POT.

Following (Chapel et al., 2020), in the primal-PGW and PGW methods, we incorporate the prior knowledge π into the definition of p and q. Thus it is sufficient to set $mass = \pi$ for primal-PGW and choose a sufficiently large value for λ in the PGW method. This configuration ensures that the mass matched in the target domain $\mathcal{Y}$ is exactly equal to π. However, in the UGW method (Séjourné et al., 2021), the setting is $p = \frac{1}{n}1_n$ and $q = \frac{1}{m}1_m$. Therefore, in each experiment, we test different parameters (ρ, ρ_2, ϵ) and select the ones that result in transported mass close to π.

Overall, all methods show improved performance in MNIST and EMNIST datasets. One possible reason for this could be the better separability of the embeddings in MNIST and EMNIST, as illustrated in Figure 3. Additionally, since primal-PGW and PGW incorporate information from r into their formulations, in many experiments, they exhibit slightly better accuracy.

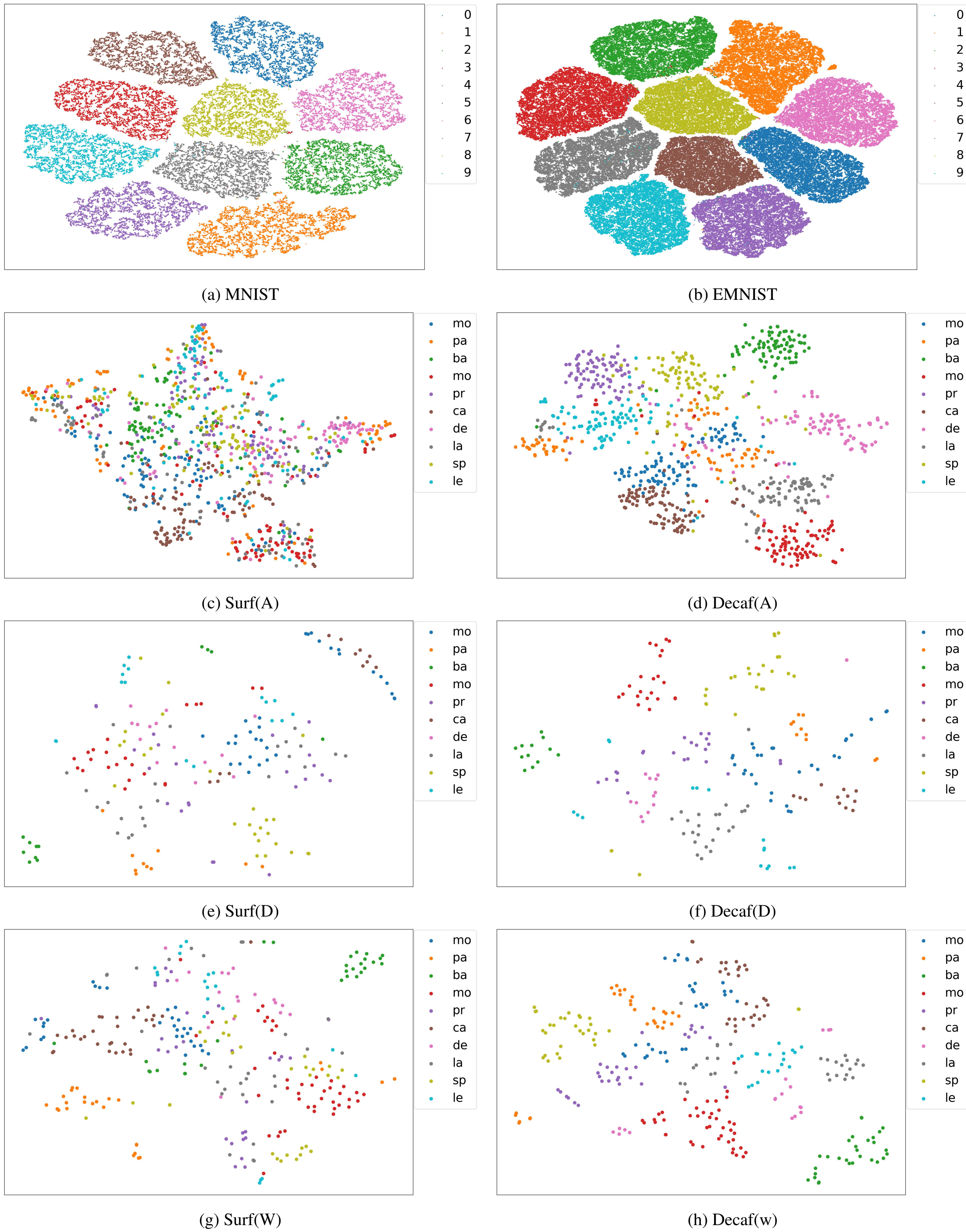

Figure 3. TSNE visulization for datasets MNIST,EMNIST,Caltech Office.