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 $X = \{x_1, \dots, x_n\}$, 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

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

  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 $1 \le p, q < \infty$.
- 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., $\langle A,B\rangle_F = \mathrm{trace}(A^TB) = \sum_{i,j}^{n,m} A_{i,j}B_{i,j}$ for all $A,B\in\mathbb{R}^{n\times m}$.
- $M\circ\gamma$: product between the tensor M and the matrix γ.
- ∇: gradient.
- $[1:n]=\{1,\dots,n\}$.
- 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

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

$$= \inf_{\gamma\in\Gamma_{\leq}(\mu,\nu)}\int_{(X\times Y)^2} L(d_X^q(x,x'),d_Y^q(y,y'))d\gamma(x,y)d\gamma(x',y') + \lambda\left(|\mu^{\otimes 2}-\gamma_1^{\otimes 2}|+|\nu^{\otimes 2}-\gamma_2^{\otimes 2}|\right). \tag{40}$$

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

  $$\begin{aligned}
  (I) &= |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A \times A) \\
  &= |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_1 \times A_1) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_2 \times A_2) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_1 \times A_2) + |\mu^{\otimes 2} - \gamma_1^{\otimes 2}|(A_2 \times A_1) \\
  &= (\gamma_1^{\otimes 2} - \bar{\gamma}_1^{\otimes 2})(A_1 \times A_1) + (\mu^{\otimes 2} - \bar{\gamma}_1^{\otimes 2})(A_2 \times A_2) + \\
  &\quad + |\bar{\gamma}_1 \otimes \mu - \gamma_1 \otimes \bar{\gamma}_1|(A_1 \times A_2) + |\mu \otimes \bar{\gamma}_1 - \bar{\gamma}_1 \otimes \gamma_1|(A_2 \times A_1) \\
  &= (\gamma_1^{\otimes 2} - \bar{\gamma}_1^{\otimes 2})(A_1 \times A_1) + (\mu^{\otimes 2} - \bar{\gamma}_1^{\otimes 2})(A_2 \times A_2) + 2|\bar{\gamma}_1 - \gamma_1|(A_1) \cdot |\mu - \bar{\gamma}_1|(A_2) \\
  &= |\gamma_1^{\otimes 2} - \bar{\gamma}_1^{\otimes 2}|(A \times A) + |\mu^{\otimes 2} - \bar{\gamma}_1^{\otimes 2}|(A \times A) + \underbrace{2|\bar{\gamma}_1 - \gamma_1|(A_1) \cdot |\mu - \bar{\gamma}_1|(A_2)}_{\geq 0}.
  \end{aligned}$$

- *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, $0 \leq \gamma \mid_B \leq \gamma$, we have $\gamma \mid_B \in \Gamma_{\leq}(\mu, \nu)$. 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, $\gamma^{\otimes 2} - (\gamma')^{\otimes 2} \geq \gamma^{\otimes 2}|_A$. Also,

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

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

That is, for any $\gamma \in \Gamma_{\leq}(\mu, \nu)$, 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 $A := \{(x, y, x', y') : L(d_X^q(x, x'), d_Y^q(y, y'))^2 \geq 2\lambda\}$. 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 $PGW_\lambda(\mathbb{X}, \mathbb{Y}) \geq 0$, 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 $|\nu| \neq |\mu|$. For convenience, suppose $|\nu|^2 \leq |\mu|^2 - \epsilon$ for some $\epsilon > 0$. Then, for each $\gamma \in \Gamma_{\leq}(\mu, \nu)$, we have $|\gamma^{\otimes 2}| \leq |\nu|^2 \leq |\mu|^2 - \epsilon$, and so

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

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

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

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 $\gamma \in \Gamma_{\leq}(\mu, \nu)$ (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, $(GW_g(\cdot,\cdot))^{1/2}$ satisfies the triangle inequality (see Lemma F.1 below for its proof). Therefore, we have

$$(PGW_\lambda(\mathbb{X}, \mathbb{Y}))^{1/2} \leq (PGW_\lambda(\mathbb{X}, \mathbb{S}))^{1/2} + (PGW_\lambda(\mathbb{S}, \mathbb{Y}))^{1/2}.$$

□

Lemma F.1. *Consider the generalized GW problem (52) for three give mm-spaces $\mathbb{S}, \mathbb{X}, \mathbb{Y}$. Then, $(GW_g(\cdot,\cdot))^{1/2}$ satisfies the triangle inequality*

$$(GW_g(\mathbb{X}, \mathbb{Y}))^{1/2} \leq (GW_g(\mathbb{X}, \mathbb{S}))^{1/2} + (GW_g(\mathbb{S}, \mathbb{Y}))^{1/2}.$$

Proof. 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 $L^2((\hat{X} \times \hat{S} \times \hat{Y})^2, \sigma^{\otimes 2})$. □

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

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

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:

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

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,

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

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

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

(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

$$\begin{aligned}
\tilde{g}_K = \min_{1\leq k\leq K} g_k &\leq \frac{\max\{2L_1, 4s\sqrt{nm}\,|\max_{i,j,i',j'}\{M_{i,j,i',j'}\} - 2\lambda|\}}{\sqrt{K}} \\
&\leq 2\frac{\max\{L_1, s\sqrt{nm}\,\left|(\max_{i,i'}\{C^X_{i,i'}\})^2 + (\max_{j,j'}\{C^Y_{j,j'}\})^2 - 4\lambda\right|\}}{\sqrt{K}} \qquad \text{(if } M \text{ is as in (21))}
\end{aligned}$$

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 $X = \{x_1, \dots x_n\}$ and $Y = \{y_1, \dots, y_m\}$ 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%

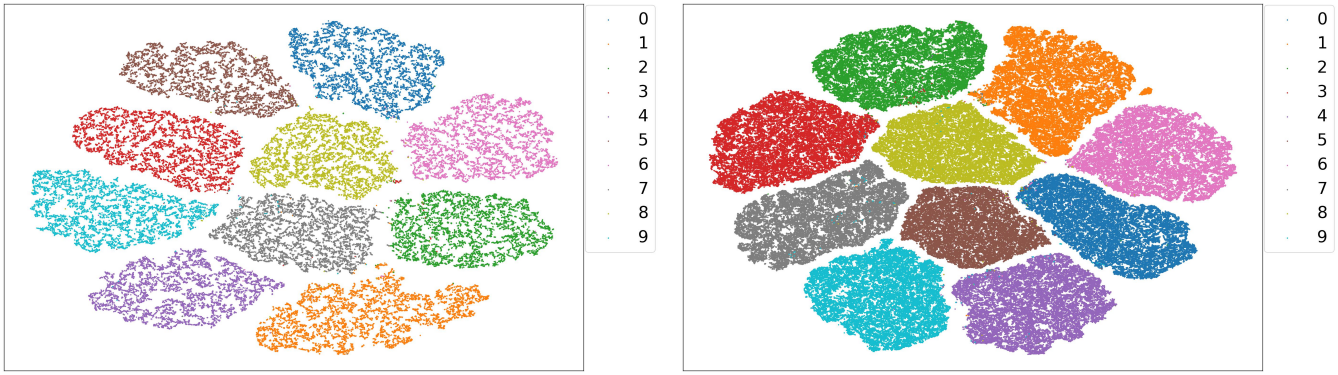
Table 3. In this table, we present the accuracy comparison of the primal-PGW, UGW, and the proposed PGW method in this paper. In "Init" column, the first entry is the name of initialization method. The second entry is its accuracy. The prior distribution $\pi = p(l = 1)$ is set to be 0.2 in all the experiment. 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.

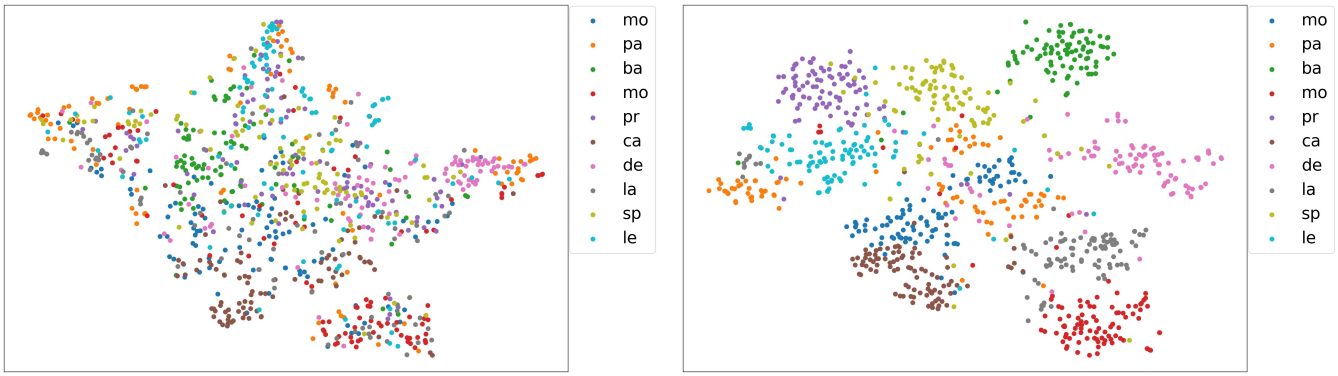
In Table 1 and 3, we present the accuracy results for the primal-PGW, UGW, and the proposed PGW methods when using three different initialization methods: POT, FLB-UOT, and FLB-POT.

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 authors set $p = \frac{1}{n}1_n$ 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.

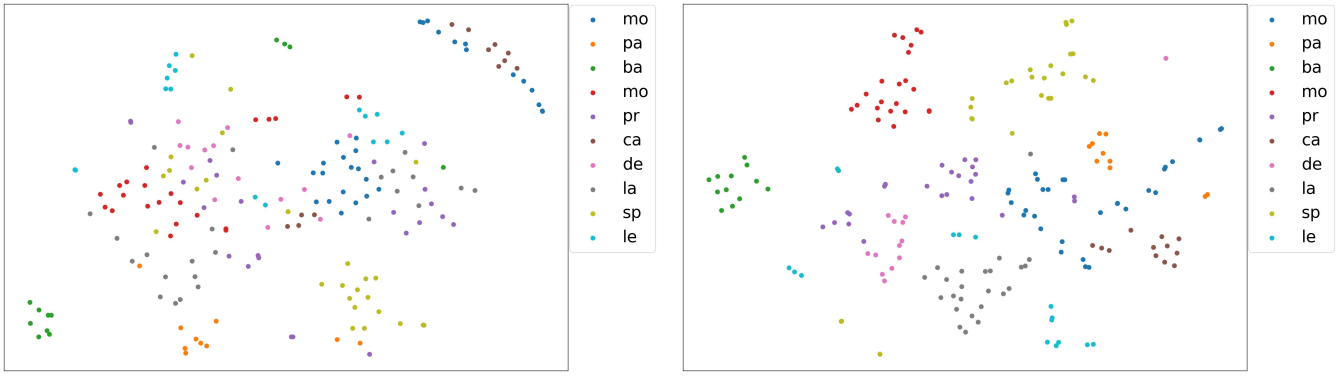
Overall, all methods show improved performance in MNIST and EMNIST datasets. One possible reason for this could be the better separability of the embeddings in MNIST and EMNIST, as illustrated in Figure 3. Additionally, since primal-PGW and PGW incorporate information from r into their formulations,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



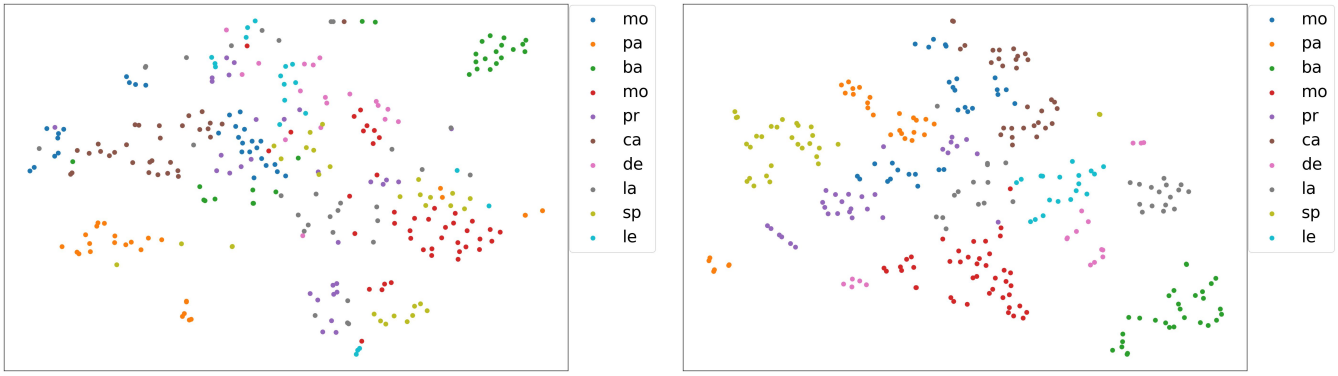
(a) MNIST (b) EMNIST



(c) Surf(A) (d) Decaf(A)



(e) Surf(D) (f) Decaf(D)



(g) Surf(W) (h) Decaf(w)

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