Supplementary Notes

SpaOT: Fused Partial Gromov-Wasserstein Optimal-Transport Background, Algorithms, and Proofs

Contents

A Notation and Abbreviations.

Abbreviations

- OT: Optimal Transport problem; see (1).
- POT: Partial Optimal Transport problem; see (3).
- GW: Gromov–Wasserstein problem; see (6).
- PGW: Partial Gromov–Wasserstein problem; see (9).
- $MPGW$: Mass-Constrained Partial Gromov–Wasserstein; see (10).
- FW: Frank–Wolfe algorithm Frank et al. (1956).
- $EFUGW(\mathbb{X}, \mathbb{Y})$: entropic fused unbalanced GW objective (cf. (81)).
- $EFPGW(\mathbb{X}, \mathbb{Y})$: entropic fused *partial* GW objective (cf. (83)).
- $LB\text{-}FPGW_\lambda(\mathbb{X}, \mathbb{Y})$: relaxed lower-bound formulation (cf. (84)).
- $EFMPGW(\mathbb{X}, \mathbb{Y})$: entropic fused mass-constrained PGW (cf. (87)); $LB\text{-}EFMPGW$ its relaxation.
- $c(x, y)$: feature cost; d_X, d_Y: intra-space distances; $|d_X - d_Y|^2$: structural discrepancy.

Sets, Spaces, and Indices

- $\mathbb{R}^d$: Euclidean space; $\mathbb{R}_+$: nonnegative reals.
- $X, Y \subset \mathbb{R}^d$: non-empty, convex, compact sets (default setting).
- $X^2 = X^{\otimes 2} = X \times X$.
- $[1:n] = \{1, \ldots, n\}$.
- $r, p, q \in [1, \infty)$: real exponents used in costs/metrics.

Vector, Norms and Basic Operators

- $\| \cdot \|$: Euclidean norm
- $|\mu|$: total mass of a measure μ. That is $|\mu| = \mu(\Omega)$, where μ is defined on Ω.
- $\| \cdot \|_{\mathrm{TV}}, \| \cdot \|_{\mathrm{TV}}$: total variation norm. In particular, given a signed measure $\mu = \mu_+ - \nu_-$, where positive measures μ_+, μ_- are the unique Jordan measure decomposition of μ. Then

$$\|\mu\|_{TV} = |\mu_+ + \mu_v|$$

- $\langle A, B \rangle = \mathrm{tr}(A^\top B)$: Frobenius inner product.
- 1_n, $1_{n \times m}$, $1_{n \times m \times n \times m}$: all-ones vector, matrix, and tensor.
- $\mathbf{1}_E$: indicator of a measurable set E, $\mathbf{1}_E(z) = 1$ if $z \in E$, else 0.
- ∇: gradient.

Measures and Pushforwards

- $\mathcal{M}_+(X)$: finite nonnegative Radon measures on X;
- $\mathcal{P}_2(X)$: probability measures on X with finite second moment.
- $|\mu| = \mu(X)$: total mass (TV-norm) of μ.
- $\mu \le \sigma$: measure domination, i.e., $\mu(B) \le \sigma(B)$ for all Borel $B \subseteq X$.
- $\mu^{\otimes 2} = \mu \otimes \mu$: product measure.
- $\mu(\phi) := \langle \phi, \mu \rangle := \int \phi(x)\, d\mu(x)$.
- $T_\# \sigma$: pushforward of σ by measurable $T\colon X \to Y$, i.e., $T_\#\sigma(A) = \sigma(T^{-1}(A))$.

Metric-Measure (mm) Spaces and Distance Matrices

- $\mathbb{X} = (X, d_X, \mu)$, $\mathbb{Y} = (Y, d_Y, \nu)$: mm-spaces.
- For discrete $\mathbb{X}$ with $X = \{x_i\}_{i=1}^n$, define $C^X \in \mathbb{R}^{n\times n}$ by $C^X_{i,i'} = d_X^q(x_i, x_{i'})$; similarly C^Y.
- $\mathbb{X} \sim \mathbb{Y}$: mm-space equivalence if they have equal total mass and $GW_q^p(\mathbb{X}, \mathbb{Y}) = 0$.

Couplings and Partial Couplings

- $\Gamma(\mu, \nu) := \{\gamma \in \mathcal{P}_2(X \times Y) :\ \gamma_1 = \mu,\ \gamma_2 = \nu\}$.
- Discrete weights: $\mathrm{p} = [p_1^X, \dots, p_n^X]^\top$, $\mathrm{q} = [q_1^Y, \dots, q_m^Y]^\top$, $\quad |\mathrm{p}| = \sum_i p_i, \quad \mathrm{p} \le \mathrm{p}'$ if $p_j \le p_j'$ $\forall j$.
- $\Gamma(\mathrm{p}, \mathrm{q}) = \{\gamma \in \mathbb{R}_+^{n\times m} :\ \gamma\, 1_m = \mathrm{p},\ \gamma^\top 1_n = \mathrm{q}\}$.
- $\Gamma_\le(\mu, \nu) := \{\gamma \in \mathcal{M}_+(X \times Y) :\ \gamma_1 \le \mu,\ \gamma_2 \le \nu\}$.
- $\Gamma_\le(\mathrm{p}, \mathrm{q}) := \{\gamma \in \mathbb{R}_+^{n\times m} :\ \gamma\, 1_m \le \mathrm{p},\ \gamma^\top 1_n \le \mathrm{q}\}$.
- $\gamma, \gamma_1, \gamma_2$: joint and marginal measures; in discrete form $\gamma \in \mathbb{R}_+^{n\times m}$, $\gamma_1 = \gamma 1_m$, $\gamma_2 = \gamma^\top 1_n$.
- $\pi_1 : X \times Y \to X$, $\pi_2 : X \times Y \to Y$.
- $\pi_{1,2} : S \times X \times Y \to X \times Y$, $(s, x, y) \mapsto (x, y)$; similarly $\pi_{0,1}, \pi_{0,2}$ for other coordinate pairs.

Optimal Transport Problems

- $c : X \times Y \to \mathbb{R}_+$: lower semicontinuous cost for (partial) OT.
- $OT(\mu, \nu)$: classical OT; $W_p(\mu, \nu)$: p-Wasserstein distance; $POT_\lambda(\mu, \nu)$: partial OT with parameter $\lambda > 0$; see (1), (2), (3).
- $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $D : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$: GW loss and scalar distance.
- $GW^L(\cdot, \cdot)$: GW objective with loss L; $d^q_{GW,r}$: GW with $L(a, b) = |a - b|^q$; see (6).
- $PGW_\lambda(\cdot, \cdot)$: partial GW objective; see (9).
- $C(\gamma; \lambda, \mu, \nu) := \gamma^{\otimes 2}\big(L(d_X^q, d_Y^q)\big) + \lambda\big(|\mu|^2 + |\nu|^2 - 2|\gamma|^2\big)$: PGW transport cost for $\gamma \in \Gamma_\le(\mu, \nu)$.
- $UGW_\lambda(\mathbb{X}, \mathbb{Y})$: unbalanced GW; $FUGW_\lambda(\mathbb{X}, \mathbb{Y})$: fused unbalanced GW.
- $d_{FGW}(\mathbb{X}, \mathbb{Y})$: fused GW distance; $d^p_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y})$: fused partial GW distance.

Discrete Tensorized Forms

- Discrete measure setting: $\mu = \sum_{i=1}^{n} \mathrm{p}_i \delta_{x_i}, \nu = \sum_{j=1}^{m} \mathrm{q}_j \delta_{y_j}$.
- $M \in \mathbb{R}^{n\times m\times n\times m}$ with $M_{i,j,i',j'} = L(C^X_{i,i'}, C^Y_{j,j'})$; $M^\top$ swaps index pairs: $M^\top_{i,j,i',j'} = M_{i',j',i,j}$. In default, L is L2 norm square, and we denote $M = \|C_X - C_Y\|^2$
- For fused-partial formulations (with weight ω_2), we set $\hat{\lambda} := \lambda/\omega_2$ and write
$$(M - 2\hat{\lambda})_{i,j,i',j'} := M_{i,j,i',j'} - 2\hat{\lambda}.$$
- $M \circ \gamma \in \mathbb{R}^{n\times m}$ with $[M \circ \gamma]_{i,j} = \sum_{i',j'} M_{i,j,i',j'} \gamma_{i',j'}$.
- $\langle \cdot, \cdot \rangle_F$: Frobenius inner product on $\mathbb{R}^{n\times m}$.

Frank-Wolfe Optimization and Algorithmic Symbols

- $\mathcal{L}$: objective functional for $PGW_\lambda(\cdot, \cdot)$.
- $\alpha \in [0, 1]$: line-search step size.
- $\gamma^{(1)}$: initialization; $\gamma^{(k)}$, $\gamma^{(k)\prime}$: transport plans before/after Step 1 in the k-th FW iteration.
- $G_{C,M} = \omega_1 C + \omega_2 \tilde{M} + M^\top \circ \gamma, \quad G = 2\hat{M} \circ \hat{\gamma}$: gradients in two FW variants.
- $a, b, c \in \mathbb{R}$: coefficients defined in (59) (For FPGW, replace M by $M - 2\hat{\lambda}$).
- $MPGW_\rho$ and $\Gamma^\rho_\leq(\mu, \nu)$: mass-constrained partial GW and its feasible set; see (10), (5).

Entropic / Sinkhorn Notation

- $\epsilon > 0$: entropic regularization parameter.
- $D_\phi(\cdot\|\cdot)$: ϕ-divergence; special cases below.
- $H(\mu\|\nu) = \bar{H}(\mu\|\nu) + |\nu| - |\mu|$ with $\bar{H}(\mu\|\nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu$ when $\mu \ll \nu$, $+\infty$ otherwise.
- $D_{PTV}(\mu\|\nu) = \begin{cases} \|\mu - \nu\|_{\mathrm{TV}} = |\nu - \mu| & \text{if } \mu \leq \nu, \\ +\infty & \text{otherwise.} \end{cases}$
- $\Gamma_\leq(\mu, \nu)$, $\Gamma^\rho_\leq(\mu, \nu)$: partial feasible sets (mass-unconstrained and mass-constrained with $|\gamma| = \rho$).
- $|\gamma| = \gamma(X \times Y)$, $|\mu| = \mu(X)$, $|\nu| = \nu(Y)$: total mass.
- $\mathcal{L}(\gamma)$: fused GW functional (feature + structure + penalties) used in EFUGW/EFPGW.
- $c_\pi(x, y) = \frac{1}{2}\omega_1 d(x, y) + \omega_2 \left[L(d^r_X, d^r_Y) \circ \pi\right](x, y) + \epsilon\, \bar{H}(\pi\|\mu \otimes \nu)$: π-conditioned cost (cf. Prop. 10).
- $[L(d_X, d_Y) \circ \pi](x, y) = \int_{X\times Y} L(d^r_X(x, x'), d^r_Y(y, y'))\, d\pi(x', y')$.
- $K = \exp(-c/\epsilon)$: Gibbs kernel for feature cost; elementwise exponential.
- $u \in \mathbb{R}^n_+$, $v \in \mathbb{R}^m_+$: Sinkhorn scaling vectors.

- $\odot$, $\oslash$: elementwise (Hadamard) product and division.
- $\mathrm{diag}(a)$: diagonal matrix with vector a on the diagonal.
- $\mathrm{Proj}^{KL}_{\mathcal{C}_i}$: Bregman (KL) projection onto constraint set $\mathcal{C}_i$ (cf. Alg. 2).
- $\mathcal{C}_1 = \{\gamma \geq 0 : \ \gamma_2 \leq q\}, \ \mathcal{C}_2 = \{\gamma \geq 0 : \ \gamma_1 \leq p\}, \ \mathcal{C}_3 = \{\gamma \geq 0 : \ |\gamma| = \rho\}$: partial-marginal and mass constraints.
- $\gamma^{\otimes 2}$, $(\mu \otimes \nu)^{\otimes 2}$: product measures on $(X \times Y)^2$.
- $\pi, \gamma \in \mathcal{M}_+(X \times Y)$: current and updated couplings in alternating minimization; $\rho \in [0, \min(|p|, |q|)]$ target mass.
- Stopping criteria: norms/duality gap on (u, v) or fixed-point tolerance on γ.

Graph-Specific Notation

- $G = (V, E)$: graph with nodes $V = \{v_1, \dots, v_N\}$ and edges $E \subset V^2$.
- $d_V : V^2 \to \mathbb{R}$: structural distance (default: shortest-path distance).
- $f : V \to \mathcal{F}$: node feature map; $\mathcal{F}$ is the feature space.
- $wl : \mathcal{F} \to S^H$: Weisfeiler–Lehman feature map Vishwanathan et al. (2010) for discrete $\mathcal{F}$; S finite alphabet, $H \in \mathbb{N}$.
- $d_{\mathcal{F}}$: feature-space metric. Default: Euclidean if $\mathcal{F} = \mathbb{R}^d$; if $\mathcal{F}$ is discrete, Hamming on $wl(\mathcal{F})$:

$$d_{\mathcal{F}}\big(f(v_1), f(v_2)\big) := \sum_{h=1}^{H} \delta\big(wl(f(v_1))_h \neq wl(f(v_2))_h\big).$$

It is a lower semi-continuous function.

B Background of optimal transport problems

The Optimal Transport (OT) problems for $\mu, \nu \in \mathcal{P}(\Omega)$, with transportation cost $c(x, y) : \Omega \times \Omega \to \mathbb{R}_+$ being a lower-semi continuous function is defined as:

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

and where $\Gamma(\mu, \nu)$ denotes the set of all joint probability measures on $\Omega^2 := \Omega \times \Omega$ with marginals μ, ν, i.e., $\gamma_1 := \pi_{1\#}\gamma = \mu, \gamma_2 := \pi_{2\#}\gamma = \nu$, where $\pi_1, \pi_2 : \Omega^2 \to \Omega$ are the canonical projections $\pi_1(x, y) := x, \pi_2(x, y) := y$. A minimizer for (1) always exists Villani (2009, 2021) and when $c(x, y) = \|x - y\|^p$, for $p \geq 1$, it defines a metric on $\mathcal{P}(\Omega)$, which is referred to as the "p-Wasserstein distance":

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

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

$$POT(\mu, \nu; \lambda) := \inf_{\gamma \in \mathcal{M}_+(\Omega^2)} \gamma(c) + \lambda(|\mu - \gamma_1| + |\nu - \gamma_2|), \tag{3}$$

where, in general, $|\sigma|$ denotes the total variation norm of a measure σ, i.e., $|\sigma| := \sigma(\Omega)$. The constraint $\gamma \in \mathcal{M}_+(\Omega^2)$ in (3) can be further restricted to $\gamma \in \Gamma_\leq(\mu, \nu)$:

$$\Gamma_\leq(\mu, \nu) := \{\gamma \in \mathcal{M}_+(\Omega^2) : \gamma_1 \leq \mu, \gamma_2 \leq \nu\},$$

denoting $\gamma_1 \leq \mu$ if for any Borel set $B \subseteq \Omega$, $\gamma_1(B) \leq \mu(B)$ (respectively, for $\gamma_2 \leq \nu$) Figalli (2010). Roughly speaking, the linear penalization indicates that if the classical transportation cost exceeds 2λ, it is better to create/destroy' mass (see Bai et al. (2023b) for further details). In addition, the above formulation has an equivalent form, to distinguish them, we call it "mass-constraint partial optimal transport problem" (MPOT):

$$MOPT(\mu, \nu; \rho) := \inf_{\gamma \in \Gamma^\rho_\leq(\mu, \nu)} \gamma(c), \tag{4}$$

where the feasible set is defined as

$$align\Gamma^\rho_\leq(\mu, \nu) := \{\gamma \in \Gamma_\leq(\mu, \nu) : \; |\gamma| \geq \rho\}, \qquad \rho \in [0, \min(|\mu|, |\nu|)]. \tag{5}$$

C Background:Gromov–Wasserstein Problems

C.1 Gromov–Wasserstein, unbalanced GW, and partial GW

We briefly recall three closely related formulations.

(1) Gromov–Wasserstein (GW). Given two probability mm-spaces $\mathbb{X} = (X, d_X, \mu)$ and $\mathbb{Y} = (Y, d_Y, \nu)$, and a nonnegative loss L, the GW objective is

$$GW(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \Gamma(\mu, \nu)} \gamma^{\otimes 2}\big(L(d_X^r(\cdot, \cdot), d_Y^r(\cdot, \cdot))\big), \tag{6}$$

where

$$\Gamma(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) : \; \gamma_1 = \mu, \; \gamma_2 = \nu\}. \tag{7}$$

(2) Unbalanced GW (UGW). When masses may differ, unbalanced GW relaxes hard marginal constraints by penalizing deviations via f-divergences. The unbalanced Gromov–Wasserstein (UGW) problem is proposed by Séjourné et al. (2021); Chapel et al. (2020); Bai et al. (2024) and is defined as

$$UGW_\lambda(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \mathcal{M}_+(X \times Y)} \langle L(d_X^r, d_Y^r), \gamma^{\otimes 2}\rangle + \lambda\big(D_{\phi_1}(\gamma_1^{\otimes 2} \parallel \mu^{\otimes 2}) + D_{\phi_2}(\gamma_2^{\otimes 2} \parallel \nu^{\otimes 2})\big), \tag{8}$$

where D_{ϕ_1}, D_{ϕ_2} are f-divergences and can be distinct.

(3) Partial GW (PGW) and mass-constrained PGW (MPGW). An alternative relaxation is to restrict to *partial couplings* $\gamma \in \Gamma_{\leq}(\mu, \nu)$ (or impose a transported-mass constraint), in addition, D_ϕ is A total derivative. Two commonly used formulations are

$$PGW(\mathbb{X}, \mathbb{Y}) = \inf_{\gamma \in \Gamma_{\leq}(\mu,\nu)} \gamma^{\otimes 2}(L(d_X^r, d_Y^r)) + \lambda\big(|\mu^{\otimes 2} - \gamma_1^{\otimes 2}| + |\nu^{\otimes 2} - \gamma_2^{\otimes 2}|\big), \tag{9}$$

$$MPGW(\mathbb{X}, \mathbb{Y}) = \inf_{\gamma \in \Gamma_{\leq}^{\rho}(\mu,\nu)} \gamma^{\otimes 2}(L(d_X^r, d_Y^r)). \tag{10}$$

C.2 Related works: fused Gromov–Wasserstein problems

Fused Gromov–Wasserstein (FGW). When each point/node also has a feature cost C across spaces, fused GW combines a linear feature-matching term with the quadratic GW structure term Titouan et al. (2019); Vayer et al. (2020):

$$FGW(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \Gamma(\mu,\nu)} \omega_1 \langle C, \gamma \rangle + \omega_2 \langle L(d_X^r, d_Y^r), \gamma^{\otimes 2} \rangle. \tag{11}$$

Fused unbalanced GW (FUGW). When masses may differ, fused unbalanced GW relaxes the marginal constraints by replacing them with f-divergence penalties. The fused-unbalanced Gromov–Wasserstein (FUGW), proposed by Thual et al. (2022), is

$$FUGW_\lambda(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \mathcal{M}_+(X \times Y)} \omega_1 \langle C, \gamma \rangle + \omega_2 \langle L(d_X^r, d_Y^r), \gamma^{\otimes 2} \rangle + \lambda\big(D_{\phi_1}(\gamma_1^{\otimes 2} \parallel \mu^{\otimes 2}) + D_{\phi_2}(\gamma_2^{\otimes 2} \parallel \nu^{\otimes 2})\big). \tag{12}$$

When the f-divergence terms D_{ϕ_1}, D_{ϕ_2} are KL divergences, Thual et al. (2022) propose a Sinkhorn solver.

C.3 Fused partial GW (FPGW) and relation to FUGW

We define the fused-partial objectives used throughout this supplementary material. For convenience, we set $\hat{\lambda} := \lambda/\omega_2$ (when $\omega_2 > 0$).

$$FPGW_\lambda(\mathbb{X}, \mathbb{Y}) = \inf_{\gamma \in \mathcal{M}_+(X \times Y)} \omega_1 \gamma(C) + \omega_2 \gamma^{\otimes 2}(L(d_X^r, d_Y^r)) + \lambda\big(|\mu^{\otimes 2} - \gamma_1^{\otimes 2}|_{TV} + |\nu^{\otimes 2} - \gamma_2^{\otimes 2}|_{TV}\big), \tag{13}$$

$$FMPGW_\rho(\mathbb{X}, \mathbb{Y}) = \inf_{\gamma \in \Gamma_{\leq}^{\rho}(\mu,\nu)} \omega_1 \gamma(C) + \omega_2 \gamma^{\otimes 2}(L(d_X^r, d_Y^r)) \tag{14}$$

Theorem 1. *We have the following:*

(1) When $C(x, y) \geq 0, \omega_2 > 0$ and $L(\mathrm{r}_1, \mathrm{r}_2) \geq 0$, the problem (13) *can be simplified as*

$$FPGW_\lambda(\mathbb{X}, \mathbb{Y}) = \inf_{\gamma \in \Gamma_{\leq}(\mu,\nu)} \omega_1 \gamma(C) + \omega_2 \gamma^{\otimes 2}(L(d_X^r, d_Y^r) - 2\hat{\lambda}) + \lambda(|\mu|^2 + |\nu|^2). \tag{15}$$

(2) The problems (13), (15) *and* (14) *admit a minimizer.*

(3) When $C(x, y) = \|x - y\|^q$, $L(\cdot_1, \cdot_2) = |\cdot_1 - \cdot_2|^q$ and $\omega_2, \lambda > 0$, (15) *induces a semi-metric; for $q = 1$ it is a metric.*

We now record a useful relation between FUGW and FPGW; this is equivalent to statement (1) in Theorem 1.

Proposition 1. *When D_{ϕ_1}, D_{ϕ_2} are total variation, then we have:*

$$\begin{aligned} FUGW_\lambda(\mathbb{X},\mathbb{Y}) &= FPGW_\lambda(\mathbb{X},\mathbb{Y}) \\ &= \inf_{\gamma\in\Gamma_\leq(\mu,\nu)} \omega_1\langle C,\gamma\rangle + \omega_2\langle L(d_X^r,d_Y^r),\gamma^{\otimes 2}\rangle + \lambda(|\mu|^2+|\nu|^2-2|\gamma|^2). \end{aligned}$$

Equivalently speaking, one can restrict the searching space from $\mathcal{M}_+(X\times Y)$ to $\Gamma_\leq(\mu,\nu)$.

Proof. Choose $\gamma\in\mathcal{M}_+(X\times Y)$ and let $\gamma_{Y|X}(\cdot|x)$ denote the conditional measure of γ given $x\in X$. Let $\gamma'=\gamma_{Y|X}(\cdot|X)\cdot(\gamma_1\wedge\mu)$. From the definition, we have $\gamma'\leq\gamma,\gamma_1'\leq\mu$. From Inequality (22) in Bai et al. (2023c), we have

$$\omega_2\langle\|d_X-d_Y\|^p,\gamma'^{\otimes 2}\rangle+\lambda(\|\mu-\gamma_1'\|_{TV}+\|\nu-\gamma_2'\|_{TV})\leq\omega_2\langle\|d_X-d_Y\|^p,\gamma^{\otimes 2}\rangle+\lambda(\|\mu-\gamma_1\|_{TV}+\|\nu-\gamma_2\|_{TV}).$$

Since $C\geq 0,\gamma'\leq\gamma$, we have $\langle C,\gamma'\rangle\leq\langle C,\gamma\rangle$. Therefore, we have

$$\begin{aligned} &\omega_1\langle C,\gamma'\rangle+\omega_2\langle\|d_X-d_Y\|^p,\gamma'^{\otimes 2}\rangle+\lambda(\|\mu-\gamma_1'\|_{TV}+\|\nu-\gamma_2'\|_{TV}) \\ &\leq\omega_1\langle C,\gamma\rangle+\omega_2\langle\|d_X-d_Y\|^p,\gamma^{\otimes 2}\rangle+\lambda(\|\mu-\gamma_1\|_{TV}+\|\nu-\gamma_2\|_{TV}). \end{aligned}$$

That is, based on γ, we can construct γ' such that $\gamma_1'\leq\mu\wedge\gamma_1,\gamma'\leq\gamma$ and the γ' admits smaller cost. Based on the same process, based on γ', we can construct γ'' such that $\gamma''\leq\gamma'$, $\gamma''\leq\nu$. That is $\gamma''\in\Gamma_\leq(\mu,\nu)$ and admits smaller cost than γ'. Therefore, we can restrict the searching space for (12) in this case from $\mathcal{M}_+(X\times Y)$ to $\Gamma_\leq(X\times Y)$. □

Remark 1. In Figure 1, we illustrate the transportation plans of various methods, including optimal transport (OT), unbalanced optimal transport (UOT), partial optimal transport (POT), ... and fused-Partial Gromov–Wasserstein (FPGW).

Due to the balanced mass setting, OT, GW, and fused GW match all points from the source shape to the target shape. In contrast, the Sinkhorn entropic regularization in UOT, UGW, and FUGW allows for mass splitting, resulting in a small amount of mass between the grey shape and the pink points. PGW and fused PGW achieve the desired one-to-one matching, where the straight-line segment in the source shape (red points) corresponds to the straight-line segment in the target shape (grey points). However, since PGW does not account for the spatial location of points, there is a 50% chance of obtaining "anti-identity" matching as demonstrated in Figure 1.

C.4 Existence of minimizers for fused PGW

In this section, we establish basic well-posedness properties for fused PGW.

Proposition 2. *Given mm-spaces $\mathbb{X}=(X,d_X,\mu),\mathbb{Y}=(Y,d_Y,\nu)$, where X,Y are compact sets, the fused-PGW problems* (15) *and* (14) *admit minimizers. That is, we can replace* inf *by* min *in the formulations.*

Note, this proposition is equivalent to the statement (2) in Theorem 1.

We first introduce the following statements:

Proposition 3. *The set $\Gamma_\leq(\mu,\nu)$ and $\Gamma_\leq^\rho(\mu,\nu)$ are weakly compact, convex sets, where $\rho\in[0,\min(|\mu|,|\nu|)]$.*

Proof. By Liu et al. (2023) Lemma B.1, we have $\Gamma_\leq(\mu,\nu)$ is a weakly compact set. By e.g. Caffarelli & McCann (2010); Bai et al. (2023b), we have $\Gamma_\leq(\mu,\nu)$ is convex.

It remains to show the compactness and convexity of the set $\Gamma_\leq^\rho(\mu,\nu)$.

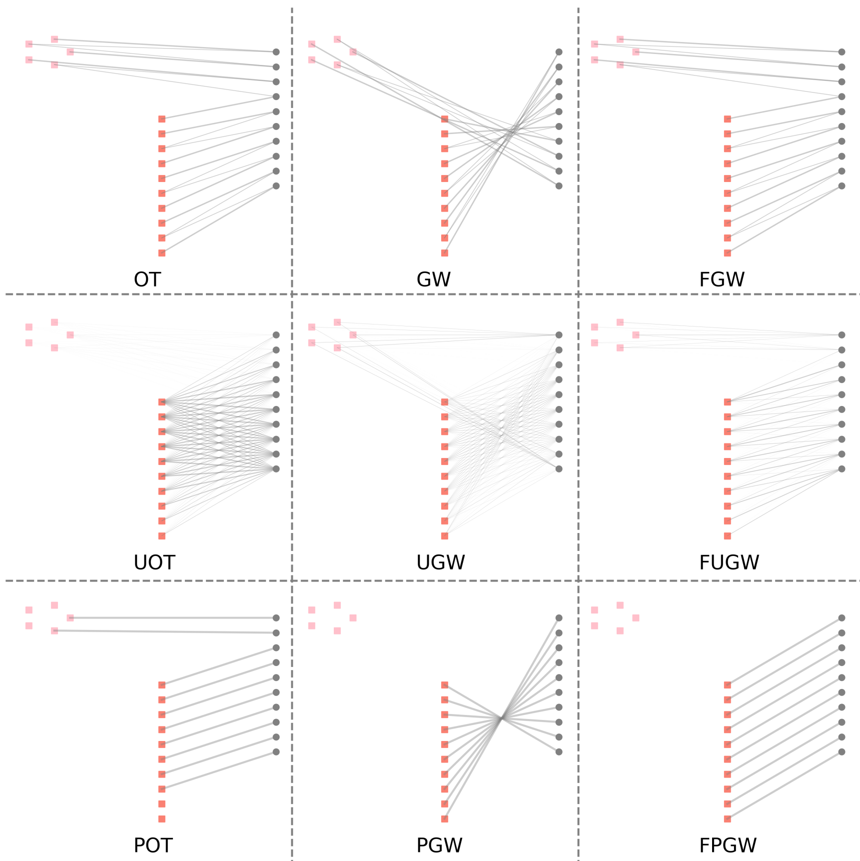


Figure 1: In this toy example, we illustrate the relationships and differences between OT, Unbalanced OT, Partial OT, GW, Unbalanced GW, Partial GW, fused GW, fused unbalanced GW, and Fused Partial GW. The source shape consists of the union of the pink and red points, and the target shape is the grey shape. The objective is to establish a reasonable match between the two shapes.

Pick $\gamma^1, \gamma^2 \in \Gamma^\rho_\leq(\mu,\nu)$ and $\omega \in [0,1]$ and let $\gamma = (1-\omega)\gamma^1 + \omega\gamma^2$. We have $\gamma \in \Gamma_\leq(\mu,\nu)$ by convexity. In addition,

$$|(1-\omega)\gamma^1 + \omega\gamma^2| = (1-\omega)|\gamma^1| + \omega|\gamma^2| \geq \rho,$$

thus $\gamma \in \Gamma^\rho_\leq(\mu,\nu)$.

Pick a sequence $(\gamma^n)_{n=1}^\infty \subset \Gamma^\rho_\leq(\mu,\nu) \subset \Gamma_\leq(\mu,\nu)$. By compactness of $\Gamma_\leq(\mu,\nu)$, there exists a subsequence $(\gamma^{n_k})_{k=1}^\infty$ that converges weakly to $\gamma^* \in \Gamma_\leq(\mu,\nu)$. It remains to show $\gamma^* \in \Gamma^\rho_\leq(\mu,\nu)$. Indeed,

$$|\gamma^*| = \langle 1_{X\times Y}, \gamma^*\rangle = \lim_{k\to\infty} \langle 1_{X\times Y}, \gamma^{n_k}\rangle = \lim_{k\to\infty} |\gamma^{n_k}|.$$

Since $|\gamma^{n_k}| \geq \rho$ for all k, we have $|\gamma^*| \geq \rho$. □

Lemma 1. *Suppose X and Y are compact sets. Let $Z = (X \times Y)$ and $d_Z((x^1,y^1),(x^2,y^2)) = d_X(x^1,x^2) + d_Y(y^1,y^2)$. Suppose $\phi : Z^2 \to \mathbb{R}$ is Lipschitz continuous and $C : Z \to \mathbb{R}$ is bounded. Then the following problem admits a minimizer:*

$$\inf_{\gamma\in\mathcal{M}} \langle \phi, \gamma^{\otimes 2}\rangle + \langle C, \gamma\rangle. \tag{16}$$

where $\mathcal{M}$ is a non-empty weakly compact set.

Proof. Pick a weakly convergent sequence $(\gamma^n)_{n=1}^\infty \subset \mathcal{M}$ with $\gamma^n \xrightarrow{w} \gamma^* \in \mathcal{M}$.

From Lemma C.3 in Bai et al. (2024), we have

$$\langle \phi, (\gamma^n)^{\otimes 2}\rangle \to \langle \phi, \gamma^*\rangle.$$

By weak convergence, $\langle C, \gamma^n\rangle \to \langle C, \gamma^*\rangle$. Hence the objective is lower semicontinuous along convergent subsequences, and compactness yields existence of a minimizer. □

Proof of Proposition 2. From Lemma C.1 in Bai et al. (2024), the map $((x^1,y^1),(x^2,y^2)) \mapsto |d^r_X(x^1,x^2) - d^r_Y(y^1,y^2)|^q$ is Lipschitz. In addition, $C : X \times Y$ is bounded. From Proposition 3, the feasible sets are weakly compact. Applying Lemma 1 yields existence of a minimizer. □

D Metric property of Fused PGW

In this section, we discuss the proof of Theorem 1 (3). We will discuss the details in the following subsections. The related conclusion can be treated as an extension of Theorem 6.3 in Titouan et al. (2019) and Theorem 3.1 in Vayer et al. (2020).

D.1 Rotation/translation invariance

Proposition 4 (Rotation/translation invariance of FPGW). *Assume $X \subset \mathbb{R}^{d_x}$ and $Y \subset \mathbb{R}^{d_y}$ are embedded in Euclidean spaces. Let $T_X : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ and $T_Y : \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}$ be rigid motions of the form*

$$T_X(x) := R_X x + t_X, \qquad R_X \in \mathbb{R}^{d_x\times d_x},\ R_X^\top R_X = I,\ t_X \in \mathbb{R}^{d_x}, \tag{17}$$

$$T_Y(y) := R_Y y + t_Y, \qquad R_Y \in \mathbb{R}^{d_y\times d_y},\ R_Y^\top R_Y = I,\ t_Y \in \mathbb{R}^{d_y}. \tag{18}$$

Assume the cross-feature loss $C(x,y)$ is rotation/translation invariant, i.e.,

$$C(T_X(x), T_Y(y)) = C(x,y), \qquad \forall (x,y) \in X \times Y, \tag{19}$$

and that the intra-domain dissimilarities d_X, d_Y are rotation/translation invariant. Define the transformed spaces $\widetilde{\mathbb{X}} = (T_X(X), d_X, (T_X)_\# \mu)$ and $\widetilde{\mathbb{Y}} = (T_Y(Y), d_Y, (T_Y)_\# \nu)$. Then

$$d_{FPGW,\lambda}(\widetilde{\mathbb{X}}, \widetilde{\mathbb{Y}}) = d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}). \tag{20}$$

Proof. Given any feasible coupling $\gamma \in \Gamma_\leq(\mu, \nu)$, set $\widetilde{\gamma} := (T_X \times T_Y)_\# \gamma$. Then $\widetilde{\gamma} \in \Gamma_\leq((T_X)_\# \mu, (T_Y)_\# \nu)$ and $|\widetilde{\gamma}| = |\gamma|$. By the assumed invariances of C and of d_X, d_Y, every term in the FPGW objective evaluated at $\widetilde{\gamma}$ equals the corresponding term evaluated at γ. Taking the infimum over couplings yields the claim. □

Remark 2. The same argument applies to d_{FGW} and to the fused unbalanced GW (FUGW): since their objectives are built from the same rotation/translation invariant ingredients (the cross-feature term C and the intra-domain dissimilarities d_X, d_Y), they are also invariant under rotations and translations.

D.2 Background: Isometry, fused-GW semi-metric.

Given $\mathbb{X} = (X, d_X, \mu), \mathbb{Y} = (Y, d_Y, \nu)$, we say X, Y are equivalent, noted $X \sim Y$ if and only if:

There exists a mapping $\phi : X \to Y$ such that

- $\phi_\# \mu = \nu$
- $d_X(x, x') = d_Y(\phi(x), \phi(x')), \forall x, x' \in X.$

Such a function ϕ is called **measure preserving isometry**.

In addition, in the formulation of fused-GW (15), we set $d(x, y) = \|x - y\|^q$ and $L(\mathrm{r}_1, \mathrm{r}_2) = |\mathrm{r}_1 - \mathrm{r}_2|^q$. The reduced formulation is called "fused-Gromov Wasserstein distance" Vayer et al. (2020):

$$d_{FGW}(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{(X \times Y)^2} \omega_1 \frac{\|x - y\|^q}{|\gamma|} + \omega_2 |d_X^r(x, x') - d_Y^r(y, y')|^q d\gamma(x, y) d\gamma(x', y'). \tag{21}$$

and it defines a metric where the above equivalence relation induces the identity.

Inspired by this formulation, we introduce the **Fused Partial GW metric**:

$$d^p_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \Gamma_\leq(\mu,\nu)} \int_{(X \times Y)^2} \omega_1 \frac{\|x - y\|^q}{|\gamma|} + \omega_2 |d_X^r(x, x') - d_Y^r(y, y')|^q + \lambda \left(\frac{|\mu|^2}{|\gamma|^2} + \frac{|\nu|^2}{|\gamma|^2} - 2 \right) d\gamma(x, y) d\gamma(x', y'). \tag{22}$$

Remark 3. We adapt the convention $\frac{1}{0} \cdot 0 = 1$, thus the above formulation is well-defined. In particular, when $|\gamma| = 0$, i.e., γ is zero measure, the above integration is defined by

$$\begin{aligned}
&\int_{(X \times Y)^2} \omega_1 \frac{\|x - y\|^q}{|\gamma|} + \omega_2 |d_X^r(x, x') - d_Y^r(y, y')|^q + \lambda \left(\frac{|\mu|^2}{|\gamma|^2} + \frac{|\nu|^2}{|\gamma|^2} - 2 \right) d\gamma(x, y) d\gamma(x', y') \\
&= \lambda(|\mu|^2 + |\nu|^2) \qquad (23)\\
&= \lim_{|\gamma| \searrow 0} \int_{(X \times Y)^2} \omega_1 \frac{\|x - y\|^q}{|\gamma|} + \omega_2 |d_X^r(x, x') - d_Y^r(y, y')|^q + \lambda \left(\frac{|\mu|^2}{|\gamma|^2} + \frac{|\nu|^2}{|\gamma|^2} - 2 \right) d\gamma(x, y) d\gamma(x', y').
\end{aligned}$$

Remark 4. By Proposition 2, the above problem admits a minimizer.

Next, we introduce the formal statement of Theorem 1 (2).

Theorem 2. *Define space for mm-spaces,*

$$\mathcal{G} = \{\mathbb{X} = (X, d_X, \mu), X \subset \mathbb{R}^d, X \text{ is compact}; d_X \text{ is a metric}; \mu \in \mathcal{M}_+(X)\}.$$

Then Fused PGW (22) *defines a semi-metric in quotient space* $\mathcal{G}/\sim$. *In particular:*

1. $d_{FPGW,\lambda}(\cdot,\cdot)$ *is non-negative and symmetric.*
2. *Suppose* $\omega_2, \lambda > 0$, *then* $d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) = 0$ *iff* $\mathbb{X} \sim \mathbb{Y}$.
3. *If* $\omega_2 > 0$, *for* $q \geq 1$, *we have*

$$d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) \leq 2^{q-1}(d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Z}) + d_{FPGW,\lambda}(\mathbb{Y}, \mathbb{Z})). \tag{24}$$

In particular, when $q = 1$, *fused-PGW satisfies the triangle inequality.*

D.3 Proof of the (1)(2) in Theorem 2

For statement (1), by definition, for each $\gamma \in \Gamma_\leq(\mu, \nu)$, we have

$$\int_{X\times Y} \|x-y\|^q d\gamma, \int_{(X\times Y)^2} \|d_X^r(x,x') - d_Y^r(y,y')\|^q d\gamma^{\otimes 2} \geq 0.$$

Thus, $d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) \geq 0$.

In addition,

$$\begin{aligned}
&d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) \\
&= \inf_{\gamma\in\Gamma_\leq(\mu,\nu)} \int_{(X\times Y)^2} \omega_1\|x-y\|^q + \omega_2|d_X^r(x,x') - d_Y^r(y,y')|^q)^p d\gamma(x,y)d\gamma(x',y') + \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma|^2) \\
&= \inf_{\gamma\in\Gamma_\leq(\nu,\mu)} \int_{(Y\times X)^2} \omega_1\|y-x\|^q + \omega_2|d_Y^r(y,y') - d_X^r(x,x')|^q)^p d\gamma(y,x)d\gamma(y',x') + \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma|^2) \\
&= d_{FPGW,\lambda}(\mathbb{Y}, \mathbb{X}).
\end{aligned}$$

And we prove the symmetric.

For statement (2), suppose a measure-preserving isometry $\phi : X \to Y$ exists, then we have $|\mu| = |\nu|$. Let $\gamma = (\mathrm{id} \times \phi)_\# \mu$, we have $\gamma \in \Gamma(\mu, \nu) \subset \Gamma_\leq(\mu, \nu)$. Thus $|\gamma| = |\mu| = |\nu|$.

Furthermore,

$$\begin{aligned}
&d^p_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) \\
&\leq \int_{(X\times Y)^2} \omega_1\|x-y\|^q + \omega_2|d_X^r(x,x') - d_Y^r(y,y')|^q d\gamma(x,y)d\gamma(x',y') + \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma|^2) \\
&= \int_{(X\times Y)^2} \omega_1 \underbrace{\|x-\phi(x)\|^q}_{0} + \omega_2 \underbrace{|d_X^r(x,x') - d_Y^r(\phi(x),\phi(x'))|^q}_{0} d\mu(x)d\mu(x') + \lambda(\underbrace{|\mu|^2 + |\nu|^2 - 2|\gamma|^2}_{0}) \\
&= 0.
\end{aligned} \tag{25}$$

For the other direction, suppose $d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) = 0$. We have two cases:

Case 1: $|\mu| = |\nu| = 0$. We've done.

Case 2: $|\mu| > 0$ or $|\nu| > 0$.

By the Proposition 2, a minimizer for problem (22) exists. Choose one minimizer, denoted as γ^*.

First, we claim $|\mu| = |\nu| = |\gamma^*|$.

Indeed, assume the above equation is not true. For convenience, we suppose $|\mu| < |\nu|$. Then $|\gamma| \leq |\mu| < |\nu|$. We have

$$0 = d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) \geq \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma^*|^2) > \lambda(|\mu|^2 - |\gamma^*|^2) \geq 0, \tag{26}$$

since $\lambda > 0$. Thus we have a contradiction.

Since $\gamma^* \in \Gamma_{\leq}(\mu, \nu)$, we have $\gamma^* \in \Gamma(\mu, \nu)$. Thus, we have

$$\begin{aligned}
0 &\leq d_{FPGW}(\mathbb{X}, \mathbb{Y}) \\
&\leq \int_{(X\times Y)^2} \omega_1 \frac{\|x-y\|^q}{|\gamma|} + \omega_2 |d_X^r(x,x') - d_Y^r(y,y')|^q \, d\gamma^*(x,y) d\gamma^*(x',y') \\
&= \int_{(X\times Y)^2} \omega_1 \frac{\|x-y\|^q}{|\gamma|} + \omega_2 |d_X^r(x,x') - d_Y^r(y,y')|^q + \lambda \left(\frac{|\mu|^2}{|\gamma|^2} + \frac{|\nu|^2}{|\gamma|^2} - 2 \right) d\gamma^*(x,y) d\gamma^*(x',y') \\
&= d_{FPGW}(\mathbb{X}, \mathbb{Y}) \\
&= 0.
\end{aligned}$$

That is, $d_{FPGW}(\mathbb{X}, \mathbb{Y}) = 0$. Since $\omega_2 > 0$, by Proposition 5.2 in Vayer et al. (2020) or Theorem 6.4 in Titouan et al. (2019), there exists measure preserving isometry $\phi : X \to Y$ and thus we have $\mathbb{X} \sim \mathbb{Y}$ and we complete the proof.

D.4 Proof of statement (3) in Theorem 2.

Choose mm-spaces $\mathbb{S} = (S, d_S, \sigma), \mathbb{X} = (X, d_X, \mu), \mathbb{Y} = (Y, d_Y, \nu)$. In this section, we will prove the triangle inequality

$$d_{FPGW,\lambda}(\mathbb{X}, \mathbb{Y}) \leq d_{FPGW,\lambda}(\mathbb{S}, \mathbb{X}) + d_{FPGW,\lambda}(\mathbb{S}, \mathbb{Y}).$$

First, introduce auxiliary points $\hat{\infty}_0, \hat{\infty}_1, \hat{\infty}_2$ and set

$$\begin{cases} \hat{S} &= S \cup \{\hat{\infty}_0, \hat{\infty}_1, \hat{\infty}_2\}, \\ \hat{X} &= X \cup \{\hat{\infty}_0, \hat{\infty}_1, \hat{\infty}_2\}, \\ \hat{Y} &= Y \cup \{\hat{\infty}_0, \hat{\infty}_1, \hat{\infty}_2\}. \end{cases}$$

Define $\hat{\sigma}, \hat{\mu}, \hat{\nu}$ as follows:

$$\begin{cases} \hat{\sigma} &= \sigma + |\mu|\delta_{\hat{\infty}_1} + |\nu|\delta_{\hat{\infty}_2}, \\ \hat{\mu} &= \mu + |\sigma|\delta_{\hat{\infty}_0} + |\nu|\delta_{\hat{\infty}_2}, \\ \hat{\nu} &= \nu + |\sigma|\delta_{\hat{\infty}_0} + |\mu|\delta_{\hat{\infty}_1}. \end{cases} \tag{27}$$

Next, we define $d_{\hat{S}} : \hat{S}^2 \to \mathbb{R} \cup \{\infty\}$ as follows:

$$d_{\hat{S}}(s, s') = \begin{cases} d_S(s, s') & \text{if } (s, s') \in S^2, \\ \infty & \text{elsewhere.} \end{cases} \tag{28}$$

Note, $d_{\hat{S}}(\cdot, \cdot)$ is not a rigorous metric in $\hat{S}$ since we allow $d_{\hat{S}} = \infty$, $d_{\hat{X}}, d_{\hat{Y}}$ are defined similarly.

Then, we define the following mapping $L_\lambda : (\mathbb{R} \cup \{\infty\}) \times (\mathbb{R} \cup \{\infty\}) \to \mathbb{R}_+$:

$$L_\lambda^q(\mathrm{r}_1, \mathrm{r}_2) = \begin{cases} |\mathrm{r}_1 - \mathrm{r}_2|^q & \text{if } \mathrm{r}_1, \mathrm{r}_2 < \infty, \\ \lambda/\omega_2 & \text{if } \mathrm{r}_1 = \infty, \mathrm{r}_2 < \infty \text{ or vice versa,} \\ 0 & \text{if } \mathrm{r}_1 = \mathrm{r}_2 = \infty; \end{cases} \tag{29}$$

and mapping $\hat{D} : \mathbb{R}^d \cup \{\hat{\infty}_i : i \in [0:2]\} \to \mathbb{R}$:

$$\hat{D}^q(x,y) := \begin{cases} \|x-y\|^q & \text{if } x, y \in \mathbb{R}^d \\ 0 & \text{elsewhere.} \end{cases} \tag{30}$$

Finally, we define the following mappings:

$$\begin{aligned}
&\Gamma_{\leq}(\sigma,\mu) \ni \gamma^{01} \mapsto \hat{\gamma}^{01} \in \Gamma(\hat{\sigma},\hat{\mu}),\\
&\hat{\gamma}^{01} := \gamma^{01} + (\sigma - \gamma^{01}_1) \otimes \delta_{\hat{\infty}_0} + \delta_{\hat{\infty}_1} \otimes (\mu - \gamma^{01}_2) + |\gamma|\delta_{\hat{\infty}_1,\hat{\infty}_0} + |\nu|\delta_{\hat{\infty}_2,\hat{\infty}_2};\\
&\Gamma_{\leq}(\sigma,\nu) \ni \gamma^{02} \mapsto \hat{\gamma}^{02} \in \Gamma(\hat{\sigma},\hat{\nu}),\\
&\hat{\gamma}^{02} := \gamma^{02} + (\sigma - \gamma^{02}_1) \otimes \delta_{\hat{\infty}_0} + \delta_{\hat{\infty}_2} \otimes (\nu - \gamma^{02}_2) + |\gamma|\delta_{\hat{\infty}_2,\hat{\infty}_0} + |\mu|\delta_{\hat{\infty}_1,\hat{\infty}_1};\\
&\Gamma_{\leq}(\mu,\nu) \ni \gamma^{12} \mapsto \hat{\gamma}^{12} \in \Gamma(\hat{\mu},\hat{\nu}),\\
&\hat{\gamma}^{12} := \gamma^{12} + (\mu - \gamma^{12}_1) \otimes \delta_{\hat{\infty}_1} + \delta_{\hat{\infty}_2} \otimes (\nu - \gamma^{12}_2) + |\gamma|\delta_{\hat{\infty}_2,\hat{\infty}_1} + |\mu|\delta_{\hat{\infty}_0,\hat{\infty}_0}.
\end{aligned} \tag{31}$$

Remark 5. It is straightforward to verify that the above mappings are well-defined. In addition, we can observe that, for each $\gamma^{01} \in \Gamma_{\leq}(\sigma,\mu), \gamma^{02} \in \Gamma_{\leq}(\sigma,\nu), \gamma^{12} \in \Gamma_{\leq}(\mu,\nu)$,

$$\hat{\gamma}^{01}(\{\hat{\infty}_2\} \times X) = \hat{\gamma}^{01}(S \times \{\hat{\infty}_2\}) = 0, \tag{32}$$

$$\hat{\gamma}^{02}(\{\hat{\infty}_1\} \times Y) = \hat{\gamma}^{02}(S \times \{\hat{\infty}_1\}) = 0, \tag{33}$$

$$\hat{\gamma}^{12}(\{\hat{\infty}_0\} \times Y) = \hat{\gamma}^{12}(X \times \{\hat{\infty}_0\}) = 0.$$

Based on these concepts, we define the following fused-GW variant problem:

$$\hat{d}_{FGW,\lambda}(\hat{\mathbb{X}},\hat{\mathbb{Y}}) := \inf_{\hat{\gamma}\in\Gamma(\hat{\mu},\hat{\nu})} \int_{(\hat{X}\times\hat{Y})^2} \omega_1 \frac{1}{c}\hat{D}^q(x,y) + \omega_2 L^q_\lambda(d^r_{\hat{X}}(x,x'), d^r_{\hat{Y}}(y,y'))d\hat{\gamma}(x,y)d\hat{\gamma}(x',y'). \tag{34}$$

where constant $c = |\sigma| + |\mu| + |\nu|$.

Proposition 5. *For each $\gamma^{12} \in \Gamma(\mu,\nu)$, construct $\hat{\gamma}^{12} \in \Gamma(\hat{\mu},\hat{\nu})$, we have:*

$$\begin{aligned}
&\int_{(X\times Y)^2} \left(\omega_1 \frac{\|x-y\|^q}{|\gamma^{12}|} + \omega_2 \left|d^r_X(x,x') - d^r_Y(y,y')\right|^q + \lambda\left(\frac{|\mu|}{|\gamma^{12}|} + \frac{|\nu|}{|\gamma^{12}|} - 2\right)\right) d(\gamma^{12})^{\otimes 2}\\
&= \int_{(\hat{X}\times\hat{Y})^2} \left(\omega_1 \frac{D^q(x,y)}{c} + \omega_2 L^q_\lambda(d_{\hat{X}}(x,x'), d_{\hat{Y}}(y,y'))\right) d(\hat{\gamma}^{12})^{\otimes 2}
\end{aligned} \tag{35}$$

Furthermore, we have:

$$d_{FPGW,\lambda}(\mathbb{X},\mathbb{Y}) = \hat{d}_{FGW,\lambda}(\hat{\mathbb{X}},\hat{\mathbb{Y}}). \tag{36}$$

Proof. We have:

$$\begin{aligned}
&\int_{(X\times Y)^2} \omega_1 \frac{\|x-y\|^q}{|\gamma^{12}|}, d(\gamma^{12})^{\otimes 2}\\
&= \int_{(X\times Y)} \omega_1 \|x-y\|^q d\gamma^{12}\\
&= \int_{(\hat{X}\times\hat{Y})} \omega_1 D^q(x,y) d\hat{\gamma}^{1,2}\\
&= \int_{(\hat{X}\times\hat{Y})^2} \omega_1 \frac{D^p(x,y)}{|\hat{\gamma}^{12}|} d(\hat{\gamma}^{12})^{\otimes 2}\\
&= \int_{(\hat{X}\times\hat{Y})^2} \omega_1 \frac{1}{c} D^p(x,y), d(\hat{\gamma}^{12})^{\otimes 2}.
\end{aligned} \tag{37}$$

In addition, by Bai et al. (2024) Proposition D.3. We have

$$\begin{aligned}&\int_{(X\times Y)^2}\omega_2|d^r_X(x,x')-d^r_Y(y,y')|^q d(\gamma^{12})^{\otimes 2}+\lambda(|\mu|^2+|\nu|^2-2|\gamma^{12}|^2)\\&=\int_{(\hat X\times\hat Y)^2}\omega_2 D^q_\lambda(d^r_{\hat X}(x,x'),d^r_{\hat Y}(y,y'))d(\hat\gamma^{12})^{\otimes 2}.\end{aligned}\tag{38}$$

Combining the above two equalities, we prove (35).

Now, we prove the second equality. Note, if we merge the three auxiliary points, i.e., $\hat\infty_1=\hat\infty_2=\hat\infty_3$. The optimal value for the above problem (34) is unchanged.

As we merged the points $\hat\infty_1,\hat\infty_2,\hat\infty_3$, by Bai et al. Bai et al. (2023b), the mapping $\gamma^{12}\mapsto\hat\gamma^{12}$ defined in (31) is a bijection. Thus, by (35), we have $\gamma^{12}\in\Gamma_\leq(\mu,\nu)$ is optimal in (22) iff $\hat\gamma^{12}$ is optimal in (34) and we complete the proof. □

Lemma 2. *Choose $\gamma^{01}\in\Gamma_\leq(\sigma,\mu),\gamma^{02}\in\Gamma_\leq(\sigma,\nu),\gamma^{12}\in\Gamma_\leq(\mu,\nu)$ and construct $\hat\gamma^{01},\hat\gamma^{02},\hat\gamma^{12}$. Then there exists $\hat\gamma\in\mathcal{M}_+(\hat S\times\hat X\times\hat Y)$ such that:*

$$(\pi_{0,1})_\#\hat\gamma=\hat\gamma^{01},\tag{39}$$

$$(\pi_{0,2})_\#\hat\gamma=\hat\gamma^{02},\tag{40}$$

$$\gamma(A_i)=0,\forall i=0,1,2 \textit{ where } A_i=\{\hat\infty_i\}\times X\times Y.\tag{41}$$

Proof. By *gluing lemma* (see Lemma 5.5 Santambrogio (2015)), there exists $\hat\gamma\in\mathcal{M}_+(\hat S\times\hat X\times\hat Y)$, such that (39),(40) are satisfied. For the third property, we have:

$$\begin{aligned}&\hat\gamma(A_0)\leq\hat\gamma(\{\infty_0\}\times\hat X\times\hat Y)=\hat\sigma(\{\infty_0\})=0\qquad\text{by definition (27) of }\hat\sigma\ ,\\&\hat\gamma(A_1)\leq\hat\gamma(\{\infty_1\}\times\hat X\times Y)=\hat\gamma^{02}(\{\hat\infty_1\times Y\})=0\qquad\text{by (33)},\\&\hat\gamma(A_2)\leq\hat\gamma(\{\infty_2\}\times X\times\hat Y)=\hat\gamma^{01}(\{\hat\infty_2\times X\})=0\qquad\text{by (32)}.\end{aligned}$$

And we complete the proof. □

Now, we demonstrate the proof of triangle inequality for fused-PGW distance (22).

Proof of Theorem 2 (3). Note, by the Proposition 5, the triangle inequality for fused-PGW distance (22) is equivalent to show

$$\hat d_{FGW,\lambda}(\hat{\mathbb{X}},\hat{\mathbb{Y}})\leq 2^{q-1}(\hat d_{FGW,\lambda}(\hat{\mathbb{S}},\hat{\mathbb{X}})+\hat d_{FGW,\lambda}(\hat{\mathbb{S}},\hat{\mathbb{Y}})).\tag{42}$$

Choose optimal transportation plans $\gamma^{01},\gamma^{02},\gamma^{12}$ for fused PGW problems $d_{FPGW,r,\lambda}(\mathbb{S},\mathbb{X}),d_{FPGW,r,\lambda}(\mathbb{S},\mathbb{Y})$ and $d_{FPGW,r,\lambda}(\mathbb{X},\mathbb{Y})$ respectively. We construct the corresponding $\hat\gamma^{01},\hat\gamma^{02},\hat\gamma^{12}$. By the Proposition 5, $\hat\gamma^{01},\hat\gamma^{02},\hat\gamma^{12}$ are optimal.

Choose $\hat\gamma$ in lemma 2. We have:

$$\begin{aligned}&\hat d_{FGW,\lambda}(\mathbb{X},\mathbb{Y})\\&=\int_{(\hat X\times\hat Y)^2}\omega_1\frac{D^q(x,y)}{c}+\omega_2L^q_\lambda(d_{\hat X}(x,x'),d_{\hat Y}(y,y'))d\hat\gamma^{12}(x,y)d\hat\gamma^{12}(x',y')\\&\leq\int_{(\hat S\times\hat X\times\hat Y)^2}\omega_1\frac{D^q(x,y)}{c}+\omega_2L^q_\lambda(x,y)d\hat\gamma(s,x,y)d\hat\gamma(s',x',y')\\&=\underbrace{\int_{(\hat S\times\hat X\times\hat Y)}\omega_1D^q(x,y)d\hat\gamma(s,x,y)}_{A}+\underbrace{\int_{(\hat S\times\hat X\times\hat Y)^2}\omega_2L^q_\lambda(x,y)d\hat\gamma(s,x,y)d\hat\gamma(s',x',y')}_{B}.\end{aligned}$$

To bound A, we consider the case $D(x,y) > D(s,x) + D(s,y)$ for some $(s,x,y) \in \hat{S} \times \hat{X} \times \hat{Y}$. By definition of D, we have $s \in \{\hat{\infty}_i : i \in [0:2]\}, x \in X, y \in Y$. That is, $(s,x,y) \in \bigcup_{i=0}^2 A_i$. By lemma 2, we have $\hat{\gamma}(\bigcup_{i=0}^2 A_i) = 0$. That is, this case has a measure of 0.

Thus, we have

$$\begin{aligned}
A &\leq \int_{\hat{S}\times\hat{X}\times\hat{Y}} \omega_1 (D(s,x) + D(s,y))^q d\hat{\gamma}(s,x,y) \\
&\leq \int_{\hat{S}\times\hat{X}\times\hat{Y}} \omega_1 (2^{q-1} D(s,x) + 2^{q-2} D(s,y)) d\hat{\gamma}(s,x,y) \\
&= \int_{\hat{S}\times\hat{X}\times\hat{Y}} \omega_1 2^{q-1} D(s,x) d\hat{\gamma}(s,x,y) + \int_{\hat{S}\times\hat{X}\times\hat{Y}} \omega_1 2^{q-2} D(s,y) d\hat{\gamma}(s,x,y) \\
&= 2^{q-1} \int_{\hat{S}\times\hat{X}} \omega_1 \frac{D(s,x)}{c} d\hat{\gamma}^{01}(s,x) + 2^{q-1} \int_{\hat{S}\times\hat{Y}} \omega_1 \frac{D(s,y)}{c} d\hat{\gamma}^{02}(s,y). \qquad (43)
\end{aligned}$$

where the second inequality follows from the fact

$$(a+b)^q \leq 2^{q-1} a^q + 2^{q-2} b^q, \forall a, b \geq 0. \qquad (44)$$

Now we bounde the term B. Proposition D.4 in Bai et al. (2024), we have

$$\begin{aligned}
B &\leq \int_{(\hat{S}\times\hat{X}\times\hat{Y})^2} \omega_2 (L_\lambda(s,x) + L_\lambda(s,y))^q d\hat{\gamma}(s,x,y) d\hat{\gamma}(s',x',y') \\
&\leq \int_{(\hat{S}\times\hat{X}\times\hat{Y})^2} \omega_2 \left(2^{q-1} L_\lambda^q(s,x) + 2^{q-1} L_\lambda^q(s,y)\right) d\hat{\gamma}(s,x,y) d\hat{\gamma}(s',x',y') \\
&= 2^{q-1} \int_{(\hat{S}\times\hat{X})^2} \omega_2 L_\lambda^q(s,x) d\hat{\gamma}^{01}(s,x) d\hat{\gamma}^{01}(s',x') + 2^{q-1} \int_{(\hat{S}\times\hat{Y})^2} \omega_2 L_\lambda^q(s,y) d\hat{\gamma}^{02}(s,y) d\hat{\gamma}^{02}(s',y').
\end{aligned} \qquad (45)$$

where the second inequality holds from (44). Combining (43) and (45), we prove the inequality (42) and we complete the proof. □

D.5 Future's direction.

Note, the general version for (the p-th power of) fused-Gromov Wasserstein distance is defined by

$$d_{FGW,p}^p(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int \left(\omega_1 \|x-y\|^q + |d_X(x,x') - d_Y(y,y')|^q\right)^p d\gamma^{\otimes 2}. \qquad (46)$$

Inspired by the above formulation, we propose the following generalized fused-partial Gromov Wasserstein distance:

$$d_{FGW,p}^p(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int \left(\omega_1 \frac{\|x-y\|^q}{|\gamma|} + |d_X(x,x') - d_Y(y,y')|^q + \lambda(\frac{|\mu|^2}{|\gamma|^2} + \frac{|\nu|^2}{|\gamma|^2} - 2)\right)^p d\gamma^{\otimes 2}. \qquad (47)$$

The fused-PGW distance defined in (22) can be treated as a special case of this general formulation by setting $p = 1$. In our conjecture, a similar (semi-)metric property proposed in Theorem 2 holds for the above general form. We leave the theoretical study of the metric property for our future work.

E Frank Wolf Algorithm for the Fused Mass-constraint Partial Gromov Wasserstein problem

In discrete setting, the FMPGW problem (14) becomes the following:

$$FMPGW_\rho(\mathbb{X},\mathbb{Y}) = \min_{\gamma\in\gamma^\rho_\leq(p,q)} \underbrace{\omega_1\langle C,\gamma\rangle + \omega_2\langle M\circ\gamma,\gamma\rangle}_{\mathcal{L}_{C,M}} \tag{48}$$

where $\Gamma^\rho_\leq(\mathrm{p},\mathrm{q}) := \{\gamma\in\mathbb{R}^{n\times m}_+ : \gamma 1_m \leq \mathrm{p}, \gamma^\top 1_n \leq \mathrm{q}, |\gamma|\geq \rho\}$, $C = [C(x_i,y_j)]_{i\in[1:n],j\in[1:m]}$, $M_{i,j,i',j'} := L(C^X,C^Y) := [L(C^X_{i,i'},C^Y_{j,j'})]_{i,i\in[1:n],j,j'\in[1:n]}$, $M\circ\gamma = [\langle M[i,j,\cdot,\cdot],\gamma\rangle]_{i,j}$.

Similar to the Fused Gromov-Wasserstein problem, we propose the following Frank-Wolfe algorithm as a solver: The above problem will be solved iteratively. In every iteration, say k, we will adapt the following steps:

Step 1. Gradient computation.

Suppose $\gamma^{(k-1)}$ is the transportation plan in the previous iteration; it is straightforward to verify:

$$\nabla\mathcal{L}_{C,M}(\gamma) = \omega_1 C + \omega_2((M+M^\top)\circ\gamma),$$

where $M^\top = [M_{i',j',i,j}]_{i,i'\in[1:n],j,j'\in[1:m]}\in\mathbb{R}^{n\times m\times n\times m}$. Next, we aim to find $\gamma\in\Gamma^\rho_\leq(\mu,\nu)$ for the following problem:

$$\gamma^{(k)\prime} := \arg\min_{\gamma\in\Gamma^\rho_\leq(\mu,\nu)} \langle\nabla\mathcal{L}_{C,M}(\gamma^{(k-1)}),\gamma\rangle. \tag{49}$$

which is a mass-constraint partial OT problem.

Step 2. linear search algorithm. In this step, we aim to find the optimal step size $\alpha^{(k)}\in[0,1]$. In particular,

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

Let $\delta\gamma = \gamma^{(k)\prime} - \gamma^{(k-1)}$, the above problem is essentially quadratic problem:

$$\begin{aligned}
\mathcal{L}((1-\alpha)\gamma^{(k-1)} + \alpha\gamma^{(k)\prime}) &= \alpha^2 \underbrace{\langle\omega_2 M\circ\delta\gamma,\delta\gamma\rangle}_{a} \\
&+ \alpha\underbrace{\langle\omega_2(M+M^\top)\circ\gamma^{(k-1)} + \omega_1 C,\delta\gamma\rangle}_{b} \\
&+ \underbrace{\langle\omega_2 M\circ\gamma^{(k-1)} + \omega_1 C,\gamma\rangle}_{c}
\end{aligned} \tag{50}$$

and α^* is given by

$$\alpha^* = \begin{cases} 1 & \text{if } a\leq 0, a+b\leq 0, \\ 0 & \text{if } a\leq 0, a+b>0 \\ \mathrm{clip}(\frac{-b}{2a},[0,1]), & \text{if } a>0 \end{cases}. \tag{51}$$

Then $\gamma^{(k)} = (1-\alpha^*)\gamma^{(k-1)} + \alpha^*\gamma^{(k)\prime}$.

In the next section, we provide a detailed introduction to the derivation of the FW algorithms.

Algorithm 1 Frank-Wolfe Algorithm for FMPGW/FPGW

Input: $C \in \mathbb{R}^{n\times m}, C^X \in \mathbb{R}^{n\times n}, C^Y \in \mathbb{R}^{m\times m}, \mathrm{p} \in \mathbb{R}_+^n, \mathrm{q} \in \mathbb{R}_+^m$, $\omega_1 \in [0,1], \rho \in [0, \min(|\mathrm{p}|, |\mathrm{q}|)]$.
Output: $\gamma^{(final)}$
for $k = 1, 2, \ldots$ **do**
  $G^{(k)} \leftarrow \omega_1 C + \omega_2(M + M^\top) \circ \gamma^{(k)}$ // Compute gradient
  $\gamma^{(k)\prime} \leftarrow \arg\min_{\gamma \in \Gamma_\leq^\rho(\mathrm{p},\mathrm{q})} \langle G^{(k)}, \gamma \rangle_F$ // Solve the POT problem.
  Compute $\alpha^{(k)} \in [0,1]$ via (60) // Line Search
  $\gamma^{(k+1)} \leftarrow (1-\alpha^{(k)})\gamma^{(k)} + \alpha^{(k)}\gamma^{(k)\prime}$// Update γ
  if convergence, break
end for
$\gamma^{(final)} \leftarrow \gamma^{(k)}$

F Gradient computation of FW algorithms

In this section, we provide a detailed discussion of the gradient computation in fused-MPGW and fused-PGW.

F.1 Basics in Tensor product

In this section, we introduce some fundamental results for tensor computation. Suppose $M \in \mathbb{R}^{n\times m\times n\times m}$. We define the **Transportation of tensor**, denoted as $M^\top$, as:

$$M^\top_{i,j,i',j'} = M_{i',j',i,j}, \tag{52}$$

and we say M is symmetric if $M = M^\top$.

It is straightforward to verify the following:

Proposition 6. *Given tensor $M \in \mathbb{R}^{n\times m\times n\times m}$, then we have:*

- $(M^\top)^\top = M$.
- $M^\top \circ \gamma = [\sum_{i',j'} M_{i',j',i,j}\gamma_{i,j}]_{i,j\in[1:n]\times[1:m]}$.

Proof. The first item follows from the definition of $M^\top$. For the second statement, pick $(i,j) \in [1:n] \times [1:m]$, we have

$$\begin{aligned}
&\sum_{i',j'} M_{i',j',i,j}\gamma_{i',j'} \\
&= \sum_{i',j'} M^\top_{i,j,i',j'}\gamma_{i',j'} \\
&= \langle M^\top \circ \gamma \rangle.
\end{aligned}$$

□

Therefore, we have:

Proposition 7. *The gradient for $\mathcal{L}_{C,M}(\gamma)$ in Fused-PGW (14) is given by:*

$$\nabla \mathcal{L}_{C,M}(\gamma) = \omega_1 C + \omega_2(M \circ \gamma + M^\top \circ \gamma). \tag{53}$$

.

Similarly, the gradient for $\mathcal{L}_{C,M-2\hat{\lambda}}$ in fused-PGW (15) is given by:

$$\nabla\mathcal{L}_{C,M-2\hat{\lambda}}(\gamma) = \omega_1 C + \omega_2((M-2\hat{\lambda})\circ\gamma + (M-2\hat{\lambda})^\top \circ \gamma). \tag{54}$$

.

Proof. Pick $(i,j) \in [1:n]\times[1:m]$, we have:

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_{ij}} L_{C,M}(\gamma) \\
&= \frac{\partial}{\partial \gamma_{ij}} \sum_{i,j,i',j'} \omega_1 C_{i,j}\gamma_{i,j} + \omega_2 \sum_{i,j,i',j'} M_{i,j,i',j'}\gamma_{i,j}\gamma_{i',j'} \\
&= \omega_1 C_{i,j} + \omega_2(\sum_{i',j'} M_{i,j,i',j'}\gamma_{i,j}\gamma_{i',j'} + \sum_{i',j'} M_{i',j',i,j}\gamma_{i,j}\gamma_{i',j'})
\end{aligned}$$

Therefore, $\nabla L_{C,M}(\gamma) = \omega_1 C + \omega_2(M\circ\gamma + M^\top\circ\gamma)$ and we complete the proof. The gradient for Fused-PGW (15) can be derived similarly. □

At the end of this subsection, we discuss the computation of $M\circ\gamma$ and $M^\top\gamma$.
In general, the computation cost for $M\circ\gamma$ is n^2m^2. However, if the cost function L satisfies:

$$L(\mathrm{r}_1,\mathrm{r}_2) = f_1(\mathrm{r}_1) + f_2(\mathrm{r}_2) - h_1(\mathrm{r}_1)h_2(\mathrm{r}_2), \tag{55}$$

$$M\circ\gamma = f_1(C^X)\gamma_1 1_m^\top + 1_n\gamma_2^\top f_2(C^Y) - h_1(C^X)\gamma h_2(C^Y), \tag{56}$$

and the corresponding complexity is $\mathcal{O}(n^2+m^2)$ (see e.g. Peyré et al. (2016); Chapel et al. (2020); Bai et al. (2024) for details.)

Therefore, we have the following:

Lemma 3. *Suppose $M = [L(d_X(x_i,x_{i'}), d_Y(y_j,y_{j'}))]$, then we have:*

$$(M^\top)_{i,j,i',j'} = f_1((C^X)^\top_{i,i'}) + f_2((C^Y)^\top_{j,j'}) - h_1((C^X)_{i,i'})h_2((C^Y)_{j,j'}).$$

It directly implies:

$$M^\top\circ\gamma = f_1((C^X)^\top)\gamma_1 1_m^\top + 1_n\gamma_2^\top f_2((C^Y)^\top) - h_1((C^X)^\top)\gamma h_2((C^Y)^\top).$$

Proof. We have:

$$\begin{aligned}
M^\top_{i,j,i',j'} &= M_{i',j',i,j} \\
&= f_1(C^X_{i',i}) + f_2(C^Y_{j',j}) - h_1(C^X_{i',i})h_2(C^Y_{j',j})) \\
&= f_1((C^X)^\top_{i',i}) + f_2((C^Y)^\top_{j',j}) - h_1((C^X)^\top_{i',i})h_2((C^Y)^\top_{j',j}).)
\end{aligned}$$

And we complete the proof. □

F.2 Gradient in fused-MPGW.

As we discussed in previous section, after iteration $k-1$, the gradient is given by $\nabla\mathcal{L}(\gamma) = \omega_1 C + \omega_2(M\circ\gamma + M^\top\circ\gamma)$, where $M_{i,j,i',j'} = (L(C^X_{i,i'}, C^Y_{j,j'})) \in \mathbb{R}_+^{n\times m\times n\times m}$. Furthermore, the computational complexity for $M\circ\gamma := [\langle M[i,j,:,:],\gamma\rangle]$ can be improved to be $\mathcal{O}(n^2+m^2)$.

Based on the gradient, we aim to solve the following to update the transportation plan in iteration k:

$$\gamma^{(k)\prime} = \arg\min_{\gamma\in\Gamma^\rho_\leq(\mu,\nu)} \langle \underbrace{\omega_1 C + \omega_2(M\circ\gamma^{(k-1)} + M^\top\gamma^{(k-1)})}_{\mathcal{C}}, \gamma\rangle. \tag{57}$$

The above problem is essentially the partial OT problem, where the cost function is defined by $\mathcal{C}$.

F.3 Gradient of Fused-PGW.

Similarly, after iteration $k-1$, the gradient of cost $\mathcal{L}_{C,M-2\hat{\lambda}}$ in FPGW with respect to γ is given by

$$\nabla\mathcal{L}(\gamma) = \omega_1 C + \omega_2((M-2\hat{\lambda})\circ\gamma^{(k-1)} + (M^\top - 2\hat{\lambda})\circ\gamma^{(k-1)}).$$

We aim to solve the following problem:

$$\begin{aligned}
&\min_{\gamma\in\Gamma_\leq(\mu,\nu)} \langle \omega_1 C + \omega_2((M-2\hat{\lambda})\circ\gamma^{(k-1)} + (M^\top - 2\hat{\lambda})\circ\gamma^{(k-1)}),\gamma\rangle \\
&= \min_{\gamma\in\Gamma_\leq(\mu,\nu)} \langle \underbrace{\omega_1 C + \omega_2(M\circ\gamma^{(k-1)} + M^\top\circ\gamma^{(k-1)})}_{\mathcal{C}},\gamma\rangle + \lambda|\gamma^{(k-1)}|(|\mu|+|\nu|-2|\gamma|) - \underbrace{\lambda|\gamma^{(k-1)}|(|\mu|+|\nu|)}_{\text{constant}}.
\end{aligned}$$

Note, if we ignore the constant term, the above problem is the partial OT problem and can be solved by Bonneel et al. (2011) or Bai et al. (2023b).

G Line search algorithm

The line search for Fused-MPGW is defined as

$$\min_{\alpha\in[0,1]} \mathcal{L}((1-\alpha)\gamma^{(k-1)} + \alpha\gamma^{(k)\prime}), \tag{58}$$

where $\mathcal{L}(\gamma) = \omega_1\langle C,\gamma\rangle + \omega_2\langle M,\gamma^{\otimes 2}\rangle$. Let $\delta\gamma = \gamma^{(k)\prime} - \gamma^{(k-1)}$. The above problem is essentially a quadratic problem with respect to α:

$$\begin{aligned}
&\mathcal{L}((1-\alpha)\gamma^{(k-1)} + \alpha\gamma^{(k)\prime}) \\
&= \mathcal{L}(\gamma^{(k-1)} + \alpha\delta\gamma) \\
&= \omega_1\langle C, (\gamma^{(k-1)} + \alpha\delta\gamma)\rangle + \omega_2\langle M\circ(\gamma^{(k-1)} + \alpha\delta\gamma), (\gamma^{(k-1)} + \alpha\delta\gamma)\rangle \\
&= \alpha^2 \underbrace{\omega_2\langle M\circ\delta\gamma,\delta\gamma\rangle}_{a} + \alpha \underbrace{\langle \omega_2 M\circ\gamma^{(k-1)} + \omega_1 C, \delta\gamma\rangle + \langle \omega_2 M\circ\delta\gamma, \gamma^{(k-1)}\rangle}_{b} + \underbrace{\langle \omega_2 M\circ\gamma^{(k-1)} + \omega_1 C, \gamma^{(k-1)}\rangle}_{c},
\end{aligned} \tag{59}$$

and α^* is given by

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

Next, we simplify the term b in the above formula. We first introduce the following lemma:

Lemma 4. *Choose $\gamma^1, \gamma^2 \in \mathbb{R}^{n\times m}, M \in \mathbb{R}^{n\times m\times n\times m}$, we have:*

$$\langle M \circ \gamma^1, \gamma^2\rangle = \langle M^\top \circ \gamma^2, \gamma^1\rangle. \tag{61}$$

Proof. We have

$$\begin{aligned}\langle M \circ \gamma^1, \gamma^2\rangle &= \sum_{i,j}\sum_{i',j'} M_{i,j,i',j'}\gamma^1_{i',j'}\gamma^2_{i,j} \\ &= \sum_{i',j'}\sum_{i,j} M^\top_{i',j',i,j}\gamma^2_{i,j}\gamma^1_{i,j} \\ &= \langle M^\top \circ \gamma^2, \gamma^1\rangle.\end{aligned} \tag{62}$$

□

Therefore, the term b can be further simplified as

$$\begin{aligned}b &= \omega_1\langle C, \delta\gamma\rangle + \omega_2(\langle M\circ\gamma^{(k-1)}, \delta\gamma\rangle + \langle M^\top \circ \gamma^{(k-1)}, \delta\gamma\rangle) \\ &= \langle \underbrace{\omega_1 C + M\circ\gamma^{(k-1)} + M^\top\circ\gamma^{(k-1)}}_{\nabla L_{C,M}(\gamma)}, \delta\gamma\rangle.\end{aligned} \tag{63}$$

That is, we can directly adapt the gradient obtained before to compute b and improve the computation efficiency.

Remark 6. When we select quadratic cost, i.e., $L(\mathrm{r}_1, \mathrm{r}_2) := |\mathrm{r}_1 - \mathrm{r}_2|^2$, we can set $f_1(\mathrm{r}_1) = \mathrm{r}_1^2, f_2(\mathrm{r}_2) = \mathrm{r}_2^2, h_1(\mathrm{r}_1) = 2\mathrm{r}_1, h_2(\mathrm{r}_2) = \mathrm{r}_2$, then $L(\mathrm{r}_1, \mathrm{r}_2)$ satisfies (55). Furthermore, suppose the problem is in the balanced fused-GW setting, i.e., $\rho = |\mu| = |\nu| = 1$. We have:

$$\begin{aligned}a &= \omega_2\langle M\circ\delta\gamma, \delta\gamma\rangle \\ &= \omega_2\langle f_1(C^X)\underbrace{\delta\gamma_1}_{0_n}1_m^\top + \underbrace{1_n\delta\gamma_2^\top}_{0_m} f_2(C^Y) - h_1(C^X)\gamma h_2(C^Y), \delta\gamma\rangle \\ &= -2\omega_2\langle C^X\delta\gamma C^Y, \delta\gamma\rangle.\end{aligned} \tag{64}$$

Similarly,

$$\begin{aligned}b &= \langle \omega_2(M + M^\top)\circ\gamma^{(k-1)} + \omega_1 C, \delta\gamma\rangle \\ &= \omega_1\langle C, \delta\gamma\rangle + \omega_2\langle (M+M^\top)\circ\gamma^{(k-1)}, \delta\gamma\rangle \\ &= \omega_1\langle C, \delta\gamma\rangle + \omega_2\langle (M+M^\top)\circ\delta\gamma, \gamma^{(k-1)}\rangle \\ &= \omega_1\langle C, \delta\gamma\rangle - 2\omega_2\langle C^X\delta\gamma C^Y, \gamma^{(k-1)}\rangle - 2\omega_2\langle (C^X)^\top\delta\gamma(C^Y)^\top, \gamma^{(k-1)}\rangle + \omega_2\langle c_{C^X,C^Y}, \gamma^{(k-1)}\rangle \\ &\quad + \omega_2\langle c_{(C^X)^\top,(C^Y)^\top}, \gamma^{(k-1)}\rangle.\end{aligned} \tag{65}$$

where $c_{C^X,C^Y} = f_1(C^X)\mathrm{p}1_m^\top + 1_m\mathrm{q}^\top f_2(C^Y)$. Thus, the above formulations recover the line search algorithm (see algorithm 2) in Peyré et al. (2016).

Next, we discuss the line search step in fused-PGW (11).

Replacing M by $M - 2\hat{\lambda}$ in (59), the solution is obtained by (60).

H Convergence Analysis

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.

H.1 Fused-MPGW

Consider the minimization problems

$$\min_{\gamma\in\Gamma^{\rho}_{\leq}(\mathrm{p},\mathrm{q})} \mathcal{L}_{C,M}(\gamma) := \omega_1\langle C,\gamma\rangle + \omega_2\langle M\circ\gamma,\gamma\rangle$$

that corresponds to the discrete fused-PGW problem (15).

The objective functions

$$\gamma \mapsto \mathcal{L}_{\hat{M}}(\gamma) = \omega_1\langle C,\gamma\rangle + \omega_2\langle M\circ\gamma,\gamma\rangle,$$

is non-convex in general. However, the constraint set $\Gamma^{\rho}_{\leq}(p,q)$ are convex and compact on $\mathbb{R}^{n\times m}$ (see Proposition 3.)

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

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

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 (66) 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-Wolfe gap:

$$g_K := \min_{1\leq k\leq K} g_k \leq \frac{\max\{2L_1, \mathrm{Lip}\cdot(\mathrm{diam}(\Gamma^{\rho}_{\leq}(\mathrm{p},\mathrm{q})))^2\}}{\sqrt{K}}, \tag{67}$$

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 sub-optimal bound for the initialization $\gamma^{(1)}$ of the algorithm; Lip is the Lipschitz constant of function $\gamma\mapsto\nabla\mathcal{L}_{\tilde{M}}$; and

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

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 in $\tilde{g}_K$ 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.

Next, we will continue to simplify the upper bound (67). We first introduce the following fundamental results:

Lemma 5. *In the discrete setting, we have*

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

Proof. Choose $\gamma,\gamma'\in\Gamma^{\rho}_{\leq}(p,q)$. We apply the property

$$(a-b)^2 \leq 2a^2+2b^2, \quad \forall a,b\in\mathbb{R}. \tag{69}$$

and obtain

$$
\begin{aligned}
\|\gamma-\gamma'\|_F^2 &= \sum_{i,j}^{n,m} |\gamma_{i,j}-\gamma'_{i,j}|^2 \\
&\leq \sum_{i,j}^{n,m} 2|\gamma_{i,j}|^2 + 2|\gamma'_{i,j}|^2 \\
&\leq 2\left[\left(\sum_{i,j}^{n,m} \gamma_{i,j}\right)^2 + \left(\sum_{i,j}^{n,m} \gamma'_{i,j}\right)^2\right] \\
&= 2(|\gamma|^2 + |\gamma'|^2) \\
&= 2(\rho^2+\rho^2) = 4\rho^2,
\end{aligned} \tag{70}
$$

and thus, we complete the proof. □

Lemma 6. *The Lipschitz constant term in* (67) *can be bounded as follows:*

$$
Lip \leq \omega_2 nm \max(|M|)^2. \tag{71}
$$

In particular, when $L(\mathrm{r}_1,\mathrm{r}_2)=|\mathrm{r}_1-\mathrm{r}_2|^p$ where $p\geq 1$, we have $Lip \leq nm(2^{p-1}C^X + 2^{p-1}C^Y)^2$.

Proof. Pick $\gamma,\gamma' \in \Gamma_{\leq}(\mathrm{p},\mathrm{q})$ we have,

$$
\begin{aligned}
&\|\nabla\mathcal{L}_M(\gamma) - \nabla\mathcal{L}_{C,M}(\gamma')\|_F^2 \\
&= \|(\omega_1 C + \omega_2(M\circ\gamma + M^\top\circ\gamma)) - (\omega_1 C + \omega_2(M\circ\gamma' + M^\top\circ\gamma'))\|_F^2 \\
&= 2\omega_2^2\|M\circ(\gamma-\gamma')\|_F^2 + 2\omega_2^2\|M^\top\circ(\gamma-\gamma')\|_F^2 \\
&= 2\omega_2^2\sum_{i,j}\left(\left[M\circ(\gamma-\gamma')\right]_{i,j}\right)^2 + 2\omega_2^2\sum_{i,j}\left(\left[M^\top\circ(\gamma-\gamma')\right]_{i,j}\right)^2 \\
&= 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'} M_{i,j,i',j'}(\gamma_{i',j'}-\gamma'_{i',j'})\right)^2 + 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'} M^\top_{i,j,i',j'}(\gamma_{i',j'}-\gamma'_{i',j'})\right)^2 \\
&\leq 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'} |M_{i,j,i',j'}||\gamma_{i',j'}-\gamma'_{i',j'}|\right)^2 + 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'} |M^\top_{i,j,i',j'}||\gamma_{i',j'}-\gamma'_{i',j'}|\right)^2 \\
&\leq 4\omega_2^2 \underbrace{\max(M)^2}_{A} \cdot \underbrace{\sum_{i,j}^{n,m}\left(\sum_{i',j'}^{n,m}|\gamma_{i',j'}-\gamma'_{i',j'}|\right)^2}_{B}
\end{aligned}
$$

For the second term, we have:

$$
\begin{aligned}
&\sum_{i,j}^{n,m}\left(\sum_{i',j'}^{n,m}|\gamma_{i',j'}-\gamma'_{i',j'}|\right)^2 \\
&\leq \sum_{i,j}^{n,m}\left(nm\sum_{i',j'}^{n,m}\left|\gamma_{i',j'}-\gamma'_{i',j'}\right|^2\right) \\
&\leq n^2m^2\|\gamma-\gamma'\|_F^2
\end{aligned} \tag{72}
$$

For the first term, when $L(\mathrm{r}_1, \mathrm{r}_2) = |\mathrm{r}_1 - \mathrm{r}_2|^p$, from the inequality (44), we have:

$$A = \max M \leq \max\{2^{p-1}(C^X)^p + 2^{p-1}(C^Y)^p\}^2.$$

where

$$\max\{2^{p-1}(C^X)^p + 2^{p-1}(C^Y)^p\} := \max\{\max_{i,i',j',j'} 2^{p-1}(C^X_{i,i'})^p + 2^{p-1}(C^Y_{j,j'})^p\}.$$

Thus we obtain

$$\mathrm{Lip} \leq \frac{\max(2^{p-1}(C^X)^p + 2^{p-1}(C^Y)^p)nm\|\gamma - \gamma'\|_F}{\|\gamma - \gamma'\|_F} = \omega_2 nm \max(2^{p-1}(C^X)^p + 2^{p-1}(C^Y)^p).$$

and we complete the proof. □

Combined the above two lemmas, we derive the convergence rate of the Frank-Wolfe gap (66):

Proposition 8. *When $L(\mathrm{r}_1, \mathrm{r}_2) = |\mathrm{r}_1 - \mathrm{r}_2|^2$ in the PGW problem, the Frank-Wolfe gap of algorithm 1, defined in* (66) *at iteration k satisfies the following:*

$$g_k \leq \frac{\max\left\{2L_1, 4\omega_2\rho^2 \cdot nm(\max\{2(C^X)^2 + 2(C^Y)^2\})\right\}}{\sqrt{k}}. \tag{73}$$

Proof. The proof directly follows from the upper bounds (68),(71) and the inequality (67). □

Remark 7. Note, if the cost function in PGW is defined by $|\mathrm{r}_1 - \mathrm{r}_2|^p$ for some $p \neq 2$, it is straightforward to verify that the upper bound of g_k is obtained by replacing the term $\max((C^X)^2 + (C^Y)^2)$ should be replaced by

$$\max_{i,j,i',j'} \left[2^{p-1}((C^X)^p + (C^Y)^p)\right].$$

Remark 8. From the proposition (8), to achieve an $\epsilon-$accurate solution, the required number of iterations is

$$\frac{\max\left\{2L_1, 4\rho\omega_2 \cdot n^2m^2 \max(\{2(C^X)^2 + 2(C^Y)^2\})\right\}^2}{\epsilon^2}.$$

Remark 9. Note, when $\omega_2 = 1$, this convergence rate is constant to the convergence rate of the FW algorithm for MPGW in Chapel et al. (2020). In addition, the convergence rate is independent of C. The main reason is that the part $\omega_1\langle C, \gamma\rangle$ is linear with respect to γ. Thus, this part does not contribute to the Lipschitz of the mapping $\gamma \mapsto \nabla_M(\mathcal{L})$.

H.2 Convergence of Fused-PGW

Similar to the above, in this subsection, we discuss the convergence rate for algorithm 2. Suppose γ is a solution for the fused-PGW problem (15). In iteration k, we define the gap between γ and the approximation $\gamma^{(k)}$ and the Frank-Wolfe gap:

$$g_k := \max_{\gamma \in \Gamma_{\leq}(p,q)} \langle \nabla \mathcal{L}_{C,M-2\hat{\lambda}}(\gamma^{(k)}), \gamma^{(k)} - \gamma\rangle_F. \tag{74}$$

$$g_K := \min_{1 \leq k \leq K} g_k. \tag{75}$$

Thus, the convergence rate for the FW algorithm for the fused-PGW problem (15) can be bounded by the following proposition:

Proposition 9. *When $L(\mathrm{r}_1, \mathrm{r}_2) = |\mathrm{r}_1 - \mathrm{r}_2|^2$ in the Fused-PGW problem, the Frank-Wolfe gap of algorithm 1,* (66) *at iteration k satisfies the following:*

$$g_K \leq \frac{\max\left\{2L_1, 4\omega_2 \min^2(|p|,|q|)\cdot nm(\max\{2(C^X)^2+2(C^Y)^2, 2\hat{\lambda}\})\right\}}{\sqrt{k}}, \tag{76}$$

where $\max(2(C^X)^2+2(C^Y)^2,2\hat{\lambda}) = \max\{\max_{i,i',j,j'}(2(C^X_{i,i'})^2+2(C^Y_{j,j'})^2, 2\hat{\lambda}\}$.

From Theorem 1 in Lacoste-Julien (2016), we have:

$$g_K \leq \frac{\max\{2L_1, \mathrm{Lip}\cdot(\mathrm{diam}(\Gamma_{\leq}(p,q)))\}}{\sqrt{K}}, \tag{77}$$

where Lip is the Lipschitz constant of function $\gamma \to \nabla_M \mathcal{L}$ (with respect to Fubini norm).

By Bai et al. (2024),

$$\mathrm{diam}(\Gamma_{\leq}(p,q)) \leq 2\min(|p|,|q|). \tag{78}$$

Next, we bound the Lipschitz constant.

Lemma 7. *The Lipschitz constant in* (77) *can be bounded as follows:*

$$\mathrm{Lip} \leq \omega_2 nm \max|M - 2\hat{\lambda}|^2. \tag{79}$$

When $L(\mathrm{r}_1,\mathrm{r}_2) = |\mathrm{r}_1 - \mathrm{r}_2|^p$, we have:

$$\mathrm{Lip} \leq \omega_2 nm(\max(2^{p-1}(C^X)^2 + 2^{p-1}(C^Y)^p, 2\hat{\lambda}))^2. \tag{80}$$

Proof. We have:

$$\begin{aligned}
&\|\nabla\mathcal{L}_{C,M-2\hat{\lambda}}(\gamma) - \nabla\mathcal{L}_{C,M-2\hat{\lambda}}(\gamma')\|_F^2 \\
&= \|(\omega_1 C + \omega_2((M-2\hat{\lambda})\circ\gamma + (M^\top - 2\hat{\lambda})\circ\gamma)) - (\omega_1 C + \omega_2((M-2\hat{\lambda})\circ\gamma' + (M^\top-2\hat{\lambda})\circ\gamma'))\|_F^2 \\
&= 2\omega_2^2\|(M-2\hat{\lambda})\circ(\gamma-\gamma')\|_F^2 + 2\omega_2^2\|(M^\top-2\hat{\lambda})\circ(\gamma-\gamma')\|_F^2 \\
&= 2\omega_2^2\sum_{i,j}\left(\left[(M-2\hat{\lambda})\circ(\gamma-\gamma')\right]_{i,j}\right)^2 + 2\omega_2^2\sum_{i,j}\left(\left[(M^\top-2\hat{\lambda})\circ(\gamma-\gamma')\right]_{i,j}\right)^2 \\
&= 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'}(M_{i,j,i',j'}-2\hat{\lambda})(\gamma_{i',j'}-\gamma'_{i',j'})\right)^2 + 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'}(M^\top_{i,j,i',j'}-2\hat{\lambda})(\gamma_{i',j'}-\gamma'_{i',j'})\right)^2 \\
&\leq 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'}|M_{i,j,i',j'}-2\hat{\lambda}||\gamma_{i',j'}-\gamma'_{i',j'}|\right)^2 + 2\omega_2^2\sum_{i,j}\left(\sum_{i',j'}(|M^\top_{i,j,i',j'}-2\hat{\lambda}||\gamma_{i',j'}-\gamma'_{i',j'}|\right)^2 \\
&\leq 4\omega_2^2\underbrace{\max(|M-2\hat{\lambda}|)^2}_{A}\cdot\underbrace{\sum_{i,j}^{n,m}\left(\sum_{i',j'}^{n,m}|\gamma_{i',j'}-\gamma'_{i',j'}|\right)^2}_{B}.
\end{aligned}$$

Term B can be bounded by (72). Thus, we obtain the upper bound (79). Furthermore, when $L(\mathrm{r}_1, \mathrm{r}_2) = |\mathrm{r}_1 - \mathrm{r}_2|^p$, from inequality (44), we have:

$$\max(|M - 2\hat{\lambda}|) \leq \max(\max_{i,i'} 2^{p-1}(C^X)^p_{i,i'} + \max_{j,j'} 2^{p-1}(C^Y)^p_{j,j'}, 2\hat{\lambda}) := \max(2^{p-1}(C^X)^{p-1}, 2^{p-1}(C^Y)^p, 2\hat{\lambda})$$

and we obtain the bound (80). □

Proof of Proposition 9. Combining Lemma 6, (78) and (77), we obtain the upper bound (73) and complete the proof. □

I Sinkhon Algorithm for Fused Partial Gromov Wasserstein problem

For convenience, in this section we default to setting

$$L(d_X^r, d_Y^R) = \|d_X - d_Y\|^2,$$

while all propositions, algorithms and proofs extend without loss of generality to a generic loss function $L(d_X^r, d_Y^r)$.

I.1 Sinkhorn Algorithm for FPGW metric

Given mm-spaces $\mathbb{X} = (X, d_X, \mu), \mathbb{Y} = (Y, d_Y, \nu)$, the general unbalanced Fused-Gromov Wasserstein setting, the entropic problem is defined as the following:

$$\min_{\gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \mathcal{L}(\gamma) + \epsilon H(\gamma^{\otimes 2} \parallel (\mu \otimes \nu)^{\otimes 2}) \tag{81}$$

$$\mathcal{L}(\gamma) := \omega_1 \int_{X \times Y} d(x, y) d\gamma + \omega_2 \int_{X^2 \times Y^2} |d_X - d_Y|^2 d\gamma^{\otimes 2} + \lambda(D_\phi(\gamma_1^{\otimes 2} \parallel \mu^{\otimes 2}) + D_\phi(\gamma_2^{\otimes 2} \parallel \nu^{\otimes 2}))$$

where D_ϕ is $\phi-$divergence, H is the (negative) relative entropy

$$H(\mu \parallel \nu) = \begin{cases} \int \ln(\frac{d\mu}{d\nu}) d\mu & \text{if } \mu \ll \nu \\ +\infty & \text{elsewhere} \end{cases}.$$

In the Fused Partial GW setting, D_ϕ becomes

$$D_{PTV}(\mu \parallel \nu) := \begin{cases} |\mu - \nu|_{TV} = |\nu - \mu| & \text{if } \mu \leq \nu \\ +\infty & \text{elsewhere} \end{cases}. \tag{82}$$

Note, from the definition of (82), we can restrict the searching space for γ to $\Gamma_\leq(\mu, \nu)$:

$$EFPGW(\mathbb{X}, \mathbb{Y}) := \min_{\gamma \in \Gamma_\leq(\mu, \nu)} \mathcal{L}(\gamma) + \epsilon H(\gamma^{\otimes 2} \parallel (\mu \otimes \nu)^{\otimes 2}) \tag{83}$$

$$\mathcal{L}(\gamma) = \omega_1 \langle c, \gamma \rangle + \omega_2 \langle |d_X - d_Y|^2 \otimes \gamma^{\otimes 2} \rangle + \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma|^2)$$

Problem (83) can be further relaxed as:

$$\min_{\gamma, \pi \in \Gamma_\leq(\mu, \nu)} \mathcal{F}(\gamma, \pi) + \epsilon H(\pi \otimes \gamma \parallel (\mu \otimes \nu)^{\otimes 2}) \tag{84}$$

$$\mathcal{F}(\gamma, \pi) := \omega_1 \langle d(x, y), \frac{\gamma + \pi}{2} \rangle + \omega_2 \langle |d_X - d_Y|^2, \gamma \otimes \pi \rangle + \lambda(|\mu|^2 + |\nu|^2 - 2|\gamma||\pi|)$$

It is clear $\mathcal{F}(\gamma,\gamma)=\mathcal{F}(\gamma)$. Thus, (84) $\leq$ (83), and we denote (84) as $LB-FPGW_\lambda(\mathbb{X},\mathbb{Y})$ (lower bound of Fused Partial Gromov Wasserstein). Essentially, the Sinkhorn algorithm aims to solve $LB-FPGW$.

We first introduce the following fundamental proposition. Note, a similar version can be found in (Séjourné et al., 2021, Proposition 4):

Proposition 10. *Given a fixed $\pi\in\Gamma_\leq(\mu,\nu)$, considering the problem:*

$$\min_{\gamma\in\Gamma_\leq(\mu,\nu)}\mathcal{F}(\pi,\gamma)+\epsilon H(\pi\otimes\gamma\parallel(\mu\otimes\nu)^{\otimes 2}),$$

it is equivalent to solve the following entropic optimal partial transport problem:

$$\min_{\gamma\in\Gamma_\leq(\mu,\nu)}\int_{X\times Y}c_\pi(x,y)d\gamma+\lambda|\pi|(|\mu|+|\nu|-2|\gamma|)+\epsilon|\pi|H(\gamma\parallel\mu\otimes\nu) \tag{85}$$

where

$$c_\pi(x,y)=\frac{1}{2}\omega_1 d(x,y)+\omega_2[|d_X-d_Y|^2\circ\pi](x,y)+\epsilon H(\pi\parallel\mu\otimes\nu)$$
$$[|d_X-d_Y|^2\circ\pi](x,y)=\int_{X\times Y}|d_X(x,x')-d_Y(y,y')|d\pi(x',y')$$

Proof. Fix $\pi\in\Gamma_\leq(\mu,\nu)$ in (84). Since $\pi,\gamma\leq\mu\otimes\nu$, we have $\pi,\gamma\ll\mu\otimes\nu$, thus we have:

$$\begin{aligned}
&H(\pi\otimes\gamma\parallel(\mu\otimes\nu)^{\otimes 2})\\
&=\int_{(X\times Y)^2}\ln\left(\frac{d\pi d\gamma}{d\mu\otimes\nu\cdot d\mu\otimes\nu}\right)d\mu d\gamma+(|\mu||\nu|)^2-|\pi||\gamma|\\
&=\int_{X\times Y}\left[\int_{X\times Y}\ln\left(\frac{d\pi}{d\mu\otimes\nu}\right)d\pi\right]d\gamma+\int_{X\times Y}\left[\int_{X\times Y}\ln\left(\frac{d\gamma}{d\mu\otimes\nu}\right)d\gamma\right]d\pi+(|\mu||\nu|)^2-|\pi||\gamma|\\
&=\int_{X\times Y}d\pi H(\gamma\parallel\mu\otimes\nu)+\int_{X\times Y}d\gamma H(\pi\parallel\mu\otimes\nu)+(|\mu||\nu|)^2-|\pi||\gamma|\\
&=|\pi|H(\gamma\parallel\mu\otimes\nu)+|\gamma|H(\pi\parallel\mu\otimes\nu)+((|\mu||\nu|)^2-|\pi||\mu||\nu|).
\end{aligned}\tag{86}$$

We obtain:

$$\begin{aligned}
&\mathcal{F}(\pi,\gamma)+\epsilon H(\gamma\times\pi\parallel(\mu\otimes\nu)^{\otimes 2})\\
&=\omega_1\langle d,\frac{\gamma+\pi}{2}\rangle+\omega_2\langle|d_X-d_Y|^2,\gamma\otimes\pi\rangle+\lambda(D_{PTV}(\gamma_1\otimes\pi_1\parallel\mu^{\otimes 2})+D_{PTV}(\gamma_2\otimes\mu_2\parallel\nu^{\otimes 2}))\\
&\quad+\epsilon H(\gamma\times\pi\parallel(\mu\otimes\nu)^{\otimes 2})\\
&=\omega_1\langle d,\frac{\gamma+\pi}{2}\rangle+\omega_2\langle|d_X-d_Y|^2,\gamma\otimes\pi\rangle+\lambda(|\mu|^2+|\nu|^2-2|\gamma||\pi|)+\epsilon H(\gamma\times\pi\parallel(\mu\otimes\nu)^{\otimes 2})\\
&=\underbrace{\frac{1}{2}\omega_1\langle d,\pi\rangle+\lambda(|\mu|^2+|\nu|^2)+\epsilon[(|\mu||\nu|)^2-|\pi||\mu||\nu|]}_{constant}\\
&\quad+\langle\frac{1}{2}\omega_1 d+\omega_2|d_X-d_Y|^2\circ\pi+\epsilon H(\pi\parallel\mu\otimes\nu),\gamma\rangle-2\lambda|\pi||\gamma|)+\epsilon|\pi|H(\gamma\parallel\mu\otimes\nu)
\end{aligned}$$

where we use the fact (86) and

$$|\gamma|\epsilon H(\pi\parallel\mu\otimes\nu)=\langle\epsilon H(\pi\parallel\mu\otimes\nu),\gamma\rangle.$$

If we ignore the constant part, the remaining part is exactly the entropic partial OT (up to a constant $\lambda|\pi|(|\mu|+|\nu|)$), and we complete the proof. $\square$

Note, the partial OT problem (85) can be solved by the classical Sinkhorn algorithm. See e.g. Chizat et al. (2018a); Bai (2024).

Algorithm 1: Sinkhorn Partial OT

Input : $c, \epsilon, \lambda, \mathrm{p}, \mathrm{q}$
Output: γ
Initialize $u = 1_n, v = 1_m,\ K = e^{-c/\epsilon} pq^\top$
for $l = 1, 2, \ldots$ **do**
  $u = \min(\frac{\mathrm{p}}{Kv}, e^{\lambda/\epsilon})$
  $v = \min(\frac{\mathrm{q}}{K^\top u}, e^{\lambda/\epsilon})$
  If (u, v) converge, break
$\gamma \leftarrow (u_i K_{ij} v_j)_{ij}$

I.2 Sinkhorn for the fused Mass-constraint Gromov Wasserstein problem.

Similar to the preious section, in the entropic regularziation setting, the fused mass-constraint Gromov Wasserstein problem is defined as:

$$EFMPGW(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma \in \Gamma_{\leq}^{\rho}(\mu,\nu)} \omega_1 \langle c, \gamma \rangle + \omega_2 \langle |d_X - d_Y|, \gamma^{\otimes 2} \rangle + \epsilon H(\gamma^{\otimes 2} \parallel (\mu \otimes \nu)^{\otimes 2}) \tag{87}$$

and it can be relaxed as

$$\begin{aligned} LB - EFMPGW(\mathbb{X}, \mathbb{Y}) := \inf_{\gamma, \pi \in \Gamma_{\leq}^{\rho}(\mu,\nu)} & \underbrace{\omega_1 \langle c, \frac{1}{2}(\gamma + \pi) \rangle + \omega_2 \langle |d_X - d_Y|, \gamma^{\otimes 2} \rangle}_{\mathcal{F}(\gamma,\pi)} \\ & + \epsilon H(\gamma \otimes \pi \parallel (\mu \otimes \nu)^{\otimes 2}) \end{aligned} \tag{88}$$

Note, since $\frac{1}{2}(\gamma + \pi) \in \Gamma_{\leq}^{\rho}(\mu, \nu)$, and $\mathcal{F}(\gamma, \gamma) = \mathcal{F}(\gamma)$, thus, we have

$$LB - EFMPGW(\mathbb{X}, \mathbb{Y}) \leq EFMPGW(\mathbb{X}, \mathbb{Y}).$$

Similar to the previous section, we have the following property:

Proposition 11. *Given a fixed $\pi \in \Gamma_{\leq}^{\rho}(\mu, \nu)$, considering the problem:*

$$\min_{\gamma \in \mathcal{M}_+(X \times Y)} \mathcal{F}(\pi, \gamma) + \epsilon H(\pi \otimes \gamma \parallel (\mu \otimes \nu)^{\otimes 2}),$$

it is equivalent to solving the following entropic optimal partial transport problem:

$$\min_{\gamma \in \Gamma_{\leq}^{\rho}(\mu,\nu)} \int_{X \times Y} c_\pi(x, y) d\gamma + \epsilon H(\gamma \parallel \mu \otimes \nu) \tag{89}$$

where c_π is defined in Proposition 10.

Proof. Similar to the proof in Proposition 10, given $\pi \in \Gamma_{\leq}^{\rho}(\mu, \nu)$, we have

Algorithm 2 Sinkhorn Algorithm for FMPGW

Input: $C \in \mathbb{R}^{n\times m}, C^X \in \mathbb{R}^{n\times n}, C^Y \in \mathbb{R}^{m\times m}, p \in \mathbb{R}^n_+, q \in \mathbb{R}^m_+$, $\omega_2 \in [0,1], \rho \in [0, \min(|p|,|q|)]$.
Output: γ
for $k = 1, 2, \ldots$ **do**
$\pi \leftarrow \gamma$
Solve the Sinkhorn partial OT problem (89) via algorithm (2):

$$\gamma \leftarrow \min_{\gamma \in \Gamma^\rho_\leq(\mu,\nu)} \int_{X\times Y} c_\pi(x,y) d\gamma + \epsilon\rho H(\gamma \parallel \mu \otimes \nu)$$

Fix γ and solve the similar Sinkhorn partial OT problem (89) via algorithm (2):

$$\pi \leftarrow \min_{\pi \in \Gamma^\rho_\leq(\mu,\nu)} \int_{X\times Y} c_\gamma(x,y) d\pi + \epsilon\rho H(\pi \parallel \mu \otimes \nu)$$

Break if $\pi \approx \gamma$
end for

$$\begin{aligned}
&\mathcal{F}(\pi,\gamma) + \epsilon H(\gamma \otimes \pi \parallel (\mu\otimes\nu)^{\otimes 2}) \\
&= \omega_1 \langle d, \frac{\gamma+\pi}{2}\rangle + \omega_2 \langle |d_X - d_Y|^2, \gamma\otimes\pi\rangle + \epsilon H(\gamma\times\pi \parallel (\mu\otimes\nu)^{\otimes 2}) \\
&= \underbrace{\frac{1}{2}\omega_1\langle d,\pi\rangle + \epsilon[(|\mu||\nu|)^2 - |\pi||\mu||\nu|]}_{constant} \\
&\quad + \langle \underbrace{\frac{1}{2}\omega_1 d + \omega_2 |d_X - d_Y|^2 \circ \pi + \epsilon H(\pi \parallel \mu\otimes\nu)}_{c_\pi}, \gamma\rangle + \epsilon \underbrace{|\pi|}_{=\rho} H(\gamma \parallel \mu\otimes\nu)
\end{aligned}$$

and we complete the proof. □

The Entropic partial OT problem (89) can be solved by the corresponded Sinkhorn algorithm Benamou et al. (2015); Bai (2024), thus we can derive the Sinkhorn algorithm problem for the relaxed fused-mass constraint Gromov Wasserstein problem:

where

$$\begin{aligned}
\mathcal{C}_1 &= \{\gamma \in \mathbb{R}^{n\times m}_+ : \gamma_2 \leq q\} \\
\mathcal{C}_2 &= \{\gamma \in \mathbb{R}^{n\times m}_+ : \gamma_1 \leq p\} \\
\mathcal{C}_3 &= \{\gamma \in \mathbb{R}^{n\times m}_+ : |\gamma| = \rho\}
\end{aligned}$$

$$\mathrm{Proj}^{KL}_{\mathcal{C}_i}(\gamma) := \min_{\gamma^i \in \mathcal{C}_i} KL(\gamma^i \parallel \gamma) = \begin{cases} \mathrm{diag}(\min(\frac{p}{\gamma_2}, 1_n))\gamma \\ \gamma \mathrm{diag}(\min(\frac{q}{\gamma^\top 1_n}, 1_m)) \\ \gamma \frac{\eta}{\|\gamma\|} \end{cases}$$

J Fused Partial Gromov Wasserstein Barycenter

In discrete setting, suppose we have K mm-spaces $\mathbb{X}^k = (X^k \subset \mathbb{R}^d, d_{X^k}, \mu^k := \sum_{i=1}^{n^k} p^k_i \delta_{x^k_i})$ for $k = 1, \ldots, K$ and a fixed pmf function $p \in \mathbb{R}^n_+$ where $n \in \mathbb{N}$. Note for each $k \in [1:K]$, let

Algorithm 2: Sinkhorn Mass-constraint Partial OT

Input : p, q, ρ, c

Output: γ

Initialization

$K = e^{-c/\epsilon} pq^\top$

for i=1,2,3 **do**

  $\xi^i \leftarrow 1_{n\times m}$

$\gamma^{(0)} \leftarrow K \frac{\rho}{\|K\|}$

Main loop

$k = 0$ **for** $l = 0, 1, 2, \ldots$ **do**

  for i=1,2,3 **do**

    $k \leftarrow k + 1$

    $\gamma^{(k)} \leftarrow \texttt{Proj}^{KL}_{\mathcal{C}_i}(\gamma^{(k-1)} \odot \xi^i)$

    $\xi^i \leftarrow \xi^i \odot \frac{\gamma^{(k-1)}}{\gamma^{(k)}}$

  Break if $\gamma^{(k)}$ converges

$C^k = [d^r_{X^k}(x^k_i, x^k_{i'})]_{i,i' \in [1:n^k]} \in \mathbb{R}^{n^k \times n^k}$ and let $X^k = [x^k_1, \ldots x^k_{n^k}]^\top \in \mathbb{R}^{n_k \times d}$, we have (X^k, p^k, C^k) can represent the space $\mathbb{X}^k$. Thus, for convenience, we use the convention $\mathbb{X}^k = (X^k, p^k, C^k)$ to denote the corresponsded mm-space.

Then, given pmf function $\mathrm{p} \in \mathbb{R}^n_+$ where $n \in \mathbb{N}$, and $\beta_1, \ldots \beta_K \geq 0$ with $\sum_{k=1}^K \beta_k = 1$, the fused-Gromov Wasserstein barycenter problem Titouan et al. (2019) is defined as:

$$\min_{C \in \mathbb{R}^{n\times n}, X \in \mathbb{R}^{n \times d}} \beta_i FGW_L(\mathbb{X}, \mathbb{X}^k), \text{ where } \mathbb{X} = (X, \mathrm{p}, C), \tag{90}$$

where we adapt notation $FGW_L(\cdot,\cdot)$ to simplify the notation $FGW(\cdot,\cdot)$ since r has been incorporated into matrix C^k.

Inspired by this work, we present the following fused Partial GW barycenter problems.

J.1 Fused-MPGW barycenter.

Choose values $\rho_1, \rho_2, \ldots, \rho_K$ where $\rho_k \in [0, \min(|p|, |p^k|)]$ for each k. In addition, choose $(\omega^k_1, \omega^k_2) \in [0,1]$ for $k = 1, 2, \ldots K$ such that $\omega^k_1 + \omega^k_2 = 1, \forall k$. The fused mass-constrained Partial GW barycenter problem is defined as:

$$\begin{aligned} &\min_{C \in \mathbb{R}^{n\times n}, X \in \mathbb{R}^{n \times d}} \sum_{k=1}^K \beta_k FMPGW_{L,\rho_k}(\mathbb{X}, \mathbb{X}^k), \\ &= \min_{\substack{C \in \mathbb{R}^{n\times n} \\ X \in \mathbb{R}^{n\times d}}} \min_{\substack{\gamma^k \in \Gamma^{\rho_k}_{\leq}(p, p^k) \\ k \in [1:K]}} \sum_{k=1}^K \beta_k \left(\omega^k_1 \langle D(X, X^k), \gamma^k \rangle + \omega^k_2 \langle L(C, C^k)(\gamma^k)^{\otimes 2} \rangle \right), \end{aligned} \tag{91}$$

where $\mathbb{X} = (X, p, C)$, $\mathbb{X}^k = (X^k, p^k, C^k)$, and $D(X, X^K) = [\|x_i - x^k_j\|^2]_{i \in [1:n], j \in [1:n^k]} \in \mathbb{R}^{n \times n_k}$.

Note that the above problem is convex with respect to (C, X) when γ^k is fixed for each k. However, it is not convex with respect to γ^k for each k. Similar to classical fused-GW, it can be solved iteratively by updating (C, X) and $(\gamma^k)^K_{k=1}$ alternatively in each iteration.

Step 1. Given (C, X), we update $(\gamma^k)_{k=1}^K$. Note, when (C, X) is fixed, each optimal $(\gamma^k)^*$ is given by

$$(\gamma^k)^* = \arg \min_{\gamma^k \in \Gamma_{\leq}^{\rho_k}(p, p^k)} \omega_1^k \langle D(X, X^k), \gamma^k \rangle + \omega_2^k \langle L(C, C^k), (\gamma^k)^{\otimes 2} \rangle,$$

which is a solution for the fused partial GW problem $F - MPGW_{\rho^k}(\mathbb{X}, \mathbb{X}^k)$.

Step 2. Given γ^k, update (C, X).

Suppose γ^k is given for each k, the objective function in (90) becomes:

$$\begin{aligned}
&\min_{C \in \mathbb{R}^{n\times n}, X \in \mathbb{R}^{n\times d}} \sum_{k=1}^K \beta_k \left(\omega_1 \langle D(X, X^k), \gamma^k \rangle + \omega_2 \langle L(C, C^k), (\gamma^k)^{\otimes 2} \rangle \right) \\
&= \omega_1 \underbrace{\min_{X \in \mathbb{R}^{n\times d}} \sum_{k=1}^K \beta_k \langle D(X, X^k), \gamma^k \rangle}_{A} + \omega_2 \underbrace{\min_{C \in \mathbb{R}^{n\times n}} \sum_{k=1}^K \beta_k \langle L(C, C^k), (\gamma^k)^{\otimes 2} \rangle}_{B}.
\end{aligned}$$

Problem B admits the solution (we refer Bai et al. (2024) Section M for details). In particular, if L satisfies (55), f_1, h_1 are differentiable, then

$$C = \left(\frac{f_1'}{h_1'} \right)^{-1} \left(\frac{\sum_k \beta_k \gamma^k h_2(C^k)(\gamma^k)^\top}{\sum_k \beta_k \gamma_1^k (\gamma_1^k)^\top} \right). \tag{92}$$

In particular, when $L(\mathrm{r}_1, \mathrm{r}_2) = |\mathrm{r}_1 - \mathrm{r}_2|^2$, the above formula becomes:

$$C = \frac{\sum_k \beta_k \gamma^k C^k (\gamma^k)^\top}{\sum_k \beta_k \gamma_1^k (\gamma_1^k)^\top}.$$

J.2 Solving the subproblem A.

We first introduce the barycentric projection:

For each (X^k, p^k, γ^k), the barycentric projection Bai et al. (2023a) is defined by

$$\hat{x}_i^k = \begin{cases} \frac{1}{\gamma_1^k[i]} \sum_{j=1}^{n_k} \gamma_{i,j}^k x_j^k & \text{if } \gamma_1^k[i] = \sum_{j=1}^{n_k} \gamma_{i,j}^k > 0, \\ x_i & \text{elsewhere.} \end{cases} \tag{93}$$

Note, if $|\gamma^k| = |p|, \forall k$, by Cuturi & Doucet (2014) (Eq 8), the optimal X is given by

$$X = [x_1, \ldots x_n]^\top, x_i = \frac{\sum_{k=1}^K \beta_k \gamma_1^k[i] \hat{x}_i^k}{\sum_{k=1}^K \beta_i \gamma_1^k[i]}.$$

In this subsection, we will extend the above result to the general case.

Proposition 12. *Any matric X satisfies the following is a solution for problem A.*

$$X = [x_1, \ldots, x_n]^\top, x_i = \frac{\sum_{k=1}^K \beta_k \gamma_1^k[i] \hat{x}_i^k}{\sum_{k=1}^K \beta_i \gamma_1^k[i]} \qquad \textit{if } \sum_{k=1}^K \beta_k \gamma_1^k[i] > 0. \tag{94}$$

Note, for each i, we use convention $\frac{0}{0} = 0$ if $\sum_{k=1}^K \beta_i \gamma_1^k[i] = 0$.

Proof. Problem A can be written in terms of barycentric projection (93). In particular, let $\mathcal{D}_k := \{i : \sum_{i,j} \gamma^k_{i,j} > 0\}$ and $\mathcal{D} = \bigcup_{k=1}^K \mathcal{D}_k$. We have:

$$\begin{aligned} A &= \min_{X \in \mathbb{R}^{n\times d}} \sum_{k=1}^K \beta_k \sum_{i=1}^n \gamma_1^k[i](\|x_i - \hat{x}_i^k\|^2) \\ &= \min_{X \in \mathbb{R}^{n\times d}} \sum_{i=1}^n \sum_{k=1}^K \beta_k \gamma_1^k[i](\|x_i - \hat{x}_i^k\|^2) \\ &= \sum_{i=1}^n \min_{x_i \in \mathbb{R}^d} \sum_{k=1}^K \beta_k \gamma_1^k[i](\|x_i - \hat{x}_i^k\|^2). \end{aligned}$$

For each i, we have two cases:

Case 1: $\sum_{k=1}^K \beta_k \gamma_1^k[i] > 0$. Then the optimal x_i is given by the weighted average vector:

$$\frac{\sum_{k \in D_i} \beta_k \gamma_1^k[i] \hat{x}_i^k}{\sum_{i \in D_k} \beta_k \gamma_1^k[i]}.$$

Case 2: $\sum_{k=1}^K \beta_k \gamma_1^k[i] = 0$. The problem becomes

$$\min_{x_i \in \mathbb{R}^d} 0,$$

and there is no requirement for x_i. And we complete the proof.

□

Remark 10. In practice, the above formulation can be described by matrices. In particular, let $\hat{X}^k = [\hat{x}_1^k, \ldots \hat{x}_n^k]^\top$, we have

$$\hat{X}^k = \frac{\gamma^k X^k}{\gamma_1^k}, \gamma_1^k = \gamma^k 1_m.$$

Then we have (94) becomes

$$X = \frac{\sum_{k=1}^K \beta_k \gamma_1^k 1_d^\top \odot \hat{X}}{\sum_{k=1}^K \beta_k \gamma_1^k},$$

where $\odot$ denotes the element-wise multiplication; the notation $\frac{A}{B}$ where $A \in \mathbb{R}^{n\times d}, B \in \mathbb{R}^n$ denotes the following

$$\frac{A}{B} = [A[:,1]/B_1, \ldots, A[:,n]/B_n]^\top,$$

and we use $\frac{0}{0} = 0$ if any element in the denominator is 0.

J.3 Fused-PGW barycenter

Similar to the previous subsection, we define and derive the solution for the fused-Partial GW problem.

Given $\lambda_1, \ldots \lambda_K \geq 0, p \in \mathbb{R}_+^n$, the fused partial GW barycenter problem is defined as:

$$\min_{C\in\mathbb{R}^{n\times n},X\in\mathbb{R}^{n\times d}}\sum_{k=1}^{K}\beta_k FPGW_{L,\lambda_i}(\mathbb{X},\mathbb{X}^k), \text{where } \mathbb{X}=(X,p,C)$$

$$=\min_{C\in\mathbb{R}^{n\times n},X\in\mathbb{R}^{n\times d}}\min_{\gamma^k\in\Gamma_\leq(p,p^k):k\in[1:K]}\sum_{k=1}^{K}\omega_1^k\langle D(X,X^k),\gamma^k\rangle$$
$$+\omega_2^k\langle L(C,C^k),(\gamma^k)^{\otimes 2}\rangle+\lambda_k(|p^k|^2+|p|^2-2|\gamma^k|^2). \tag{95}$$

Similar to the fused MPGW barycenter problem, the above problem can be solved iteratively. In each iteration, we have two steps:

Step 1. Given X, D, update γ^k. Note finding each γ^k is essentially solving the fused-PGW problem:

$$\gamma^k=\arg\min_{\gamma\in\Gamma_\leq(p,p^k)}\omega_1^k\langle D(X,X^k),\gamma\rangle+\omega_2^k\langle L(C,C^k),(\gamma^{\otimes 2})\rangle+\lambda_k(|p^k|^2+|p|^2-2|\gamma^k|^2)$$

And it can be solved by algorithm 1.

Step 2. Given $\gamma^k : k\in[1:K]$, update C, X.

In this case, (95) becomes

$$\omega_2^k\underbrace{\min_{C\times\mathbb{R}^{n\times n}}\langle L(C,C^k)-2\lambda_k,(\gamma^k)^{\otimes 2}\rangle}_{A}+\omega_1^k\underbrace{\min_{X\in\mathbb{R}^{n\times d}}\langle D(X,X^k),\gamma^k\rangle}_{B}. \tag{96}$$

Optimal C for subproblem A is (92) by [Proposition M.2 Bai et al. (2024)]. Optimal X for subproblem B is given by (94) by the Proposition 12.

K Relation between FGW, FPGW and FMPGW.

In this section, we briefly discuss the relation between FGW, Fused-PGW problem (15), and the fused-MPGW (14).

First, we introduce the following "equivalent relation" between FPGW and FMPGW. It can be treated as the generalization of Proposition L.1. in Bai et al. (2024) in the fused-PGW formulation.

Proposition 13. *Given mm-spaces $\mathbb{X}=(X,d_X,\mu),\mathbb{Y}=(Y,d_Y,\nu)$, $r\geq 1,\lambda>0$. Suppose γ^* is a minimizer for FPGW problem $FPGW_\lambda(\mathbb{X},\mathbb{Y})$, then γ^* is also a minimizer for Fused-MPGW problem $FPGW_\rho(\mathbb{X},\mathbb{Y})$ where $\rho=|\gamma^*|$.*

Proof. Pick $\gamma\in\Gamma_\leq^\rho(\mu,\nu)\subset\Gamma_\leq(\mu,\nu)$. Since γ^* is optimal for $FPGW_\lambda(\mathbb{X},\mathbb{Y})$, we have:

$$\omega_1\langle C,\gamma^*\rangle+\omega_2\langle L,(\gamma^*)^{\otimes 2}\rangle+\lambda(|\mu|^2+|\nu|^2-2|\gamma^*|^2)$$
$$\leq\omega_1\langle C,\gamma\rangle+\omega_2\langle L,\gamma^{\otimes 2}\rangle+\lambda(|\mu|^2+|\nu|^2-2|\gamma|^2).$$

Combine it with the fact $|\gamma|=|\gamma^*|=\rho$, we complete the proof. □

Next, we discuss the relation between FGW and FMPGW problems.

Proposition 14. *Under the setting of Proposition 13, suppose $|\mu|=|\nu|$, then*

$$FMPGW_\rho(\mathbb{X},\mathbb{Y})=FGW_\rho(\mathbb{X},\mathbb{Y}).$$

Proof. In this case, $\Gamma^{\rho}_{\leq}(\mu,\nu)=\Gamma(\mu,\nu)$ and we complete the proof. $\square$

Similarly, in the extreme case, the FPGW problem can also recover the FGW problem. We first introduce the following lemma:

Lemma 8. *Suppose μ,ν are supported in compact set and $|\mu|,|\nu|>0$. Let $X=supp(\mu),Y=supp(\nu)$, and $\max L(d_X^r,d_Y^r):=\max_{x,x'\in X,y,y'\in Y}L(d_X^r(x,x'),d_Y^r(y,y')),\max C:=\max_{x\in X,y\in Y}C(x,y)$.*

Suppose

$$2\lambda\geq\omega_1\frac{1}{\min(|\mu|,|\nu|)}\max(C)+\omega_2\max L(d_X^r,d_Y^r) \tag{97}$$

there exists a solution, denoted as γ^*, for $FPGW_\lambda(\mathbb{X},\mathbb{Y})$ such that

$$|\gamma^*|=\min(|\mu|,|\nu|). \tag{98}$$

If we replace "$\geq$" by "$>$" in the inequality (97)*, then every solution of $FPGW_\lambda(\mathbb{X},\mathbb{T})$ satisfies* (98)*.*

Proof. For convenience, we suppose $|\mu|\leq|\nu|$. Suppose (97) holds. Choose an optimal $\gamma\in\Gamma_{\leq}(\mu,\nu)$. By lemma E.1. in Bai et al. (2024), there exists $\gamma'\in\Gamma_{\leq}(\mu,\nu)$ such that $\gamma\leq\gamma'$ with $|\gamma'|=|\mu|$, i.e. $\gamma'_1=\mu$.

We have:

$$\begin{aligned}
&\int_{X\times Y}\omega_1 C(x,y)d\gamma'+\int_{(X\times Y)^2}\omega_2 L(d_X^r,d_Y^r)d\gamma'^{\otimes 2}+\lambda(|\mu|^2+|\nu|^2-2|\gamma'|^2)\\
&-\int_{X\times Y}\omega_1 C(x,y)d\gamma+\int_{(X\times Y)^2}\omega_2 L(d_X^r,d_Y^r)d\gamma^{\otimes 2}+\lambda(|\mu|^2+|\nu|^2-2|\gamma|^2)\\
&=\int_{X\times Y}\omega_1 C(x,y)d(\gamma'-\gamma)+\int_{(X\times Y)^2}\omega_2 L(d_X^r,d_Y^r)-2\lambda d(\gamma'^{\otimes 2}-\gamma^{\otimes 2})\\
&\leq\omega_1\max(C)|\gamma'-\gamma|+\omega_2(\max(L(d_X^r,d_Y^r))-2\lambda)|\gamma'^{\otimes 2}-\gamma^{\otimes 2}|\\
&=\underbrace{\left(\omega_1\frac{\max(C)}{|\gamma+\gamma'|}+\omega_2\max(L(d_X^r,d_Y^r))-2\lambda\right)}_{A}\underbrace{|\gamma'^{\otimes 2}-\gamma^{\otimes 2}|}_{B}\\
&\leq 0,
\end{aligned} \tag{99}$$

where (99) follows by the fact $A\leq 0,B\geq 0$. Thus we have γ' is optimal.

Now we suppose

$$2\lambda>\omega_1\frac{1}{\min(|\mu|,|\nu|)}\max(C)+\omega_2\max L(d_X^r,d_Y^r).$$

We assume there exists an optimal γ such that $|\gamma|<|\mu|$.

If we can find optimal γ with $|\gamma|<|\mu|$, then $A<0$ and $B>0$ and we have (99) <0. That is, γ' admits a smaller cost. It is a contradiction since γ is optimal, and we complete the proof. $\square$

Remark 11. If $\min(|\mu|,|\nu|)=0$, $\Gamma_{\leq}(\mu,\nu)=\{\mathbf{0}\}$ where $\{\mathbf{0}\}$ is the zero measure. In this case, (98) is automatically satisfied, and there is no requirement for λ.

Based on this lemma, the FPGW can recover FGW in the extreme case:

Proposition 15. *Under the setting of Proposition 14, suppose λ satisfies* (97)*, then*

$$FPGW_\lambda(\mathbb{X},\mathbb{Y}) = FGW(\mathbb{X},\mathbb{Y}).$$

Proof. By Lemma 8, there exists optimal γ for $FPGW_\lambda(\mathbb{X},\mathbb{Y})$ with $|\gamma| = |\mu| = |\nu|$. By Proposition 13, we have γ is optimal for $FMPGW_\rho(\mathbb{X},\mathbb{Y})$ where $\rho = |\mu| = |\nu|$. By Proposition 14, we have γ is optimal for $FGW(\mathbb{X},\mathbb{Y})$. Thus, γ and we complete the proof. □

Remark 12 (Summary of relations between FPGW, FMPGW, FGW, and OT). Due to the inherent non-convexity of Gromov–Wasserstein objectives, it is not rigorous to claim that FGW and FPGW are fully equivalent. Nevertheless, the following informal relations help clarify their conceptual connections:

Relation between λ in FPGW and ρ in FMPGW. Both parameters control the amount of transported mass. In the FPGW formulation, increasing λ encourages more mass to be transported, while in FMPGW, ρ directly specifies the transported mass fraction. Hence, λ in FPGW can be viewed as the Lagrange multiplier associated with the transported-mass constraint in FMPGW.

Relation between FPGW and FGW. When $|\mu| = |\nu|$ and λ is sufficiently large, FPGW reduces to FGW. In this case, the partial-mass penalty vanishes, and the formulation coincides with the standard fused Gromov–Wasserstein distance.

Relation between FPGW and OT. In the same setting where FPGW recovers FGW, by choosing $\omega_1 = 1$ and $\omega_2 = 0$, FPGW simplifies to the classical Optimal Transport problem.

These relationships summarize how the proposed FPGW framework unifies and generalizes several existing OT-based formulations.

K.1 Comparison for FPGW, FUGW, Sinkhorn and Frank-Wolfes

Remark 13 (FPGW vs. FUGW: Total Variation vs. KL Divergence).

(1) Theory. We show that FPGW admits a (semi-)metric, while the metric property of FUGW remains unclear. In the graph classification experiments, we also do not observe an advantage of FUGW over FPGW (see Appendix N).

(2) Algorithms. FPGW can be optimized by both a Frank–Wolfe (FW) solver and a Sinkhorn-based solver, whereas FUGW can only be solved via the Sinkhorn algorithm.

(3) Computational aspects. The KL divergence used in FUGW is nonlinear, while the total variation (TV) penalty used in FPGW is linear. We hypothesize that this difference contributes to faster convergence for FPGW even under entropic regularization. For example, in graph classification, **FUGW (Sinkhorn)** requires 100–400 seconds, whereas **FPGW (Sinkhorn)** completes in 7–40 seconds (see the updated PDF).

Remark 14 (FW vs. Sinkhorn: Transportation Cost vs. Partial Matching Plan).

(1) Transportation cost. When the downstream task requires the *transportation cost itself* (e.g., as a kernel matrix in graph classification), the Frank-Wolfe (FW) algorithm provides more accurate estimates. The Sinkhorn solver, on the other hand, requires a sufficiently large entropic regularization weight to prevent numerical instability (NaN errors), which in turn biases the objective and blurs the transport cost. Therefore, we use **FPGW (FW)** in the graph classification experiments, where it yields higher accuracy than **FUGW (Sinkhorn)** (see Table 4). For example, on the *Protein* dataset, FPGW (FW) achieves approximately 69–72% accuracy, while FUGW

(Sinkhorn) gives 68–70%. On the *Synthetic* dataset, FPGW (FW) attains 94–97% versus 45–60% for FUGW (Sinkhorn).

(2) Partial matching plan. When the task requires identifying a *partial matching plan* (e.g., in graph matching problems), the objectives of GW, PGW, FGW, and FPGW are non-convex, and the FW algorithm can converge to suboptimal local minima. In this case, introducing an entropic regularization term improves the optimization landscape and enhances numerical stability. Therefore, for partial matching tasks, we employ the **Sinkhorn-based** solver, which provides more stable convergence and reliable partial correspondences.

K.2 Wall-clock Comparison Experiment

In this section, we present a wall-clock time comparison among **FPGW**, **Sinkhorn-FPGW**, **FMPGW** (from PythonOT Flamary et al. (2021)), and **Sinkhorn-FMPGW** (from PythonOT).

Dataset. Given two empirical measures $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^2$ and $Y = \{y_1, \ldots, y_m\} \subset \mathbb{R}^2$ with $x_i \sim$ i.i.d. $\mathcal{N}(0, I_2)$ and $y_j \sim$ i.i.d. $\mathcal{N}(0, I_2)$, we construct the distance matrices

$$C_X = [\|x_i - x_{i'}\|^2]_{i,i'=1}^n, \tag{100}$$

$$C_Y = [\|y_j - y_{j'}\|^2]_{j,j'=1}^m. \tag{101}$$

We test problem sizes $n \in \{10, 100, 200, 400, 800, 1000, 2000, 5000, 10000\}$, and for each n, we choose m uniformly at random from the interval $[1.1n, 1.5n]$.

Experiment and parameter settings. For FPGW in Eq. 15, we test the cases $\lambda \in \{0.1, 1, 10\}$. For FMPGW and Sinkhorn-FMPGW, the mass parameter is set to $|\gamma|$, where γ is obtained from FPGW (see Proposition 13).

Each experiment is repeated $k = 3$ times and we report the average wall-clock time. We set $\omega_2 = 0.5$ and the Sinkhorn regularization parameter $\epsilon = 0.1$.

For the FW-based methods, the linear subproblem is solved by the C++ solver "emd" from PythonOT, with a maximum number of iterations

$$\max\{1000,\ 100(n + m)\}. \tag{102}$$

This prevents non-convergence inside an FW iteration.

The Sinkhorn solvers for FPGW-Sinkhorn and FMPGW-Sinkhorn are implemented in NumPy, with maximum iterations

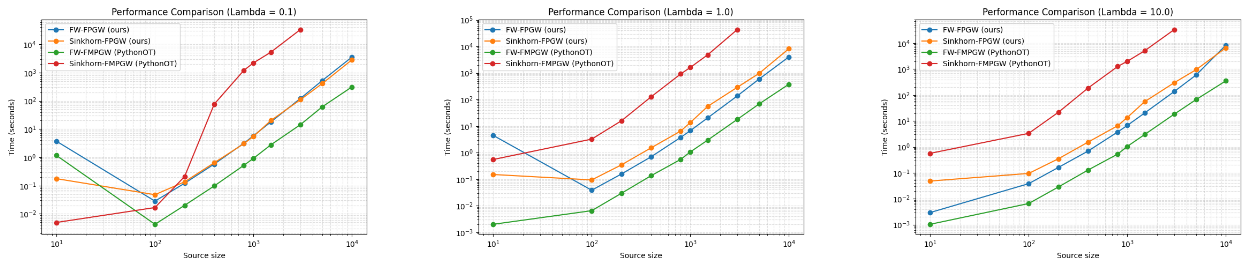
$$\max\{200,\ 10(n + m)\}. \tag{103}$$

The convergence tolerance for all methods is set to 10^{-8}.

Performance Analysis. The log-log wall-clock comparison is shown in Figure 2. FMPGW-FW, FPGW-FW, and FPGW-Sinkhorn have similar computational cost, with FPGW-FW being slightly faster than FMPGW-FW. Both FW algorithms are faster than FPGW-Sinkhorn because they use a C++ linear programming solver.

Sinkhorn-FMPGW is extremely slow. When $n \geq 2000$, it requires about two days to solve a single pair, while the other methods take only two to three hours. This is because the partial OT Sinkhorn update in FMPGW-Sinkhorn follows the algorithm of Benamou et al. Benamou et al. (2015), which needs six matrix operations per iteration, while FPGW-Sinkhorn requires only two (see Chizat et al. Chizat et al. (2018a)).

In summary, although FPGW and FMPGW are theoretically related, in practice both the FW and Sinkhorn solvers for FPGW are faster than those for FMPGW, with the difference being especially large for the Sinkhorn variants.



(a) $\lambda = 0.1$ (b) $\lambda = 1$ (c) $\lambda = 10$

Figure 2: We evaluate problem sizes $n = 10, \ldots, 10^4$ for **FMPGW–FW (PythonOT)**, **FPGW–FW (ours)**, and **FPGW–Sinkhorn (ours)**. The **FMPGW–Sinkhorn (PythonOT)** solver becomes extremely slow when $n \geq 2000$; for example, it requires *approximately two days* to solve a single pair, whereas the other methods finish within *2–3 hours*. Therefore, for FMPGW–Sinkhorn we report results only for $n = 10, \ldots, 2000$.

References

Yikun Bai. Sinkhorn algorithms and linear programming solvers for optimal partial transport problems. *arXiv preprint arXiv:2407.06481*, 2024.

Yikun Bai, Ivan Vladimir Medri, Rocio Diaz Martin, Rana Shahroz, and Soheil Kolouri. Linear optimal partial transport embedding. In *International Conference on Machine Learning*, pp. 1492–1520. PMLR, 2023a.

Yikun Bai, Bernhard Schmitzer, Matthew Thorpe, and Soheil Kolouri. Sliced optimal partial transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13681–13690, 2023b.

Yikun Bai, Huy Tran, Steven B Damelin, and Soheil Kolouri. Partial transport for point-cloud registration. *arXiv preprint arXiv:2309.15787*, 2023c.

Yikun Bai, Rocio Diaz Martin, Abihith Kothapalli, Hengrong Du, Xinran Liu, and Soheil Kolouri. Partial gromov-wasserstein metric. *arXiv preprint arXiv:2402.03664*, 2024.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pp. 1–12, 2011.

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

Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018a.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018b.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.

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

Alessio Figalli and Nicola Gigli. 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.

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

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

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

Xinran Liu, Yikun Bai, Huy Tran, Zhanqi Zhu, Matthew Thorpe, and Soheil Kolouri. Ptlp: Partial transport l^p distances. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Benedetto Piccoli and Francesco Rossi. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211:335–358, 2014.

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

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

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in Neural Information Processing Systems*, 35:21792–21804, 2022.

Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal transport for structured data with application on graphs. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/titouan19a.html`.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

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

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

S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.