Learning Geometry Preserving Optimal Transport Plan via Convex Relaxation

Junhyoung Chung*
Department of Statistics
Seoul National University
Seoul 08826, Republic of Korea
junhyoung0534@gmail.com

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

Optimal transport (OT) is a foundational tool for comparing probability measures by accounting for the cost of moving mass across spaces [Villani et al., 2008]. Beyond classical settings where distributions live in Euclidean domains, modern applications in statistics, machine learning, and computer vision routinely involve signals supported on non-Euclidean or structured spaces (e.g., meshes, graphs, manifolds) together with auxiliary features (descriptors). In such problems, meaningful alignment must jointly respect (i) feature correspondence and (ii) the geometry of the underlying domain.

A popular way to incorporate geometry is the Gromov–Wasserstein (GW) framework, which matches distributions by comparing within-domain pairwise distances. More recently, the fused GW (FGW) discrepancy [Vayer et al., 2020] blends feature matching with GW, offering a flexible interpolation. However, the GW term is intrinsically non-convex in the coupling, leading to non-unique minimizers and algorithmic sensitivity to initialization. As a consequence, generic solvers often converge only to stationary points, which complicates both optimization and statistical analysis.

This paper. We introduce a *convex* formulation of fused optimal transport (FOT) that preserves domain geometry via a kernelized operator discrepancy. Concretely, we replace the non-convex GW component by a convex surrogate built from the distance–kernel operator $D^{\kappa}_{\mathbb{P}_X}$ acting on $L^2(S, \mathbb{P}_X)$, where $K(x, x') = \kappa(d_S(x, x'))$ with a bounded continuous monotone κ. For a coupling $\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)$ with conditional expectation operator T_π, our structural penalty is the squared Hilbert–Schmidt norm

$$\left\| D^{\kappa}_{\mathbb{P}_X} T_\pi \;-\; T_\pi D^{\kappa}_{\mathbb{P}_Y} \right\|^2_{\mathrm{HS}},$$

which vanishes exactly when T_π commutes with the distance–kernel operators. Intuitively, this enforces that *distance potentials* are preserved under the coupling, encouraging near-isometries at the level of kernelized geometry. Coupled with a standard feature cost, the overall objective

$$\mathcal{L}(\pi) = (1-\alpha)\, \mathbb{E}_{(X,Y)\sim\pi}\left[\|f(X) - f(Y)\|_2^2\right] + \frac{\alpha}{2} \left\| D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y} \right\|^2_{\mathrm{HS}}$$

is *convex in* π (Theorem 1). At the sample level this yields a convex quadratic program ((2)) that we solve with a projection-free Frank–Wolfe (FW) method and, when desired, post-process via a linear-assignment (LAP) projection to obtain a hard matching (algorithm 1).

*https://junhyoung-chung.github.io/

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Geometric fidelity. We show that the structural penalty is *isometry-consistent*: if π is induced by a measurable injective map T that preserves d_S almost everywhere, then the penalty vanishes; conversely, under a strictly monotone κ, vanishing penalty implies distance preservation and hence identifies isometries (Proposition 2). Thus our convex surrogate targets the same structure-preserving maps sought by GW/FGW, but without the non-convex landscape.

Statistical and computational guarantees. Because the objective is convex in the coupling, global optima can be obtained with standard first-order methods, enabling clean finite-sample analysis. We provide a basic consistency result that decomposes the error into an *optimization* part (controlled by FW iterations) and a *statistical* part governed by the empirical OT approximation of the marginals (Theorem 2). This separates algorithmic accuracy from sample complexity and clarifies how geometry (via D^κ) interacts with feature noise.

Contributions.

1. **Convex FOT objective.** We propose a kernel-operator surrogate for the GW term and prove convexity of the overall fused objective in the coupling (Theorem 1).
2. **Isometry consistency.** We establish that the surrogate vanishes precisely on isometries (and only on them under strictly monotone κ), aligning with the geometric aims of GW while avoiding non-convexity (Proposition 2).
3. **Scalable solver with guarantees.** We give a simple Frank–Wolfe algorithm for the empirical convex QP, with an optional LAP projection for hard matchings, and derive an optimization-statistical error bound (Algorithm 1, Theorem 2).
4. **Empirical validation.** On lattice-structured graph benchmarks with block-aligned features, our method smoothly trades off feature alignment and geometry preservation and exhibits stable convergence, in contrast to the concave behavior of GW along permutation trajectories (cf. the motivating example).

Organization. Section 2 formalizes the kernel-operator construction and proves convexity and isometry consistency. Section 3 presents the empirical QP, FW solver, and LAP projection together with complexity remarks and the optimization–statistical error bound. Section 4 reports simulations on lattice block models that stress-test geometric alignment under within-block ambiguity.

2 Methodology

Notations. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and (S, d_S) be a compact metric space. A measurable map $X : \Omega \to S$ is called a random element with distribution $\mathbb{P}_X := \mathbb{P} \circ X^{-1}$. Denote $L^2(S, \mathbb{P}_X)$ by the Hilbert space of real-valued, square integrable functions on S with respect to $\mathbb{P}_X$. We also introduce a feature space $M \subset \mathbb{R}^d$ which is compact, and call any one-to-one and continuous $f : S \to M$ a feature function. Throughout this study, we assume that $\mathrm{diam}(S) = \mathrm{diam}(M) = 1$, where $\mathrm{diam}(A) := \sup_{x,x' \in A} d_A(x, x')$. For probability measures $\mu_1, ..., \mu_m$ on S, denote by

$$\Pi(\mu_1, ..., \mu_m) := \{\pi \text{ on } S^m : \text{ the marginals are } \mu_1, ..., \mu_m\}$$

the set of all couplings between $\mu_1, ..., \mu_m$. Given an arbitrary coupling $\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)$, the conditional expectation operator $T_\pi : L^2(S, \mathbb{P}_Y) \to L^2(S, \mathbb{P}_X)$ is defined as $(T_\pi g)(x) := \mathbb{E}[g(Y) \mid X = x]$ for any $g \in L^2(S, \mathbb{P}_Y)$. Lastly, we say a measurable map $T : S \to S$ pushes forward μ to ν if $\mu(T^{-1}(A)) = \nu(A)$ for all $A \in \mathcal{B}(S)$, where $\mathcal{B}(S)$ is the Borel σ-algebra of S. We denote $T_\# \mu = \nu$ if T pushes forward μ to ν.

Fused Gromov-Wasserstein Discrepancy. For $0 \leq \alpha \leq 1$ and a feature function f, Vayer et al. [2020] propose the following optimization problem:

$$\inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} (1-\alpha)\mathbb{E}_{(X,Y)\sim\pi}\left[\|f(X) - f(Y)\|_2^2\right] + \alpha \mathbb{E}_{\substack{(X,Y)\sim\pi \\ (X',Y')\sim\pi}} \left[\left|d_S(X, X') - d_S(Y, Y')\right|^2\right]. \tag{1}$$

The first term enforces feature-wise alignment via f, while the second encourages structural consistency under the spatial metric d_S. When $\alpha = 0$, the problem reduces to classical quadratic OT; when $\alpha = 1$, it coincides with the Gromov–Wasserstein setting emphasizing relational geometry.

Proposition 1 (Existence of a minimizer)**.** *For each* $0 \leq \alpha \leq 1$, (1) *admits at least one minimizer; that is,* (1) *is solvable.*

Proof. The proof can be found in Vayer et al. [2020]. □

The existence follows from standard weak compactness of the set of couplings $\Pi(\mathbb{P}_X, \mathbb{P}_Y)$ and lower semicontinuity of the objective functional. However, the minimizer of (1) is not necessarily unique, and due to the non-convexity of the second (structural) term, the optimization landscape may contain multiple local minima. Consequently, standard numerical algorithms can only guarantee convergence to stationary or locally optimal solutions, rather than the global optimum. This highlights the importance of developing a convex reformulation or an appropriate convex relaxation of the fused Gromov–Wasserstein problem to ensure computational tractability and theoretical robustness.

Proposed method. Our proposed method introduces a surrogate loss for the second term in (1), thereby ensuring that the problem is convex.

Definition 1 (Distance kernel)**.** *A function* $K : S \times S \to \mathbb{R}_+$ *is said to be a* distance kernel *if there exists a bounded by* 1*, continuous, (strictly) monotone* $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ *such that*

$$K(x, x') = \kappa\left(d_S(x, x')\right).$$

The direction of monotonicity of κ determines the geometric emphasis of the regularization term. When κ is decreasing (for example, $\kappa(t) = e^{-t}$ or $(1 + t)^{-p}$), the kernel assigns larger weights to nearby pairs, thus promoting preservation of local geometric structures. Conversely, when κ is increasing but bounded (for example, $\kappa(t) = \exp(t)/(1 + \exp(t))$), it emphasizes distant pairs, discouraging global contraction and encouraging large-scale geometric alignment. Finally, the boundedness of κ ensures that K is bounded and that the operator $D^{\kappa}_{\mathbb{P}_X}$ is Hilbert–Schmidt; importantly, the convexity result in Theorem 1 holds regardless of the monotonicity direction, which will be discussed later.

Definition 2 (Distance potential operator)**.** *Let* $(S, d_S, \mathbb{P}_X)$ *be a compact metric space. The* distance potential operator $D^{\kappa}_{\mathbb{P}_X} : L^2(S, \mathbb{P}_X) \to L^2(S, \mathbb{P}_X)$ *is defined by*

$$(D^{\kappa}_{\mathbb{P}_X} f)(x) \coloneqq \mathbb{E}\left[K(x, X) f(X)\right] = \int_S K(x, y)\, f(y)\, \mathbb{P}_X(dy), \quad \forall f \in L^2(S, \mathbb{P}_X),\ \forall x \in S.$$

The distance potential operator is a special case of a Hilbert-Schmidt operator. Intuitively, $(D^{\kappa}_{\mathbb{P}_X} f)(x)$ represents a distance-weighted average of f with respect to the point x.

Lemma 1. *For any* $\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)$*,*

$$\|D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}\|^2_{\mathrm{HS}} = \int_S \int_S \Gamma_\pi(x, y)^2 \mathbb{P}_X(dx) \mathbb{P}_Y(dy),$$

where $\Gamma^1_\pi(x, y) \coloneqq \mathbb{E}_\pi[K(x, X) \mid Y = y]$*,* $\Gamma^2_\pi(x, y) \coloneqq \mathbb{E}_\pi[K(y, Y) \mid X = x]$*, and* $\Gamma_\pi(x, y) \coloneqq \Gamma^1_\pi(x, y) - \Gamma^2_\pi(x, y)$*.*

Proof. See Appendix A.1. □

Lemma 1 proposes a convex surrogate for the Gromov-Wasserstein term in (1), which can be explicitly expressed as the squared integral of a kernel function $\Gamma_\pi(x, y)$. This kernel Γ_π measures the discrepancy between two conditional expected distances: $\Gamma^1_\pi(x, y)$ and $\Gamma^2_\pi(x, y)$. Specifically, $\Gamma^1_\pi(x, y)$ represents the expected kernel value $K(x, X)$ given $Y = y$, while $\Gamma^2_\pi(x, y)$ is the expected kernel value $K(y, Y)$ given $X = x$. The regularization term can thus be interpreted as a metric that quantifies the symmetric alignment of these "cross-spatial" kernel-weighted expectations induced by the coupling π.

Theorem 1. *For $0 \leq \alpha \leq 1$ and a feature function f, consider the optimization problem*

$$\inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} \underbrace{(1-\alpha)\mathbb{E}_{(X,Y)\sim\pi}\left[\|f(X) - f(Y)\|_2^2\right] + \frac{\alpha}{2}\|D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}\|^2_{\mathrm{HS}}}_{=:\mathcal{L}(\pi)}, \tag{2}$$

where $\|\cdot\|^2_{\mathrm{HS}}$ is a Hilbert-Schmidt norm. Then, (2) *is a convex problem.*

Proof. See Appendix A.2. □

Theorem 1 establishes that the proposed optimization problem (2) is convex with respect to the coupling π. This stems from the fact that the first term (feature-wise alignment) is linear in π, and the second regularization term is also a convex function of π. The latter holds because the map $\pi \mapsto T_\pi$ is affine, and the squared Hilbert-Schmidt norm $\|\cdot\|^2_{\mathrm{HS}}$ is a convex function; their composition thus preserves convexity (as detailed in Appendix A.2). Consequently, this problem formulation circumvents the computational challenges arising from non-convexity, which are inherent to the original Fused Gromov-Wasserstein problem (1), and guarantees that a global optimum can be efficiently found.

Proposition 2. *Let $T : S \to S$ be an injective measurable map, and consider $\pi = (\mathrm{Id}, T)_{\#}\mathbb{P}_X$. Then,*

$$\|D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}\|^2_{\mathrm{HS}} = 0 \iff d_S(T(x), T(x')) = d_S(x, x'), \text{ for } \mathbb{P}_X \otimes \mathbb{P}_X\text{-a.e. } (x, x').$$

The converse holds if κ is further assumed to be strictly monotone. Moreover, if in addition T is continuous and $\mathrm{supp}(\mathbb{P}_X) = S$, then the identity $d_S(T(x), T(x')) = d_S(x, x')$ holds for all $x, x' \in S$, hence T is an isometry on S.

Proof. See Appendix A.3. □

Proposition 2 provides a crucial validation for the proposed regularization term, demonstrating its consistency with the goals of geometric structure preservation. It shows that the regularization term vanishes if the coupling π is induced by a deterministic almost-isometry T, and the converse holds when the transform is strictly monotone. This equivalence ($\iff$) is a powerful consequence of the strictly monotone property of the kernel function κ (Definition 1), which ensures that the kernel values are identical ($K(x, x') = K(T(x), T(x'))$) if and only if the underlying distances are identical ($d_S(x, x') = d_S(T(x), T(x'))$). This confirms that our convex surrogate correctly and exclusively identifies these ideal, structure-preserving maps as optimal solutions for the structural part of the problem, mimicking the behavior of the original Gromov-Wasserstein discrepancy.

Indeed, this also reveals that the objective (2) is generally not strictly convex; for instance, if multiple distinct isometries exist and they produce the same feature-matching cost, they will all be global minimizers.

Motivation. We provide a motivation for comparing our convex surrogate with the Gromov–Wasserstein penalty. Setting $K(x, x') = d_S(x, x')$, consider the following 3×3 matching problem:

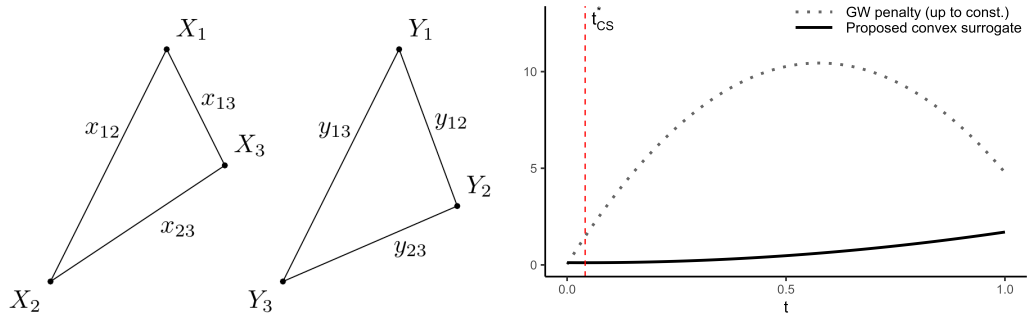
$$D_{\mathbb{P}_X} = \begin{pmatrix} 0 & x_{12} & x_{13} \\ x_{12} & 0 & x_{23} \\ x_{13} & x_{23} & 0 \end{pmatrix}, \quad D_{\mathbb{P}_Y} = \begin{pmatrix} 0 & y_{12} & y_{13} \\ y_{12} & 0 & y_{23} \\ y_{13} & y_{23} & 0 \end{pmatrix}.$$

For clarity, we consider the trajectory $3\pi(t) = tI + (1-t)P$ over $0 \leq t \leq 1$, where P is the permutation matrix obtained by switching the second and third rows of the identity matrix, that is,

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \pi(t) = \frac{1}{3}\left(tI + (1-t)P\right), \quad 0 \leq t \leq 1.$$

A straightforward calculation yields

$$\|D_{\mathbb{P}_X} T_{\pi(t)} - T_{\pi(t)} D_{\mathbb{P}_Y}\|^2_{\mathrm{HS}} = \|D_{\mathbb{P}_X}\pi(t) - \pi(t) D_{\mathbb{P}_Y}\|^2_F = \frac{1}{9}(at^2 + bt + c),$$



(a) Geometric structure of source and target samples (b) Proposed convex surrogate and GW penalty

Figure 1: Comparison of the geometric structures and associated penalty functions

where

$$
\begin{aligned}
A &= 2\left[(x_{12}-y_{12})^2+(x_{13}-y_{13})^2+(x_{23}-y_{23})^2\right],\\
B &= 2\left[(x_{13}-y_{12})^2+(x_{12}-y_{13})^2+(x_{23}-y_{23})^2\right],\\
C &= (x_{12}+x_{13}-y_{12}-y_{13})^2,\\
a &= A+B-2C, \quad b=-2B+2C, \quad c=B.
\end{aligned}
$$

Thus, when $a \neq 0$, the unique minimizer t^*_{CS} is given by

$$
t^*_{\mathrm{CS}} = \frac{B-C}{(A-C)+(B-C)}.
$$

Similarly, we can readily derive

$$
\mathbb{E}_{\substack{(X,Y)\sim\pi(t)\\(X',Y')\sim\pi(t)}}\left[\left|d_S(X,X')-d_S(Y,Y')\right|^2\right] = \text{const.} - \frac{2}{3}\mathrm{Tr}(D_{\mathbb{P}_X}\pi(t)D_{\mathbb{P}_Y}\pi(t)^\top),
$$

and

$$
\begin{aligned}
&\mathrm{Tr}(D_{\mathbb{P}_X}\pi(t)D_{\mathbb{P}_Y}\pi(t)^\top) = Dt^2+E(1-t)^2+Ft(1-t),\\
&D = 2(x_{12}y_{12}+x_{13}y_{13}+x_{23}y_{23}),\\
&E = 2(x_{13}y_{12}+x_{12}y_{13}+x_{23}y_{23}),\\
&F = 2(x_{12}+x_{13})(y_{12}+y_{13}).
\end{aligned}
$$

Since the second derivative of the above expression is positive (equal to $8x_{23}y_{23} > 0$), the GW penalty is concave in t. Comparing the two endpoints, the minimizer t^*_{GW} can be expressed as

$$
t^*_{\mathrm{GW}} = \begin{cases} 0 & A \geq B,\\ 1 & A \leq B. \end{cases}
$$

Figure 1 visualizes the case when $A \gg B \approx 0$. While both penalties encourage small values of t, the proposed convex surrogate attains its minimum near zero, whereas the GW penalty reaches its minimum exactly at $t^*_{\mathrm{GW}} = 0$. Nevertheless, due to its concave shape, the GW objective may converge to the opposite extreme ($t = 1$) depending on the initialization.

Overall, this example highlights the difference between the proposed convex surrogate and the GW penalty. While the GW objective exhibits a concave behavior that often results in extreme solutions (corresponding to permutation-like transport plans), our convex surrogate yields a smooth and well-behaved solution. In particular, the minimizer t^*_{CS} varies continuously with the geometric discrepancy, while ensuring the convexity. This property ensures numerical stability and uniqueness of the solution, making the convex surrogate more suitable for optimization and statistical analysis, especially in complex settings where the GW penalty might fall into suboptimal solutions due to its non-convexity.

Algorithm 1 Convex Quadratic Fused Transport Plan via FW and LAP Projection

Require: Source data $\{(X_i, f(X_i))\}_{i=1}^{n_X}$, target data $\{(Y_j, f(Y_j))\}_{j=1}^{n_Y}$, weight parameter $0 \le \alpha \le 1$, distance kernel K, max iters T

Construct matrices:
$(C_f)_{ij} \leftarrow \|f(X_i) - f(Y_j)\|_2^2, \quad (\hat{D}_X^\kappa)_{ii'} \leftarrow K(X_i, X_{i'}), \quad (\hat{D}_Y^\kappa)_{jj'} \leftarrow K(Y_j, Y_{j'})$

Initialize $\pi^{(0)}$

for $t = 0, ..., T-1$ **do**

  Calculate the gradient $\nabla \mathcal{L}_{n_X n_Y}(\pi^{(t)})$ in (3)

  Take $\tilde{\pi}^{(t)} \leftarrow \arg\min_{\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)} \mathrm{Tr}(\nabla \mathcal{L}_{n_X n_Y}(\pi^{(t)})^\top \pi)$

  $\pi^{(t+1)} \leftarrow (1-\gamma_t)\pi^{(t)} + \gamma_t \tilde{\pi}^{(t)}$ for some $0 < \gamma_t < 1$

end for

$\hat{\pi} \leftarrow \pi^{(T)}$

Optional (LAP): $\hat{P} \leftarrow \arg\max_{P \in \mathcal{P}} \mathrm{Tr}(P^\top \hat{\pi})$, where $\mathcal{P}$ is as defined in (5)

Return: $\hat{\pi}$ (and optionally $\hat{P}$)

3 Algorithm

Suppose that we have $(X_i, f(X_i))$ for $i = 1, ..., n_X$ as source data and $(Y_j, f(Y_j))$ for $j = 1, ..., n_Y$ as target data. Denote $\hat{\mathbb{P}}_X$ and $\hat{\mathbb{P}}_Y$ by the empirical distributions of X and Y, respectively. Then, the empirical version of (2) corresponds to

$$\inf_{\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)} (1-\alpha)\mathbb{E}_{(X,Y)\sim\pi}\left[\|f(X) - f(Y)\|_2^2\right] + \frac{\alpha}{2}\|D_{\hat{\mathbb{P}}_X}^\kappa T_\pi - T_\pi D_{\hat{\mathbb{P}}_Y}^\kappa\|_{\mathrm{HS}}^2.$$

In fact, the above problem can be written as a convex quadratic program (and reduces to a linear program when $\alpha = 0$):

$$\min_{\pi} \underbrace{(1-\alpha)\mathrm{Tr}\left(C_f^\top \pi\right) + \frac{\alpha}{2n_X n_Y}\left\|n_Y \hat{D}_X^\kappa \pi - n_X \pi \hat{D}_Y^\kappa\right\|_F^2}_{=:\mathcal{L}_{n_X n_Y}(\pi)}$$
$$\text{s.t.} \quad \pi \mathbf{1}_{n_Y} = \frac{1}{n_X}\mathbf{1}_{n_X}, \quad \pi^\top \mathbf{1}_{n_X} = \frac{1}{n_Y}\mathbf{1}_{n_Y}, \quad \pi \ge 0, \tag{3}$$

where $\pi \in \mathbb{R}_+^{n_X \times n_Y}$, $(C_f)_{ij} = \|f(X_i) - f(Y_j)\|_2^2$, $(\hat{D}_X^\kappa)_{ii'} = K(X_i, X_{i'})$, and $(\hat{D}_Y^\kappa)_{jj'} = K(Y_j, Y_{j'})$.

As (3) is a convex quadratic program, numerous standard optimization algorithms are available to find its global minimizer. In this paper, however, we focus on the Frank-Wolfe (FW) algorithm, also known as the conditional gradient (CG) method. The FW algorithm is particularly well-suited for this problem due to its "projection-free" nature. Unlike projected gradient methods that require a potentially costly projection back onto the feasible set $\Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)$ at each iteration, the FW algorithm only requires solving a linear minimization problem over this same set. This linear subproblem (or "Linear Minimization Oracle") is often computationally simpler and more efficient to solve than the full projection, making the FW algorithm an attractive choice for optimization problems over the transport polytope.

In many practical applications, it is often more desirable to find a deterministic transport map (or hard assignment) from X to Y, rather than the soft coupling $\hat{\pi}$ found by (3). While the optimal solution $\hat{\pi}$ is not guaranteed to be deterministic, we can obtain such a map by solving the following linear assignment problem (LAP):

$$\hat{P} \coloneqq \arg\max_{P \in \mathcal{P}} \mathrm{Tr}(P^\top \hat{\pi}), \tag{4}$$

where $\hat{\pi}$ is an optimal solution to (3) and $\mathcal{P}$ denotes the set of deterministic assignment matrices. Specifically,

$$\mathcal{P} \coloneqq \{P \in \{0,1\}^{n_X \times n_Y} : P\mathbf{1}_{n_Y} \le \mathbf{1}_{n_X},\ P^\top \mathbf{1}_{n_X} \le \mathbf{1}_{n_Y}\}. \tag{5}$$

This LAP seeks the hard assignment P that best aligns with the optimal soft coupling $\hat{\pi}$ and can be solved efficiently using standard methods like the Hungarian algorithm.

Consistency We first state some technical assumptions to provide a convergence rate for our proposed method.

Assumption 1.

(A1) (S, d_S, η) *is a* d*-dimensional compact Riemannian manifold.*

(A2) The densities of X *and* Y*, defined by* $p_X = d\mathbb{P}_X/d\eta$ *and* $p_Y = d\mathbb{P}_Y/d\eta$*, are strictly bounded above and below:*

$$0 < p_{\min} \leq p_X, p_Y \leq p_{\max} < \infty.$$

(A3) For each $n_X, n_Y \geq 1$*, no sample is replicated, that is, all* $X_1, ..., X_{n_X}$ *and* $Y_1, ..., Y_{n_Y}$ *are distinct, respectively.*

(A4) f *is* L_f*-Lipschitz continuous.*

Theorem 2. *Let* $\hat{\pi}$ *be the solution of Algorithm 1 with* $\alpha > 0$*. Suppose that Assumption 1 holds. Then, for some constant* $C > 0$*,*

$$\left| \mathcal{L}_{n_X n_Y}(\hat{\pi}) - \inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} \mathcal{L}(\pi) \right| \leq \underbrace{\frac{8\alpha\, n_{\max}^2}{n_{\min}(T+1)}}_{\text{Optimization error}} + C \underbrace{\left(W_2^{d_S}(\mathbb{P}_X, \hat{\mathbb{P}}_X) + W_2^{d_S}(\mathbb{P}_Y, \hat{\mathbb{P}}_Y) \right)}_{\text{Statistical error}}.$$

4 Numerical Experiments

This section provides numerical experiments to empirically validate the main results: (i) the proposed algorithm asymptotically returns a global minimizer of (2); and (ii) it provides a compelling estimator that strikes a balance between minimizing transport costs in feature space and preserving geometry.

4.1 Empirical Evidence

4.2 Comparison to the FGW Method

In this section, we aim to demonstrate the performance of our proposed algorithm (2) compared to the FGW method (1) with graph data.

Data generation. We conduct a simulation study generating 100 instances of lattice-structured objects with clusters. Each simulated instance consists of an undirected graph with $B \in \{2, 3, 4\}$ latent blocks and $N = B^{2B-1}$ nodes. To generate ground-truth correspondences with controlled structural similarity, we first sample a lattice block model whose within-block connectivity is organized as a two-dimensional grid. Specifically, nodes within each block are arranged on a nearly square lattice and connected via von Neumann (4-neighbor) wiring, while edges between different blocks are added independently with probability $p_{\text{out}} = 0.02$.

To construct the paired graphs (G_X, G_Y), we first generate G_X from the above model, and then apply a block-wise permutation P_{true} that shuffles node identities within each block and permutes the block order according to

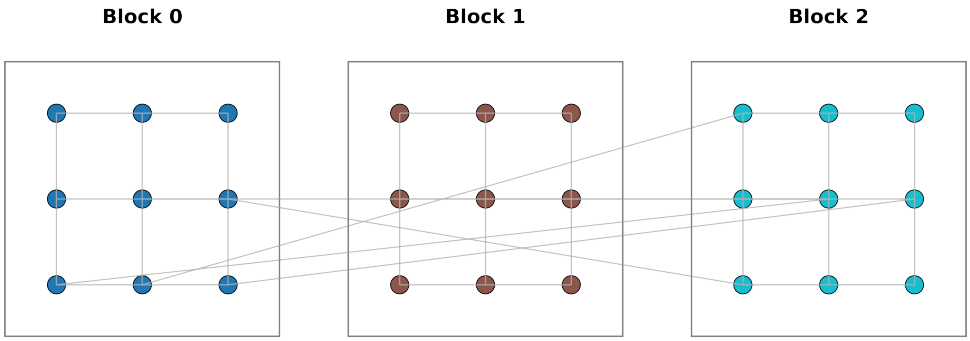
$$b_{\text{target}} \equiv b_{\text{source}} + 1 \pmod{B}.$$

The resulting $G_Y = P_{\text{true}}^\top G_X P_{\text{true}}$ thus represents an isomorphism of G_X but with latent correspondences obscured both within and across blocks. Figure 2 shows an example for simulated lattice block models used in the experiments.

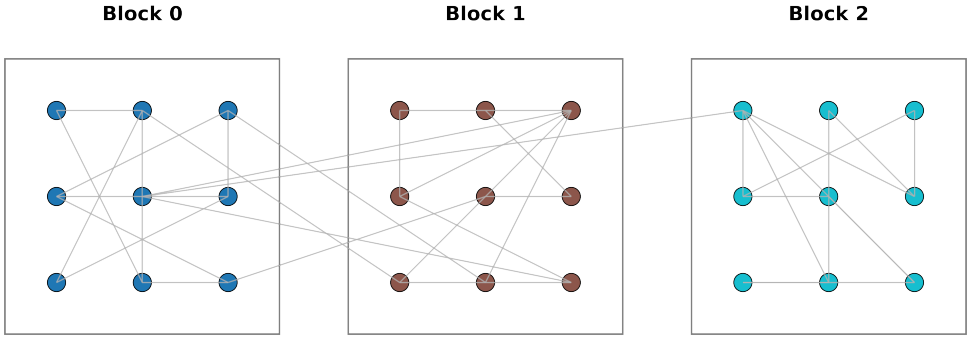
Each node i is further assigned a feature vector $f_i \in \mathbb{R}^B$ drawn from a B-dimensional Gaussian aligned with the block labels: if node i belongs to the b-th cluster,

$$f(i) \sim N(e_b, \sigma^2 I_B), \quad e_b = [0, ...0, \underset{b\text{-th}}{1}, 0, ..., 0]^\top,$$

where $\sigma^2 \in \{0, 0.001, 1\}$, so that features are discriminative at the block level but nearly indistinguishable within each block, especially when σ^2 is small.



(a) Source graph G_X consisting of three lattice-structured blocks ($B = 3$, $N = 27$).



(b) Target graph G_Y obtained by applying a within-block shuffling and a cyclic block permutation to G_X.

Figure 2: Illustration of simulated lattice block models. The target graph preserves the same global structure as the source but obscures node-wise correspondences both within and across blocks.

Implementation details. The pairwise feature cost is defined by the squared Euclidean distance $C_f(i, j) = \|f(i) - f(j)\|_2^2$, as specified in Algorithm 1. To encode structural geometry, we consider three complementary graph-based distances: (i) the geodesic distance, (ii) a diffusion distance based on the random-walk heat kernel, and (iii) an RKHS distance induced by the symmetric heat kernel. For pair comparison, we set $\kappa(x) = x$ for the proposed algorithm.

The geodesic distance d_G measures the length of the shortest path between nodes i and j. The diffusion distance $d_{\mathrm{diff},t}$ is computed from the random-walk heat kernel $K_{\mathrm{rw},t} = \exp(-tL_{\mathrm{rw}})$, which captures multi-step diffusion connectivity. The RKHS distance $d_{\mathrm{rkhs},t}$ is defined from the symmetric heat kernel $K_{\mathrm{sym},t} = \exp(-tL_{\mathrm{sym}})$ and reflects smooth, global interactions. In all experiments we fix the diffusion scale to $t = 1$ and approximate $\exp(-tL)$ by a second-order Taylor expansion for computational efficiency.

Unless otherwise stated, we set the fusion weight to $\alpha \in \{0, 0.5, 1\}$ for all methods and fix the number of iterations to $T = 100$. We report performance using $\hat{P}$ obtained by solving the LAP. To study sensitivity to the starting point, we consider two initializations of the transport plan: (i) a fully random doubly-stochastic initialization, and (ii) the outer product of empirical marginals (uniform in our setting), i.e., $\pi^{(0)} = \hat{\mathbb{P}}_X \otimes \hat{\mathbb{P}}_Y$. For the FGW baseline we use the `gromov.fused_gromov_wasserstein` function from the `POT` Python library.

Results. Table 1 shows the results when both methods are initialized by the independent product coupling $\hat{\mathbb{P}}_X \otimes \hat{\mathbb{P}}_Y$. Under this initialization, both the proposed algorithm and the FGW solver converge to the correct correspondence across all settings. This behavior can be understood in light of the illustration provided in Figure 1: although the GW objective is non-convex and typically exhibits

σ^2	$\alpha = 0$	Geodesic				Diffusion				RKHS			
		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 1$	
		Ours	FGW	**Ours**	FGW	**Ours**	FGW	**Ours**	FGW	**Ours**	FGW	**Ours**	FGW
0	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 1: Mean node-level matching accuracy over 10 runs ($\hat{\mathbb{P}}_X \otimes \hat{\mathbb{P}}_Y$ initialization only). Each pair of columns reports the results for the proposed algorithm (**Ours**) and the FGW method under different α values and distances. The $\alpha = 0$ column represents the pure optimal transport, which is shared by both methods.

σ^2	$\alpha = 0$	Geodesic				Diffusion				RKHS			
		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 1$	
		Ours	FGW	**Ours**	FGW	**Ours**	FGW	**Ours**	FGW	**Ours**	FGW	**Ours**	FGW
0	0.01	1.00	1.00	1.00	0.92	1.00	0.27	1.00	0.12	1.00	0.09	1.00	0.04
1	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	0.14	1.00	1.00	1.00	0.04

Table 2: Mean node-level matching accuracy over 10 runs (random initialization only). Each pair of columns reports the results for the proposed algorithm (**Ours**) and the FGW method under different α values and distances. The $\alpha = 0$ column represents the pure optimal transport, which is shared by both methods.

a concave trajectory that drives the solution toward one of its extreme points, the initialization here happens to lie symmetrically between competing optima. Consequently, the solver descends directly to the correct permutation rather than being trapped in an opposite extreme. From a theoretical perspective, this is a rather intriguing observation—despite the inherent non-convexity of the FGW formulation, the outer-product initialization empirically leads to globally optimal transport plans in these perfectly isomorphic graph settings. We leave a deeper understanding of this phenomenon and its potential connections to the geometry of the FGW landscape as an interesting direction for future research. In contrast, our convex formulation arrives at the same solution deterministically, without relying on such a fortuitous starting configuration.

Table 2 summarizes the results when the transport plan is initialized randomly. Not surprisingly, our algorithm achieves perfect node-level recovery across all α and kernel choices, maintaining its accuracy even when the feature information is uninformative within each block ($\sigma^2 = 0$). The FGW solver, however, exhibits large performance drops—most notably under diffusion and RKHS distances—confirming its vulnerability to initialization. These results empirically support the theoretical insight from Figure 1: the non-convex GW penalty tends to fall into one of several extreme local minima, producing nearly permutation-like but incorrect couplings, whereas the convex surrogate provides a smooth, well-behaved objective that guarantees convergence to a unique global solution. This robustness to initialization demonstrates the principal practical advantage of our convex approach over classical FGW.

References

J. Kitagawa, Q. Mérigot, and B. Thibert. Convergence of a newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21(9):2603–2651, 2019.

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

F. Santambrogio. Optimal transport for applied mathematicians. 2015.

T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

A Appendix

A.1 Proof for Lemma 1

Proof. For $g \in L^2(S, \mathbb{P}_Y)$, observe that

$$(D^{\kappa}_{\mathbb{P}_X} T_\pi g)(x) = \mathbb{E}\Big[K(x, X)\mathbb{E}\left[g(Y) \mid X\right]\Big] = \int_S \Gamma^1_\pi(x, y) g(y) \mathbb{P}_Y(dy),$$
$$(T_\pi D^{\kappa}_{\mathbb{P}_Y} g)(x) = \mathbb{E}\Big[(D^{\kappa}_{\mathbb{P}_Y} g)(Y) \mid X = x\Big] = \int_S \Gamma^2_\pi(x, y) g(y) \mathbb{P}_Y(dy).$$

Since K is bounded, $D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}$ is a well-defined Hilbert-Schmidt operator with a kernel $\Gamma_\pi(x, y)$:

$$\|D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}\|^2_{\mathrm{HS}} = \int_S \int_S \Gamma_\pi(x, y)^2 \mathbb{P}_X(dx)\mathbb{P}_Y(dy) \leq 4.$$

□

A.2 Proof for Theorem 1

Proof. Refer to Lemma 1 to confirm that the operator $D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}$ is well-defined and Hilbert–Schmidt for any $\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)$. To establish convexity of (2), it suffices to show that

$$\pi \;\mapsto\; \|D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}\|^2_{\mathrm{HS}}$$

is convex on $\Pi(\mathbb{P}_X, \mathbb{P}_Y)$.

Let $\pi_1, \pi_2 \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)$ and $t \in [0, 1]$, and define $\pi_t = t\pi_1 + (1 - t)\pi_2$. Since both π_1 and π_2 share the same marginals $\mathbb{P}_X$ and $\mathbb{P}_Y$, the disintegration theorem ensures that their corresponding conditional kernels satisfy

$$k_t(\cdot|x) = t\, k_1(\cdot|x) + (1 - t)\, k_2(\cdot|x) \quad \text{for } \mathbb{P}_X\text{-a.e. } x \in S.$$

That is, the conditional distribution of Y given $X = x$ under π_t is the convex combination of the conditional distributions under π_1 and π_2. Consequently, for any $g \in L^2(S, \mathbb{P}_Y)$,

$$\begin{aligned}(T_{\pi_t} g)(x) &= \int g(y)\, k_t(dy|x) = t \int g(y)\, k_1(dy|x) + (1 - t) \int g(y)\, k_2(dy|x) \\ &= t\,(T_{\pi_1} g)(x) + (1 - t)\,(T_{\pi_2} g)(x),\end{aligned}$$

which shows that T_{π_t} depends affinely on π, i.e.,

$$T_{\pi_t} = t\, T_{\pi_1} + (1 - t)\, T_{\pi_2}.$$

Because $D^{\kappa}_{\mathbb{P}_X}$ and $D^{\kappa}_{\mathbb{P}_Y}$ are linear operators, it follows that

$$D^{\kappa}_{\mathbb{P}_X} T_{\pi_t} - T_{\pi_t} D^{\kappa}_{\mathbb{P}_Y} = t\,(D^{\kappa}_{\mathbb{P}_X} T_{\pi_1} - T_{\pi_1} D^{\kappa}_{\mathbb{P}_Y}) + (1 - t)\,(D^{\kappa}_{\mathbb{P}_X} T_{\pi_2} - T_{\pi_2} D^{\kappa}_{\mathbb{P}_Y}).$$

Denoting $A_i := D^{\kappa}_{\mathbb{P}_X} T_{\pi_i} - T_{\pi_i} D^{\kappa}_{\mathbb{P}_Y}$ $(i = 1, 2)$ and $A_t := D^{\kappa}_{\mathbb{P}_X} T_{\pi_t} - T_{\pi_t} D^{\kappa}_{\mathbb{P}_Y}$, we have $A_t = tA_1 + (1 - t)A_2$. Since $\|\cdot\|^2_{\mathrm{HS}}$ is convex, we obtain

$$\|A_t\|^2_{\mathrm{HS}} = \|tA_1 + (1 - t)A_2\|^2_{\mathrm{HS}} \leq t\,\|A_1\|^2_{\mathrm{HS}} + (1 - t)\,\|A_2\|^2_{\mathrm{HS}}.$$

Therefore, $\pi \mapsto \|D^{\kappa}_{\mathbb{P}_X} T_\pi - T_\pi D^{\kappa}_{\mathbb{P}_Y}\|^2_{\mathrm{HS}}$ is convex on $\Pi(\mathbb{P}_X, \mathbb{P}_Y)$. □

A.3 Proof for Proposition 2

Proof. Since $Y = T(X)$ almost surely, for any $(x, y) \in \mathrm{supp}(\pi)$, we have

$$\Gamma_\pi^1(x,y) = \mathbb{E}[K(x,X) \mid Y = y] = \kappa\left(d_S\big(x, T^{-1}(y)\big)\right),$$
$$\Gamma_\pi^2(x,y) = \mathbb{E}[K(y,Y) \mid X = x] = \kappa\left(d_S\big(y, T(x)\big)\right).$$

Hence, for such (x, y),

$$\Gamma_\pi(x,y) = \Gamma_\pi^1(x,y) - \Gamma_\pi^2(x,y) = \kappa\left(d_S\big(x, T^{-1}(y)\big)\right) - \kappa\left(d_S\big(y, T(x)\big)\right).$$

Now, setting $y = T(x')$ gives

$$\Gamma_\pi\big(x, T(x')\big) = \kappa\left(d_S\big(x, x'\big)\right) - \kappa\left(d_S\big(T(x), T(x')\big)\right).$$

($\Leftarrow$) If $d_S(T(x), T(x')) = d_S(x, x')$ holds for $\mathbb{P}_X \otimes \mathbb{P}_X$-almost every (x, x'), substituting this into the above expression yields $\Gamma_\pi(x, T(x')) = 0$ for $\mathbb{P}_X \otimes \mathbb{P}_X$-almost every (x, x'). Consequently, $\Gamma_\pi = 0$ holds for π-almost every (x, y), which implies that the Hilbert–Schmidt norm is zero.

($\Rightarrow$) Now, assume that κ is strictly monotone. If $\|D^\kappa_{\mathbb{P}_X} T_\pi - T_\pi D^\kappa_{\mathbb{P}_Y}\|^2_{\mathrm{HS}} = 0$, then $\Gamma_\pi = 0$ holds for $\mathbb{P}_X \otimes \mathbb{P}_Y$-almost every (x, y), and in particular, for pairs $(x, T(x'))$ with respect to $\mathbb{P}_X \otimes \mathbb{P}_X$-almost every (x, x'). By the strict monotone property of κ, $d_S(T(x), T(x')) = d_S(x, x')$ holds for $\mathbb{P}_X \otimes \mathbb{P}_X$-almost every (x, x').

Finally, if T is continuous and $\mathrm{supp}(\mathbb{P}_X) = S$, then the function

$$F(x, x') := K(x, x') - K(T(x), T(x'))$$

is continuous and vanishes on the dense set $\mathrm{supp}(\mathbb{P}_X) \times \mathrm{supp}(\mathbb{P}_X)$. By continuity, $F \equiv 0$ on $S \times S$. Again by the strict monotonicity of κ, $d_S(T(x), T(x')) = d_S(x, x')$ holds for all $x, x' \in S$, implying that T is an isometry on S. □

A.4 Proof for Theorem 2

Proof. One key of consistency is to calculate the convergence rate of empirical distributions. To this end, we first introduce the Wasserstein-1 distance.

Definition 3 (Wasserstein-p distance)**.** *Let (S, d_S) be a compact metric space. For probability measures μ and ν on S, we define*

$$W_p^S(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{S \times S} d_S(x,y)^p \pi(dx, dy) \right)^{1/p}.$$

Lemma 2. *Consider the sequence of probability measures $\{\mu_n : n \geq 1\}$ and μ on S. Then,*

$$W_p^S(\mu_n, \mu) \to 0 \ \ as \ \ n \to \infty \iff \mu_n \xrightarrow{w} \mu \ \ as \ \ n \to \infty,$$

where $\mu_n \xrightarrow{w} \mu$ denotes the weak convergence.

Proof. See details in Villani et al. [2008]. □

First, note that

$$\begin{aligned}
&\left| \mathcal{L}_{n_X n_Y}(\hat{\pi}) - \inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} \mathcal{L}(\pi) \right| \\
&\leq \left| \mathcal{L}_{n_X n_Y}(\hat{\pi}) - \inf_{\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)} \mathcal{L}_{n_X n_Y}(\pi) \right| + \left| \inf_{\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)} \mathcal{L}_{n_X n_Y}(\pi) - \inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} \mathcal{L}(\pi) \right| \\
&=: E_1 + E_2.
\end{aligned}$$

We first provide the bound for E_1, which is the optimization error.

Lemma 3. *Let $\{\pi^{(t)} : t \geq 0\}$ be the sequence of iterates generated by Algorithm 1 with*

$$\gamma_t = \frac{2}{t+2},$$

for each t. Then, for any $t \geq 1$,

$$\mathcal{L}_{n_X n_Y}(\pi^{(t)}) - \inf \mathcal{L}_{n_X n_Y} \leq \frac{2\alpha}{n_{\min}(t+1)} \left(\|\hat{D}_X^\kappa\|_{\mathrm{op}} + \|\hat{D}_Y^\kappa\|_{\mathrm{op}} \right)^2.$$

Proof. See Appendix A.5. □

The proof for Lemma 3 is a standard technique to show the convergence of the FW algorithm. Considering that $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_\infty$,

$$\|\hat{D}_X^\kappa\|_{\mathrm{op}} \leq \max_i \sum_{j=1}^{n_X} [\hat{D}_X^\kappa]_{ij} \leq n_X, \quad \|\hat{D}_Y^\kappa\|_{\mathrm{op}} \leq n_Y.$$

This gives that

$$\left(\|\hat{D}_X^\kappa\|_{\mathrm{op}} + \|\hat{D}_Y^\kappa\|_{\mathrm{op}} \right)^2 \leq 4 n_{\max}^2.$$

Thus,

$$E_1 \leq \frac{8\alpha\, n_{\max}^2}{n_{\min}(T+1)}.$$

To control the bound of E_2, we introduce a useful lemma known as the gluing lemma.

Lemma 4 (Gluing lemma)**.** *Let $(\mathcal{X}_i, \mu_i)$, $i = 1, 2, 3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3), then one can construct a triple of random elements (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same distribution as (X_1, X_2) and (Z_2, Z_3) has the same distribution as (Y_2, Y_3).*

Proof. See details in Villani et al. [2008]. □

To use the gluing lemma, we first define two projection couplings $Q_X \in \Pi(\mathbb{P}_X, \hat{\mathbb{P}}_X)$ and $Q_Y \in \Pi(\hat{\mathbb{P}}_Y, \mathbb{P}_Y)$, which will be specified later.

For an arbitrary $\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)$, we can construct a coupling $\Xi \in \Pi(\mathbb{P}_X, \hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y, \mathbb{P}_Y)$ such that

$$\Xi(dx, d\hat{x}, d\hat{y}, dy) = \underbrace{Q_Y(dy \mid \hat{y})}_{\hat{\mathbb{P}}_Y \to \mathbb{P}_Y} \underbrace{\pi(d\hat{y} \mid \hat{x})}_{\hat{\mathbb{P}}_X \to \hat{\mathbb{P}}_Y} \underbrace{Q_X(dx \mid \hat{x})}_{\hat{\mathbb{P}}_X \to \mathbb{P}_X} \hat{\mathbb{P}}_X(d\hat{x}), \tag{6}$$

owing to the gluing lemma. Then, define

$$\Phi_n(\pi)(dx, dy) \coloneqq \int_{(\hat{x}, \hat{y}) \in S \times S} Q_X(dx \mid \hat{x})\, Q_Y(dy \mid \hat{y})\, \pi(d\hat{y} \mid \hat{x})\, \hat{\mathbb{P}}_X(d\hat{x}),$$

which makes clear that $\Phi_n(\pi) : \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y) \to \Pi(\mathbb{P}_X, \mathbb{P}_Y)$ is a well-defined function that maps the coupling on $\Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)$ to that on $\Pi(\mathbb{P}_X, \mathbb{P}_Y)$.

Let $\pi_n^* \in \arg\min_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} \mathcal{L}_{n_X n_Y}(\pi)$. Observe that

$$
\begin{aligned}
&\left|\mathcal{L}_{n_X n_Y}(\pi_n^*) - \inf_{\pi\in\Pi(\mathbb{P}_X,\mathbb{P}_Y)} \mathcal{L}(\pi)\right| \\
&\le |\mathcal{L}_{n_X n_Y}(\pi_n^*) - \mathcal{L}(\Phi_n(\pi_n^*))| + \left|\mathcal{L}(\Phi_n(\pi_n^*)) - \inf_{\pi\in\Pi(\mathbb{P}_X,\mathbb{P}_Y)} \mathcal{L}(\pi)\right| \\
&\le \sup_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} |\mathcal{L}_{n_X n_Y}(\pi) - \mathcal{L}(\Phi_n(\pi))| \\
&\qquad + |\mathcal{L}(\Phi_n(\pi_n^*)) - \mathcal{L}_{n_X n_Y}(\pi_n^*)| + \left|\mathcal{L}_{n_X n_Y}(\pi_n^*) - \inf_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} \mathcal{L}(\Phi_n(\pi))\right| \\
&\qquad + \left|\inf_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} \mathcal{L}(\Phi_n(\pi)) - \inf_{\pi\in\Pi(\mathbb{P}_X,\mathbb{P}_Y)} \mathcal{L}(\pi)\right| \\
&\le 3 \underbrace{\sup_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} |\mathcal{L}_{n_X n_Y}(\pi) - \mathcal{L}(\Phi_n(\pi))|}_{=:\mathrm{I}_n} + \underbrace{\left|\inf_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} \mathcal{L}(\Phi_n(\pi)) - \inf_{\pi\in\Pi(\mathbb{P}_X,\mathbb{P}_Y)} \mathcal{L}(\pi)\right|}_{=:\mathrm{II}_n}.
\end{aligned}
$$

For the second inequality, we use the fact that $\mathcal{L}_{n_X n_Y}(\pi_n^*) = \inf_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} \mathcal{L}_{n_X n_Y}(\pi)$ and that $|\inf_x f(x) - \inf_x g(x)| \le \sup_x |f(x) - g(x)|$.

Now, define

$$
Q_X \in \operatorname*{arg\,min}_{Q\in\Pi(\mathbb{P}_X,\hat{\mathbb{P}}_X)} \mathbb{E}_{(X,\hat{X})\sim Q}\left[d_S(X,\hat{X})^2\right], \quad Q_Y \in \operatorname*{arg\,min}_{Q\in\Pi(\hat{\mathbb{P}}_Y,\mathbb{P}_Y)} \mathbb{E}_{(\hat{Y},Y)\sim Q}\left[d_S(Y,\hat{Y})^2\right]. \quad (7)
$$

Then, by Definition 3, we get

$$
\begin{aligned}
W_2^{d_S}(\mathbb{P}_X,\hat{\mathbb{P}}_X) &= \left(\mathbb{E}_{(X,\hat{X})\sim Q_X}\left[d_S(X,\hat{X})^2\right]\right)^{1/2}, \\
W_2^{d_S}(\hat{\mathbb{P}}_Y,\mathbb{P}_Y) &= \left(\mathbb{E}_{(\hat{Y},Y)\sim Q_Y}\left[d_S(Y,\hat{Y})^2\right]\right)^{1/2}.
\end{aligned}
$$

We first look into I_n. For brevity, let $c_f(x,y) \coloneqq \|f(x) - f(y)\|_2^2$. Then,

$$
\begin{aligned}
|c_f(x,y) - c_f(x',y')| &\le \left(\|f(x) - f(y)\|_2 + \|f(x') - f(y')\|_2\right)\left(\|f(x) - f(y)\|_2 - \|f(x') - f(y')\|_2\right) \\
&\le 2\left(\|f(x) - f(x')\|_2 + \|f(y) - f(y')\|_2\right).
\end{aligned}
$$

The last inequality uses the fact that $\mathrm{diam}(M) = 1$. Note that this confirms that c_f is a bounded and 2-Lipschitz function with respect to the metric

$$
d_f((x,y),(x',y')) \coloneqq \|f(x) - f(x')\|_2 + \|f(y) - f(y')\|_2.
$$

Thus, by the duality formula of the Wasserstein-1 distance, for an arbitrary $\pi \in \Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)$,

$$
\left|\mathbb{E}_{(X,Y)\sim\pi}\left[c_f(X,Y)\right] - \mathbb{E}_{(X',Y')\sim\Phi_n(\pi)}\left[c_f(X',Y')\right]\right| \le 2W_1^{d_f}(\pi,\Phi_n(\pi)).
$$

Here $W_1^{d_f}(\pi,\Phi_n(\pi))$ can be decomposed as follows:

$$
\begin{aligned}
W_1^{d_f}(\pi,\Phi_n(\pi)) &= \inf_{\gamma\in\Pi(\pi,\Phi_n(\pi))} \mathbb{E}_{((X,Y),(X',Y'))\sim\gamma}\left[\|f(X) - f(X')\|_2 + \|f(Y) - f(Y')\|_2\right] \\
&\le \mathbb{E}_{(X',X,Y,Y')\sim\Xi}\left[\|f(X) - f(X')\|_2 + \|f(Y) - f(Y')\|_2\right] \\
&= \mathbb{E}_{(X',X)\sim Q_X}\left[\|f(X) - f(X')\|_2\right] + \mathbb{E}_{(Y,Y')\sim Q_Y}\left[\|f(Y) - f(Y')\|_2\right] \\
&\le L_f\left(\mathbb{E}_{(X',X)\sim Q_X}\left[d_S(X,X')\right] + \mathbb{E}_{(Y,Y')\sim Q_Y}\left[d_S(Y,Y')\right]\right) \\
&\le L_f\left(W_2^{d_S}(\mathbb{P}_X,\hat{\mathbb{P}}_X) + W_2^{d_S}(\mathbb{P}_Y,\hat{\mathbb{P}}_Y)\right).
\end{aligned}
$$

The second inequality is from (6).

We now analyze the convex surrogate term:

$$\begin{aligned}
&\left| \|D^{\kappa}_{\hat{\mathbb{P}}_X} T_\pi - T_\pi D^{\kappa}_{\hat{\mathbb{P}}_Y}\|^2_{\mathrm{HS}} - \|D^{\kappa}_{\mathbb{P}_X} T_{\Phi_n(\pi)} - T_{\Phi_n(\pi)} D^{\kappa}_{\mathbb{P}_Y}\|^2_{\mathrm{HS}} \right| \\
&= \left| \int_S \int_S \Gamma_\pi(x,y)^2 \hat{\mathbb{P}}_X(dx)\hat{\mathbb{P}}_Y(dy) - \int_S \int_S \Gamma_{\Phi_n(\pi)}(x,y)^2 \mathbb{P}_X(dx)\mathbb{P}_Y(dy) \right| \\
&\leq \underbrace{\left| \int_S \int_S (\Gamma_\pi(x,y)^2 - \Gamma_{\Phi_n(\pi)}(x,y)^2) \hat{\mathbb{P}}_X(dx)\hat{\mathbb{P}}_Y(dy) \right|}_{=:(A)} \\
&\qquad + \underbrace{\left| \int_S \int_S \Gamma_{\Phi_n(\pi)}(x,y)^2 (\hat{\mathbb{P}}_X(dx)\hat{\mathbb{P}}_Y(dy) - \mathbb{P}_X(dx)\mathbb{P}_Y(dy)) \right|}_{=:(B)}.
\end{aligned}$$

(A) can be further decomposed as

$$\begin{aligned}
(A) &\leq 2 \int_S \int_S \left| \Gamma_\pi(x,y) - \Gamma_{\Phi_n(\pi)}(x,y) \right| \hat{\mathbb{P}}_X(dx)\hat{\mathbb{P}}_Y(dy) \\
&= \frac{2}{n_X n_Y} \sum_{i,j} \left| \Gamma_\pi(X_i, Y_j) - \Gamma_{\Phi_n(\pi)}(X_i, Y_j) \right|.
\end{aligned}$$

Observe that

$$\begin{aligned}
&\left| \Gamma_\pi(X_i, Y_j) - \Gamma_{\Phi_n(\pi)}(X_i, Y_j) \right| \\
&\leq \left| \mathbb{E}_\pi(K(X_i, X) \mid Y = Y_j) - \mathbb{E}_{\Phi_n(\pi)}(K(X_i, X) \mid Y = Y_j) \right| \\
&\qquad + \left| \mathbb{E}_\pi(K(Y_j, Y) \mid X = X_i) - \mathbb{E}_{\Phi_n(\pi)}(K(Y_j, Y) \mid X = X_i) \right|.
\end{aligned}$$

By the symmetry, it suffices to analyze the first term in the last inequality. Note that

$$\mathbb{E}_\pi(K(X_i, X) \mid Y = Y_j) = n_Y \sum_{i'=1}^{n_X} \pi_{i'j} K(X_i, X_{i'}).$$

To discuss $\mathbb{E}_{\Phi_n(\pi)}(K(X_i, X) \mid Y = Y_j)$, we first state a useful lemma:

Lemma 5. *Let Q_X be defined as in* (7)*, and suppose that Assumption 1 holds. Then, for all $n_X \geq 1$, there exists a collection of measurable sets $\{V_i : i = 1, \ldots, n_X\}$ forming a Laguerre diagram such that*

$$V_i = \left\{ x \in S : \frac{1}{2} d_S(x, X_i)^2 - \psi_i \leq \frac{1}{2} d_S(x, X_j)^2 - \psi_j,\ \forall j \right\}, \quad \mathbb{P}_X(V_i) = \frac{1}{n_X},$$

for some $\psi = (\psi_1, \ldots, \psi_{n_X}) \in \mathbb{R}^{n_X}$. Then, the optimal coupling admits the form

$$Q_X = \sum_{i=1}^{n_X} \mathbb{P}_X \lfloor_{V_i} \otimes \delta_{X_i}, \quad \mathbb{P}_X \lfloor_{V_i}(dx) = I_{V_i}(x)\, \mathbb{P}_X(dx),$$

where I_{V_i} is the indicator function of V_i.

Proof. See details in Peyré et al. [2019], Kitagawa et al. [2019], Santambrogio [2015]. □

By Lemma 5, we have

$$\begin{aligned}
\Phi_n(\pi)(dx \mid Y_j) &= n_Y \sum_{i=1}^{n_X} Q_X(dx \mid X_i) \pi_{ij} \\
&= n_Y \sum_{i=1}^{n_X} \pi_{ij} \Big(n_X \mathbb{P}_X \lfloor_{V_i}(dx) \Big),
\end{aligned}$$

which gives

$$\mathbb{E}_{\Phi_n(\pi)}(K(X_i, X) \mid Y = Y_j) = n_Y \sum_{i'=1}^{n_X} \pi_{i'j} \int_{V_{i'}} K(X_i, x) \Big(n_X \mathbb{P}_X(dx) \Big).$$

Then,

$$\begin{aligned}
&\left| \Gamma_\pi(X_i, Y_j) - \Gamma_{\Phi_n(\pi)}(X_i, Y_j) \right| \\
&\leq n_X n_Y \sum_{i'=1}^{n_X} \pi_{i'j} \int_{V_{i'}} |K(X_i, x) - K(X_i, X_{i'})| \, \mathbb{P}_X(dx) \\
&\qquad\qquad + n_X n_Y \sum_{j'=1}^{n_Y} \pi_{ij'} \int_{W_{j'}} |K(Y_j, y) - K(Y_j, Y_{j'})| \, \mathbb{P}_Y(dy),
\end{aligned}$$

where the second term in the last inequality is obtained by similar technique.

Finally, we get

$$\begin{aligned}
(A) &\leq 2 \sum_{i,j,i'} \pi_{i'j} \int_{V_{i'}} |K(X_i, x) - K(X_i, X_{i'})| \, \mathbb{P}_X(dx) \\
&\leq \frac{2}{n_X} \sum_{i,i'} \int_{V_{i'}} |K(X_i, x) - K(X_i, X_{i'})| \, \mathbb{P}_X(dx) + \frac{2}{n_Y} \sum_{j,j'} \int_{W_{j'}} |K(Y_j, y) - K(Y_j, Y_{j'})| \, \mathbb{P}_Y(dy) \\
&\leq 2L_\kappa \sum_{i'=1}^{n_X} \int_{V_{i'}} d_S(x, X_{i'}) \mathbb{P}_X(dx) + 2L_\kappa \sum_{j'=1}^{n_Y} \int_{W_{j'}} d_S(y, Y_{j'}) \mathbb{P}_Y(dy) \\
&\leq 2L_\kappa \left(\sum_{i'=1}^{n_X} \int_{V_{i'}} d_S(x, X_{i'})^2 \mathbb{P}_X(dx) \right)^{1/2} + 2L_\kappa \left(\sum_{j'=1}^{n_Y} \int_{W_{j'}} d_S(y, Y_{j'})^2 \mathbb{P}_Y(dy) \right)^{1/2} \\
&\leq 2L_\kappa \left(W_2^{d_S}(\mathbb{P}_X, \hat{\mathbb{P}}_X) + W_2^{d_S}(\mathbb{P}_Y, \hat{\mathbb{P}}_Y) \right).
\end{aligned}$$

We now move onto (B):

$$\begin{aligned}
(B) \leq 2 &\left| \int_S \left(\int_S \Gamma_{\Phi_n(\pi)}(x, y) \hat{\mathbb{P}}_Y(dy) \right) (\hat{\mathbb{P}}_X - \mathbb{P}_X)(dx) \right| \\
&+ 2 \left| \int_S \left(\int_S \Gamma_{\Phi_n(\pi)}(x, y) \hat{\mathbb{P}}_x(dx) \right) (\hat{\mathbb{P}}_Y - \mathbb{P}_Y)(dy) \right|.
\end{aligned}$$

Similarly, we only bound the first term. Let

$$\gamma(x) \coloneqq \int_S \Gamma_{\Phi_n(\pi)}(x, y) \hat{\mathbb{P}}_Y(dy).$$

Then, it follows that

$$\begin{aligned}
\left| \int_S \left(\int_S \Gamma_{\Phi_n(\pi)}(x, y) \hat{\mathbb{P}}_Y(dy) \right) (\hat{\mathbb{P}}_X - \mathbb{P}_X)(dx) \right| &= \left| \mathbb{E}_{\hat{X} \sim \hat{\mathbb{P}}_X}[\gamma(\hat{X})] - \mathbb{E}_{X \sim \mathbb{P}_X}[\gamma(X)] \right| \\
&= \left| \int_S (\gamma(\hat{x}) - \gamma(x)) Q_X(dx, d\hat{x}) \right| \\
&\leq \int_S |\gamma(\hat{x}) - \gamma(x)| \, Q_X(dx, d\hat{x}).
\end{aligned}$$

Now we show that γ is Lipschitz on $\mathrm{supp}(Q_X)$. □

A.5 Proof for Lemma 3

Proof. First, we obtain the smoothness of $\mathcal{L}_{n_X n_Y}$. For brevity, denote $A \coloneqq n_Y \hat{D}_X^\kappa$ and $B \coloneqq n_X \hat{D}_Y^\kappa$. Observe that A and B are symmetric. Hence, the gradient of $\nabla \mathcal{L}_{n_X n_Y}(\pi)$ can be calculated as

$$\alpha^{-1}(n_X n_Y) \nabla \mathcal{L}_{n_X n_Y}(\pi) = A^2 \pi + \pi B^2 - 2A\pi B + \text{const.},$$

where the constant does not depend on π. Thus, the triangle inequality and the Cauchy-Schwarz inequality yield that

$$\begin{aligned}\alpha^{-1}(n_X n_Y)\left\|\nabla\mathcal{L}_{n_X n_Y}(\pi_1)-\nabla\mathcal{L}_{n_X n_Y}(\pi_2)\right\|_F &\le \left(\|A\|_{\mathrm{op}}+\|B\|_{\mathrm{op}}\right)^2\|\pi_1-\pi_2\|_F\\ &\le \left(n_Y\|\hat{D}_X^\kappa\|_{\mathrm{op}}+n_X\|\hat{D}_Y^\kappa\|_{\mathrm{op}}\right)^2\|\pi_1-\pi_2\|_F\\ &\le n_{\max}^2\left(\|\hat{D}_X^\kappa\|_{\mathrm{op}}+\|\hat{D}_Y^\kappa\|_{\mathrm{op}}\right)^2\|\pi_1-\pi_2\|_F.\end{aligned}$$

Thus, we get

$$\left\|\nabla\mathcal{L}_{n_X n_Y}(\pi_1)-\nabla\mathcal{L}_{n_X n_Y}(\pi_2)\right\|_F \le \alpha\cdot\frac{n_{\max}}{n_{\min}}\left(\|\hat{D}_X^\kappa\|_{\mathrm{op}}+\|\hat{D}_Y^\kappa\|_{\mathrm{op}}\right)^2\|\pi_1-\pi_2\|_F. \tag{8}$$

What remains is to show the convergence of the FW algorithm. Define

$$\beta \coloneqq \alpha\cdot\frac{n_{\max}}{n_{\min}}\left(\|\hat{D}_X^\kappa\|_{\mathrm{op}}+\|\hat{D}_Y^\kappa\|_{\mathrm{op}}\right)^2.$$

Let $\pi_n^* \in \arg\min_\pi \mathcal{L}_{n_X n_Y}(\pi)$, whose existence is guaranteed by the convexity. Then,

$$\begin{aligned}\mathcal{L}_{n_X n_Y}(\pi^{(t+1)})-\mathcal{L}_{n_X n_Y}(\pi^{(t)}) &\le \mathrm{Tr}\left(\nabla\mathcal{L}_{n_X n_Y}(\pi^{(t)})^\top(\pi^{(t+1)}-\pi^{(t)})\right)+\frac{\beta}{2}\|\pi^{(t+1)}-\pi^{(t)}\|_F^2\\ &= \gamma_t\mathrm{Tr}\left(\nabla\mathcal{L}_{n_X n_Y}(\pi^{(t)})^\top(\tilde{\pi}^{(t)}-\pi^{(t)})\right)+\frac{\beta}{2}\|\pi^{(t+1)}-\pi^{(t)}\|_F^2\\ &\le \gamma_t\mathrm{Tr}\left(\nabla\mathcal{L}_{n_X n_Y}(\pi^{(t)})^\top(\pi_n^*-\pi^{(t)})\right)+\frac{\beta}{2}\|\pi^{(t+1)}-\pi^{(t)}\|_F^2\\ &\le \gamma_t\left(\inf\mathcal{L}_{n_X n_Y}-\mathcal{L}_{n_X n_Y}(\pi^{(t)})\right)+\frac{\beta}{2}\|\pi^{(t+1)}-\pi^{(t)}\|_F^2.\end{aligned}$$

The first inequality comes from (8); the third inequality is due to the definition of $\tilde{\pi}^{(t)}$; and the last inequality is from the convexity of $\mathcal{L}_{n_X n_Y}$.

Considering that

$$\|\pi\|_F^2=\sum_{ij}\pi_{ij}^2\le(\max_{i,j}\pi_{ij})\sum_{ij}\pi_{ij}=\max_{i,j}\pi_{ij}\le\frac{1}{n_{\max}},$$

we obtain

$$\|\pi^{(t+1)}-\pi^{(t)}\|_F=\gamma_t\|\tilde{\pi}^{(t)}-\pi^{(t)}\|_F\le 2\gamma_t\sqrt{\frac{1}{n_{\max}}}.$$

Then, it follows that

$$\begin{aligned}\mathcal{L}_{n_X n_Y}(\pi^{(t+1)})-\inf\mathcal{L}_{n_X n_Y} &\le (1-\gamma_t)\left(\mathcal{L}_{n_X n_Y}(\pi^{(t)})-\inf\mathcal{L}_{n_X n_Y}\right)+2\beta\gamma_t^2\cdot\frac{1}{n_{\max}}\\ &= \frac{t}{t+2}\left(\mathcal{L}_{n_X n_Y}(\pi^{(t)})-\inf\mathcal{L}_{n_X n_Y}\right)+\frac{2\beta}{(t+2)^2}\cdot\frac{1}{n_{\max}}.\end{aligned}$$

By this inequality, we have

$$\mathcal{L}_{n_X n_Y}(\pi^{(1)})-\inf\mathcal{L}_{n_X n_Y}\le\frac{\beta}{2}\cdot\frac{1}{n_{\max}}\le\frac{2\beta}{2}\cdot\frac{1}{n_{\max}}.$$

Therefore, using the mathematical induction, we arrive at

$$\mathcal{L}_{n_X n_Y}(\pi^{(t+1)})-\inf\mathcal{L}_{n_X n_Y}\le\left(\frac{t}{t+2}\cdot\frac{2\beta}{t+1}+\frac{2\beta}{(t+2)^2}\right)\cdot\frac{1}{n_{\max}}\le\frac{2\beta}{t+2}\cdot\frac{1}{n_{\max}}.$$

□