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) provides a powerful mathematical framework for comparing probability measures by quantifying the minimal cost of transporting mass from one distribution to another. In recent years, OT has found wide applications in statistics, machine learning, and computer vision, where distributions often lie on non-Euclidean or structured domains. However, in many real-world problems, each observation possesses both spatial and feature information—for example, geometric shapes with embedded descriptors, or spatially indexed random fields with associated features. In such settings, it is desirable to align not only the feature embeddings but also the underlying spatial structures.

To address this, we consider a *fused optimal transport* (FOT) formulation, which simultaneously accounts for feature similarity and spatial coherence through a kernel-weighted coupling cost. This formulation generalizes both the classical quadratic OT and the Gromov–Wasserstein (GW) transport, providing a flexible interpolation between them. The rest of this section introduces the formal setup, notation, and basic existence results for the fused optimal transport plan.

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.

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

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

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)**.** *For a fixed bandwidth $h > 0$, a function $K_h : S \times S \to \mathbb{R}_+$ is said to be a* distance kernel *if there exists a bounded, continuous, (strictly) monotone $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$K_h(x, x') = \frac{1}{h}\kappa\left(\frac{d_S(x,x')}{h}\right).$$

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. In both cases, the bandwidth h controls the spatial scale of this emphasis—small h focuses on fine local structures, while large h smooths over broader spatial relationships. Finally, the boundedness of κ ensures that K_h is bounded and that the operator D_X^κ 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 $D_X^\kappa : L^2(S, \mathbb{P}_X) \to L^2(S, \mathbb{P}_X)$ *is defined by*

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

The distance potential operator is a special case of a Hilbert-Schmidt operator. Intuitively, $(D_X^\kappa f)(x)$ represents a distance-weighted average of f with respect to the point x.

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

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

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

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 $K_h(x, X)$ given $Y = y$, while $\Gamma^2_\pi(x, y)$ is the expected kernel value $K_h(y, Y)$ 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*

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

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,*

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

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 and only if the coupling π is induced by a deterministic almost-isometry T. 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 ($K_h(x, x') = K_h(T(x), T(x'))$) 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.

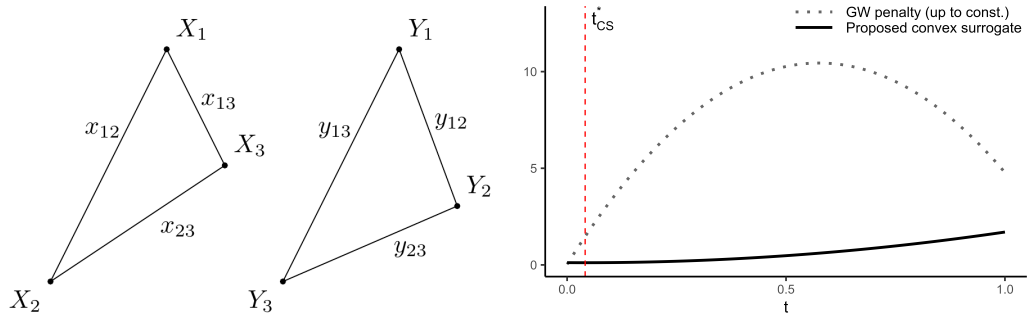
Indeed, this also reveals that the objective (2) is generally not strictly convex; for instance, if multiple distinct isometries exist and they produce the same feature-matching cost, they will all be global minimizers.

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

$$D_X = \begin{pmatrix} 0 & x_{12} & x_{13} \\ x_{12} & 0 & x_{23} \\ x_{13} & x_{23} & 0 \end{pmatrix}, \quad D_Y = \begin{pmatrix} 0 & y_{12} & y_{13} \\ y_{12} & 0 & y_{23} \\ y_{13} & y_{23} & 0 \end{pmatrix}.$$

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) 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

A straightforward calculation yields

$$\|D_X T_{\pi(t)} - T_{\pi(t)} D_Y\|_{\mathrm{HS}}^2 = \|D_X \pi(t) - \pi(t) D_Y\|_F^2 = \frac{1}{9}(at^2 + bt + c),$$

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

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

and

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

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_h, max iters T

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

Initialize $\pi^{(0)} \leftarrow \hat{\mathbb{P}}_X \otimes \hat{\mathbb{P}}_Y$

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 projection): $\hat{P} \leftarrow \arg\max_{P \in \mathcal{P}} \mathrm{Tr}(P^\top \hat{\pi})$, 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

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

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

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

Theorem 2. *Let $\hat{\pi}$ be the solution of Algorithm 1 with $\alpha > 0$. For any $\varepsilon > 0$, choose the maximum number of iteration T as*

$$T + 1 \geq \frac{16\alpha(\sup \kappa)^2}{\varepsilon h^2} \cdot \frac{n_{\max}^2}{n_{\min}}.$$

Then,

$$\mathbb{P}\left(\mathcal{L}_{n_X n_Y}(\hat{\pi}) - \inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} \mathcal{L}(\pi) > \varepsilon\right) \leq ...$$

References

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

$$(D_X^\kappa T_\pi g)(x) = \mathbb{E}\Big[K_h(x, X)\mathbb{E}\left[g(Y) \mid X\right]\Big] = \int_S \Gamma_\pi^1(x, y) g(y) \mathbb{P}_Y(dy),$$
$$(T_\pi D_Y^\kappa g)(x) = \mathbb{E}\Big[(D_Y^\kappa g)(Y) \mid X = x\Big] = \int_S \Gamma_\pi^2(x, y) g(y) \mathbb{P}_Y(dy).$$

Since K_h is bounded, $D_X^\kappa T_\pi - T_\pi D_Y^\kappa$ is a well-defined Hilbert-Schmidt operator with a kernel $\Gamma_\pi(x, y)$:

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

□

A.2 Proof for Theorem 1

Proof. Refer to Lemma 1 to confirm that the operator $D_X^\kappa T_\pi - T_\pi D_Y^\kappa$ 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

$$\pi \;\mapsto\; \|D_X^\kappa T_\pi - T_\pi D_Y^\kappa\|_{\mathrm{HS}}^2$$

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_X^κ and D_Y^κ are linear operators, it follows that

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

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

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

Therefore, $\pi \mapsto \|D_X^\kappa T_\pi - T_\pi D_Y^\kappa\|_{\mathrm{HS}}^2$ 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

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

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

Hence, for such (x, y),

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

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

$$\Gamma_\pi\big(x, T(x')\big) = \frac{1}{h}\left[\kappa\left(\frac{d_S\big(x, x'\big)}{h}\right) - \kappa\left(\frac{d_S\big(T(x), T(x')\big)}{h}\right)\right].$$

($\Leftarrow$) 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 $\|D_X^\kappa T_\pi - T_\pi D_Y^\kappa\|_{\mathrm{HS}}^2 = 0$, 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

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

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-1 distance)**.** *Let (S, d_S) be a compact metric space. For probability measures μ and ν on S, we define*

$$W_1^S(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{S\times S} d_S(x, y)\pi(dx, dy).$$

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

$$W_1^S(\mu_n, \mu) \to 0 \;\; as \;\; n \to \infty \iff \mu_n \xrightarrow{w} \mu \;\; as \;\; n \to \infty,$$

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

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

First, note that

$$\mathbb{P}\left(\mathcal{L}_{n_X n_Y}(\hat{\pi}) - \inf_{\pi\in\Pi(\mathbb{P}_X,\mathbb{P}_Y)} \mathcal{L}(\pi) > \varepsilon\right)$$
$$\leq \mathbb{P}\left(\left|\mathcal{L}_{n_X n_Y}(\hat{\pi}) - \inf_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} \mathcal{L}_{n_X n_Y}(\pi)\right| > \frac{\varepsilon}{2}\right)$$
$$+ \ \mathbb{P}\left(\left|\inf_{\pi\in\Pi(\hat{\mathbb{P}}_X,\hat{\mathbb{P}}_Y)} \mathcal{L}_{n_X n_Y}(\pi) - \inf_{\pi\in\Pi(\mathbb{P}_X,\mathbb{P}_Y)} \mathcal{L}(\pi)\right| > \frac{\varepsilon}{2}\right)$$
$$=: E_1 + E_2.$$

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. With this result, E_1 can be bounded above as follows:

$$\mathbb{P}\left(|\mathcal{L}_{n_X n_Y}(\hat{\pi}) - \inf \mathcal{L}_{n_X n_Y}| > \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(\left(\|\hat{D}_X^\kappa\|_{\mathrm{op}} + \|\hat{D}_Y^\kappa\|_{\mathrm{op}}\right)^2 > \frac{\varepsilon n_{\min}(T+1)}{4\alpha}\right).$$

Now, considering that $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_\infty$,

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

This gives that

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

Thus, if we choose

$$T + 1 \geq \frac{16\alpha(\sup\kappa)^2}{\varepsilon h^2} \cdot \frac{n_{\max}^2}{n_{\min}},$$

we obtain $E_1 = 0$.

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

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

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 $|\inf_x f(x) - \inf_x g(x)| \leq \sup_x |f(x) - g(x)|$.

Now, define

$$Q_X \in \underset{\pi \in \Pi(\mathbb{P}_X, \hat{\mathbb{P}}_X)}{\arg\min} \mathbb{E}_{(X,\hat{X}) \sim Q} \left[\|f(X) - f(\hat{X})\|_2 \right], \quad Q_Y \in \underset{\pi \in \Pi(\hat{\mathbb{P}}_Y, \mathbb{P}_Y)}{\arg\min} \mathbb{E}_{(\hat{Y},Y) \sim Q} \left[\|f(Y) - f(\hat{Y})\|_2 \right].$$

Then, by Definition 3, we get

$$\begin{aligned}
W_1^{\|\cdot\|_2}(f_\# \mathbb{P}_X, f_\# \hat{\mathbb{P}}_X) &= \mathbb{E}_{(X,\hat{X}) \sim Q_X} \left[\|f(X) - f(\hat{X})\|_2 \right], \\
W_1^{\|\cdot\|_2}(f_\# \hat{\mathbb{P}}_Y, f_\# \mathbb{P}_Y) &= \mathbb{E}_{(\hat{Y},Y) \sim Q_Y} \left[\|f(Y) - f(\hat{Y})\|_2 \right].
\end{aligned}$$

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

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

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 $\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)$,

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

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

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

The second inequality is from (6).

We now analyze the convex surrogate term. For $\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)$,

$$\begin{aligned}
&\left| \|D_X^\kappa T_\pi - T_\pi D_Y^\kappa\|_{\mathrm{HS}}^2 - \|D_X^\kappa T_{\Phi_n(\pi)} - T_{\Phi_n(\pi)} D_Y^\kappa\|_{\mathrm{HS}}^2 \right| \\
&\leq \|D_X^\kappa (T_\pi + T_{\Phi_n(\pi)}) - (T_\pi + T_{\Phi_n(\pi)}) D_Y^\kappa\|_{\mathrm{HS}} \|D_X^\kappa (T_\pi - T_{\Phi_n(\pi)}) - (T_\pi - T_{\Phi_n(\pi)}) D_Y^\kappa\|_{\mathrm{HS}} \\
&\leq 2(\|D_X^\kappa\|_{\mathrm{op}} + \|D_Y^\kappa\|_{\mathrm{op}})^2 \|T_\pi - T_{\Phi_n(\pi)}\|_{\mathrm{HS}} \\
&\leq \frac{4 \sup \kappa}{h} \|T_\pi - T_{\Phi_n(\pi)}\|_{\mathrm{HS}}.
\end{aligned}$$

The first inequality is from the Cauchy-Schwarz inequality, and the last two inequalities are direct results from Schur's test, which gives $\|T_\pi\|_{\mathrm{HS}} = \|T_{\Phi_n(\pi)}\|_{\mathrm{HS}} \leq 1$ and $\|D_X^\kappa|_{\mathrm{op}} \leq \sup \kappa / h$.

Observe that

$$(T_{\Phi_n(\pi)} g)(x) = \int_S g(y) \Phi_n(\pi)(dy \mid x) = \int_S g(y) R_X(d\hat{x} \mid x) \pi(d\hat{y} \mid \hat{x}) Q_Y(dy \mid \hat{y}),$$

where $R_X(d\hat{x} \mid x)$ satisfies $Q_X(dx, d\hat{x}) = R_X(d\hat{x} \mid x)\mathbb{P}_X(dx)$. By defining

$$(S_Y g)(\hat{y}) \coloneqq \int_S g(y) Q_Y(dy \mid \hat{y}), \quad (U_X \ell)(x) \coloneqq \int_S \ell(\hat{x}) R_X(d\hat{x} \mid x),$$

it follows that $T_{\Phi_n(\pi)} = U_X T_\pi S_Y$. Thus,

$$\begin{aligned}
\|T_\pi - T_{\Phi_n(\pi)}\|_{\mathrm{HS}} &\leq \|T_\pi - U_X T_\pi S_Y\|_{\mathrm{HS}} \\
&\leq \|(I - U_X) T_\pi + U_X T_\pi (I - S_Y)\|_{\mathrm{HS}} \\
&\leq \left(\|I - U_X\|_{\mathrm{op}} + \|I - S_Y\|_{\mathrm{op}} \right).
\end{aligned}$$

Now, consider an arbitrary $\ell : S \to \mathbb{R}$ such that

$$|\ell(x) - \ell(x')| \leq \|f(x) - f(x')\|_2, \quad \|\ell\|_\infty \leq 1.$$

Its existence is clear since $\ell(x) \coloneqq [f(x)]_1$ satisfies the above constraints. Then,

$$|(I - U_X)\ell(x)| = \left| \int_S (\ell(x) - \ell(\hat{x})) R_X(d\hat{x} \mid x) \right| \leq L \int_S \|f(x) - f(\hat{x})\|_2 R_X(d\hat{x} \mid x).$$

This gives that

$$\int_S |(I - U_X)\ell(x)| \mathbb{P}_X(dx) \leq \mathbb{E}_{(X,\hat{X})\sim Q_X} \left[\|f(X) - f(\hat{X})\|_2 \right] = W_1^{\|\cdot\|_2}(f_{\#}\mathbb{P}_X, f_{\#}\hat{\mathbb{P}}_X).$$

Moreover, we also have

$$\begin{aligned}
\left(\int_S |(I - U_X)\ell(x)|^2 \mathbb{P}_X(dx) \right)^{1/2} &\leq \|(I - U_X)\ell\|_\infty^{1/2} \left(\int_S |(I - U_X)\ell(x)| \mathbb{P}_X(dx) \right)^{1/2} \\
&\leq \sqrt{2} \sqrt{W_1^{\|\cdot\|_2}(f_{\#}\mathbb{P}_X, f_{\#}\hat{\mathbb{P}}_X)},
\end{aligned}$$

which implies that $\|I - U_X\|_{\mathrm{op}} \leq \sqrt{2}\sqrt{W_1^{\|\cdot\|_2}(f_{\#}\mathbb{P}_X, f_{\#}\hat{\mathbb{P}}_X)}$. We can analogously get $\|I - S_Y\|_{\mathrm{op}} \leq \sqrt{2}\sqrt{W_1^{\|\cdot\|_2}(f_{\#}\mathbb{P}_Y, f_{\#}\hat{\mathbb{P}}_Y)}$.

Since $\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)$ is arbitrary,

$$\mathrm{I}_n \leq$$

□

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

$$\left\|\nabla \mathcal{L}_{n_X n_Y}(\pi_1) - \nabla \mathcal{L}_{n_X n_Y}(\pi_2)\right\|_F \le \alpha \cdot \frac{n_{\max}}{n_{\min}} \left(\|\hat{D}_X^\kappa\|_{\mathrm{op}} + \|\hat{D}_Y^\kappa\|_{\mathrm{op}}\right)^2 \|\pi_1 - \pi_2\|_F. \tag{7}$$

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 (7); 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

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

□