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 $\operatorname{diam}(S) = \operatorname{diam}(M) = 1$, where $\operatorname{diam}(A) := \sup_{x,x' \in A} d_A(x, x')$. For two probability measures μ, ν on S, denote by

$$\Pi(\mu, \nu) := \{\pi \text{ on } S \times S : \text{ the marginals are } \mu \text{ and } \nu\}$$

the set of all couplings between μ and ν. 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 potential operator)**.** *Let* $(S, d_S, \mathbb{P}_X)$ *be a compact metric space. The* distance potential operator $D_X : L^2(S, \mathbb{P}_X) \to L^2(S, \mathbb{P}_X)$ *is defined by*

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

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

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

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

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

Proof. See Appendix A.1. □

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 $\|\cdot\|_{\mathrm{HS}}^2$ is a convex function; their composition thus preserves convexity (as detailed in Appendix A.1). 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.

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

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

where $K_1^\pi(x,y) := \mathbb{E}_\pi[d_S(x,X) \mid Y = y]$, $K_2^\pi(x,y) := \mathbb{E}_\pi[d_S(y,Y) \mid X = x]$, *and* $\Gamma^\pi(x,y) := K_1^\pi(x,y) - K_2^\pi(x,y)$.

Proof. See Appendix A.2. □

Corollary 1 provides an explicit expression for the regularization term from Theorem 1, reformulating it as the squared integral of a kernel function $\Gamma^\pi(x, y)$. This kernel Γ^π measures the discrepancy between two conditional expected distances: $K_1^\pi(x, y)$ and $K_2^\pi(x, y)$. Specifically, $K_1^\pi(x, y)$ represents the average distance from x to the X's coupled with y (given $Y = y$), while $K_2^\pi(x, y)$ is the average distance from y to the Y's coupled with x (given $X = x$). The regularization term can thus be interpreted as a metric that quantifies the symmetric alignment of these "cross-spatial" average distances induced by the coupling π.

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

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

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 if the coupling π is induced by a deterministic almost-isometry T, which represents a perfect alignment of the metric structures, our regularization term vanishes completely. This confirms that our convex surrogate correctly identifies such 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.

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

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

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

$$\min_{\pi} \underbrace{\mathrm{Tr}\left(C_f^\top \pi\right) + \frac{\lambda}{2n_X n_Y}\left\|n_Y \hat{D}_X \pi - n_X \pi \hat{D}_Y\right\|_F^2}_{=:\mathcal{L}(\pi)}$$

$$\text{s.t.} \quad \pi \mathbf{1}_{n_Y} = \frac{1}{n_X}\mathbf{1}_{n_X}, \quad \pi^\top \mathbf{1}_{n_X} = \frac{1}{n_Y}\mathbf{1}_{n_Y}, \quad \pi \geq 0, \tag{3}$$

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

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

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 $\{(Y_j, f(Y_j))\}_{j=1}^{n_Y}$, regularization parameter $\lambda \geq 0$, max iters $T$
Construct matrices:
        $(C_f)_{ij} \leftarrow \|f(X_i) - f(Y_j)\|_2^2, \quad (\hat{D}_X)_{ii'} \leftarrow d_S(X_i, X_{i'}), \quad (\hat{D}_Y)_{jj'} \leftarrow d_S(Y_j, Y_{j'})$
Initialize $\pi^{(0)} \leftarrow \frac{1}{n_X n_Y} \mathbf{1}_{n_X} \mathbf{1}_{n_Y}^\top$
for $t = 0, ..., T$ do
    Calculate the gradient $\nabla \mathcal{L}(\pi^{(t)})$ in (3)
    Take $\tilde{\pi}^{(t)} \leftarrow \arg\min_{\pi \in \Pi(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Y)} \operatorname{Tr}(\nabla \mathcal{L}(\pi^{(t)})^\top \pi)$
    $\pi^{(t+1)} \leftarrow (1 - \gamma_t)\pi^{(t)} + \gamma_t \tilde{\pi}^{(t)}$ for some $0 < \gamma_t < 1$
end for
$\hat{\pi} \leftarrow \pi^{(T)}$
Optional (LAP projection): $\hat{P} \leftarrow \arg\max_{P \in \mathcal{P}} \operatorname{Tr}(P^\top \hat{\pi})$, where $\mathcal{P}$ is as defined in (5)
Return: $\hat{\pi}$ (and optionally $\hat{P}$)
```

$\hat{\pi}$ is not guaranteed to be deterministic, we can obtain such a map by solving the following linear assignment problem (LAP):

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

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

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

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.

References

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

A Appendix

A.1 Proof for Theorem 1

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

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

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 and D_Y are linear operators, it follows that

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

Denoting $A_i := D_X T_{\pi_i} - T_{\pi_i} D_Y$ $(i = 1, 2)$ and $A_t := D_X T_{\pi_t} - T_{\pi_t} D_Y$, 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, $\pi \mapsto \|D_X T_\pi - T_\pi D_Y\|_{\mathrm{HS}}^2$ is convex on $\Pi(\mathbb{P}_X, \mathbb{P}_Y)$. □

A.2 Proof for Corollary 1

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

$$\begin{aligned}(D_X T_\pi g)(x) &= \mathbb{E}\Big[d_S(x, X)\mathbb{E}\left[g(Y) \mid X\right]\Big] = \int_S K_1^\pi(x, y) g(y) \mathbb{P}_Y(dy),\\ (T_\pi D_Y g)(x) &= \mathbb{E}\Big[(D_Y g)(Y) \mid X = x\Big] = \int_S K_2^\pi(x, y) g(y) \mathbb{P}_Y(dy).\end{aligned}$$

Since $\sup_{x,y\in S} d_S(x, y) \le 1$, $D_X T_\pi - T_\pi D_Y$ is a well-defined Hilbert-Schmidt operator with a kernel $\Gamma^\pi(x, y)$:

$$\|D_X T_\pi - T_\pi D_Y\|_{\mathrm{HS}}^2 = \int_S \int_S \Gamma^\pi(x, y)^2 \mathbb{P}_X(dx) \mathbb{P}_Y(dy) \le 1.$$

□

A.3 Proof for Proposition 2

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

$$\begin{aligned}K_1^\pi(x, y) &= \mathbb{E}[d_S(x, X) \mid Y = y] = d_S\big(x, T^{-1}(y)\big),\\ K_2^\pi(x, y) &= \mathbb{E}[d_S(y, Y) \mid X = x] = d_S\big(y, T(x)\big).\end{aligned}$$

Hence, for such (x, y),

$$\Gamma^\pi(x, y) = K_1^\pi(x, y) - K_2^\pi(x, y) = d_S\big(x, T^{-1}(y)\big) - d_S\big(y, T(x)\big).$$

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

$$\Gamma^\pi\big(x, T(x')\big) = d_S(x, x') - d_S\big(T(x), T(x')\big).$$

($\Rightarrow$) Suppose that $\|D_X T_\pi - T_\pi D_Y\|_{\mathrm{HS}}^2 = 0$. Then $\Gamma^\pi = 0$ holds for π-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'). Hence $d_S(T(x), T(x')) = d_S(x, x')$ holds for $\mathbb{P}_X \otimes \mathbb{P}_X$-almost every (x, x').

($\Leftarrow$) Conversely, 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'). Therefore, $\Gamma^\pi = 0$ holds for π-almost every (x, y), which implies that the Hilbert–Schmidt norm is zero.

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

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

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