Relative Translation Invariant Wasserstein Distance

Anonymous authors
Paper under double-blind review

Abstract

Motivated by the Bures-Wasserstein distance, we introduce a new family of *relative translation invariant Wasserstein distances*, denoted (RW_p), as an extension of the classical Wasserstein distances W_p for $p \in [1, +\infty)$. We establish that RW_p defines a valid metric and demonstrate that this type of metric is more robust to perturbation than the classical Wasserstein distances. A bi-level algorithm is designed to compute the general RW_p distances between arbitrary discrete distributions. Additionally, when $p = 2$, we show that the optimal coupling solutions are invariant under distributional translation in discrete settings, and we further propose two algorithms, the RW_2-Sinkhorn algorithm and RW_2-LP algorithm, to improve the numerical stability of computing W_2 distances and the optimal coupling solutions. Finally, we conduct three experiments to validate our theoretical results and algorithms. The first two experiments report that the RW_2-Sinkhorn algorithm and RW_2-LP algorithm can significantly reduce the numerical errors compared to standard algorithms. The third experiment shows that RW_p algorithms are computationally scalable and applicable to the retrieval of similar thunderstorm patterns in practical applications.

1 Introduction

Optimal transport (OT) theory provides a rigorous and interpretable framework for measuring discrepancies between probability distributions. Due to its strong theoretical foundations and flexibility, OT has become one of central tools in modern machine learning. It has found wide-ranging applications in domain adaptation (Courty et al., 2017), generative modeling—most notably in Wasserstein GANs (Arjovsky et al., 2017)—and evaluation metrics such as the Fréchet Inception Distance (FID) (Heusel et al., 2017). In addition, OT also played an important role in distributionally robust learning, including regression (Shafieezadeh-Abadeh et al., 2015; Chen & Paschalidis, 2018) and Markov decision processes (Yu et al., 2023), as well as in object tracking and matching using graph neural networks (Grand-Clement & Kroer, 2021; Sarlin et al., 2019).

Although many computational methods, such as linear programming–based solvers (Villani, 2009; Peyré & Cuturi, 2019) and the Sinkhorn algorithm (Cuturi, 2013), have been developed to compute optimal transport accurately, practical data settings can still lead to a loss of precision. Measurement errors and systematic perturbations are inevitable in real-world settings. When two distributions are very close, it becomes difficult to distinguish whether an observed discrepancy reflects intrinsic differences in the underlying data or arises from exogenous factors such as sensor noise, calibration drift, or other systematic biases. While modern OT methods can accurately quantify distributional differences, their sensitivity to such perturbations may lead to instability in downstream tasks and hinder robust performance in practice. As a result, it is natural to raise the following question:

Can we design a new metric, along with an efficient algorithm, that is robust to systematic perturbations while still capturing intrinsic differences between probability distributions?

To answer this question, we introduce a new family of *relative translation invariant Wasserstein distances*, denoted (RW_p), as an extension of the classical Wasserstein distances W_p for $p \in [1, +\infty)$. Compared with the classical Wasserstein distances, these new distances are more robust to systematic perturbations and global translational shifts. We propose an efficient algorithm to compute the general RW_p distances. In the special case $p = 2$, we show that the optimal transport coupling solution is invariant under any relative translation.

Building on this property, we further develop two adaptive algorithms, RW_2-Sinkhorn algorithm and RW_2-LP algorithm, to reduce computational errors and improve numerical stability. Finally, we conduct three experiments to validate the effectiveness of the proposed framework. The first two experiments demonstrate that the RW_2-Sinkhorn algorithm and RW_2-LP algorithm achieve significantly numerical stability compared to standard algorithms. The third experiment validates that the RW_p algorithms are computationally scalable and applicable to similar thunderstorm retrieval in real-world applications.

Contributions. The main contributions of this paper are summarized as follows:

(a) We introduce a new family of metrics, the *relative translation invariant Wasserstein* (RW_p) distances, and prove that they are true metrics and invariant to relative translations of probability distributions.

(b) We design a bi-level algorithm for efficiently computing the general RW_p distances between arbitrary discrete distributions for arbitrary $p \geq 1$.

(c) We show that the optimal coupling solutions are invariant under distributional translation in discrete settings, when $p = 2$. Based on this property, we develop two adaptive algorithms for the LP-based optimal transport algorithm and the Sinkhorn algorithm to improve the numerical stability in the computation of the W_2 distance. In particular, we show that the RW_2-Sinkhorn algorithm has more advantages in numerical stability, while the convergence rate remains the same as the standard Sinkhorn algorithm. Our experiments report that the proposed algorithms can significantly reduce numerical errors.

(d) We demonstrate the practical applications of RW_p metrics in the tasks of retrieval of similar thunderstorm patterns, showcasing their effectiveness in large-scale real-world applications.

Organization. The remainder of the paper is organized as follows. Section 2 reviews classical results in optimal transport theory and the Sinkhorn algorithm. Section 3 provides the definition of the RW_p distances and some key properties of the distances. Section 4 presents computational algorithms for the general RW_p distances and the RW_2-based algorithms for the Sinkhorn algorithm and the LP-based OT algorithm, along with an analysis of their stability and convergence rate. Finally, Section 5 provides numerical validation for the RW_2 customized Sinkhorn algorithm and the LP algorithm and demonstrates the retrieval results of similar thunderstorm patterns.

Notations. Let $\mathcal{P}_p(\mathbb{R}^d)$ denote the set of all probability distributions on $\mathbb{R}^d$ with *finite* p-th order moments. For simplicity, we let μ and ν denote a pair of source and target distributions, respectively. μ and ν are supported on finite sets $\{x_i\}_{i=1}^{n_1}$ and $\{y_j\}_{j=1}^{n_2}$, respectively, where n_1 and n_2 denote the numbers of support points. $\bar{\mu}$ and $\bar{\nu}$ are the means (mass centers) of μ and ν, respectively. Let $\mathbb{R}_*^{n_1 \times n_2}$ be the set of all $n_1 \times n_2$ matrices with non-negative entries. We use $[\mu]$ to denote the equivalence class (orbit) of μ under the translation equivalence relation in $\mathcal{P}_p(\mathbb{R}^d)$. The vector $\mathbf{1}$ denotes the all-ones vector. The operation ./ denotes the component-wise vector division. $\|\cdot\|$ denotes a norm on $\mathbb{R}^d$. $\|C\|_\infty$ denotes the value of the largest component in matrix C.

Related work. Optimal transport (OT) theory is a classical area with deep connections to probability theory, diffusion processes, and partial differential equations. For comprehensive overviews, we refer the reader to monographs (Villani, 2003; Ambrosio et al., 2005; Villani, 2009; Ollivier, 2014). A wide range of computational OT methods have been developed, including the Greenkhorn algorithm (Altschuler et al., 2017b), the network simplex method (Peyré & Cuturi, 2019).

Despite the extensive body of existing work, relatively little literature focuses on the effectiveness of the relative translation, either from a theoretical perspective or in terms of its practical benefits. The Wasserstein–Bures metric (Chen et al., 2018; K. et al., 2019; Peyré & Cuturi, 2019; Malagò et al., 2018) is the most closely related to our work; however, it is restricted to Gaussian distributions. In this paper, we extend Gaussian distributions to a much broader class of distributions and consider general p-norm metrics, thereby providing a more flexible perspective.

In addition, information geometry (ichi Amari, 2016; Liero et al., 2018; Janati et al., 2020) is also related to our work, as it offers tools for quantifying variability and structural differences between distributions.

However, important distinctions remain. Information geometry is typically formulated in terms of divergences, such as Bregman divergences or other statistical measures, whereas our approach is grounded in transport costs as the metric, resulting in a geometry that is different from information geometry.

2 Preliminaries

Before presenting our proposed method, we briefly review key concepts and formulations from classical optimal transport theory. This section establishes the technical foundation for Section 3.

2.1 Optimal Transport Theory

Optimal transport (OT) addresses the problem of finding a minimal-cost plan for transporting one probability distribution to another on a metric space. Given a cost function $c(x, y)$ and two probability measures μ and ν on $\mathbb{R}^d$, the goal is to identify a coupling of μ and ν that minimizes the total cost of moving mass from μ to ν under $c(x, y)$. While the cost function can be any non-negative function, a common and particularly useful choice is a distance-based cost of order p, such as $c(x, y) = \|x - y\|^p$, where $\|\cdot\|$ denotes a norm on $\mathbb{R}^d$ and $p \in [1, \infty)$. Under these mild conditions, the corresponding OT problem is well-defined (Villani, 2003).

Let μ be the source distribution and ν the target distribution, with $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. The optimal transport problem can be formulated as the following optimization problem.

Definition 2.1 (p-norm optimal transport problem (Villani, 2003))**.**

$$\mathrm{OT}(\mu, \nu, p) := \min_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^{2d}} \|x - y\|^p \, d\gamma(x, y), \tag{1}$$

where

$$\Gamma(\mu, \nu) = \Big\{ \gamma \in \mathcal{P}_p(\mathbb{R}^{2d}) \,\Big|\, \int_{\mathbb{R}^d} \gamma(x, y)\, dx = \nu(y),\ \int_{\mathbb{R}^d} \gamma(x, y)\, dy = \mu(x),\ \gamma(x, y) \geq 0 \Big\}.$$

Here, γ is a transport plan (or coupling), specifying how much mass is moved from source location x to target location y. The objective is to minimize the total transport cost, i.e., the overall cost of moving masses across all source–target pairs (x, y).

Building on this formulation, one obtains a family of metrics on $\mathcal{P}_p(\mathbb{R}^d)$ known as Wasserstein distances (Villani, 2009), defined directly from the optimal transport cost. It is worth noting that the norm $\|\cdot\|$ can have a different order from the order p.

Definition 2.2 (p-Wasserstein distances (Villani, 2009))**.** The p-Wasserstein distance between two probability distributions μ and ν is given by

$$W_p(\mu, \nu) := \mathrm{OT}(\mu, \nu, p)^{1/p}, \quad p \in [1, \infty).$$

The Wasserstein distance defines a true metric on $\mathcal{P}_p(\mathbb{R}^d)$, satisfying non-negativity, identity of indiscernibles, symmetry, and the triangle inequality (Villani, 2009). Moreover, it is well-defined for a broad type of probability distributions, including both discrete and continuous distributions.

In practical applications, the functional optimization in Equation equation 1 is typically reformulated as a discrete optimization problem. In this setting, the distributions μ and ν are represented by *finite* number of support points (data samples)$\{x_i\}_{i=1}^{n_1}$ and $\{y_j\}_{j=1}^{n_2}$, with associated probability masses $\{a_i\}_{i=1}^{n_1}$ and $\{b_j\}_{j=1}^{n_2}$, where n_1 and n_2 denote the number of support points, respectively.

Since both n_1 and n_2 are finite, we define a cost matrix $C \in \mathbb{R}_*^{n_1 \times n_2}$, whose entries represent the transport cost from x_i to y_j,

$$C_{ij} = \|x_i - y_j\|^p.$$

The discrete optimal transport problem can then be regarded as a linear program

$$\mathrm{OT}(\mu, \nu, p) = \min_{P \in \Pi(\mu,\nu)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} C_{ij}\, P_{ij}, \tag{2}$$

where the feasible set is

$$\Pi(\mu,\nu) = \Big\{ P \in \mathbb{R}_*^{n_1 \times n_2} \;\Big|\; P\mathbf{1} = a,\ P^\top \mathbf{1} = b \Big\}.$$

Here, P_{ij} denotes the coupling variable, representing the amount of probability mass transported from source point x_i to target point y_j. This linear programming formulation provides a tractable and widely used approach for solving discrete OT problems in practical applications.

2.2 Sinkhorn Algorithm

The discrete OT problem in Equation equation 2 is a linear program that can be solved by simplex or interior-point algorithms (Peyré & Cuturi, 2019). However, for large-scale problems, these approaches can become computationally expensive. A popular alternative exploits the special structure of the feasible set $\Pi(\mu,\nu)$ by introducing an entropy regularization term in the objective function (Cuturi, 2013). This leads to a strictly convex optimization problem whose solution can be obtained via a simple matrix scaling procedure known as the Sinkhorn algorithm.

The entropy-regularized OT problem is given by

$$\mathrm{OT}_\lambda(\mu,\nu,p) := \min_{P\in\Pi(\mu,\nu)} \sum_{i,j} C_{ij}P_{ij} + \lambda \sum_{i,j} P_{ij}(\log P_{ij} - 1),$$

where $\lambda > 0$ controls the strength of the regularization. Defining

$$K_{ij} = \exp\left(-\frac{C_{ij}}{\lambda}\right),$$

the optimal coupling can be written in the factorized form $P = \mathrm{diag}(u)\, K\, \mathrm{diag}(v)$ for some positive scaling vectors $u \in \mathbb{R}^{n_1}$ and $v \in \mathbb{R}^{n_2}$ satisfying the marginal constraints.

The Sinkhorn algorithm starts from initial vectors $u^{(0)} = v^{(0)} = \mathbf{1}$. For iteration $k \geq 0$, the updates proceed alternately as

$$u^{(k+1)} \leftarrow a./(Kv^{(k)}), \quad v^{(k+1)} \leftarrow b./(K^\top u^{(k+1)}),$$

where the division is component-wise.

Once the updates converge to (u^*, v^*), the coupling matrix P can be recovered as

$$P = \mathrm{diag}(u^*)\, K\, \mathrm{diag}(v^*).$$

It is known that as $\lambda \to 0$, the entropy-regularized solution converges to the exact optimal transport plan of the linear program (Cominetti & Martín, 1994), while for fixed $\lambda > 0$ the Sinkhorn iterations are computationally efficient and highly scalable.

3 Relative Translation Optimal Transport and RW_p Distances

In this section, we introduce the *relative translation optimal transport* (ROT) problem and define a family of distances, *relative translation invariant Wasserstein distances* (RW_p). We establish basic properties, including existence of the minimizers and RW_p defines a true metric for $p \geq 1$. Special attention is devoted to the quadratic case ($p = 2$), where an additional structure allows for a decomposition of the problem.

3.1 Relative Translation Optimal Transport and the RW_p Distance

Classical optimal transport compares two distributions in a fixed coordinate system. However, when the primary difference between two distributions is caused by a global translation of their support points, the classical OT distance may overestimate the global translation, rather than their intrinsic difference. To measure the intrinsic difference, we introduce the *relative translation optimal transport* problem, which aligns one distribution with the other through a *relative* coordinate system, rather than a fixed coordinate system.

Definition 3.1 (Relative translation optimal transport)**.** Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. The relative translation optimal transport problem is defined as

$$\mathrm{ROT}(\mu, \nu, p) := \inf_{t \in \mathbb{R}^d} \mathrm{OT}(\mu + t, \nu, p), \tag{3}$$

where $t \in \mathbb{R}^d$ is a translation vector and $(\mu + t)$ denotes the pushforward of μ under the map $x \mapsto x + t$.

This formulation introduces an outer optimization over t, while the inner optimization corresponds to the classical p-Wasserstein problem in terms of the translated distribution $\mu + t$. As a result, the ROT problem captures the minimal transport cost while dynamically aligning the two distributions.

The following proposition shows that the search domain for the optimal translation can be restricted to a compact set. Therefore, the minimizer exists, and the minimal value can be achieved.

Proposition 3.1 (Compactness and existence of minimizers)**.** *In equation 3, the search for the optimal translation t may be restricted to the following compact ball set*

$$B = \Big\{ t \in \mathbb{R}^d : \|t\| \leq 2\, W_p(\mu, \nu) \Big\}.$$

Consequently,

$$\mathrm{ROT}(\mu, \nu, p) = \min_{t \in B} \mathrm{OT}(\mu + t, \nu, p),$$

and the minimizer can be attained.

The proof is provided in Appendix A.1. The compactness ensures that the ROT problem is well-defined and avoids pathological behavior such as unbounded translations.

A quotient-space perspective. Let $\sim_T$ denote the equivalence relation on $\mathcal{P}_p(\mathbb{R}^d)$ induced by translations: we write $\mu \sim_T \mu'$ when μ' is obtained by applying a translation from μ. This relation partitions $\mathcal{P}_p(\mathbb{R}^d)$ into different equivalence classes $[\mu]$, which are the elements in the quotient space

$$\mathcal{P}_p(\mathbb{R}^d)/\!\sim_T.$$

From this perspective, the ROT problem naturally becomes an optimal transport problem on the quotient space, whose objective is to compute the minimal transport cost between two equivalence classes $[\mu]$ and $[\nu]$.

Coming from this observation, we introduce a new family of Wasserstein distances that quantify the minimal transport cost in terms of the above translation equivalence classes of probability distributions. Since the value of the ROT problem depends only on the equivalence classes themselves, and the value is actually *invariant* under relative translations, we refer to these distances as *relative translation invariant Wasserstein distances*, denoted by RW_p.

Definition 3.2 (p-relative translation invariant Wasserstein distance)**.** For $p \in [1, \infty)$, the relative translation invariant Wasserstein distance between equivalence classes $[\mu]$ and $[\nu]$ is

$$RW_p([\mu], [\nu]) := \mathrm{ROT}(\mu, \nu, p)^{1/p},$$

where any representatives μ and ν from $[\mu]$ and $[\nu]$ may be chosen.

The following theorem establishes that RW_p is a true metric.

Theorem 3.3. *For any $p \in [1, \infty)$, the function RW_p defines a real metric on the quotient space $\mathcal{P}_p(\mathbb{R}^d)/\!\sim_T$.*

The proof is provided in Appendix A.2. We remark that analogous definitions can also be made for other transformations, such as rotation (see Appendix C.1 for details). However, it is worth noting that the corresponding optimization problem for rotation is generally non-convex and difficult to solve for the global minimizers, as illustrated in the example in Appendix C.2. Because of this, we primarily focus on translation transformation in this work.

Choice of p on noise tolerance. The choice of p in the RW_p distance directly influences the sensitivity of the distance metric to noise, in a manner similar to the ℓ_p metric or W_p metric. Metrics with smaller p (e.g., RW_1) tend to be more robust to outliers and localized noise, since the cost grows linearly with displacement and therefore does not heavily penalize large but sparse deviations. In contrast, metrics with larger p (e.g., RW_2 or RW_4) amplify the influence of large transport displacements, making the metric more sensitive to outliers but simultaneously more responsive to global geometric differences in shape. Thus, different choices of p imply different notions of similarity: small p favors robustness, while large p emphasizes shape similarity. The experimental results in Subsection 5.2 are also consistent with the above analysis.

3.2 Tractability of the ROT Optimization Problem

Because of the cyclical monotonicity, the ROT problem can be solved analytically when dimension $d = 1$. When the dimension $d \geq 2$, although the ROT formulation is well-defined and admits at least one minimizer, we found the corresponding optimization problem is generally non-convex. Several non-convex examples can be found in Appendix B. This non-convexity stems from the bilinear structure of the objective function for both variables t and P in Equation equation 3. To solve the optimization problem, even though the overall problem is non-convex, the two subproblems exhibit partially convex or linear structure:

- For fixed P, the optimization is convex in the translation variable t.
- For fixed t, the optimization over P reduces to a classical linear program.

This suggests that an alternating minimization scheme—iteratively updating t and P—can be used to solve the general RW_p problem. In practice, this approach converges to the local minimizers, produces stable solutions, and the updates for t and P are computationally simple. In practice, we implement this approach with dual-simplex–based reinitialization and Armijo backtrack techniques for reducing computational time. More details and convergence discussion are provided in Subsection 4.1.

3.3 Quadratic ROT and Properties of the RW_2 Distance

As discussed previously, the ROT problem is non-convex in general and does not admit a general analytical decomposition. However, when the cost is the squared Euclidean distance, corresponding to the quadratic case $p = 2$, the problem exhibits a special structure that leads to a clear decomposition.

Proposition 3.2 (Decomposition of the quadratic ROT)**.** *For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the quadratic ROT satisfies*

$$\mathrm{ROT}(\mu, \nu, 2) = \min_{t \in \mathbb{R}^d} \mathrm{OT}(\mu + t, \nu, 2) = \mathrm{OT}(\mu, \nu, 2) - \|\bar{\mu} - \bar{\nu}\|_2^2,$$

where $\bar{\mu}$ and $\bar{\nu}$ are the means of μ and ν. In addition, in the discrete setting, the optimal coupling $P \in \mathbb{R}_*^{n_1 \times n_2}$ is invariant under any relative translation of distributions.

The proof of Proposition 3.2 is provided in Appendix A.3.

This decomposition has two important implications. First, the optimal coupling P for the classical OT problem and the ROT problem are *identical* in this case. Second, optimal coupling P is *invariant* under any relative translation of distributions. As a result, any representatives $\mu' \in [\mu]$ and $\nu' \in [\nu]$ from their respective equivalence classes can be used to solve the problem, and they will obtain the same coupling solutions. This observation provides the theoretical foundation for the RW_2 algorithms based on both LP-based optimal transport solvers and the Sinkhorn algorithm introduced in Subsection 5.1.2, leading to practical improvements in reducing computational errors. The experimental results in Section 5 further demonstrate that this decomposition can significantly reduce computational errors.

Corollary 3.4 (Decomposition of W_2 distance)**.** *For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$W_2^2(\mu, \nu) = \|\bar{\mu} - \bar{\nu}\|_2^2 + RW_2^2([\mu], [\nu]).$$

Corollary 3.4 generalizes the classical Wasserstein–Bures metric (Chen et al., 2018; K. et al., 2019; Peyré & Cuturi, 2019; Malagò et al., 2018), which was previously applicable only to Gaussian distributions. In addition, this decomposition provides an intuitive bias–variance interpretation of W_2: the displacement of the means $\|\bar{\mu} - \bar{\nu}\|_2$ corresponds to the global "bias" term, while $RW_2([\mu], [\nu])$ captures the intrinsic "variance" between the distributions.

Finally, although the above decomposition shows that the optimal translation coincides with the difference of the means when $p = 2$, this property does not necessarily hold for other orders p. One counterexample illustrating this discrepancy is provided in Appendix B.3.

4 RW_p Algorithm and RW_2 Adaptive algorithms

4.1 Algorithms for Computing RW_p Distances

We develop an efficient alternating optimization algorithm for computing RW_p distance for general $p \geq 1$. The algorithm alternates between updating the transport plan P and the translation vector t, forming a block-coordinate descent procedure that monotonically decreases the joint objective

$$\min_{t \in \mathbb{R}^d} \min_{P \in \Pi(a,b)} \sum_{i,j} P_{ij} \|x_i + t - y_j\|^p, \qquad \Pi(a,b) = \{P \geq 0 : P\mathbf{1} = a,\ P^\top \mathbf{1} = b\}.$$

Although the entire problem is non-convex, as mentioned in Subsection 3.2, each subproblem has convexity or linearity properties, allowing the overall method to remain computationally tractable.

Overview of the alternating scheme. The algorithm proceeds by fixing t and solving for the optimal coupling P, then fixing P and updating t by minimizing the reduced convex objective. Each step is computationally simple: the P-update is a linear program, while the t-update is a smooth convex minimization.

Updating the transport plan P. When the translation t is fixed, the problem reduces to a standard discrete optimal transport linear program with cost coefficients $C_{ij}(t) = \|x_i + t - y_j\|^p$. As t changes across iterations, the feasible polytope $\Pi(a, b)$ remains fixed, and only the cost matrix is updated. Thus, a previously computed coupling $P^{(k)}$, together with its associated LP basis $B^{(k)}$, remains *primal feasible* for the next iteration. This enables warm-starting the LP using a dual simplex reinitialization step, avoiding the need to recompute the entire LP basis and significantly reducing computational cost.

Updating the translation vector t. For fixed P, the reduced objective

$$F_P(t) = \sum_{i,j} P_{ij} \|x_i + t - y_j\|^p$$

is convex in t. Instead of a single gradient step, we perform a short inner loop to approximately minimize $F_P(t)$, using gradient descent with an Armijo backtracking line search to ensure sufficient decrease and stability. This "inner solve" substantially improves descent efficiency, yet remains inexpensive because each step only involves evaluating weighted residuals of the form $x_i + t - y_j$.

Geometric interpretation. The feasible region $\Pi(a, b)$ is a fixed polytope in the space. For a given translation t, the matrix $C(t)$ defines an objective hyperplane whose slope depends on the direction and magnitude of t. Updating t tilts this hyperplane, while the dual simplex step efficiently moves the solution to the new supporting face of the polytope. From this perspective, the alternating scheme repeatedly reshapes the geometry of the objective function and projects onto the polytope, tracing out a smooth descent path.

Algorithm. Algorithm 1 summarizes the procedure. We initialize t using the means' difference, then alternate between warm-started LP solves and gradient-based updates of t with Armijo backtracking. The objective decreases at every iteration.

Algorithm 1 Alternating Optimization for RW_p with Dual-Simplex-Acceleration

Require: Samples $\{x_i, a_i\}$, $\{y_j, b_j\}$, order $p \geq 1$, tolerances τ, ϵ, inner iteration cap $T_{\max}$

Initialize $t^{(0)} = \bar{\nu} - \bar{\mu}$
Solve OT with cost $C_{ij}^{(0)} = \|x_i + t^{(0)} - y_j\|^p$ to obtain $(P^{(0)}, B^{(0)})$
repeat
  Update costs $C_{ij}^{(k)} = \|x_i + t^{(k)} - y_j\|^p$
  Warm-start dual simplex to obtain $P^{(k+1)}$
  $\tilde{t}^{(0)} \leftarrow t^{(k)}$
  for $s = 0$ to $T_{\max} - 1$ **do**
    Compute gradient g_s of $F_{P^{(k+1)}}$ at $\tilde{t}^{(s)}$
    Update $\tilde{t}^{(s+1)} = \tilde{t}^{(s)} - \alpha_s g_s$ using Armijo backtracking
    if $\|\tilde{t}^{(s+1)} - \tilde{t}^{(s)}\| \leq \epsilon \max\{1, \|\tilde{t}^{(s)}\|\}$ **then break**
  $t^{(k+1)} = \tilde{t}^{(s+1)}$
  $F^{(k+1)} = \sum_{i,j} P_{ij}^{(k+1)} \|x_i + t^{(k+1)} - y_j\|^p$
until $|F^{(k+1)} - F^{(k)}|/F^{(k)} < \tau$
Output: $(t^\star, P^\star, F^\star)$

Convergence guarantee. The following proposition formalizes the descent property of the algorithm. The full proof is provided in the Appendix A.4.

Proposition 4.1 (Monotone descent and convergence)**.** *Let $p \geq 1$ and assume $c(x, y) = \|x - y\|^p$ is differentiable for $p > 1$ (or admits a subgradient for $p = 1$). Then Algorithm 1 generates a non-increasing sequence of objective values $\{F^{(k)}\}$. Every accumulation point $(t^\star, P^\star)$ satisfies the first-order optimality conditions of the RW_p problem. In addition, the warm-started dual simplex step yields locally linear convergence in the P-update when costs change small, while the inner convex t-update (with Armijo backtracking) improves stability and accelerates overall descent.*

4.2 Applications of RW_2 Decomposition to Optimal Transport Computation

The decomposition results of Theorem 3.2 and Corollary 3.4 are useful for improving OT solvers. By separating the translational component from the intrinsic coupling structure, the new optimization has the same optimal couplings while improving numerical stability and reducing computational errors. We describe its applications to both the Sinkhorn algorithm and the linear programming OT algorithm.

4.2.1 RW_2-LP Algorithm

For the LP-based OT problem, Corollary 3.4 implies that the Wasserstein cost can be separated into a translation term and a covariance term. When the objective coefficients $C_{ij} = \|x_i - y_j\|_2^2$ are extremely large, it might lead to ill-conditioned basis matrices and slower convergence. Translating the distributions can reduce the magnitude of the coefficients. Accordingly, one may translate the source distribution by $t^* = \bar{\nu} - \bar{\mu}$, compute the optimal transport plan between $(\mu + t^*, \nu)$, and then recover the full W_2 value via summation. In practice, we introduce a constant threshold $M > 1$ and apply the translation only when it substantially reduces the maximum component, thereby avoiding unnecessary translations. More details can be found in Algorithm 2.

4.2.2 RW_2-Sinkhorn Algorithm

The same improvement can also be applied to the entropically regularized OT problem. By performing the alignment conditionally controlled by threshold M, we can obtain the RW_2-Sinkhorn algorithm. In the following, the convergence rate of the Sinkhorn algorithm actually remains the same, while the numerical instability issue is reduced.

Convergence rate. Under translation t, the cost becomes $C'(t) = \|x_i + t - y_j\|^2$ and the Gibbs kernel of the Sinkhorn algorithm becomes $K' = \exp(-C'(t)/\lambda)$. We can prove that the contraction factor ρ in Hilbert's

Algorithm 2 RW_2-LP Algorithm

Require: Empirical distributions $\mu = \sum_i a_i \delta_{x_i}$, $\nu = \sum_j b_j \delta_{y_j}$; constant M

Compute $\bar{\mu}$, $\bar{\nu}$ and set $t^* = \bar{\nu} - \bar{\mu}$

Form costs $C_{ij} = \|x_i - y_j\|_2^2$ and $C'_{ij} = \|x_i + t^* - y_j\|_2^2$

if $M\|C'\|_\infty \le \|C\|_\infty$ **then** ▷ Use mean alignment only when beneficial

    $C \leftarrow C'$

Solve $\min_{P \in \Pi(a,b)} \langle C, P \rangle$ via a linear programming OT solver

Output $W_2^2(\mu, \nu) = \|\bar{\mu} - \bar{\nu}\|_2^2 + \langle C, P^* \rangle$ and the optimal plan P^*

Algorithm 3 RW_2-Sinkhorn Algorithm

Require: Measures $\mu = \sum_i a_i \delta_{x_i}$, $\nu = \sum_j b_j \delta_{y_j}$; regularizer $\lambda > 0$; tolerance $\varepsilon > 0$; constant M

Compute $\bar{\mu}$, $\bar{\nu}$ and $t^* = \bar{\nu} - \bar{\mu}$

Form costs C and C' as in Algorithm 2

if $M\|C'\|_\infty \le \|C\|_\infty$ **then**

    $C \leftarrow C'$

Initialize $K = \exp(-C/\lambda)$, and $u = v = \mathbf{1}$

repeat

    $u \leftarrow a./(Kv), \quad v \leftarrow b./(K^\top u)$

until $\max\big(\|u \odot (Kv) - a\|,\ \|v \odot (K^\top u) - b\|\big) \le \varepsilon$

$P^* = \mathrm{diag}(u) K \mathrm{diag}(v)$

Output $W_2^2(\mu, \nu) = \|t^*\|_2^2 + \langle C, P^* \rangle$

projective metric is actually invariant under any translation t (see Appendix A.5), where

$$\rho = \tanh\left(\frac{\Delta(K)}{4}\right), \qquad \Delta(K) = \frac{1}{\lambda} \max_{i,j,k,l} |C_{ik} + C_{jl} - C_{il} - C_{jk}|.$$

Thus, the aligned optimization has the same convergence rate as the original one.

Numerical stability. A principal source of numerical instability in Sinkhorn iterations is from underflow in the exponential kernel $K = \exp(-C/\lambda)$, particularly when C_{ij} is large. The translation t could alleviate this instability by conditionally reducing the magnitude of the cost coefficients,

$$C'_{ij} = \|x_i + t - y_j\|_2^2,$$

ensuring that the entries of $K' = \exp(-C'(t^*)/\lambda)$ remain well-scaled throughout the iterations. We may also measure this instability by defining the ill-condition of the matrix K as

$$\kappa(K) = \prod_{i,j} K_{ij} = \exp\Big(-\tfrac{1}{\lambda} \sum_{i,j} \|x_i + t - y_j\|_2^2 \Big).$$

By calculating the optimal condition of $\kappa(K)$, the maximizer occurs at $t = \bar{y} - \bar{x}$, which matches $t = \bar{\nu} - \bar{\mu}$ for empirical measures with uniform weights. As a result, the alignment can maximize $\kappa(K)$ and reduce kernel underflow.

Computational complexity. Altschuler et al. (2017a) show that, for precision level τ_p, time complexity of the Sinkhorn algorithm is $O\big(m^2 \|C\|_\infty^3 (\log m)\, \tau_p^{-3}\big)$, where $m = n_1 = n_2$ for simplicity. Since the alignment can reduce the value of $\|C\|_\infty$, our approach can also lower the complexity bound.

5 Experiments

We evaluate the proposed algorithms through three experiments. The first two experiments validate the numerical stability and efficiency of the RW_2-based LP and Sinkhorn algorithms. The third experiment

presents a real-world thunderstorm pattern retrieval task, illustrating the effectiveness of the general RW_p distances in large-scale applications. All experiments are conducted on a Linux workstation equipped with 64 CPU cores (Intel Core i7, 2.60 GHz), 16 GB RAM, and an NVIDIA RTX 3090 GPU.

5.1 Numerical Validation

We begin with validation tests to evaluate the numerical stability of the RW_2-based algorithms under varying dimensionality and translation magnitudes.

5.1.1 Validation of the RW_2-LP Algorithm

Setup. We consider two settings:

(1) *Same distribution with different translation*: The component of each sample in the source distribution μ is draw from $\mathcal{N}(0,1)$, and the target distribution is constructed as $\nu = \mu + t$, where t is a translation applied along the last coordinate and takes values in $\{0,1,2,4,8,16\}$. Each distribution size is 4,096 and we consider dimensions $d \in \{2,5,10\}$.

(2) *Different distributions*: The component of each sample in the distribution μ is sampled from $\mathcal{N}(0,1)$. The component of each sample in ν is drawn from the Uniform distribution $\mathcal{U}[-1,1]$ first, and then the distribution ν is translated by $t = 1$ along the last coordinate. We consider $d \in \{2,5,10\}$ with 1,024 samples for each distribution and vary the maximum iteration budget of the LP algorithm from 2^6 to 2^{16}.

We compare the RW_2-LP algorithm (Algorithm 2) with the standard LP algorithm to solve OT problems. Performance is measured in terms of the absolute error of $W_2^2(\mu,\nu)$ and the running time. Ground-truth $W_2^2(\mu,\nu)$ is given by $\|t\|_2^2$ in the setting (1) and by a high-precision LP solution in the setting (2). The LP solver is from the `ot.emd2()` function in the POT library (Flamary et al., 2021) and the threshold parameter is fixed at $M = 1$. The experiments are repeated six times for the same settings.

Results. For the first setting, Figure 1(a) shows that the RW_2-LP solver curves yield substantially lower numerical errors than the standard LP solver curves. The discrepancy between the two approaches becomes increasingly pronounced for larger translations and higher dimensions, while the three RW_2-LP solver curves completely overlap at the bottom of the plot, indicating consistent low numerical errors of the RW_2-LP formulation across dimensions. Figure 1(b) demonstrates that running time remains comparable or lower across all dimensions.

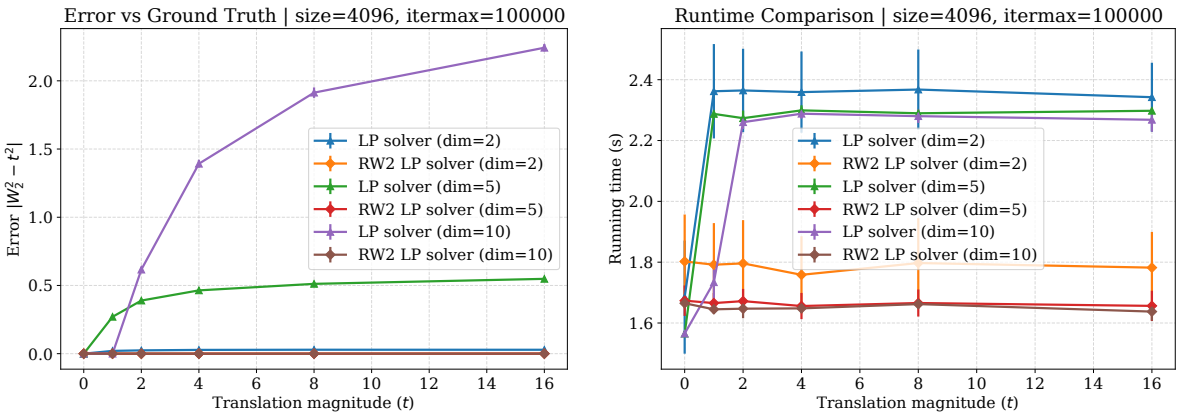
For the second setting, Figure 2(a) shows faster convergence under tight iteration budgets: the RW_2-LP formulation consistently attains lower error for all dimensions. Figure 2(b) reports that running time remains comparable or even slightly reduced. In summary, the RW_2 formulation improves both stability and efficiency in the LP-based OT solver.

5.1.2 Validation of the RW_2-Sinkhorn Algorithm

Setup. We perform one test for the Sinkhorn algorithm under a configuration similar to setting (2) in the first experiment. The component of each sample in the source distribution μ is draw from $\mathcal{N}(0,1)$, and the component of each sample in the target distribution ν is draw from $\mathcal{U}[-1,1]$, then translated by $t \in \{0,1,2,4,8,16\}$. We test dimensions $d \in \{2,5,10\}$ with 1,024 samples. We also test other pairs (Gaussian→Gaussian, Gaussian→Geometric, Gaussian→Poisson) with the same setting, and more results can be found in Appendix D.1. All these tests are repeated six times to ensure accuracy.

We compare the RW_2-Sinkhorn algorithm (Algorithm 3) with the standard Sinkhorn algorithm. Both methods use `ot.sinkhorn2()` from the POT library (Flamary et al., 2021) with regularization `reg` $= 10^{-5}$, a maximum of 1,000 iterations, and stopping threshold `stopThr`$= 10^{-5}$. The threshold parameter is fixed at $M = 1$.

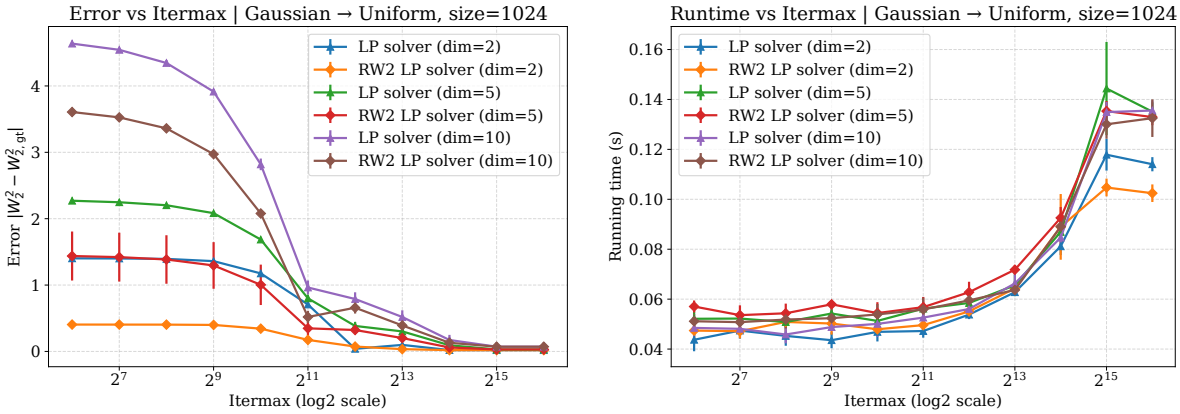
Results. Figure 3(a) shows that the RW_2-Sinkhorn algorithm achieves significantly lower numerical errors, particularly as the translation magnitude increases. The three RW_2-Sinkhorn curves completely overlap at



(a) Computation error comparison.

(b) Running time comparison.

Figure 1: LP algorithms comparison on Gaussian → Gaussian translation tasks. (a) The RW_2-LP formulation achieves consistently lower numerical errors. The three RW_2-LP solver curves completely overlap at the bottom of the plot, indicating consistent low numerical errors of the RW_2-LP formulation across dimensions. (b) Running time remains comparable or slightly lower across all dimensions.



(a) Computation error vs. iteration count.

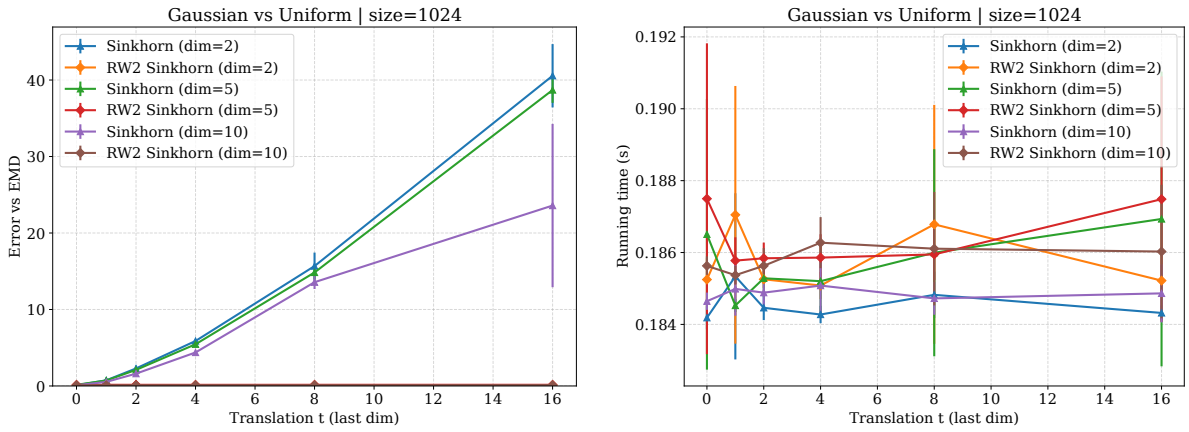
(b) Running time vs. iteration count.

Figure 2: LP algorithms comparison on Gaussian → Uniform tasks with limited iteration budgets. (a) The RW_2-LP solver achieves substantially lower error, especially under small budgets. (b) Running time remains similar or slightly lower across $d \in \{2, 5, 10\}$.

the bottom of the plot, indicating consistent low numerical errors of the RW_2-Sinkhorn formulation across dimensions. Figure 3(b) shows that the running time remains almost the same across all configurations, confirming that accuracy improvements do not incur additional running time.

5.2 Thunderstorm Pattern Retrieval with RW_p

Thunderstorm patterns are critical for airline and airport operations. Given a reference thunderstorm event, it is useful to retrieve similar historical thunderstorm events in the database. We apply general RW_p distances on the real-world thunderstorm dataset to show that general RW_p can be used to retrieve similar weather patterns.



(a) Computation error comparison.

(b) Running time comparison.

Figure 3: Sinkhorn algorithms comparison on Gaussian → Uniform tasks. (a) The RW_2-Sinkhorn algorithm achieves significantly lower errors, especially under large translations. The three RW_2-Sinkhorn curves completely overlap at the bottom of the plot, indicating consistent low numerical errors of the RW_2-Sinkhorn formulation across dimensions. (b) Running time is comparable across all settings.

Table 1: Running time for different metrics of retrieving the most similar thunderstorm snapshot from the full dataset (32,073 images) given a reference snapshot.

Metric	ℓ_2	W_2	RW_1	RW_2	RW_4
Running Time(s)	11.16	50.87	1,060.15	48.28	803.21

Dataset and preprocessing. Our data are collected radar images from MULTI-RADAR/MULTI-SENSOR SYSTEM (MRMS) (Zhang et al., 2016) focusing on a 300×300 km^2 rectangular area centered at the Dallas Fort Worth International Airport (DFW), where each pixel represents a 3×3 km^2 area. The data are updated every 10 minutes, tracking from 2014 to 2022 between March and October, including around 32,073 images having thunderstorm patterns. Vertically Integrated Liquid Density (VIL density) and reflectivity are two common measurements for assessing thunderstorm intensity, with threshold values of $3kg \cdot m^{-3}$ and $35dBZ$, respectively (Matthews & Delaura, 2010). We use reflectivity as thunderstorm measurements and use $35dBZ$ as the threshold to transform radar images to the corresponding *binary* matrices.

Thunderstorm types: We consider two types of thunderstorm events:

- **Snapshots**: single radar images representing storm patterns;
- **Sequences**: a series of consecutive snapshots representing storm evolution in a short time.

In the main text we focus on snapshot retrieval; sequence-based results are provided in Appendix D.2.

Snapshot retrieval. Given a reference thunderstorm snapshot, we compute distances to all snapshots in the dataset using RW_p for $p \in \{1, 2, 4\}$, and compare the results with the baseline distances ℓ_2 and W_2. For each metric, the top-5 most similar thunderstorm snapshots. Figure 4 illustrates one example of snapshot retrieval results.

Results and analysis. Table 1 reports the running time for each distance metric. The ℓ_2 distance is the fastest to compute. The classical Wasserstein distance W_2 incurs a substantially longer running time than

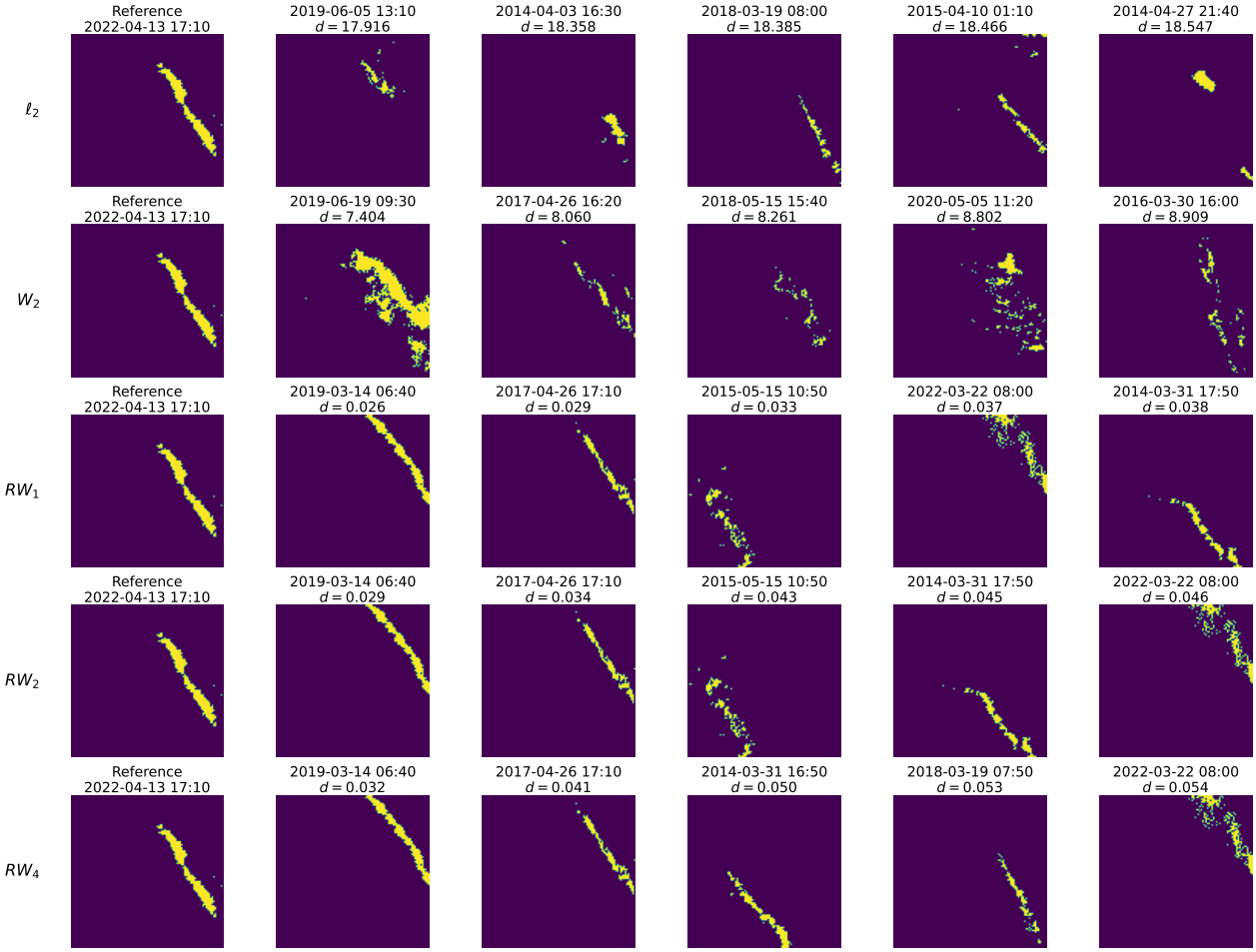


Figure 4: Top-5 retrieval results for different distance metrics using the same reference storm (leftmost column). Rows correspond to ℓ_2, W_2, and RW_p distances, $p = \{1, 2, 4\}$. Each retrieved storm is annotated with the distance to the reference. RW_p distances yield significantly more structurally consistent matches with the reference, demonstrating robustness to relative translations.

ℓ_2, reflecting the overhead of solving the full optimal transport problem. RW_2 achieves the lowest running time due to the decomposition property. In contrast, RW_1 and RW_4 require additional calculations due to alternative optimization. Overall, these results demonstrate that the proposed method is computationally feasible for large-scale applications.

Figure 4 presents the top-5 retrieval results obtained using different metrics for the same reference snapshot. The Euclidean ℓ_2 metric yields the least informative matches: the retrieved storms are sparsely distributed and poorly aligned with the reference, demonstrating strong sensitivity to spatial misalignment and a limited ability to capture structural similarity. The classical Wasserstein distance W_2 improves retrieval quality by producing storms with more coherent mass distributions; however, noticeable deformation and dispersion persist, reflecting its dependence on absolute spatial locations. In contrast, the relative Wasserstein distance RW_p consistently retrieves thunderstorm events that closely match the reference in both shape and orientation for all tested orders $p \in \{1, 2, 4\}$. In particular, RW_1 is more tolerant to outliers and local noise, whereas increasing p to 2 and 4 creates a great penalty on large transportation, leading to slightly higher distance values while preserving overall structural alignment. As shown by the third- to fifth-ranked retrievals in Figure 4, the thunderstorm event on 2022-03-22 exhibits a sparser spatial structure with more outliers than the event on 2014-03-31. Consequently, the latter achieves a smaller distance under the RW_1 metric, while

the ranking is reversed under RW_2 and RW_4. This example confirms that RW_1 is more robust to outliers and local noise, while larger p values increasingly emphasize global shape similarity.

6 Conclusions

In this paper, we introduce a novel family of distances, relative translation invariant Wasserstein (RW_p) distances, for measuring the similarity between probability distributions. Extended from the classical optimal transport framework, we show that RW_p defines a proper metric on the quotient space $\mathcal{P}_2(\mathbb{R}^d)/\sim_T$ and is invariant under relative translations. In the special case $p = 2$, the proposed distance exhibits additional structure, including a decomposition of the optimal transport formulation and translation-invariant coupling solutions. We further develop algorithms for computing general RW_p distances, and RW_2-based algorithms for both LP-based and Sinkhorn OT solvers to mitigate numerical instability and reduce computational errors. We analyze the numerical stability and computational complexity of the proposed algorithms. Finally, we validate the proposed algorithms through extensive experiments, demonstrating that our proposed algorithms significantly reduce computational errors in both LP-based and Sinkhorn OT solvers and enable practical meteorological applications in large-scale real-world settings.

References

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via greenkhorn iteration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5231–5242, 2017a.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows in metric spaces and in the space of probability measures. *Lectures in Mathematics, ETH Zurich*, 2005.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 214–223. PMLR, 2017.

Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018.

Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. Optimal transport in the space of gaussian measures. *IEEE Transactions on Automatic Control*, 63(9):2913–2928, 2018.

Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67:169–187, 1994.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Julien Grand-Clement and Christian Kroer. First-order methods for wasserstein distributionally robust MDP. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 2010–2019. PMLR, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6629–6640, Long Beach, CA, USA, 2017.

Shun ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

Hicham Janati, Gabriel Peyré, and François-Xavier Vialard. Entropic optimal transport between unbalanced gaussian measures has a closed form. *Advances in Neural Information Processing Systems*, 2020.

Rajendra N. K., Nicolas Courty, and Rémi Flamary. Wasserstein–bures metric for gaussian measures. *SIAM Journal on Imaging Sciences*, 12(4):2311–2341, 2019.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein–bures geometry of gaussian distributions. *Information Geometry*, 1(2):137–179, 2018.

Michael Matthews and Rich Delaura. Assessment and interpretation of en route weather avoidance fields from the convective weather avoidance model. In *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, 2010.

Yann Ollivier. *Optimal Transport: Theory and Applications*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2014.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946, 2019.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Cédric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, Heidelberg, 2009.

Zhuodong Yu, Ling Dai, Shaohang Xu, Siyang Gao, and Chin Pang Ho. Fast bellman updates for wasserstein distributionally robust MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 30554–30578. Curran Associates, Inc., 2023.

Jian Zhang, Kenneth Howard, Carrie Langston, Brian Kaney, Youcun Qi, Lin Tang, Heather Grams, Yadong Wang, Stephen Cocks, Steven Martinaitis, Ami Arthur, Karen Cooper, Jeff Brogden, and David Kitzmiller. Multi-radar multi-sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 97(4):621–638, 2016.

Appendix

A Proofs

A.1 Proposition 3.1

Proof of Proposition 3.1. Let $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Define W_p as the Wasserstein distance with cost $\|x - y\|^p$, and set $F(t) := W_p(\mu + t, \nu)$.

For all $t \in \mathbb{R}^d$,

$$W_p(\mu, \mu + t) = \|t\|.$$

By the triangle inequality,

$$F(t) \;\geq\; \big|\, W_p(\mu + t, \mu) - W_p(\mu, \nu) \,\big| = \big|\, \|t\| - W_p(\mu, \nu) \,\big|.$$

Hence if $\|t\| \geq 2\, W_p(\mu, \nu)$, then

$$F(t) \;\geq\; \|t\| - W_p(\mu, \nu) \;\geq\; W_p(\mu, \nu) = F(0).$$

So no minimizer lies outside the ball

$$B := \{\, t \in \mathbb{R}^d : \|t\| \leq 2\, W_p(\mu, \nu) \,\}.$$

Since F is lower semi-continuous in t and B is compact, F attains its minimum on B. Therefore,

$$\mathrm{ROT}(\mu, \nu, p) \;=\; \min_{\|t\| \leq 2\, W_p(\mu, \nu)} W_p(\mu + t, \nu).$$

□

A.2 Theorem 3.3

Proof of Theorem 3.3. Using the previous notations, we first verify that the translation relation $\sim_T$ is an equivalence relation on $\mathcal{P}_p(\mathbb{R}^d)$. It is reflexive, since any $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ can be translated to itself by the zero vector; symmetric, since if μ can be translated to ν, then ν can be translated back to μ; and transitive, since if μ can be translated to ν and ν to η, then μ can also be translated to η.

Hence, by the properties of equivalence relations, the quotient set $\mathcal{P}_p(\mathbb{R}^d)/\!\sim_T$ is well defined. Let $[\mu]$ denote an element of this quotient space. Based on that, $W_p(\cdot, \cdot)$ is a true metric on $\mathcal{P}_p(\mathbb{R}^d)$ (Villani, 2003), it satisfies identity, positivity, symmetry, and the triangle inequality. We now show that $RW_p(\cdot, \cdot)$ also satisfies these axioms on $\mathcal{P}_p(\mathbb{R}^d)/\!\sim_T$.

For any $[\mu], [\nu], [\eta] \in \mathcal{P}_p(\mathbb{R}^d)/\!\sim_T$:

- Identity:
$$RW_p([\mu], [\mu]) = \min_{\mu', \mu'' \in [\mu]} W_p(\mu', \mu'') = W_p(\mu', \mu') = 0.$$
- Positivity:
$$RW_p([\mu], [\nu]) = \min_{\mu' \in [\mu], \nu' \in [\nu]} W_p(\mu', \nu') \geq 0.$$
- Symmetry:
$$RW_p([\mu], [\nu]) = \min_{\mu' \in [\mu], \nu' \in [\nu]} W_p(\mu', \nu') = \min_{\nu' \in [\nu], \mu' \in [\mu]} W_p(\nu', \mu') = RW_p([\nu], [\mu]).$$

- Triangle inequality:

  Fix $\epsilon > 0$. By definition of the minimum, there exist $\mu' \in [\mu], \nu' \in [\nu]$ such that

  $$W_p(\mu', \nu') \leq RW_p([\mu], [\nu]) + \epsilon,$$

  and $\nu'' \in [\nu], \eta' \in [\eta]$ such that

  $$W_p(\nu'', \eta') \leq RW_p([\nu], [\eta]) + \epsilon.$$

  Since $\nu' \sim_T \nu''$, there exists a translation $t \in \mathbb{R}^d$ with $\nu'' = \nu' - t$. By translation invariance of W_p, we have

  $$W_p(\nu'', \eta') = W_p(\nu' - t, \eta') = W_p(\nu', \eta' + t).$$

  Thus $\eta' + t \in [\eta]$. And the triangle inequality for W_p gives

  $$W_p(\mu', \eta' + t) \leq W_p(\mu', \nu') + W_p(\nu', \eta' + t).$$

  Combining with the above bounds,

  $$\begin{aligned} RW_p([\mu], [\eta]) &\leq W_p(\mu', \eta' + t) \\ &\leq W_p(\mu', \nu') + W_p(\nu', \eta' + t) \\ &= W_p(\mu', \nu') + W_p(\nu'', \eta') \\ &\leq RW_p([\mu], [\nu]) + RW_p([\nu], [\eta]) + 2\epsilon. \end{aligned}$$

  Since $\epsilon > 0$ was arbitrary, the inequality follows.

Therefore, RW_p defines a metric on $\mathcal{P}_p(\mathbb{R}^d)/\sim_T$. □

A.3 Proof of Proposition 3.2

Proof of Proposition 3.2. We first establish the decomposition in the continuous setting and then verify the invariance of the optimal coupling in the discrete case.

Continuous case. Consider the quadratic ROT problem

$$\mathrm{ROT}(\mu, \nu, 2) = \min_{t \in \mathbb{R}^d} \min_{\gamma \in \Pi(\mu + t, \nu)} \int_{\mathbb{R}^{2d}} \|x + t - y\|_2^2 \, d\gamma(x, y).$$

Expanding the square yields

$$\begin{aligned} \int \|x + t - y\|_2^2 \, d\gamma &= \int \left(\|x - y\|_2^2 + \|t\|_2^2 + 2t \cdot (x - y) \right) d\gamma \\ &= \int \|x - y\|_2^2 \, d\gamma + \|t\|_2^2 + 2t \cdot \int (x - y) \, d\gamma. \end{aligned} \tag{4}$$

For any $\gamma \in \Pi(\mu, \nu)$, the marginal conditions imply

$$\int x \, d\gamma = \bar{\mu}, \qquad \int y \, d\gamma = \bar{\nu}.$$

Thus, $\int (x - y) \, d\gamma = \bar{\mu} - \bar{\nu}$. Substituting into equation 4 gives

$$\int \|x + t - y\|_2^2 d\gamma = \int \|x - y\|_2^2 d\gamma + \|t\|_2^2 + 2t \cdot (\bar{\mu} - \bar{\nu}).$$

Thus,

$$\mathrm{ROT}(\mu, \nu, 2) = \mathrm{OT}(\mu, \nu, 2) + \min_{t \in \mathbb{R}^d} \left(\|t\|_2^2 + 2t \cdot (\bar{\mu} - \bar{\nu}) \right),$$

and the strictly convex term is minimized at $t^* = \bar{\nu} - \bar{\mu}$, yielding

$$\mathrm{ROT}(\mu, \nu, 2) = \mathrm{OT}(\mu, \nu, 2) - \|\bar{\mu} - \bar{\nu}\|_2^2.$$

Discrete case (invariance of the optimal coupling). Let $\mu = \sum_{i=1}^{n_1} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{n_2} b_j \delta_{y_j}$, and let $P \in \Pi(a, b)$ be a coupling matrix. For fixed t, the discrete ROT objective is

$$\sum_{i,j} P_{ij} \, \|x_i + t - y_j\|_2^2.$$

Expanding the square gives

$$\sum_{i,j} P_{ij} \|x_i + t - y_j\|_2^2 = \sum_{i,j} P_{ij} \|x_i - y_j\|_2^2 + \|t\|_2^2 \sum_{i,j} P_{ij} + 2t \cdot \sum_{i,j} P_{ij}(x_i - y_j). \tag{5}$$

Using the marginal constraints,

$$\sum_{i,j} P_{ij} = 1, \qquad \sum_{i,j} P_{ij} x_i = \bar{\mu}, \qquad \sum_{i,j} P_{ij} y_j = \bar{\nu},$$

hence $\sum_{i,j} P_{ij}(x_i - y_j) = \bar{\mu} - \bar{\nu}$. Substituting into equation 5, we obtain

$$\sum_{i,j} P_{ij} \|x_i + t - y_j\|_2^2 = \sum_{i,j} P_{ij} \|x_i - y_j\|_2^2 + \|t\|_2^2 + 2t \cdot (\bar{\mu} - \bar{\nu}),$$

where the additional terms depend only on t and not on P. Thus, for any fixed t, the minimizers over $P \in \Pi(a, b)$ of

$$P \mapsto \sum_{i,j} P_{ij} \|x_i + t - y_j\|_2^2$$

coincide exactly with those of the original quadratic OT objective

$$P \mapsto \sum_{i,j} P_{ij} \|x_i - y_j\|_2^2.$$

Hence, the optimal coupling is invariant under any relative translation.

Combining the continuous decomposition with the discrete invariance establishes the proposition. □

A.4 Proof of Proposition 4.1

Proof. Recall that the objective of the RW_p problem is

$$F(t, P) = \sum_{i,j} P_{ij} \, \|x_i + t - y_j\|^p,$$

with $p \geq 1$. The feasible set $\Pi(a, b)$ is convex and compact, and for any fixed coupling P, the map $t \mapsto F(t, P)$ is convex (strictly convex when $p > 1$).

Step 1: Monotone descent. Each iteration consists of two substeps.

(a) P-update. For fixed $t^{(k)}$, the coupling is updated by solving the linear program

$$P^{(k+1)} = \arg \min_{P \in \Pi(a,b)} F(t^{(k)}, P),$$

which gives

$$F(t^{(k)}, P^{(k+1)}) \leq F(t^{(k)}, P^{(k)}).$$

The warm-started dual simplex step used in Algorithm 1 preserves this monotone decrease.

(b) t-update. For fixed $P^{(k+1)}$, the algorithm performs an inner gradient-based minimization of the convex function

$$t \mapsto F(t, P^{(k+1)}),$$

using Armijo backtracking to choose the step size. Therefore, each inner step satisfies the sufficient decrease condition

$$F(t^{(k+1)}, P^{(k+1)}) \leq F(t^{(k)}, P^{(k+1)}).$$

Combining the two substeps yields the global descent property

$$F(t^{(k+1)}, P^{(k+1)}) \leq F(t^{(k)}, P^{(k)}), \qquad \forall k \geq 0,$$

so $\{F^{(k)}\}$ is a non-increasing sequence bounded below by 0, and therefore convergent.

Step 2: Existence of accumulation points. Because $\Pi(a, b)$ is compact and $F(\,\cdot\,, P)$ is coercive in t for each P, the sequence $\{t^{(k)}\}$ remains bounded. Thus, the sequence $\{(t^{(k)}, P^{(k)})\}$ admits at least one accumulation point $(t^\star, P^\star)$.

Step 3: Stationarity of accumulation points. For each k, we have the optimality relation

$$P^{(k+1)} = \arg\min_{P \in \Pi(a,b)} F(t^{(k)}, P),$$

and the Armijo-based inner loop ensures that $t^{(k+1)}$ satisfies a first-order decrease condition for the convex problem $\min_t F(t, P^{(k+1)})$. Passing to the limit along any convergent subsequence and using continuity of F and of its gradient (or subgradient) in t, we obtain

$$0 \in \partial_t F(t^\star, P^\star), \qquad P^\star \in \arg\min_{P \in \Pi(a,b)} F(t^\star, P).$$

Hence, $(t^\star, P^\star)$ satisfies the first-order optimality conditions of the RW_p problem.

Step 4: Conclusion. The alternating scheme produces a monotone sequence of objective values converging to $F^\star$, and every accumulation point of the iterates is a stationary point of the non-convex RW_p problem. The dual simplex re-initialization ensures locally linear progress in the P-update when the cost perturbation is small, while the Armijo-controlled inner t-update guarantees stable and accelerated descent.

□

A.5 Invariance of the Sinkhorn Convergence Rate under Translation

The following proposition formalizes the invariance of the Sinkhorn convergence rate under translation. In particular, it establishes that translating the input distributions does not affect the contraction constant of the Sinkhorn operator in Hilbert's projective metric, even though it may significantly improve numerical conditioning.

Proposition A.1 (Translation invariance of the Hilbert–metric contraction)**.** *Let $C_{ij} = \|x_i - y_j\|_2^2$ denote the quadratic cost matrix, and let*

$$K = \exp\left(-\frac{C}{\lambda}\right), \qquad \lambda > 0,$$

be the associated Gibbs kernel used in the Sinkhorn algorithm. For any translation vector $t \in \mathbb{R}^d$, define the translated cost

$$C'_{ij} = \|x_i + t - y_j\|_2^2, \qquad K' = \exp\left(-\frac{C'}{\lambda}\right).$$

Then the projective diameter of K,

$$\Delta(K) = \frac{1}{\lambda} \sup_{i,j,k,l} |C_{ik} + C_{jl} - C_{il} - C_{jk}|,$$

satisfies $\Delta(K') = \Delta(K)$. Consequently, the Hilbert metric contraction factor

$$\rho = \tanh\left(\frac{\Delta(K)}{4}\right),$$

and hence, the geometric convergence rate of the Sinkhorn iterations is invariant under translation of the input distributions.

Proof. We begin by expanding the translated cost function:

$$C'_{ij} = \|x_i + t - y_j\|_2^2 = \|x_i - y_j\|_2^2 + 2t \cdot (x_i - y_j) + \|t\|_2^2.$$

Substituting into the expression defining the projective diameter yields

$$\begin{aligned} C'_{ik} + C'_{jl} - C'_{il} - C'_{jk} = (C_{ik} + 2t{\cdot}(x_i - y_k) + \|t\|_2^2) + (C_{jl} + 2t{\cdot}(x_j - y_l) + \|t\|_2^2) \\ - (C_{il} + 2t{\cdot}(x_i - y_l) + \|t\|_2^2) - (C_{jk} + 2t{\cdot}(x_j - y_k) + \|t\|_2^2). \end{aligned}$$

The constant terms $\|t\|_2^2$ cancel since they appear twice with positive and twice with negative sign. The linear terms in t also cancel, as

$$(x_i - y_k) + (x_j - y_l) - (x_i - y_l) - (x_j - y_k) = 0.$$

Therefore,

$$C'_{ik} + C'_{jl} - C'_{il} - C'_{jk} = C_{ik} + C_{jl} - C_{il} - C_{jk},$$

implying that $\Delta(K') = \Delta(K)$.

As a consequence, the contraction factor $\rho = \tanh(\Delta(K)/4)$ and the corresponding geometric convergence rate of the Sinkhorn algorithm remains unchanged under any translation $t \in \mathbb{R}^d$. Although translation affects the numerical scaling of the kernel entries and thereby their conditioning, it leaves the convergence rate invariant. □

B Examples of ROT Problems

In this section, we present three examples to show that solving the ROT problem for $p \geq 1$ is challenging in general. The first two examples demonstrate the non-convexity of the ROT formulation in a two-dimensional setting, and the third example shows that the optimal solution t does *not* necessarily always be the difference between the mass centers when the order $p \neq 2$.

Throughout this section, the underlying cost function is assumed to be

$$c(x, y) = \|x - y\|_p^p.$$

B.1 Non-convexity with respect to the translation variable t

Consider a two-dimensional setting where the source and target distributions μ and ν are supported on

$$\{x_i = (\cos\frac{2i\pi}{3}, \sin\frac{2i\pi}{3}),\ i = 1, 2, 3\}, \qquad \{y_j = (-\cos\frac{2j\pi}{3}, -\sin\frac{2j\pi}{3}),\ j = 1, 2, 3\},$$

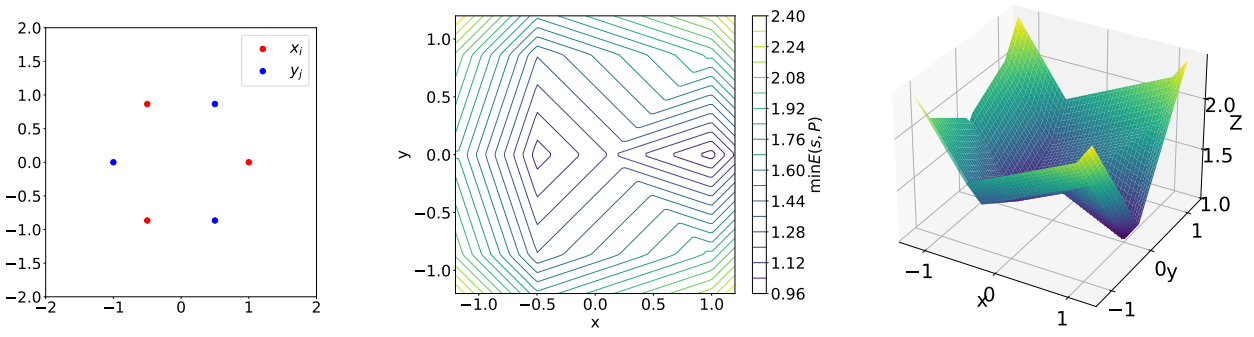
each with equal masses. These configurations are illustrated in Figure 5(a).

We focus on the effect of the translation variable t on the objective function and the transport plan P attains its minimal value for each fixed t. In other words, we study the function

$$t \ \mapsto\ \min_P \sum_{i,j} P_{ij} \, \|x_i + t - y_j\|_p^p.$$

Figures 5(b)–(c) show the corresponding contour and surface plots for $p = 1$, both clearly indicating that this function is non-convex with respect to t.

Using the same μ and ν, we also examine other exponents $p \in \{1.2, 4, 10\}$. The corresponding contour and surface plots are shown in Figure 6, indicating that non-convexity in t persists for a range of p values.



(a) Distributions μ and ν.

(b) Contour plot of the function $\min_P \sum_{i,j} P_{ij} \|x_i + t - y_j\|$ with respect to t.

(c) Surface plot of the function $\min_P \sum_{i,j} P_{ij} \|x_i + t - y_j\|$ with respect to the translation vector $t = (x, y)$.

Figure 5: Contour and surface plots of $\min_P \sum_{i,j} P_{ij} \|x_i + t - y_j\|$ showing non-convexity in t.

B.2 Non-convexity with respect to the coupling variable P

Next, using the same source and target distributions as in the first example and setting $p = 1$, we examine the effect of the variable P on the objective function via fixing the translation vector t to be its minimal value for each fixed P. In other words, we consider the function

$$F_1(P) = \min_t \sum_{i,j} P_{ij} \|x_i + t - y_j\|,$$

and show that $F_1(P)$ is non-convex with respect to the variable P.

Since the dimension of P is high, plotting the contour of $F_1(P)$ directly is not possible. Instead, we demonstrate non-convexity by exhibiting two transport plans P_1 and P_2 such that the interpolated function value is strictly smaller than the function value at the interpolated transport plan, which violates the convexity property. Therefore, the function $F_1(P)$ is non-convex.

Consider two feasible transport plans:

$$P_1 = \tfrac{1}{3}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad P_2 = \tfrac{1}{3}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

The optimal translations are $t_* P_1 = (1, 0)$ and $t_* P_2 = (-\frac{1}{2}, 0)$, therefore, $F_1(P_1) = 1$ and $F_1(P_2) = \frac{1}{2} + \frac{\sqrt{3}}{3}$. However, for their midpoint, $\frac{1}{2}(P_1 + P_2)$, we obtain

$$F_1\big(\tfrac{1}{2}(P_1 + P_2)\big) = 1 + \tfrac{\sqrt{3}}{6} > \tfrac{1}{2}(F_1(P_1) + F_1(P_2))\,,$$

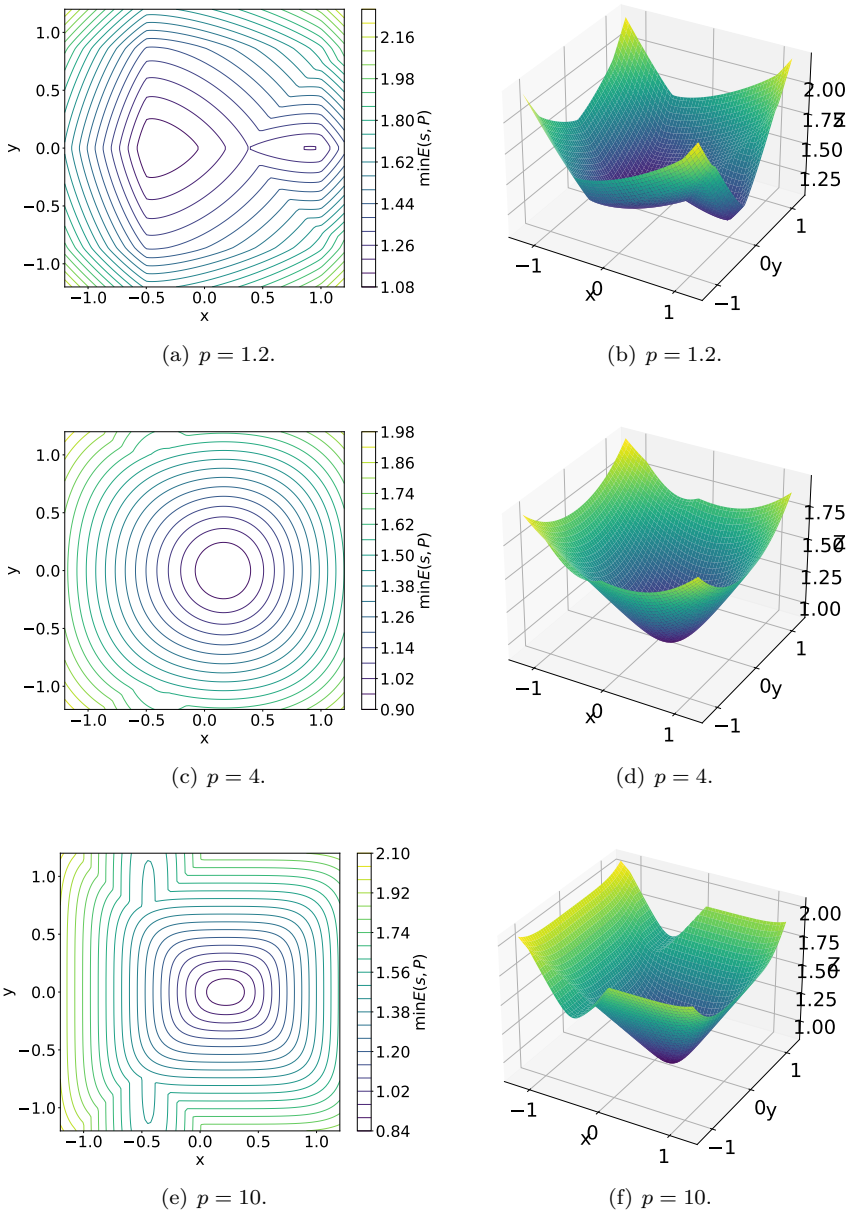
which shows that $F_1(P)$ is non-convex in P.

In summary, the above examples show that the convexity of the ROT problem cannot be guaranteed in general, especially in high-dimensional or non-quadratic cases.

B.3 Optimal translation versus mean difference

We now show that for $p \neq 2$, the optimal translation minimizing

$$\min_t \sum_{i,j} P_{ij} \|x_i + t - y_j\|_p^p$$



(a) $p = 1.2$.

(b) $p = 1.2$.

(c) $p = 4$.

(d) $p = 4$.

(e) $p = 10$.

(f) $p = 10$.

Figure 6: Contour and surface plots of the function $\min_P \sum_{i,j} P_{ij} \|x_i + t - y_j\|_p^p$ showing non-convexity in t for $p \in \{1.2, 4, 10\}$.

does not necessarily coincide with the difference between the mean vectors of the two distributions.

Consider a two-dimensional example where

$$\{x_1 = (3, 0),\ x_2 = (0, 0),\ x_3 = (0, 3)\}, \qquad \{y_1 = (-3, 0),\ y_2 = (0, 0),\ y_3 = (0, -3)\},$$

each with equal masses. For $p = 1$, the mean vectors are $\bar{\mu} = (0, 0)$ and $\bar{\nu} = (0, 0)$. Using their difference as the translation gives

$$W_1(\mu, \nu) = \tfrac{1}{3}(3 + 3 + 6) = 4.$$

However, when translating the source distribution by $t_0 = (-3, -3)$, we obtain

$$W_1(\mu + t_0, \nu) = \tfrac{1}{3}(3 + 3) = 2 < 4.$$

Therefore, for $p \neq 2$, the optimal translation minimizing the ROT cost is not the same as the difference between the mean vectors of μ and ν.

C Rotation Equivalence

In addition to the translation relation, an extension of the ROT framework is to consider the *rotation relation* on the space of probability measures. This perspective allows us to explore the geometric structure of distributions up to rigid-body rotations or reflections, and to study the computational behavior of the resulting optimization problem.

C.1 Rotation equivalence and induced quotient metric

Let $O(n)$ denote the orthogonal group of $\mathbb{R}^d$, consisting of all rotations and reflections. We define an equivalence relation $\sim_R$ on $\mathcal{P}_p(\mathbb{R}^d)$ as follows:

$$\mu \sim_R \nu \quad \Longleftrightarrow \quad \exists R \in O(n) \text{ such that } \nu = R_{\#}\mu.$$

The quotient space under this relation is denoted by

$$\mathcal{Q}_R := \mathcal{P}_p(\mathbb{R}^d)/\sim_R,$$

where each element $[\mu]_R$ represents the *orbit* of μ under all rotations and reflections.

Definition C.1 (Rotation-invariant Wasserstein distance)**.** For any two equivalence classes $[\mu]_R, [\nu]_R \in \mathcal{Q}_R$, we define

$$W_p^{(R)}([\mu]_R, [\nu]_R) := \inf_{R \in O(n)} W_p(\mu,\, R_{\#}\nu).$$

The following result establishes that this construction yields a valid metric on the quotient space.

Proposition C.1 (Well-defined metric on the rotation quotient space)**.** *The function $W_p^{(R)}$ defines a real metric on $\mathcal{Q}_R$. Moreover, the infimum in the definition is attained.*

Proof. First, $W_p^{(R)}$ is well-defined since for any μ', ν' in the same equivalence classes as μ, ν, there exist $R_0, S_0 \in O(n)$ such that $\mu' = R_{0\#}\mu$ and $\nu' = S_{0\#}\nu$. Using the invariance of W_p under orthogonal transformations,

$$\inf_R W_p(\mu', R_{\#}\nu') = \inf_R W_p(R_{0\#}\mu, (RS_0)_{\#}\nu) = \inf_Q W_p(\mu, Q_{\#}\nu),$$

where $Q = R_0^{-1}RS_0$. so the value is independent of the representatives.

The properties of a metric follow directly:

- *Non-negativity and symmetry:* Inherited from W_p, since $R \mapsto R^{-1}$ is a bijection on $O(n)$.
- *Identity of indiscernibles:* If $W_p^{(R)}([\mu]_R, [\nu]_R) = 0$, then there exists a sequence $R_k \in O(n)$ with $W_p(\mu, (R_k)_{\#}\nu) \to 0$. Compactness of $O(n)$ ensures a convergent subsequence $R_{k_\ell} \to R^*$, and continuity of the pushforward implies $\mu = R^*_{\#}\nu$. Thus $[\mu]_R = [\nu]_R$. The converse is immediate.
- *Triangle inequality:* For any μ, ν, η and $R, S \in O(n)$,

  $$W_p(\mu, (RS)_{\#}\eta) \le W_p(\mu, R_{\#}\nu) + W_p(R_{\#}\nu, (RS)_{\#}\eta) = W_p(\mu, R_{\#}\nu) + W_p(\nu, S_{\#}\eta).$$

  Taking the infimum over R, S yields

  $$W_p^{(R)}([\mu]_R, [\eta]_R) \le W_p^{(R)}([\mu]_R, [\nu]_R) + W_p^{(R)}([\nu]_R, [\eta]_R).$$

Finally, since $O(n)$ is compact and $R \mapsto W_p(\mu, R_{\#}\nu)$ is continuous, the infimum is attained. Hence, $W_p^{(R)}$ defines a real metric on $\mathcal{Q}_R$. □

C.2 Non-convexity of the optimization over rotations

Although the rotation-induced distance $W_p^{(R)}$ defines a valid metric on the quotient space $\mathcal{Q}_R$, the underlying optimization problem over rotations is generally non-convex. The following example illustrates this behavior even in a simple two-dimensional case.

Proposition C.2 (Non-convexity of $W_p(\mu, R_\#\nu)$ with respect to rotation)**.** *Consider the cost function $c(x, y) = \|x - y\|_p^p$ with $p \geq 1$. Let the source and target distributions μ and ν be defined on $\mathbb{R}^2$ as*

$$\mu = \tfrac{1}{3}\sum_{k=0}^{2} \delta_{u_k}, \qquad \nu = \tfrac{1}{3}\sum_{k=0}^{2} \delta_{u_k},$$

where $u_k = (\cos(2\pi k/3), \sin(2\pi k/3))$ for $k = 0, 1, 2$. For each rotation matrix R_θ with angle $\theta \in [0, 2\pi)$, define

$$f(\theta) := W_p^p\big(\mu, (R_\theta)_\#\nu\big).$$

Then $f(\theta)$ is a non-convex function on $[0, 2\pi)$, and it possesses three disconnected global minimizers.

Proof. Since both μ and ν have equal discrete masses, the optimal coupling matches the three support points under cyclic permutations $\sigma_m(k) = k + m \bmod 3$, for $m \in \{0, 1, 2\}$. For any fixed permutation σ_m, the transport cost is

$$\frac{1}{3}\sum_{k=0}^{2} \|u_k - R_\theta u_{\sigma_m(k)}\|_p^p = \big(2 - 2\cos(\theta - \tfrac{2\pi m}{3})\big)^{p/2},$$

since for unit vectors a, b separated by an angle δ, $\|a - b\|_2^2 = 2 - 2\cos\delta$. Therefore, taking the minimum over the three possible permutations yields

$$f(\theta) = \min_{m\in\{0,1,2\}} \big(2 - 2\cos(\theta - \tfrac{2\pi m}{3})\big)^{p/2}.$$

It follows that $f(\theta)$ achieves its minimum value $f(\theta^*) = 0$ at $\theta^* \in \{0, 2\pi/3, 4\pi/3\}$, corresponding to perfect rotational alignment. Between any two minima (e.g., between 0 and $2\pi/3$), f attains a local maximum at $\theta = \pi/3$, where

$$f(\pi/3) = \big(2 - 2\cos(\pi/3)\big)^{p/2} = 1.$$

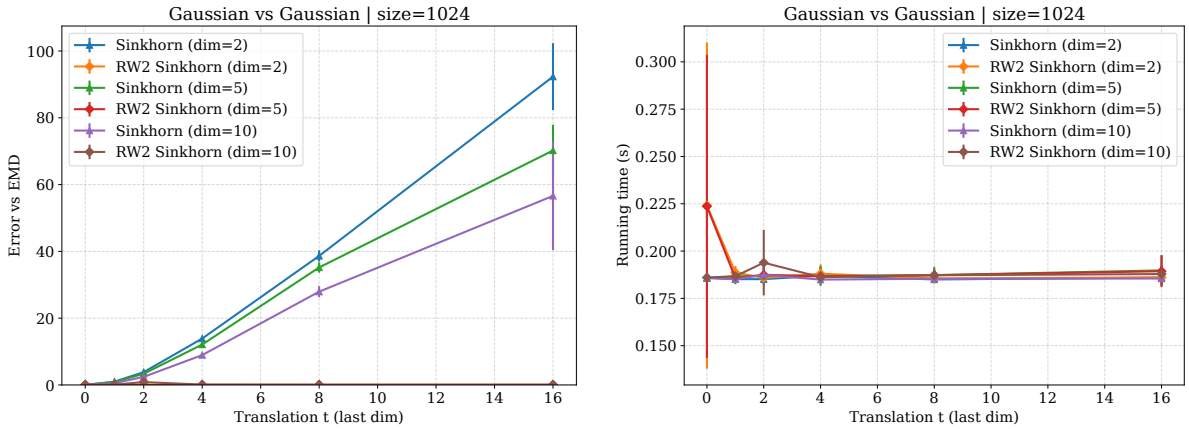
Thus, f has a periodic multi-well structure with three disconnected global minima separated by higher-cost regions. Hence, $f(\theta)$ is non-convex, even in this simple discrete case. □

This example shows that, while the rotation-invariant Wasserstein distance $W_p^{(R)}$ defines a well-posed metric on the quotient space $\mathcal{Q}_R$, the corresponding optimization problem is generally non-convex. Consequently, finding the optimal rotation R^* that minimizes $W_p(\mu, R_\#\nu)$ may require nonconvex optimization strategies or carefully designed initialization schemes.

D Additional Experiment Results

D.1 Additional Experiments for the Adaptive RW_2-Sinkhorn Algorithm

To further assess the RW_2-Sinkhorn algorithm, we evaluate its performance on three additional source–target distribution pairs: (1) Gaussian → Gaussian, (2) Gaussian → Geometric, and (3) Gaussian → Poisson. All tests use 1,024 samples, translation magnitude ranges over $\{0, 2, 4, 8, 16\}$, and dimension $d \in \{2, 5, 10\}$, following the setup in Subsection 5.1.2. We report the numerical error (relative to LP ground truth) and the running time.



(a) Computation error vs. translation magnitude.

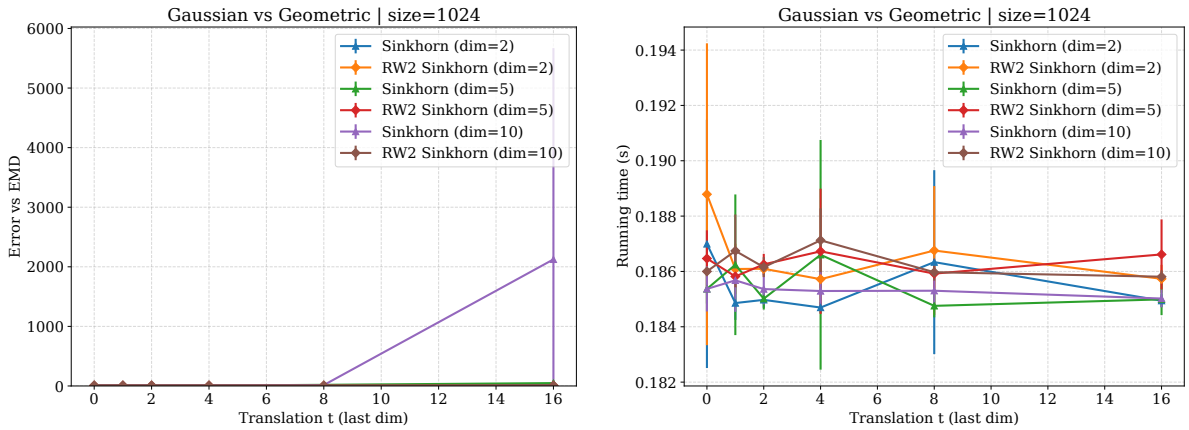
(b) Running time vs. translation magnitude.

Figure 7: Gaussian → Gaussian experiment. (a) The standard Sinkhorn algorithm becomes increasingly unstable as translation grows, producing large numerical errors. The three RW_2-Sinkhorn curves completely overlap at the bottom of the plot, indicating consistent low numerical errors of the RW_2-Sinkhorn formulation across dimensions. (b) Running time remains nearly identical across dimensions and translation magnitudes.

Gaussian → Gaussian. In this setting, the standard Sinkhorn algorithm becomes increasingly unstable as the translation magnitude grows, resulting in noticeable growth in numerical error (Figure 7(a)). In contrast, the RW_2 version consistently has much lower error across all dimensions and translation levels. The three RW_2-Sinkhorn curves completely overlap at the bottom of the plot, indicating consistent low numerical errors of the RW_2-Sinkhorn formulation across dimensions. Figure 7(b) shows that the running time of the RW_2-Sinkhorn algorithm remains nearly identical to that of the standard Sinkhorn method.

Gaussian → Geometric. For this task, both the standard and RW_2-Sinkhorn algorithms achieve comparable numerical accuracy across all tested dimensions and translation magnitudes (Figure 8(a)). The five curves, except for the Sinkhorn method at dimension 10, completely overlap at the bottom of the plot. Similarly, the running time of both methods remains nearly identical (Figure 8(b)), showing that the adaptive step does not cause extra running time.

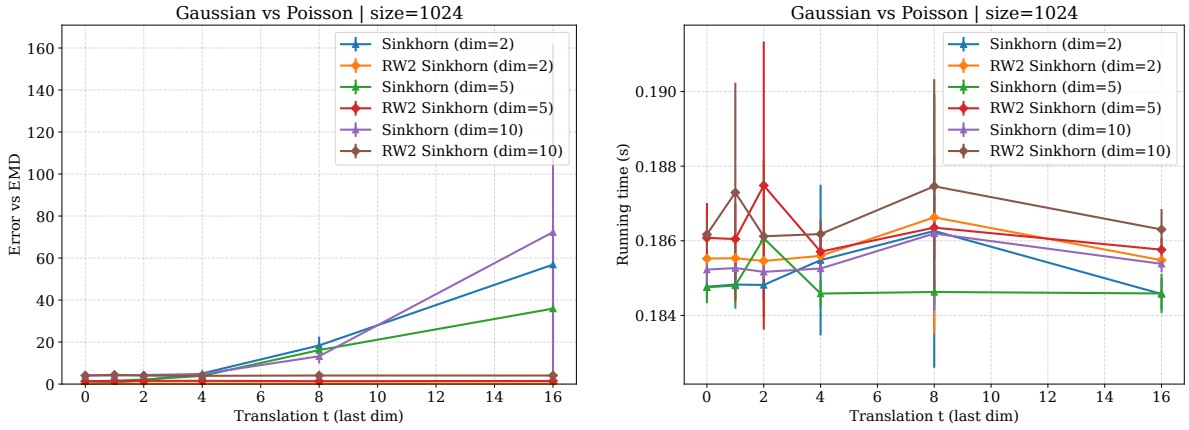
Gaussian → Poisson. The Poisson target introduces heavier-tailed behavior and greater distributional mismatch. As shown in Figure 9(a), the standard Sinkhorn algorithm becomes increasingly unstable as translation grows, with large error spikes appearing especially in higher dimensions. In contrast, the RW_2-Sinkhorn algorithm has consistently lower error across all translation magnitudes. The three RW_2-Sinkhorn curves completely overlap at the bottom of the plot, indicating consistent low numerical errors of the RW_2-Sinkhorn formulation across dimensions. Running time remains essentially unchanged between the two approaches (Figure 9(b)), showing that the improved stability comes at no additional computational cost.



(a) Computation error vs. translation magnitude.

(b) Running time vs. translation magnitude.

Figure 8: Gaussian → Geometric experiment. (a) Both standard Sinkhorn and RW_2-Sinkhorn algorithm achieve comparable numerical accuracy. The five curves, except for the Sinkhorn algorithm at dimension 10, completely overlap at the bottom of the plot. (b) Running time remains nearly identical across dimensions and translation magnitudes.



(a) Computation error vs. translation magnitude.

(b) Running time vs. translation magnitude.

Figure 9: Gaussian → Poisson experiment. (a) The standard Sinkhorn algorithm becomes increasingly unstable as translation grows, producing large numerical errors. The three RW_2-Sinkhorn curves almost overlap at the bottom of the plot, indicating consistent low numerical errors of the RW_2-Sinkhorn formulation across dimensions. (b) Running time remains essentially unchanged, showing that the adaptive RW_2 step improves stability without extra computational time.

D.2 Sequence retrieval experimental results for Section 5.2

Sequence settings. In addition to single-snapshot retrieval, we extend our evaluation to the retrieval of *thunderstorm sequences*, where each sequence consists of a temporally ordered series of thunderstorm snapshots. In our experiments, each sequence spans one hour and contains six consecutive snapshots. The similarity between two sequences is measured as the average of the metric distances computed between corresponding snapshot pairs. Under this setting, we report the top-3 sequence retrieval results for all five metrics considered in Section 5.2, namely ℓ_2, W_2, and RW_p with $p \in \{1, 2, 4\}$. To avoid redundant retrievals, the selected sequences are constrained to originate from distinct calendar dates (year–month–day).

Sequence retrieval results. Figures 10–14 present the sequence retrieval results for the five metrics. For each metric, the results are visualized using four rows: the first row corresponds to the same reference thunderstorm sequence, while the second through fourth rows show the top three retrieved sequences identified by the corresponding metric. All retrieved sequences originate from calendar dates that are distinct from one another and from that of the reference sequence, ensuring temporal diversity in the retrieved results.

Across the five metrics, clear differences in retrieval behavior can be observed. The classical W_2 distance tends to favor sequences that are spatially close to the reference, even when their internal storm structures and temporal evolution differ. In contrast, the proposed RW_p distances consistently emphasize similarity in storm morphology and evolution patterns, leading to retrieved sequences that are visually more coherent with the reference sequence. Among the proposed variants, RW_2 provides the most balanced performance, yielding sequences that closely match both the shape and temporal progression of the reference thunderstorm. These observations are consistent with the snapshot-based retrieval results reported in Section 5.2 and further demonstrate the robustness of the proposed RW_p distances for spatio-temporal retrieval tasks.

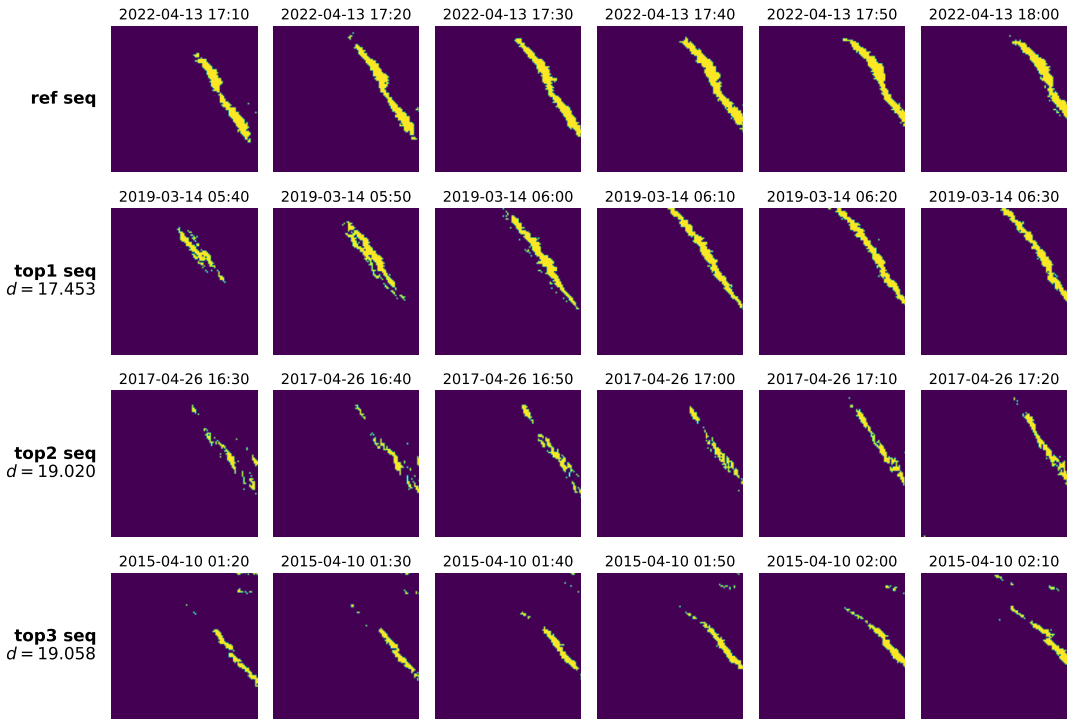


Figure 10: Sequence retrieval results using the ℓ_2 distance. The first row shows the reference thunderstorm sequence consisting of six consecutive snapshots. The second to fourth rows show the top three retrieved sequences. Each column corresponds to a 10-minute interval, and timestamps are shown above each frame.

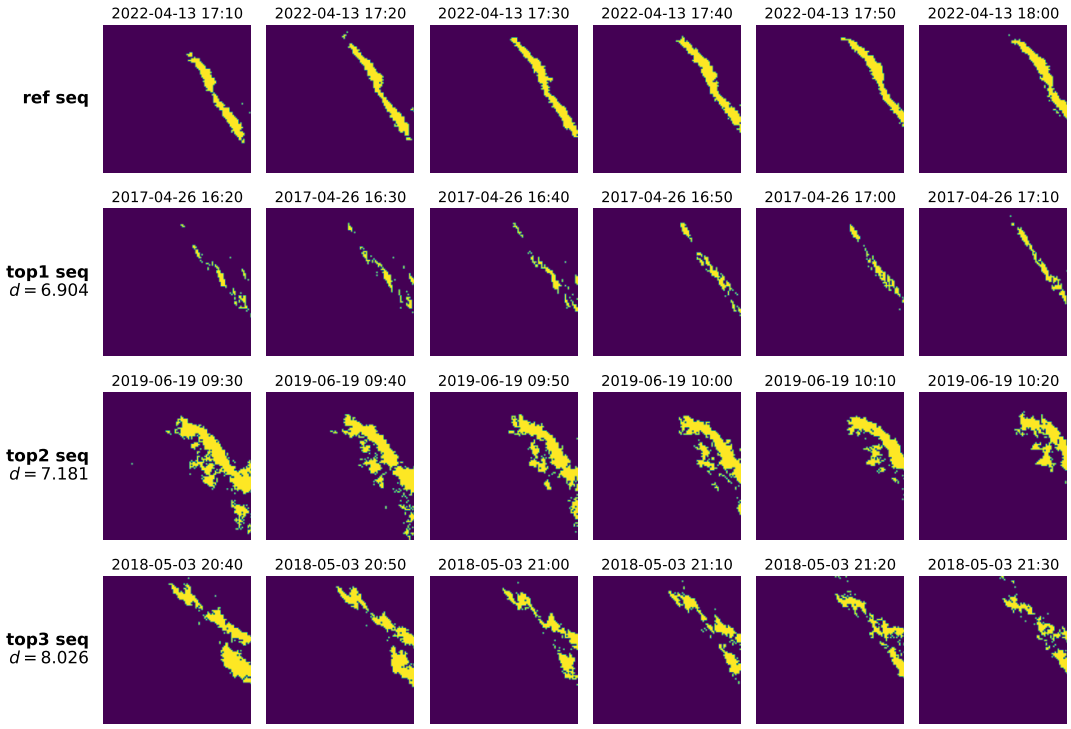


Figure 11: Sequence retrieval results using the W_2 distance. While the retrieved sequences are spatially close to the reference, their internal storm structures may differ. The first row is the reference sequence, followed by the top three retrieved sequences from distinct calendar dates.

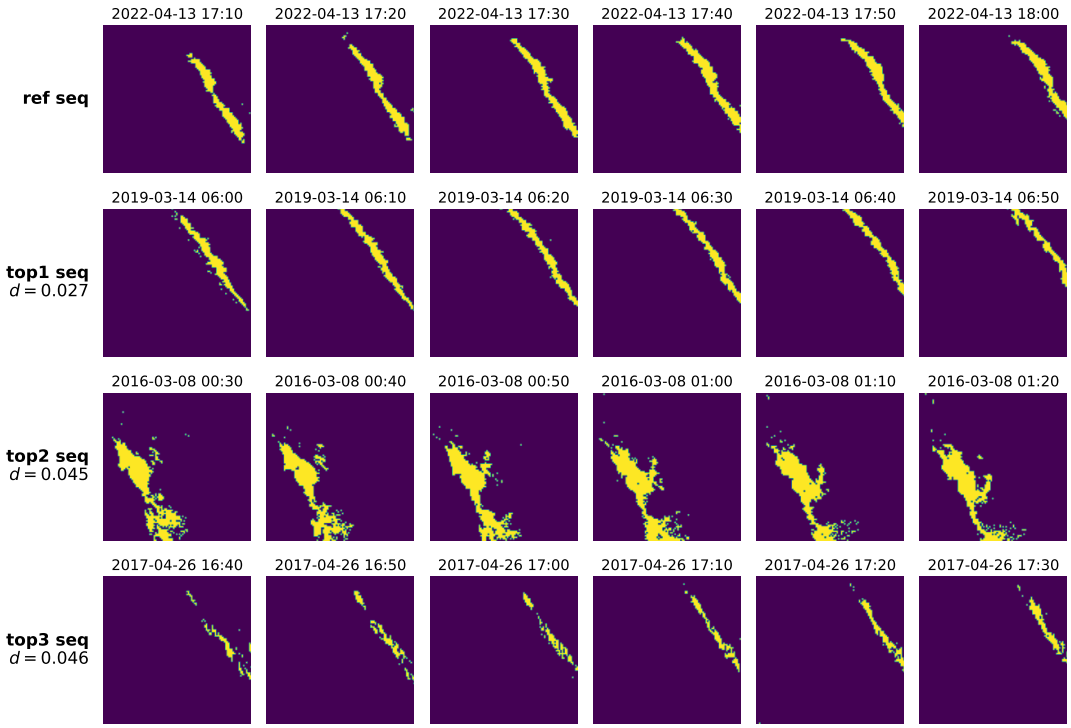


Figure 12: Sequence retrieval results using the RW_1 distance. Compared to W_2, the retrieved sequences exhibit improved consistency in storm morphology and evolution patterns.

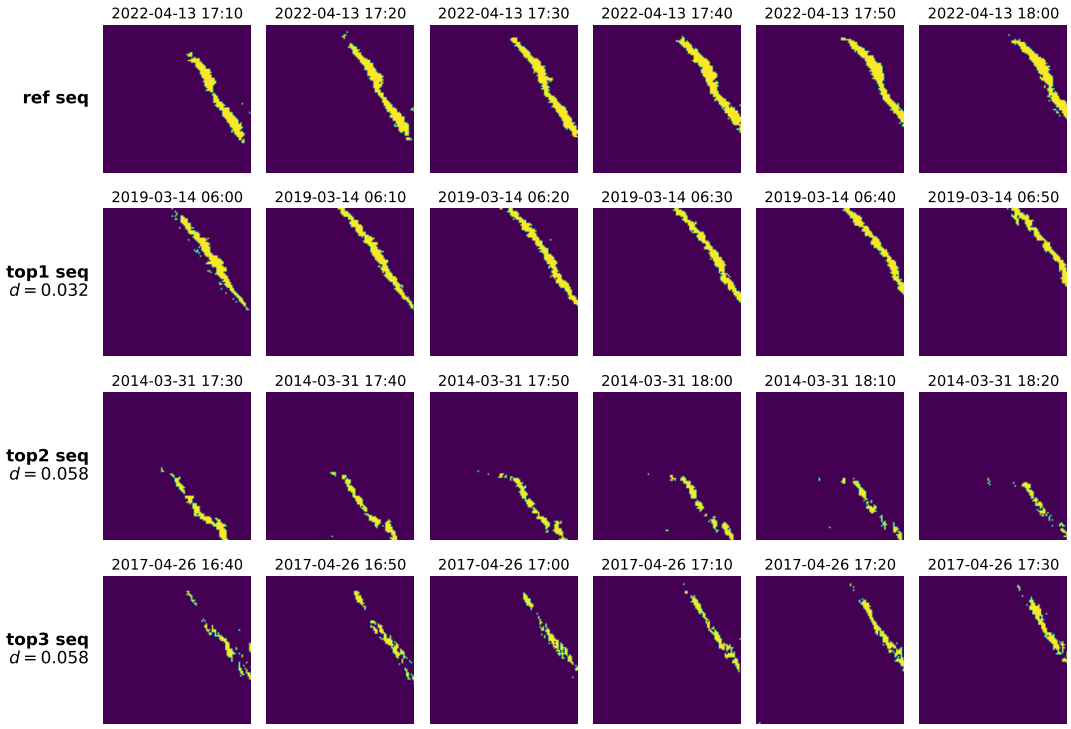


Figure 13: Sequence retrieval results using the RW_2 distance. The retrieved sequences closely match the reference sequence in both spatial structure and temporal evolution. This metric provides the most balanced performance among all considered distances.

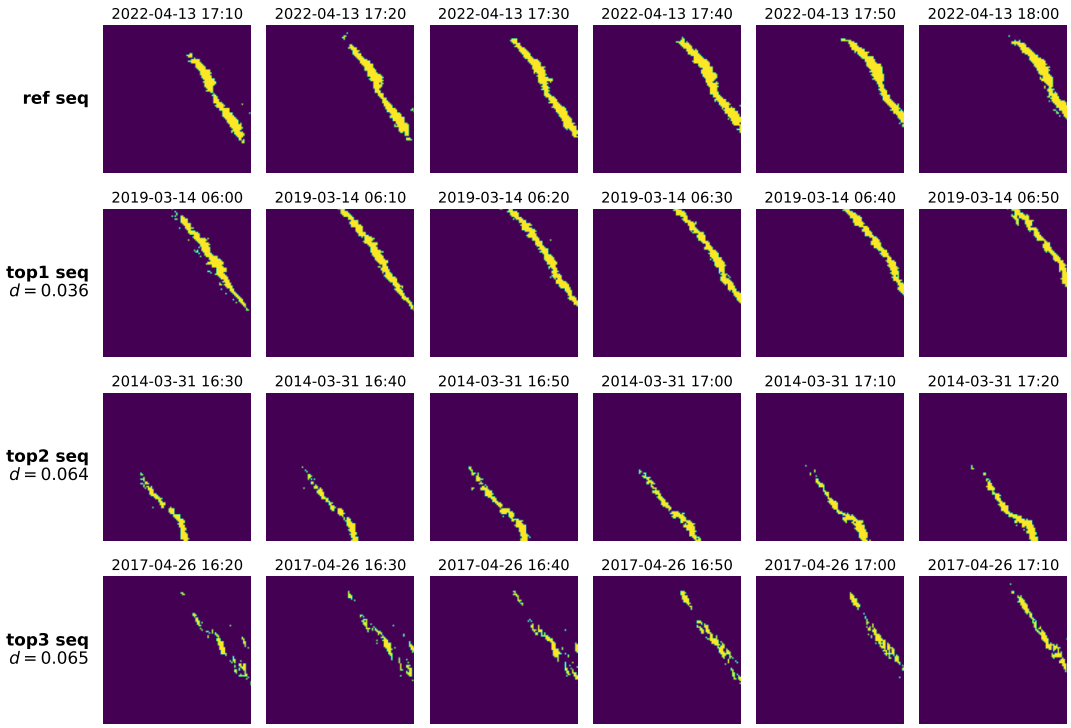


Figure 14: Sequence retrieval results using the RW_4 distance. Although slightly more sensitive to outliers, RW_4 still preserves the overall storm evolution patterns across time.