Received: 23 December 2021 | Revised: 22 March 2022 | Accepted: 13 April 2022

DOI: 10.1002/wics.1587

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Projection-based techniques for high-dimensional optimal transport problems

Jingyi Zhang[1] | **Ping Ma**[2] | **Wenxuan Zhong**[2] | **Cheng Meng**[3]

[1]Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

[2]Department of Statistics, University of Georgia, Athens, Georgia, USA

[3]Center of Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Correspondence
Cheng Meng, Center of Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China.
Email: chengmeng@ruc.edu.cn

Funding information
National Institutes of Health, Grant/Award Number: R01GM122080; National Natural Science Foundation of China, Grant/Award Number: 12101606; National Science Foundation, Grant/Award Numbers: DMS-1903226, DMS-1925066, DMS-2124493

Edited by: Henry Horng-Shing Lu, Commissioning Editor and David W. Scott, Review Editor and Co-Editor-in-Chief

Abstract

Optimal transport (OT) methods seek a transformation map (or plan) between two probability measures, such that the transformation has the minimum transportation cost. Such a minimum transport cost, with a certain power transform, is called the Wasserstein distance. Recently, OT methods have drawn great attention in statistics, machine learning, and computer science, especially in deep generative neural networks. Despite its broad applications, the estimation of high-dimensional Wasserstein distances is a well-known challenging problem owing to the curse-of-dimensionality. There are some cutting-edge projection-based techniques that tackle high-dimensional OT problems. Three major approaches of such techniques are introduced, respectively, the slicing approach, the iterative projection approach, and the projection robust OT approach. Open challenges are discussed at the end of the review.

This article is categorized under:

Statistical and Graphical Methods of Data Analysis > Dimension Reduction

Statistical Learning and Exploratory Methods of the Data Sciences > Manifold Learning

KEYWORDS

curse of dimensionality, dimension reduction, optimal transport, Wasserstein distance

1 | INTRODUCTION

Consider the resource allocation problem as shown in Figure 1. Suppose that we have n mines mining iron ore and m factories. Each factory has a certain demand for the iron ore that the mines produce. We assume that the total amount of the iron ore produced by the mines equals the total demand for which in the factories. The goal is to move all the iron ore from mines to factories, such that the total transport cost is minimized, under the condition that the demand for every factory could be successfully met.

In the 18th century, French mathematician Gaspard Monge (1746–1818) first formulated such a resource allocation problem as a mathematical problem. In particular, he regarded the resources and the demands as two probability measures, denoted by μ and ν, respectively. Of interest is to seek a transport map (or plan) between μ and ν with the minimum transport cost. Such a minimum transport cost, with a certain power transform, is called the Wasserstein distance between μ and ν. The problem of finding such a transport map (or plan) is called the optimal transport (OT) problem

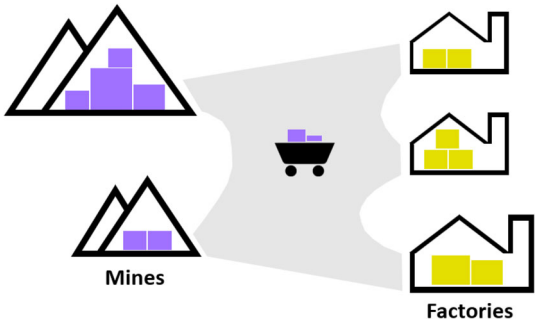

FIGURE 1 An illustration of the resource allocation problem

and has long been studied in mathematics and operational research (Villani, 2008). Formal definitions for the transport map, the transport plan, and the Wasserstein distance are provided in Section 2.

During recent decades, OT methods have been reinvigorated in a remarkable proliferation of modern data science applications. Many statistical and machine learning problems can be recast as finding the OT map between two measures. For example, the deep generative models aim to seek a transformation that maps a fixed distribution, for example, the standard Gaussian or uniform distribution, to the underlying population distribution of the observed sample (Arjovsky et al., 2017; Y. Chen et al., 2018; Goodfellow et al., 2014; N. Lei et al., 2019; Meng et al., 2019). Another example is the problem of domain adaptation, which aims to learn a well-trained model from a source data distribution and transfer this model to adopt a target data distribution (Courty et al., 2016; Muzellec & Cuturi, 2019). Due to the flexibility in practical applications, OT methods have recently drawn great attentions in various machine learning tasks, for example, density estimation (Canas & Rosasco, 2012; Weed & Berthet, 2019), dictionary learning (Cazelles et al., 2018; Rolet et al., 2016; Schmitz et al., 2018; Seguy & Cuturi, 2015), clustering (Flamary et al., 2018; Lin, Ho, et al., 2020; Meng, Yu, et al., 2020; Staib et al., 2017), kernel methods (Carriere et al., 2017; Jagarlapudi & Jawanpuria, 2020), fair machine learning (Black et al., 2020; Gordaliza et al., 2019), structural data analysis (Xu, 2020; Xu et al., 2021; Xu et al., 2022; Xu, Luo, & Carin, 2019; Xu, Luo, Zha, & Duke, 2019), and generative models (An et al., 2020; Meng et al., 2019). In addition, OT methods are widely applied in statistics, for example, two-sample testing (Ramdas et al., 2017), statistical inference (Bigot et al., 2019; Del Barrio et al., 2019; Klatt et al., 2020; Kroshnin et al., 2021; Tameling & Munk, 2018; Zemel et al., 2019), and regression analysis (Hütter & Rigollet, 2021; Janati et al., 2020; Rigollet & Weed, 2019). Last but not least, OT methods also find many applications in computer science, for example, natural language processing (Alaux et al., 2018; L. Chen et al., 2019; Grave et al., 2019; Singh et al., 2020; Z. Wang et al., 2020; Xu et al., 2018; Yurochkin et al., 2019), computer vision (Alvarez-Melis et al., 2018; Feydy et al., 2017; Seguy et al., 2018; Solomon et al., 2016; W. Wang et al., 2021), computer graphics (Cui et al., 2019; Lavenant et al., 2018), deep learning (Adler & Lunz, 2018; Arjovsky et al., 2017; Hashimoto et al., 2016; Lim et al., 2020; Montavon et al., 2016; W. Wang et al., 2021), as well as other domain sciences (Dai Yang et al., 2020; Del Barrio et al., 2020; Schiebinger et al., 2019; Tong et al., 2020). We refer to Peyré et al. (2019), Panaretos and Zemel (2019), and Zhang et al. (2020) for recent reviews.

Although OT methods find a large number of applications in practice, the estimation of the empirical optimal transport plan (OTP) and the corresponding Wasserstein distance suffers from the "curse-of-dimensionality" in high-dimensional spaces (Fournier & Guillin, 2015; Panaretos & Zemel, 2019). Suppose that we observe two d-dimensional samples of size n. It was first shown in Dudley (1969) that, in general cases such that the measure is absolutely continuous with respect to (w.r.t.) Lebesgue measure, the empirical Wasserstein distance between these two samples converges to its population counterpart roughly at the rate of $O\left(n^{-1/d}\right)$ when $d \geq 2$; see Fournier and Guillin (2015), Panaretos and Zemel (2019), Weed and Bach (2019), J. Lei et al. (2020) and the reference therein for further discussion. Such a convergence rate is in a sense disappointing, as it indicates that the empirical Wasserstein distance hardly converges to its population counterpart when the dimension d is moderate or large. Fortunately, Weed and Berthet (2019) showed that the convergence rate of the empirical Wasserstein distance could be improved if the "implicit dimension" k is much smaller than d. Here, we call two d-dimensional measures μ and ν have implicit dimension k if they differ only on a k-dimensional subspace. In such cases, loosely speaking, the convergence rate of the empirical Wasserstein distance could be improved from $O\left(n^{-1/d}\right)$ to $O\left(n^{-1/k}\right)$. Therefore, finding an informative low-dimensional subspace is essential for obtaining accurate estimations of Wasserstein distance and OTPs in high-dimensional space.

Following this line of thinking, there has been a large number of studies dedicated to developing projection-based methods for OT problems in recent decades. These methods can be roughly categorized into three classes, that is, the slicing approach, the iterative projection approach, and the projection robust OT approach. In this article, we will present the details of these three main approaches. We will also introduce the variants of these approaches as well as their applications. The rest of the article is organized as follows. We start in Section 2 by introducing the essential background of the OT problem. We then roughly introduce three classes of projection-based techniques that tackle high-dimensional OT problems. Details of the slicing approach and its variants, the iterative projection approach, and the projection robust OT approach are provided in Sections 3, 4, and 5, respectively. Section 6 summarizes the review and discusses some open areas.

2 | PROBLEM FORMULATION

In this section, we first introduce the problem setup of the OT problems. Next, two popular formulations of such problems are presented, that is, the Monge formulation and the Kantorovich formulation. The Monge formulation is more intuitive; however, it suffers from certain limitations in practice. Such limitation can be overcome by the Kantorovich formulation. Finally, we introduce the notion of Wasserstein distance, followed by the computational issues of solving the OT problems.

2.1 | OT problems and Monge formulation

Different from the resource allocation problem, now let us consider another example to motivate the OT problem. Suppose that we want to move a large pile of sand using a shovel. The goal is to construct a particular shape, say a sandcastle, using the sand. Naturally, we wish the total "effort" to be as small as possible. The effort, intuitively, can be regarded as the "work" in the sense of physical, that is, the product of force and displacement. Here, the sand and the sandcastle can be regarded as two probability measures, denoted by μ and ν, respectively. The process of constructing the sandcastle using the sand, roughly speaking, can be regarded as applying a "transport map" on μ that transports μ to ν. Note that such a transport map may not have to be a one-to-one map, as will be detailed later. Among all possible transport maps, the goal is to find the one with the minimum transport cost, as we wish to minimize the total effort. The problem of finding such a transport map is called the OT problem.

Mathematically, one can formulate the OT problem as follows. Considering the set of all Borel probability measures in $\mathbb{R}^d$, denoted by $\mathcal{P}(\mathbb{R}^d)$, and let

$$\mathcal{P}_2(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \,|\, \int \|x\|^2 \mathrm{d}\mu(x) < \infty \right\}.$$

Let μ and ν be two measures such that $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Let # denote the push-forward operator, such that for any measurable $\Omega \subset \mathbb{R}^d$, one has $\phi_{\#}(\mu)(\Omega) = \mu\left(\phi^{-1}(\Omega)\right)$. A measureable map $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is called the measure-preserving map between μ and ν if $\phi_{\#}(\mu) = \nu$ and $\phi_{\#}^{-1}(\nu) = \mu$. Let $\Phi(\mu, \nu)$ be the set of all such measure-preserving maps. Among all the maps in $\Phi(\mu, \nu)$, the optimal transport map (OTM) is defined as

$$\phi^{\dagger} \coloneqq \underset{\phi \in \Phi(\mu,\nu)}{\operatorname{arginf}} \int_{\mathbb{R}^d} c(x, \phi(x)) \mathrm{d}\mu(x). \tag{1}$$

Here, $c(\cdot, \cdot)$ is the cost function, and one popular choice of which is the squared Euclidean cost, that is, $c(x, y) = \|x - y\|^2$. Throughout this review, we mainly focus on the OT problems w.r.t. this cost. Solving the OT problem w.r.t other choices of cost functions, especially the concave ones, is still an active research area. We refer to Villani (2008) for further details of such problems, which are beyond the scope of this review.

Applying the squared Euclidean cost to Equation (1) yields

$$\phi^{\dagger} := \underset{\phi \in \Phi(\mu,\nu)}{\operatorname{arginf}} \int_{\mathbb{R}^d} \|x-\phi(x)\|^2 \mathrm{d}\mu(x). \tag{2}$$

Equations (1) and (2) are usually called the Monge formulation of the OT problem, and their solution $\phi^{\dagger}$ is called the OTM or the Monge map. One limitation of the Monge formulation, however, is that its solution may not exist. For example, consider the scenario when ν is the standard Gaussian distribution, and μ is the delta-Dirac distribution, that is, the Dirac measure on a single point. Under such a scenario, there does not exist a many-to-one map that maps μ to ν, and thus the set Φ is empty.

2.2 | Kantorovich formulation

To overcome the aforementioned limitation, Kantorovich (1942) reformulated the OT problem as finding a certain joint distribution of μ and ν instead of finding a transport map between these two. In particular, Kantorovich (1942) considered a family of the joint distribution of μ and ν, termed as the "coupling" π, such that two particular marginal distributions of π are equal to μ and ν, respectively. Let Π be the set of all such couplings, which can be defined as

$$\Pi(\mu,\nu) = \left\{\pi \in \mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right) \ s.t. \ \forall \ \text{Borelset} \ A,B \subset \mathbb{R}^d, \ \pi\left(A \times \mathbb{R}^d\right) = \mu(A), \ \pi\left(\mathbb{R}^d \times B\right) = \nu(B)\right\}. \tag{3}$$

Kantorovich stated that solving the OT problem (w.r.t. the squared Euclidean distance) is equivalent to finding the optimal coupling, defined as

$$\pi^* := \underset{\pi \in \Pi(\mu,\nu)}{\arg\inf} \int \|x-y\|^2 \mathrm{d}\pi(x,y). \tag{4}$$

Equation (4) is usually called the Kantorovich formulation of the OT problem, and its solution π^* is called the OTP.

The Kantorovich formulation enjoys several advantages over the Monge formulation. First, the solution of the Kantorovich formulation is a joint distribution that lies in the nonempty family Π, and thus always exists. Second, the Monge formulation is a highly nonlinear program that is relatively complicated to solve, while the Kantorovich formulation can be solved effectively using linear programming techniques. In particular, for two discrete distributions defined on n points, the Kantorovich OT problem can be solved within $O(n^3 \log n)$ computational time using classic linear programming techniques (Peyré et al., 2019). Last but not least, the Kantorovich formulation allows the mass to split from a source toward several targets and thus is more practical.

Recall the resource allocation problem at the beginning of this review. It may be unrealistic to assume that there always exists a one-to-one or many-to-one map between mines and factories, which can meet all the demands of the factories. Instead, a practical solution should allow the delivery of the iron ore from one mine to multiple factories. Such a solution can be regarded as an OTP.

Although the Kantorovich formulation is more flexible than the Monge formulation, the celebrated Brenier theorem (Brenier, 1991) ensures that in $\mathbb{R}^d$ for $p=2$, if at least one of μ and ν has a density, the Kantorovich and the Monge problems are equivalent. We refer to Remark 2.23 of Peyré et al. (2019) for more details.

2.3 | Wasserstein distance

Closely related to the OT problem is the Wasserstein distance, which is a metric that measures the discrepancy between two probability measures. Intuitively, Wasserstein distance measures the transport cost between two measures. The Wasserstein distance thus is also called the earth mover's distance in the literature (Levina & Bickel, 2001; Peyré et al., 2019). Intuitively, if we regard the OT problem as an optimization problem, then the Wasserstein distance is simply the optimal objective value with a certain power transform.

In terms of the Monge formulation, the Wasserstein distance of order 2 is defined as

$$W_2(\mu,\nu) := \left(\inf_{\phi \in \Phi(\mu,\nu)} \int_{\mathbb{R}^d} \| x - \phi(x) \|^2 \mathrm{d}\mu \right)^{1/2}. \tag{5}$$

Analogously, one can define the Wasserstein distance of order 2 w.r.t. the Kantorovich formulation as,

$$W_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \| x - y \|^2 \mathrm{d}\pi(x,y) \right)^{1/2}. \tag{6}$$

As a metric that measures the discrepancy between two probability measures, Wasserstein distance enjoys several advantages over other well-known metrics, for example, Jenson–Shannon divergence (JS-divergence) and the Kullback–Leibler divergence (KL-divergence). Consider the case when two measures μ and ν have different nonzero support, for example, μ and ν reside in low-dimensional manifolds without overlaps. In such a case, both JS-divergence and KL-divergence (w.r.t. these two measures) would give constant values, no matter how far away their nonzero supports are. In other words, such commonly-used metrics may fail to capture the discrepancy between two measures effectively. Such a limitation, which is closely related to the problem of *gradient vanishing*, is believed to be one of the main reasons why the original generative adversarial net (GAN) (Goodfellow et al., 2014) suffers from an unstable learning process. We refer to Arjovsky et al. (2017); Gulrajani et al. (2017); Tolstikhin et al. (2018) for more discussions.

Recall that the Wasserstein distance measures the discrepancy between two measures using the "transport cost." The Wasserstein distance thus is able to provide a reasonable discrepancy between two measures, no matter whether they have overlapping nonzero supports or not. Recently, a large number of studies suggested utilizing the Wasserstein distance and its variants for a more stable and robust training process in deep generative models (Deshpande et al., 2019; Kolouri et al., 2018; N. Lei et al., 2019; Tolstikhin et al., 2018).

2.4 | Computational issues

In practice, the OTP and its corresponding Wasserstein distance can be estimated by solving a linear system. Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be two probability distributions supported on a discrete set $\{\boldsymbol{x}_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in \Omega$ for $i = 1,...,n$, and $\Omega \subset \mathbb{R}^d$ is bounded. In other words, $\boldsymbol{p}$ and $\boldsymbol{q}$ are two vectors located on the simplex

$$\Delta_n := \left\{ \boldsymbol{v} \in \mathbb{R}^n : \ \sum_{i=1}^n \boldsymbol{v}_i = 1, \ \text{and} \ \boldsymbol{v}_i \geq 0, \ i = 1,...,n. \right\},$$

whose entries denote the weight of each distribution assigned to the points of $\{\boldsymbol{x}_i\}_{i=1}^n$. Let $\mathbf{1}_n$ be the all-ones vector with n elements. Let $\mathbf{C} \in \mathbb{R}^{n \times n}$ be the pair-wise distance matrix, where $\mathbf{C}_{ij} = \| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2$. Analogous to the definition of couplings in Equation (3), let $\Pi(\boldsymbol{p},\boldsymbol{q})$ denote the set of coupling matrices between $\boldsymbol{p}$ and $\boldsymbol{q}$, that is,

$$\Pi(\boldsymbol{p},\boldsymbol{q}) = \left\{ \mathbf{P} \in \mathbb{R}^{n \times n} : \ \mathbf{P}\mathbf{1}_n = \boldsymbol{p}, \ \mathbf{P}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{q} \right\}.$$

According to the Kantorovich formulation, finding the OTP between $\boldsymbol{p}$ and $\boldsymbol{q}$ is equivalent to solving the optimization problem

$$\mathbf{P}^* := \underset{\mathbf{P} \in \Pi(\boldsymbol{p},\boldsymbol{q})}{\arg\min} \langle \mathbf{P}, \mathbf{C} \rangle. \tag{7}$$

Here, $\langle \cdot, \cdot \rangle$ represents the summation of the entry-wise multiplication, such that, for any two matrix $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i=1}^n \sum_{j=1}^n \mathbf{A}_{ij} \mathbf{B}_{ij}$. The solution $\mathbf{P}^*$ in Equation (7) is usually called the optimal coupling matrix. Once the matrix $\mathbf{P}^*$ has been calculated, the Wasserstein distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ of order 2 can be simply written as $W_2(\boldsymbol{p},\boldsymbol{q}) = \langle \mathbf{P}^*, \mathbf{C} \rangle^{1/2}$.

The optimization problem in Equation (7) is a linear program with $O(n)$ linear constraints. Classic linear programming algorithms for solving such problems requiring a computational time of the order $O(n^3\log(n))$ (Peyré et al., 2019). Such a sizable computational cost hinders the broad applicability of OT methods in practice for datasets with large sample sizes. To alleviate the computation burden for OT problems, Cuturi (2013) considered a regularized variant of the minimization problem in Equation (7), written as

$$\mathbf{P}_{\eta}^{*} = \underset{\mathbf{P} \in \Pi(\boldsymbol{p},\boldsymbol{q})}{\arg\min} \left\{ \langle \mathbf{P}, \mathbf{C} \rangle - \eta^{-1} H(\mathbf{P}) \right\}. \tag{8}$$

Here, $\eta > 0$ is the regularization parameter, and $H(\mathbf{P}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{P}_{ij}\log\left(1/\mathbf{P}_{ij}\right)$ is the Shannon entropy of the matrix $\mathbf{P}$. We adopt the standard convention that $0\log(1/0) = 0$ in the Shannon entropy.

Cuturi (2013) showed that for a fixed η, the optimization problem (8) can be solved within $O(n^2\log(n))$ computational time using the Sinkhorn algorithm. In practice, a small η is associated with a more accurate estimation of the OTP and the Wasserstein distance. However, a small η also results in a longer computation time. Empirical studies showed, with the help of the Sinkhorn algorithm, the regularized OT problem can be solved reliably and efficiently in the cases when $n \approx 10^4$ (Flamary & Courty, 2017; Genevay et al., 2016). Recently, many studies are developed upon the Sinkhorn algorithm for faster calculations (Altschuler et al., 2017, 2019; Dvurechensky et al., 2018; Lin et al., 2019). For example, Altschuler et al. (2019) proposed the *Nys-sink* algorithm, which combined the Sinkhorn algorithm with the Nyström method, a popular technique for low-rank matrix decomposition (Gittens & Mahoney, 2016; Musco & Musco, 2017; S. Wang & Zhang, 2013; Williams & Seeger, 2001). They showed that the *Nys-sink* algorithm could efficiently solve the regularized OT problem of the size $n \approx 10^6$ on a single laptop. We refer to Zhang et al. (2020) for more efficient tools to solve the OT problem.

2.5 | Projection-based techniques for OT problems

Consider OT problems for two d-dimensional measures that are absolutely continuous w.r.t Lebesgue measure. The empirical Wasserstein distance between two samples of size n is shown to converge to its population counterpart roughly at the rate of $O\left(n^{-1/d}\right)$ (Dudley, 1969; Fournier & Guillin, 2015; J. Lei et al., 2020; Panaretos & Zemel, 2019; Weed & Bach, 2019). Such a convergence rate implies that when the dimension d grows, the empirical Wasserstein distance hardly converges. The estimation of the empirical OTP and the corresponding Wasserstein distance thus suffer from the "curse-of-dimensionality" in high-dimensional spaces (Fournier & Guillin, 2015; Panaretos & Zemel, 2019). Suppose these two d-dimensional measures differ only on a k-dimensional subspace, with k much smaller than d. Intuitively, the convergence rate of the empirical Wasserstein distance can be improved to $O\left(n^{-1/k}\right)$ as long as the k-dimensional subspace is properly estimated (Weed & Berthet, 2019). Motivated by such an idea, a large number of methods have been developed. These methods can be roughly categorized into three classes as follows.

- The slicing approach breaks down the problem of estimating high-dimensional Wasserstein distances into a series of subproblems, each of which solves a one-dimensional OT problem using projected samples (Bonneel et al., 2015; Pitie et al., 2005; Pitié et al., 2007; Rabin et al., 2011). The subproblems can be easily solved since one-dimensional OT problems admit closed-form solutions under mild conditions. The slicing approach is in some sense analogous to the additive model (Hastie & Tibshirani, 1990; Wood, 2017), as both approaches overcome the curse-of-dimensionality by approximating a multivariate function using the summation of a series of one-dimensional functions.
- The iterative projection approach is similar to the slicing approach in the sense that both of them utilize the closed-form solution of one-dimensional OT problems. Nevertheless, the main difference between them is that the one-dimensional OT components are independent of each other in the slicing approach, while in the iterative projection approach, these components depend on each other and are estimated sequentially. The idea of the iterative projection approach is similar to boosting (Schapire, 2003; Zhou, 2009) and projection pursuit regression (Friedman & Stuetzle, 1981; Friedman & Tukey, 1974; Huber, 1985) in the sense that searching for the next optimal component is based on the residual of previous ones.
- The projection robust OT approach assumes that the two underlying d-dimensional measures differ only in an implicit k-dimensional subspace. The goal is to seek such an implicit subspace and then estimate the empirical

Wasserstein distance using the projected samples. In practice, such an approach seeks the k-dimensional subspace that would maximize the Wasserstein distance between two measures after projection. Such a maximum quantity is usually called the projection robust Wasserstein distance.

3 | SLICING APPROACH FOR OT PROBLEMS

One attractive property of the OT problem is that the OTP and the Wasserstein distance have a closed-form expression for one-dimensional measures. In particular, considering two one-dimensional measures μ and ν, the Wasserstein distance (of order 2) between them takes the form

$$W_2(\mu,\nu)=\left(\int_0^1\left|F_\mu^{-1}(x)-F_\nu^{-1}(x)\right|^2\mathrm{d}x\right)^{1/2}, \tag{9}$$

where F_μ and F_ν are the cumulative distribution functions (CDF) w.r.t. μ and ν, respectively. Considering two equal-weighted one-dimensional samples, denoted by $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, respectively. Equation (9) indicates that the empirical Wasserstein distance between $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ can be simply calculated by first sorting both samples and then calculating the distance between the sorted samples.

Unfortunately, the closed-form solution is not available for general high-dimensional OT problems, except for special cases, for example, when both μ and ν are Gaussian distribution (Peyré et al., 2019; Villani, 2008). One natural idea for tackling high-dimensional OT problems is to break them down into a series of subproblems, each of which involves solving a one-dimensional OT problem. Such an idea motivates the slicing approach. We now start by introducing the notion of sliced-Wasserstein (SW) distance, followed by several variants of the SW distance.

3.1 | Sliced-Wasserstein distance

Intuitively, calculating the SW distances between two measures involves two steps: (1) obtain a family of one-dimensional representations for these two measures through linear projections, and (2) compute the average of the Wasserstein distance between these one-dimensional representations. More formally, let $\mathbb{S}^{d-1}=\left\{\boldsymbol{u}\in\mathbb{R}^d:\left\|\boldsymbol{u}\right\|=1\right\}$ be the d-dimensional unit sphere, where $\left\|\cdot\right\|$ represents the Euclidean norm, and $\langle\cdot,\cdot\rangle$ represents the Euclidean inner-product. For any $\boldsymbol{u}\in\mathbb{S}^{d-1}$, let $\boldsymbol{u}^*$ be the linear form w.r.t. $\boldsymbol{u}$, such that for $\boldsymbol{a}\in\mathbb{R}^d$, $\boldsymbol{u}^*(\boldsymbol{a})=\langle\boldsymbol{u},\boldsymbol{a}\rangle$. For any $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$, the SW distance of order 2 between them is defined as

$$\mathrm{SW}_2(\mu,\nu):=\left(\int_{\mathbb{S}^{d-1}}W_2^2\left(\boldsymbol{u}^*_{\#}\mu,\boldsymbol{u}^*_{\#}\nu\right)\mathrm{d}\delta(\boldsymbol{u})\right)^{1/2}, \tag{10}$$

where δ represents the uniform distribution on $\mathbb{S}^{d-1}$.

In practice, the integration in Equation (10) can be approximated using a Monte Carlo scheme. That is, one can randomly and uniformly draw a finite set of projection directions from $\mathbb{S}^{d-1}$, and replace the integral with a finite-sample average (Bonneel et al., 2015; Rabin et al., 2011). Algorithm 1 summarizes the details for approximating SW distances. Note that the for loop in this algorithm can be naturally paralleled.

Figure 2 provides a toy example for Algorithm 1. Two Gaussian distributions, μ and ν, are marked in blue and orange, respectively. Considering two projection directions, that is, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. The projected distributions of μ and ν w.r.t. these two projection directions are shown as blue and yellow curves, respectively. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote the Wasserstein distance between the projected distributions along $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, respectively. The SW distance between μ and ν thus can be approximated by $\left(\left(\mathcal{D}_1^2+\mathcal{D}_2^2\right)/2\right)^{1/2}$.

Algorithm 1 indicates that the empirical SW distance can be efficiently calculated in practice, as it utilizes the closed-form expression of the Wasserstein distance between one-dimensional measures. Indeed, the computational cost of Algorithm 1 is just $O(Ldn+Ln\log(n))$, where the number of projections L is usually set to be a constant (Bonneel et al., 2015; Deshpande et al., 2018).

ALGORITHM 1 **Sliced-Wasserstein distance estimation**

Input: $\{\boldsymbol{x}_i \sim \mu\}_{i=1}^n$, $\{\boldsymbol{y}_i \sim \nu\}_{i=1}^n$, number of slices L

Initialize $\mathcal{D} \leftarrow 0$

for $l = 1 : L$ **do**

(i) Generate a random vector $\boldsymbol{u}_l$ from $\mathbb{S}^{d-1}$

(ii) Compute $\widehat{\boldsymbol{x}}_i = \langle \boldsymbol{u}_l, \boldsymbol{x}_i \rangle$ and $\widehat{\boldsymbol{y}}_i = \langle \boldsymbol{u}_l, \boldsymbol{y}_i \rangle$ for $i = 1, \ldots, n$

(iii) Sort $\{\widehat{\boldsymbol{x}}_i\}_{i=1}^n$ and $\{\widehat{\boldsymbol{y}}_i\}_{i=1}^n$ in ascending order, denoted by $\{\widehat{\boldsymbol{x}}_{[i]}\}_{i=1}^n$ and $\left\{\widehat{\boldsymbol{y}}_{[i]}\right\}_{i=1}^n$, respectively

(iv) $\mathcal{D} \leftarrow \mathcal{D} + \sum_{i=1}^n \left|\widehat{\boldsymbol{x}}_{[i]} - \widehat{\boldsymbol{y}}_{[i]}\right|^2 / L$

end for

Output: $\mathcal{D}^{1/2}$

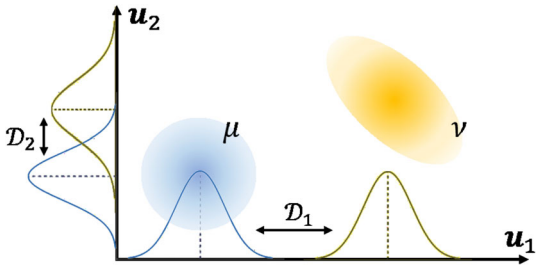

FIGURE 2 An illustration for approximating the SW distance

Besides the computational benefits, recent studies show the SW distance enjoys several elegant theoretical properties. Bonnotte (2013) showed that SW is a proper metric, and the convergence w.r.t. the SW distance implies weak convergence in compact domains. Deshpande et al. (2018) observed that the empirical SW distance of order two decreases roughly at the order $O(n^{-1/2})$ even for high-dimensional samples. Such an observation is supported by the findings in Bernton et al. (2019) and Nadjahi et al. (2019). In particular, Nadjahi et al. (2019) characterized the asymptotic distribution of the SW distances by proving a central limit theorem and establishing a convergence rate of $n^{-1/2}$. We refer to Nadjahi et al. (2020) for more theoretical analysis of SW distances. The aforementioned theoretical findings indicate that SW distance tends to be able to bypass the problem of curse-of-dimensionality. As a result, the SW distance has been an increasingly popular alternative to the Wasserstein distance; see Carriere et al. (2017), Deshpande et al. (2018), Kolouri et al. (2018), Liutkus et al. (2019); Wu et al. (2019), and Rowland et al. (2019) for some of the applications.

3.2 | Variants of SW distance

Despite the wide application, the SW distance has two major limitations. First, as the dimension d grows, the Monte Carlo procedure in Algorithm 1 requires a larger number of projections L to achieve a decent approximation of the integration. Empirically, many studies observed that if a reasonably smooth two-dimensional distribution can be approximated using L projections, then $O(L^{d-1})$ projections are required to approximate a similarly smooth d-dimensional distribution for $d \geq 2$ (Deshpande et al., 2019; Kolouri et al., 2019). Second, Algorithm 1 only focuses on linear projections, which may be ineffective for the data that lie within a manifold. Nonlinear projections or manifold learning techniques are more preferred in such cases.

To overcome such limitations, many studies have been developed to formulate novel OT metrics by generalizing the SW distance. One of such studies proposed the max-SW distance, which aims to remedy the first limitation (Deshpande et al., 2019). The idea is that, instead of using all the random projection directions generated from $\mathbb{S}^{d-1}$, one can simply pick the "best direction," along which the projected distance is maximized.

Figure 3 illustrates such a best direction using a toy example. The two-dimensional source sample and the target sample are labeled by blue dots and yellow dots, respectively. Recall that the standard slicing approach first generates a

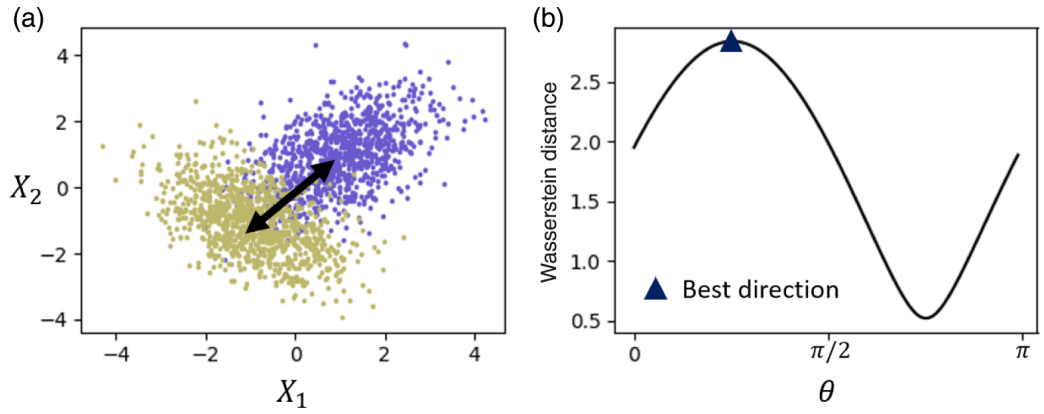

FIGURE 3 An illustration of the "best direction" used in max-SW distances. The blue dots and yellow dots in the scatter-plot represent the source and target samples, respectively. The *x*-axis and the *y*-axis of the line plots represent the projection direction θ and the corresponding Wasserstein distance between projected samples, respectively. The "best direction" is labeled as the black arrow in the scatter-plot and is marked by the triangle in the line plot

series of one-dimensional random projections, then calculates the Wasserstein distance between the projected samples w.r.t. each of the projections. Let $\theta \in [0, \pi)$ be the projection angles. The aforementioned one-dimensional Wasserstein distance can be regarded as a function of θ, denoted by $W(\theta)$, as illustrated in the right panel of Figure 3. The "best direction" defined in Deshpande et al. (2019) is simply the θ that yields the largest value of $W(\theta)$, denoted as the triangle.

More formally, following the notations in Equation (10), the max-SW distance of order 2 is defined as

$$\text{max-SW}_2(\mu,\nu) \coloneqq \max_{\boldsymbol{u} \in \mathbb{S}^{d-1}} W_2\left(\boldsymbol{u}^*_{\#}\mu, \boldsymbol{u}^*_{\#}\nu\right). \tag{11}$$

Deshpande et al. (2019) proved that the max-SW distance is a real metric and it satisfies the inequality

$$\text{SW}_2(\mu,\nu) \leq \text{max-SW}_2(\mu,\nu) \leq W_2(\mu,\nu). \tag{12}$$

In practice, however, finding the best direction $\boldsymbol{u}$ in Equation (11) is not trivial, as such a maximization problem may involve a large number of local maximums (Lin et al., 2021; Lin, Fan, et al., 2020). The authors in Deshpande et al. (2019) proposed an ad-hoc strategy to approximate this $\boldsymbol{u}$ by replacing the Wasserstein distance in the right-hand side of Equation (11) by the difference between the means of $\boldsymbol{u}^*_{\#}\mu$ and $\boldsymbol{u}^*_{\#}\nu$, respectively. In other words, they proposed to utilize the projection direction that resulted in the largest difference between the means of the projected samples. Such a direction is natural to find.

Recall that to capture the discrepancy between two measures, the SW distance essentially requires considering all possible projection directions from the unit sphere $\mathbb{S}^{d-1}$. The max-SW distance, however, considers only one of them to increase the efficiency. Despite the effectiveness, the idea of max-SW may be too greedy in some cases, as it ignores most of the projection directions from $\mathbb{S}^{d-1}$. One natural question is whether we can capture the major discrepancy between two measures by considering a relatively small number of "important" slices. Following this line of thinking, recently, Nguyen et al. (2020) proposed the distributional sliced-Wasserstein (DSW) distance. The goal is to construct a more effective sampling probability P (w.r.t. the projection directions) over the unit sphere $\mathbb{S}^{d-1}$ such that P represents how important each projection direction is. Both SW distances and the max-SW distances can be regarded as special cases of DSW distances, as SW distances take P as the uniform distribution and max-SW distances take P as the delta-Dirac distribution. Nguyen et al. (2020) proposed a relatively efficient algorithm to calculate DSW distances. In addition, they proved that, under some mild conditions, DSW distances satisfy

$$d^{-1/2}\text{max-SW}_2(\mu,\nu) \leq \text{DSW}_2(\mu,\nu) \leq \text{max-SW}_2(\mu,\nu).$$

We now consider the second limitation of SW distances that they only consider linear projections. Kolouri et al. (2019) extended the idea of the SW distances by replacing the linear projections with nonlinear ones. In particular, instead of projecting the data on a straight line, they proposed to project the data on a one-dimensional manifold using a particular nonlinear function g. The authors proposed the so-called generalized sliced-Wasserstein (GSW) distance and showed that it is indeed a well-defined metric for certain nonlinear mapping g's. Furthermore, the idea of GSW distances can be combined with the idea of max-SW distances to construct max-GSW distances. In practice, such max-GSW distances are calculated using an EM-like optimization scheme, that is, optimizing the nonlinear function g and the projection direction θ iteratively. Kolouri et al. (2019) showed the max-GSW distance could be successfully implemented in generative models and other applications. The idea of nonlinear projection is also considered in Nguyen et al. (2020) to generalize the DSW distance.

The slicing approach is also considered in a variant of the Wasserstein distance, called the tree-Wasserstein distance (Do Ba et al., 2011; Evans & Matsen, 2012; McGregor & Stubbs, 2013), which utilized the so-called tree metrics instead of the Euclidean metric in OT problems. Le et al. (2019) proposed the tree-SW distance, computed by averaging the Wasserstein distance w.r.t. random tree metrics.

4 | ITERATIVE PROJECTION APPROACH

SW distance, as a proper metric, has been extensively studied and is known to be able to overcome the problem of curse-of-dimensionality. As a result, the SW distance and its variants have been regularly used as alternatives to Wasserstein distances in many real-world applications.

Nevertheless, almost none of the existing SW-based methods can be utilized to estimate the true Wasserstein distance between the two measures. In fact, the SW distance and its variants may severely underestimate the true Wasserstein distance. Consider the case that both μ and ν are d-dimensional Gaussian distribution, that is, $\mu = N(\mathbf{m}_\mu, \Sigma_\mu)$ and $\nu = N(\mathbf{m}_\nu, \Sigma_\nu)$. It is well-known that the Wasserstein distance (of order two) between two Gaussian distributions admits a closed-form,

$$W_2^2(\mu,\nu) = \|\mathbf{m}_\mu - \mathbf{m}_\nu\|_2^2 + \text{trace}\left(\Sigma_\mu + \Sigma_\nu - 2\left(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2}\right)^{1/2}\right). \tag{13}$$

Let $\mathbf{m}_\mu = \mathbf{m}_\nu = \mathbf{0}$, $\Sigma_\mu = c \cdot \mathbf{I}_d$, and $\Sigma_\nu = \mathbf{I}_d$. Here, $\mathbf{I}_d$ is the identity matrix and $c > 0$ is a constant. In such cases, Equation (13) indicates that $W_2^2(\mu,\nu) = d(\sqrt{c}-1)^2$. We now consider the SW distance and the max-SW distance between such μ and ν. Due to the symmetry, the Wasserstein distance between the projected μ and ν w.r.t. any projection direction remains a constant. Specifically, it is easy to show that $\text{SW}_2^2(\mu,\nu) = \text{max-SW}_2^2(\mu,\nu) = (\sqrt{c}-1)^2$ in such cases. As a result, the difference between the true Wasserstein distance and the SW distance, including its variants, can be arbitrarily large as the number of dimensions d increases.

In this section, we introduce the iterative random projection approach, which can be utilized to estimate the true Wasserstein distance. Recall that the one-dimensional OT components in Algorithm 1 are independent of each other, and thus the computation of SW distances can be naturally paralleled. In the iterative projection approach, however, these components are dependent on each other and are added to the estimation process sequentially. As a byproduct, such an approach also provides an empirical OTM, which may not be obtainable using the slicing approach. The empirical OT map is of particular interest in many real-world applications, for example, color transfer and domain adaptation (Courty et al., 2016, 2017; Seguy et al., 2018). We now present the idea of the iterative random projection method, followed by its extension.

4.1 | Iterative random projection method

The iterative random projection method, also called the Radon probability density function (PDF) transformation method, was first proposed in Pitie et al. (2005) for the application of color transfer. The input of such an application is two images, each of which can be regarded as a three-dimensional sample in the RGB color space, and each pixel of the image is an observation. The goal of color transfer is to find a transport map such that the color distribution of the transformed source image follows the same color distribution of the target image. Although such a transport map does

not have to be the optimal one w.r.t. the transport cost, their proposed algorithm can be regarded as an efficient estimation method for OTM. Once the empirical OTM has been calculated, the Wasserstein distance can be easily estimated using Equation (5). Algorithm 2 summarizes the details of the iterative random projection method.

The sequential scheme adopted in Algorithm 2 is beneficial for the estimation of the Wasserstein distance and the OTM. To see this, consider a synthetic example shown in Figure 4. The goal is to find the OTM that maps the source sample to the target sample, which are marked in blue and yellow, respectively. Intuitively, if two orthogonal projection directions are used in Algorithm 2, the algorithm may converge in two steps and outputs an effective estimate of the OTM as well as the Wasserstein distance.

Instead of randomly generating the projection direction $\boldsymbol{u}'$s following the Monte Carlo scheme, one can also generate a sequence of $\boldsymbol{u}'$s with "low-discrepancy," that is, the directions that are distributed as disperse as possible on the unit sphere. The low-discrepancy sequence has been widely applied in the field of quasi-Monte Carlo and has been extensively employed for numerical integration (Owen, 2003) and subsampling in big data (T. Li & Meng, 2021; Meng et al., 2022; Meng, Xie, et al., 2020; Meng, Zhang, et al., 2020) see Lemieux (2009), Dick et al. (2013), and Glasserman (2013); Leobacher and Pillichshammer (2014) for more in-depth discussions. It is shown in Pitie et al. (2005) that using a low-discrepancy sequence of projection directions results in a potentially faster convergence rate.

4.2 | Projection pursuit Monge map

Despite the effectiveness, in practice, the iterative random projection method may face the same obstacle as the slicing approach does. In particular, for a moderate or large d, a considerable number of projection directions are required to

ALGORITHM 2 **Iterative random projection method**

Input: the source sample $\mathbf{X} \in \mathbb{R}^{n\times d}$ and the target sample $\mathbf{Y} \in \mathbb{R}^{n\times d}$

Initialize $i \leftarrow 0$, $\mathbf{X}^{[0]} \leftarrow \mathbf{X}$

repeat

(i) generate a random projection direction $\boldsymbol{u}_i \in \mathbb{S}^{d-1}$

(ii) find the one-dimensional OTM $\phi^{(i)}$ that matches $\mathbf{X}^{[i]}\boldsymbol{u}_i$ to $\mathbf{Y}\boldsymbol{u}_i$

(iii) $\mathbf{X}^{[i+1]} \leftarrow \mathbf{X}^{[i]} + \left(\phi^{(i)}\left(\mathbf{X}^{[i]}\boldsymbol{u}_i\right) - \mathbf{X}^{[i]}\boldsymbol{u}_i\right)\boldsymbol{u}_i^{\mathrm{T}}$

(iv) $i \leftarrow i+1$

until converge.

The final map is given by $\widehat{\phi}: \mathbf{X} \to \mathbf{X}^{[i]}$

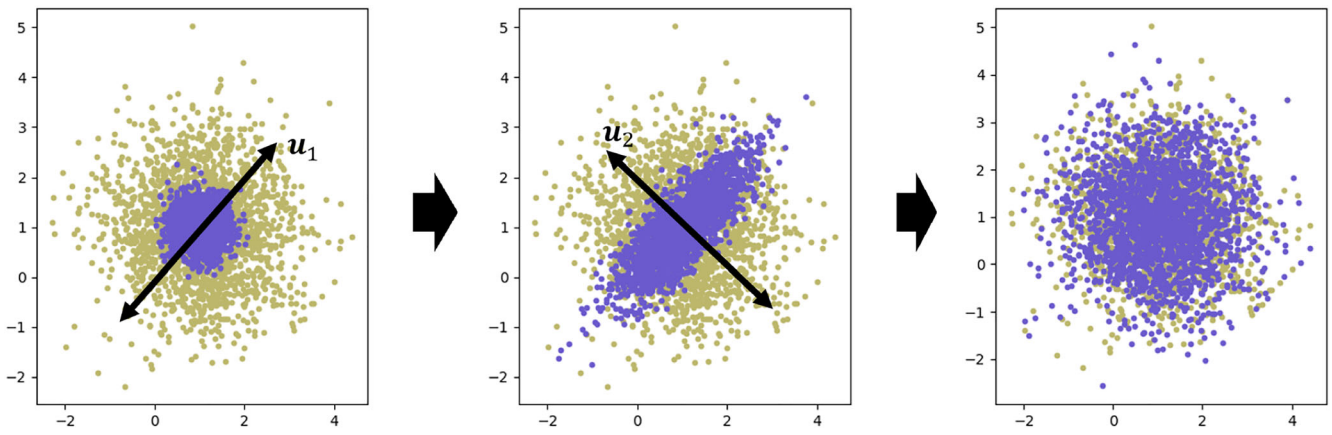

FIGURE 4 Illustration of the iteration projection method. Blue dots and yellow dots represent the source sample and the target sample, respectively. A projection direction $\boldsymbol{u}_1$ is used in the first iteration and is shown in the left panel. The middle panel shows the transformed source sample after the first iteration. We then consider a projection direction $\boldsymbol{u}_2$ that is orthogonal to $\boldsymbol{u}_1$. After two iteration steps, two samples roughly have same distribution, as shown in the right panel

explore the unit sphere $\mathbb{S}^{d-1}$, while most of them may contain only tiny information and contribute only a little to the final estimation. As a result, Algorithm 2 usually suffers from slow convergence in practice.

To overcome the limitation, Meng et al. (2019) proposed a variant of the iterative random projection method, named projection pursuit Monge map (PPMM).[1] The PPMM method combines the idea of projection pursuit regression and sufficient dimension reduction (SDR). In each iteration, PPMM seeks the "most informative" direction instead of using a randomly selected one.

At first glance, such an idea may seem similar to the idea behind the max-SW distance. Recall that the best direction defined in the max-SW distance is the one that yields the largest value of the Wasserstein distance between the projected samples. One limitation of such a definition is that the desired best direction does not admit a closed-form, and thus optimization techniques are required to find such a direction. The PPMM method, however, takes a different path to achieve the goal. In particular, PPMM provides a definition for the most informative direction, such that it admits a closed-form.

The idea behind PPMM is inspired by the SDR approach. Consider a regression problem with a univariate response T and a d-dimensional predictor Z. SDR techniques aim to reduce the dimension of Z while preserving its regression relation with T. In other words, such techniques seek a set of linear combinations of Z, say $\mathbf{B}^{\mathrm{T}}Z$ with a projection matrix $\mathbf{B} \in \mathbb{R}^{d\times k}$ $(k < d)$, such that T depends on Z only through $\mathbf{B}^{\mathrm{T}}Z$, that is,

$$T \perp\!\!\!\perp Z \mid \mathbf{B}^{\mathrm{T}}Z. \tag{14}$$

Numerous methods have developed to estimate such projection matrix $\mathbf{B}'$s, includes sliced inverse regression (SIR) (K.-C. Li, 1991), principal Hessian directions (pHd) (K.-C. Li, 1992), sliced average variance estimator (SAVE) (Cook & Weisberg, 1991), directional regression (DR) (B. Li & Wang, 2007); see B. Li (2018) for a detailed review.

Consider the problem of estimating the Wasserstein distance between a source sample and a target sample. Let Z be both samples, serving as the predictors in the regression problem. Let T be a constructed binary response variable, labeled as 0 and 1 for two samples, respectively. Under such a scenario, loosely speaking, Equation (14) indicates that, conditional on the projected variables, the probability $P(Z|T=1)$ and $P(Z|T=0)$ have the same distribution. In other words, the subspace spanned by the column vectors of $\mathbf{B}$ contains all the information of the Wasserstein distance as well as the OTP between the observed samples. As a result, such a subspace can be regarded as an "informative subspace" for the OT problem. Furthermore, intuitively, the eigenvector w.r.t. the largest eigenvalue of the projection matrix $\mathbf{B}$ can be regarded as the most informative one-dimensional projection direction.

Following this line of thinking, Meng et al. (2019) defined the most informative direction as the eigenvector w.r.t. the largest eigenvalue of $\mathbf{B}$. In practice, such $\mathbf{B}'$s can be easily estimated by existing SDR techniques. It is recommended to use the SAVE method or the DR method, as both methods utilize the first moment and the second-moment information of the observed samples to provide a closed-form estimator for $\mathbf{B}$. Let $\widehat{\mathbf{B}}$ denote the estimated projection matrix and $\boldsymbol{u}^{[1]}$ denote its eigen-vector w.r.t. the largest eigenvalue. The PPMM algorithm is very much alike Algorithm 2, except that a random direction $\boldsymbol{u}_i$ in Step (i) is replaced by the determined direction $\boldsymbol{u}_i^{[1]}/\left\|\boldsymbol{u}_i^{[1]}\right\|$, in the i-th iteration. We refer to Meng et al. (2019) for further details.

4.3 | Empirical comparison between PPMM and max-SW

Loosely speaking, the projection direction $\boldsymbol{u}^{[1]}$ determined by the PPMM method indicates that, among all possible projection directions, the projected samples respecting $\boldsymbol{u}^{[1]}$ yield the largest discrepancy between each other. Such a discrepancy reflects both the difference w.r.t. the means and the variance. Intuitively, such projected samples also yield a relatively large Wasserstein distance. One natural question remains, what is the difference between the direction determined by the PPMM method and the one calculated in the max-SW distance? Although theoretical analysis of such a question is lacking, we provide the following synthetic examples to compare these two projection directions in different settings.

Four synthetic datasets are shown in Figure 5. In each panel, the source sample and the target sample are marked by yellow and blue points, respectively. Four different settings for data generation are considered:

- In Figure 5a, the samples are generated from two Gaussian distributions with the same variance–covariance matrix and different means, respectively;

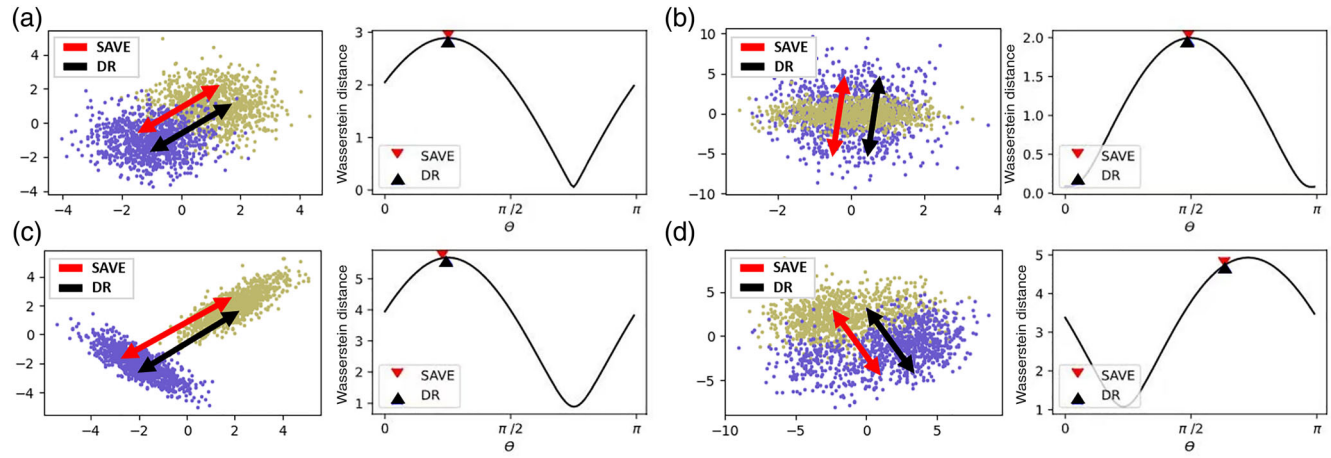

FIGURE 5 Comparison between the direction determined by the PPMM method and the one used in the max-SW distance. The directions determined by the first iteration of the PPMM algorithm, implemented with SAVE and DR, are labeled as red and black directions in the scatter plots, respectively. The *x*-axis and the *y*-axis of the line plots represent the projection direction θ and the corresponding Wasserstein distance between projected samples, respectively. In the line plots, the directions determined by the first iteration of the PPMM algorithm are denoted by triangles. Note that the "best direction" used in max-SW distances is not shown. However, by definition, such a direction is the one that yields the maximum value of the *y*-axis in line plots. We observe that in panels (a)–(c), the directions determined by PPMM are almost the same as the best direction defined in max-SW. While in panel (d), these two methods yield slightly different projection directions

- In Figure 5b, the samples are generated from two Gaussian distributions with the same mean and different variance–covariance matrices, respectively;
- In Figure 5c, the samples are generated from two Gaussian distributions with different means and different variance–covariance matrices, respectively;
- In Figure 5d, the samples are generated from two different mixture-Gaussian distributions, respectively.

In each panel, the projection directions determined by PPMM are labeled as colored directions in the scatter plots. In particular, the red ones represent the cases where the SAVE algorithm is implemented in PPMM, while for the blue ones, the DR algorithm is implemented. We observe that both methods provide projection directions that are similar to each other. Recall in Figure 3 that for each synthetic dataset, we can draw a line plot showing the projection direction θ versus the corresponding Wasserstein distance $W(\theta)$. In addition, the "best direction" defined in max-SW is the θ that yields the largest value of $W(\theta)$. In each of the line plots, the projection directions determined by PPMM are denoted by colored triangles. We observe that in panels (a)-(c), the directions selected PPMM are almost the same as the best direction defined in the max-SW distance, as the colored triangles approximately achieve the maximum value of $W(\theta)$. In panel (d), we observe that these two methods yield slightly different projection directions.

Figure 5 indicates it is possible that the PPMM method and the max-SW method pick the same projection direction in some situations. While in general cases, these two methods select different ones. Further studies are needed to quantify the difference between the selected directions w.r.t. these two methods, respectively. Recall that one major advantage of PPMM over max-SW is that the former admits a closed-form of the desired projection direction. In practice, it is not trivial to find the best direction for the max-SW distance, as such a maximization problem may involve a large number of local maximums. In contrast, PPMM provides a user-friendly iterative algorithm, and it takes $O\left(d^2 n\log(n)\right)$ time for each iteration. Empirical results show that PPMM works well when the number of iterations is in the same order of d, in which case the computational cost of the PPMM algorithm is of the order $O\left(d^3 n\log(n)\right)$ (Meng et al., 2019). The PPMM method thus has the potential to construct a computationally efficient alternative for max-SW distance in real-life applications, especially when $d \ll n$.

4.4 | Robustness of PPMM

One major concern of the PPMM approach is that both SAVE and DR methods, which PPMM highly relies on, require relatively strong moment conditions on the data (B. Li, 2018). Fortunately, PPMM is an iterative algorithm, thus, the

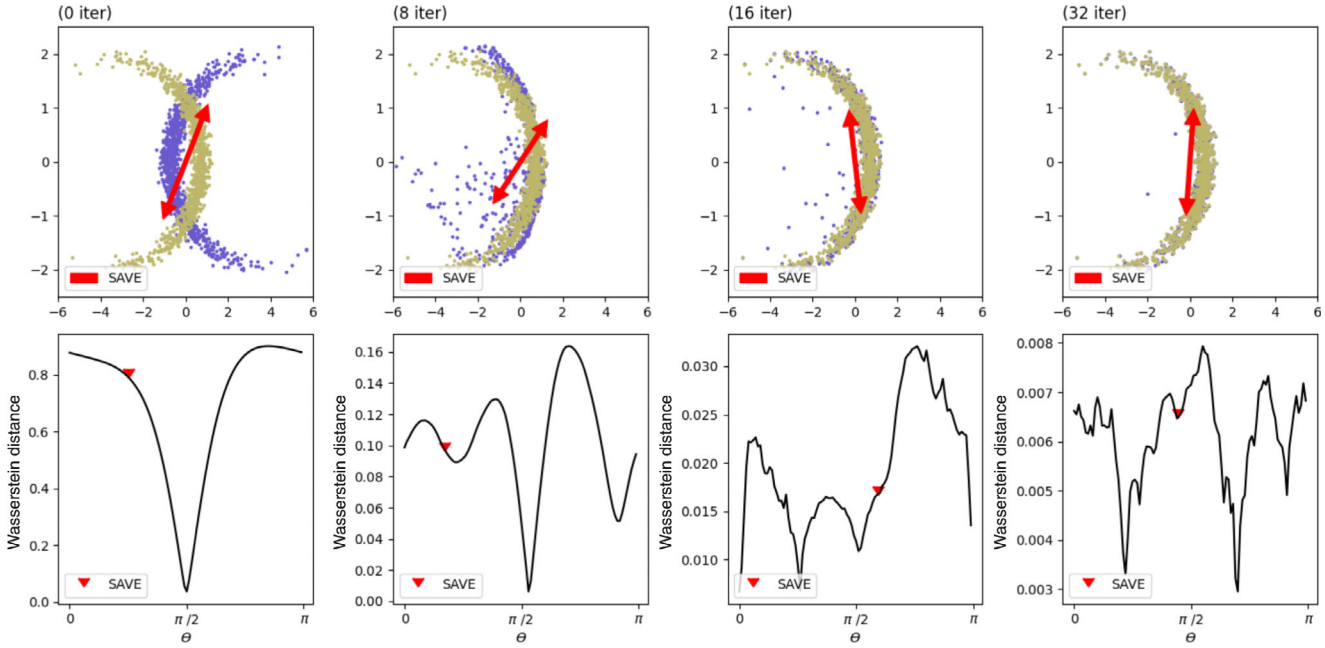

FIGURE 6 Visualization of the PPMM method. The directions determined by PPMM during the iterations are labeled as red directions in the scatter plots. These directions are denoted by triangles in the line plots. We observe that PPMM is able to provide a robust estimation of the Monge map even when the moment conditions of the dataset are dramatically violated

moment conditions may change dramatically in each iteration since the OT step in that iteration will modify the empirical distribution of the source sample. This is to say, the violation of the moment conditions may be weakened during the iterations of the PPMM algorithm. We consider a synthetic dataset to evaluate the robustness of PPMM. The two-dimensional source and target samples in this synthetic dataset are generated from two C-shaped trigonometric curves, respectively. For each sample, we first generate 300 data points and then standardize the sample such that its empirical mean equals zero and its empirical variance–covariance matrix equals the identity matrix. The synthetic dataset is visualized in Figure 6, where the source sample and the target sample are marked by yellow and blue points, respectively. The projection directions determined by PPMM[2] during the iterations are labeled as red directions in the scatter plots. These directions are also denoted by red triangles in the line plots. It is clear that such an example violates the moment conditions dramatically, and it is impossible for PPMM to find the most informative projection direction in the first iteration. However, after several iterations, we observe that the distribution of the source sample changes and the source sample fits the target sample reasonably well when the PPMM algorithm converges. Such a result indicates that PPMM is quite robust even when the moment conditions are dramatically violated.

5 | PROJECTION ROBUST OT METHODS

Encouraged by the success of max-SW, Paty and Cuturi (2019) asked whether one can gain more by seeking a "best subspace" of dimension $k \geq 2$ rather than the "best direction" considered in max-SW. In particular, Paty and Cuturi (2019) proposed to look for the k-dimensional subspace that would maximize the Wasserstein distance between two measures after projection and defined such maximum quantity as the projection robust Wasserstein (PRW) distance. Note that PRW distance is equivalent to max-SW distance when $k = 1$. Intuitively, the PRW distance could be regarded as a metric learning technique, such that the goal is to learn a better metric to quantify the cost matrix when estimating the Wasserstein distance. As a result, PRW shows an advantage over the max-SW distance when a k-dimensional ($k \geq 2$) squared Euclidean cost results in a more accurate estimation of the Wasserstein distance than the 1-dimensional counterpart.

Figure 7 illustrates the idea of the best subspace. The three-dimensional source sample and the target sample are labeled by blue dots and yellow dots, respectively. The samples share the same marginal distributions in all three dimensions with different correlations between the first two dimensions, as shown in Figure 7a. Note that a two-

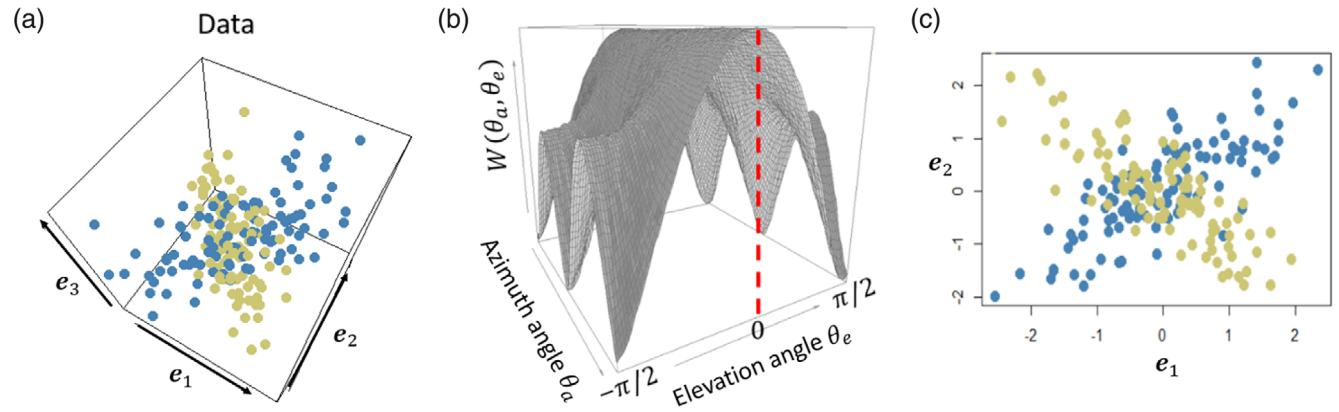

FIGURE 7 Illustration of the "best subspace." The blue dots and yellow dots in (a) represent the source and target samples, respectively. The x-axis and the y-axis of (b) represent the azimuth angle θ_a and the elevation angle θ_e of the subspace, respectively. The z-axis of (b) represents the corresponding Wasserstein distance between projected samples. The "best subspace" is associated with $\theta_e = 0$ (marked as the red dashed line) that yields the largest value of $W(\theta_a, \theta_e)$ in (b). The projected samples w.r.t. such a best subspace is shown in (c)

dimensional subspace is determined by the azimuth angle and the elevation angle. Let $\theta_a \in [-\pi/2, \pi/2)$ be the azimuth angle and $\theta_e \in [-\pi/2, \pi/2)$ be the elevation angle. Analogous to the example in Figure 3, the Wasserstein distance between the projected samples w.r.t. a two-dimensional subspace can be regarded as a function of θ_a and θ_e. We denote $W(\theta_a, \theta_e)$ to be such a function and plot it in Figure 7b. Figure 7b indicates that the pair (θ_a, θ_e) w.r.t. the best subspace is the one such that $\theta_e = 0$. As a result, the best subspace is the one that is spanned by e_1 and e_2. The projected samples w.r.t. such a best subspace is shown in Figure 7c.

The problem of finding the best subspace in OT problems is also considered in Alvarez-Melis et al. (2018); Dhouib et al. (2020); Muzellec and Cuturi (2019); Paty and Cuturi (2019). We introduce the detail of the PRW distance and its variants in this section.

5.1 | Subspace robust Wasserstein distance

We first provide some essential notations. For $k \in \{1, \ldots, d-1\}$, the Grassmannian of k-dimensional subspaces of $\mathbb{R}^d$ is defined as $\mathcal{G}_k = \{\mathcal{E} \in \mathbb{R}^d | \dim(\mathcal{E}) = k\}$. For $\mathcal{E} \in \mathcal{G}_k$, let $P_{\mathcal{E}}$ be the orthogonal projector onto the subspace $\mathcal{E}$. Analogous to the definition of max-SW distance in Equation (11), the definition of PRW distance is a natural extension of max-SW distance that considers the worst possible OT cost over all possible k dimensional projections. More formally, given two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the PRW distance of order two between μ and ν is defined as

$$\mathrm{PRW}_k(\mu, \nu) := \sup_{\mathcal{E} \in \mathcal{G}_k} W(P_{\mathcal{E}\#}\mu, P_{\mathcal{E}\#}\nu). \tag{15}$$

The "min-max" problem in Equation (15) is non-convex and falls back on a convex relaxation; thus is well-defined yet hard to solve. To overcome the obstacle, Paty and Cuturi (2019) considered a "max-min" variant of PRW distances, named subspace robust Wasserstein (SRW) distance. Intuitively, SRW distances can be regarded as a convex relaxation of PRW distances and thus tend to be easier to compute. Specifically, the SRW distance of order two is defined as

$$\mathrm{SRW}_k(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \sup_{\mathcal{E} \in \mathcal{G}_k} \left[\int \left\| P_{\mathcal{E}}(x - y)^2 \right\| \mathrm{d}\pi(x, y) \right]^{1/2}. \tag{16}$$

Paty and Cuturi (2019) showed that both PRW and SRW are indeed "distances" over $\mathcal{P}_2(\mathbb{R}^d)$ and it can be proved that $\mathrm{PRW}_k(\mu, \nu) \leq \mathrm{SRW}_k(\mu, \nu)$ for a fixed k.

Let $\Omega \in \mathbb{R}^{d \times d}$ be a symmetric positive semi-definite matrix. For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$, we write $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive semi-definite. Considering the Mahalanobis distance d^2_Ω such that

$$d_{\Omega}^2(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\Omega(\boldsymbol{x}-\boldsymbol{y}).$$

Paty and Cuturi (2019) proved that both the infimum and the supremum are achieve-able in Equation (16), and such SRW distances can be written as a min-max problem w.r.t. to d_{Ω^2}, that is,

$$\mathrm{SRW}_k(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{\mathbf{0}\preccurlyeq\Omega\preccurlyeq\mathbf{I}_d \\ \mathrm{trace}(\Omega)=k}} \left[\int d_{\Omega}^2 \mathrm{d}\pi\right]^{1/2}. \tag{17}$$

Equation (17) indicates that the problem of calculating the SRW distance can be regarded as seeking the best distance metric d_{Ω}.

Furthermore, Paty and Cuturi (2019) showed that SRW distances could be elegantly reformulated as a function of the eigendecomposition of the second-order displacement matrix $\mathbf{V}_{\pi}$. Here, the matrix $\mathbf{V}_{\pi}$ is defined as

$$\mathbf{V}_{\pi} := \int (\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}} \mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}). \tag{18}$$

In particular, considering the case when the OTM ϕ^* between two equally weighted samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ exists. In such cases, the vectors $\{\boldsymbol{x}_i-\phi^*(\boldsymbol{x}_i)\}_{i=1}^n$ are usually called the displacement vectors, and the matrix $\mathbf{V}_{\pi}$ is simply the second-order moment of all the displacements, that is, $\sum_{i=1}^n (\mathbf{x}-\phi^*(\boldsymbol{x}))(\boldsymbol{x}-\phi^*(\boldsymbol{x}))^{\mathrm{T}}/n$. Comparing Equation (18) with the definition of the Wasserstein distance in Equation (6), it is easy to notice that computing the Wasserstein distance of order two can be simply interpreted as minimizing the trace of $\mathbf{V}_{\pi}$ w.r.t. all possible OTP π's. In other words, let λ_i denote the i-th largest eigenvalue of $\mathbf{V}_{\pi}$, $i=1,...,d$, one thus has

$$W^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^{d} \lambda_i(\mathbf{V}_{\pi}).$$

Following this line of thinking, Paty and Cuturi (2019) noted that SRW distance admits a simple formulation,

$$\mathrm{SRW}_k^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^{k} \lambda_i(\mathbf{V}_{\pi}). \tag{19}$$

Motivated by Equations (17) and (19), Paty and Cuturi (2019) proposed an iterative algorithm to compute SRW distances in practice. Intuitively, the idea is that, for a given k, one can iteratively optimize between the distance metric d_{Ω} and the OTP π. More specifically, given a distance metric, one can utilize standard OT methods, for example, the Sinkhorn algorithm, to calculate the OTP π. Next, one can extract the largest k eigenvectors of such an OTP π to form a new distance metric. One then can utilize such a distance metric to calculate a new OTP, and so on so forth. Such an algorithm is called the Frank-Wolfe algorithm for regularized SRW and is summarized in Algorithm 3. We refer to Paty and Cuturi (2019) for more details about the stopping criterion of Algorithm 3.

Empirical studies show Algorithm 3 can successfully recover the informative subspace of the data and is more robust to the noise compared to the classical Wasserstein distance. For the choice of k, the authors showed SRW works well for a relatively broad range of k (Paty & Cuturi, 2019). In practice, it is recommended to select k that results in the "elbow" in the scree plot, just like how people choose the number of principal components in the principal component analysis.

5.2 | Projection robust Wasserstein distance

Lin, Fan, et al. (2020) noted that as an alternative to PRW distances, SRW distances might result in suboptimal performance in practice. To overcome the limitation, they revisited the original PRW problem (15) and showed that, despite

ALGORITHM 3 **Frank–Wolfe algorithm for regularized SRW**

Input: Empirical measures μ and ν, dimension k, regularization parameter $\eta > 0$

Initialize $\pi \leftarrow \text{reg\_OT}\left[(\mu,\nu), \text{reg} = \eta, \text{cost} = \|\cdot\|^2\right]$

$\mathbf{U}_k \in \mathbb{R}^{d\times k} \leftarrow$ top k eigenvecotrs of $\mathbf{V}_\pi$

$\Omega \leftarrow \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}$

$t \leftarrow 0$

repeat

$\pi \leftarrow \text{reg\_OT}\left[(\mu,\nu), \text{reg} = \eta, \text{cost} = d_\Omega^2\right]$

$\mathbf{U}_k \in \mathbb{R}^{d\times k} \leftarrow$ top k eigenvecotrs of $\mathbf{V}_\pi$

$\widehat{\Omega} \leftarrow \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}$

$\tau = 2/(2+t)$

$\Omega = (1-\tau)\Omega + \tau\widehat{\Omega}$

$t \leftarrow t+1$

until converge

Output: Ω, π

the hardness, such a problem could be efficiently computed in practice using Riemannian optimization. The authors proposed three practical algorithms for computing PRW distances and showed that their methods yielded better behavior than SRW distances.

Recently, Lin et al. (2021) extended the definition of the PRW distance to the so-called integral projection robust Wasserstein (IPRW) distance. The relationship between IPRW and PRW is very much similar to the relationship between SW and max-SW. In particular, instead of considering only the best k-dimensional subspace, as did in PRW, IPRW considered the averages of all possible k-dimensional subspaces. Statistical properties of both PRW distances and IPRW distances are analyzed in Lin et al. (2021). They showed that an empirical measure converges to its corresponding true measure w.r.t. PRW or IPRW distance roughly at the order of $O\left(n^{-1/k}\right)$, under certain conditions. Such results indicate that both PRW and IPRW tend to overcome the problem of curse-of-dimensionality faced by Wasserstein distances and thus are very likely to outperform Wasserstein distances in high-dimensional tasks.

6 | CONCLUSION AND FUTURE RESEARCH

In this article, we reviewed projection-based techniques in OT problems, including the slicing approach, the iterative projection approach, and the projection robust OT approach. Existing studies showed most of such methods have the potential to overcome the problem of curse-of-dimensionality. These approaches thus are widely used as an alternative for the classical Wasserstein distance and the OTPs in practice.

Most reported research efforts are exploring projection-based techniques for the classical OT problem and Wasserstein distances. In contrast, there is relatively little reported work in literature discussing such methods for other OT-related problems such as unbalanced OT (partial OT), Gromov Wasserstein distances, Wasserstein barycenter, multi-margin OT problems, among others (Peyré et al., 2019). A few exceptions are Bonneel et al. (2015) and Vayer et al. (2019), which showed that the slicing approach could be applied in constructing Wasserstein barycenters and accelerating the estimation of Gromov Wasserstein distances. More research efforts are required to develop effective and efficient projection-based algorithms for a broader scope of OT problems.

Another interesting research issue open for future investigation is to develop fast variable selection methods or screening methods for OT problems. Most projection-based methods reviewed in this article require the computational cost at least of the order $O(d^2)$ w.r.t. the dimension d. Such a sizable computational cost hinders the broad applicability of OT methods in ultra-high dimensional problems, for example, when $d \approx 10^6$ and $d \gg n$. Suppose there exist some screening methods that can exclude a large number of "irrelevant" variables using $O(d)$ computational time. It may be preferable to first apply such methods before using existing projection-based techniques. Such a two-step approach is analogous to the postselection inference, which aims at making inference conditional on the selection (Berk

et al., 2013; Lee et al., 2016; Liu et al., 2020; Taylor & Tibshirani, 2018; Tibshirani et al., 2016), and could be a possible solution for dealing with ultra-high dimensional OT problems. How to develop a valid screening method for OT problems thus is an essential topic for future investigation.

AUTHOR CONTRIBUTIONS

Jingyi Zhang: Investigation (equal); methodology (equal); software (equal); writing – original draft (equal); writing – review and editing (equal). **Ping Ma:** Funding acquisition (equal); supervision (equal). **Wenxuan Zhong:** Funding acquisition (equal); supervision (equal). **Cheng Meng:** Investigation (equal); methodology (equal); software (equal); writing – original draft (equal); writing – review and editing (equal).

FUNDING INFORMATION

This work is supported by the U.S. National Science Foundation under grants DMS-1903226, DMS-1925066, the U.S. National Institute of Health under grant R01GM122080, and the National Natural Science Foundation of China Grant No.12101606. We thank Tao Li from Renmin University of China for providing synthetic examples in Section 4.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Jingyi Zhang https://orcid.org/0000-0002-3147-8838
Ping Ma https://orcid.org/0000-0002-5728-3596
Cheng Meng https://orcid.org/0000-0002-7111-0966

ENDNOTES

[1] The code is available at https://github.com/ChengzijunAixiaoli/PPMM.

[2] The PPMM method is implemented by SAVE in this example. The PPMM implemented by DR provides similar results.

RELATED WIREs ARTICLES

Advance of the sufficient dimension reduction

REFERENCES

Adler, J., & Lunz, S. (2018). Banach Wasserstein GAN. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 6755–6764).

Alaux, J., Grave, E., Cuturi, M., & Joulin, A. (2018). Unsupervised hyper-alignment for multilingual word embeddings. *International Conference on Learning Representations*.

Altschuler, J., Bach, F., Rudi, A., & Niles-Weed, J. (2019). Massively scalable Sinkhorn distances via the Nyström. In *Advances in Neural Information Processing Systems* (pp. 4429–4439).

Altschuler, J., Weed, J., & Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems* (pp. 1964–1974).

Alvarez-Melis, D., Jaakkola, T., & Jegelka, S. (2018). Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics* (pp. 1771–1780).

An, D., Guo, Y., Lei, N., Luo, Z., Yau, S.-T., & Gu, X. (2020). AE-OT: A new generative model based on extended semi-discrete optimal transport. In *Eighth International Conference on Learning Representations (ICLR)*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning* (pp. 214–223).

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, *41*(2), 802–837.

Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, *8*(4), 657–676.

Bigot, J., Cazelles, E., & Papadakis, N. (2019). Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, *13*(2), 5120–5150.

Black, E., Yeom, S., & Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 111–121).

Bonneel, N., Rabin, J., Peyré, G., & Pfister, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, *51*(1), 22–45.

Bonnotte, N. (2013). *Unidimensional and evolution methods for optimal transportation* [Unpublished doctoral dissertation].

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, *44*(4), 375–417.

Canas, G., & Rosasco, L. (2012). Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems* (pp. 2492–2500).

Carriere, M., Cuturi, M., & Oudot, S. (2017). Sliced Wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning* (pp. 664–673).

Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., & Papadakis, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, *40*(2), B429–B456.

Chen, L., Zhang, Y., Zhang, R., Tao, C., Gan, Z., Zhang, H., Li, B., Shen, D., & Carin, L. (2019). Improving sequence-to-sequence learning via optimal transport. In *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=S1xtAjR5tX

Chen, Y., Georgiou, T. T., & Tannenbaum, A. (2018). Optimal transport for Gaussian mixture models. *IEEE Access*, *7*, 6269–6278.

Cook, R. D., & Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, *86*(414), 328–332.

Courty, N., Flamary, R., Habrard, A., & Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*.

Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(9), 1853–1865.

Cui, L., Qi, X., Wen, C., Lei, N., Li, X., Zhang, M., & Gu, X. (2019). Spherical optimal transportation. *Computer-Aided Design*, *115*, 181–193.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems* (pp. 2292–2300).

Dai Yang, K., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A. C., Shivashankar, G., & Uhler, C. (2020). Predicting cell lineages using autoencoders and optimal transport. *PLoS Computational Biology*, *16*(4), 1–20.

Del Barrio, E., Gordaliza, P., Lescornel, H., & Loubes, J.-M. (2019). Central limit theorem and Bootstrap procedure for Wasserstein's variations with an application to structural relationships between distributions. *Journal of Multivariate Analysis*, *169*, 341–362.

Del Barrio, E., Inouzhe, H., Loubes, J.-M., Matrán, C., & Mayo-Íscar, A. (2020). Optimalflow: Optimal transport approach to flow cytometry gating and population matching. *BMC Bioinformatics*, *21*(1), 1–25.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., & Schwing, A. G. (2019). Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10648–10656).

Deshpande, I., Zhang, Z., & Schwing, A. G. (2018). Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3483–3491).

Dhouib, S., Redko, I., Kerdoncuff, T., Emonet, R., & Sebban, M. (2020). A Swiss army knife for minimax optimal transport. In *International Conference on Machine Learning* (pp. 2504–2513).

Dick, J., Kuo, F. Y., & Sloan, I. H. (2013). High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, *22*, 133–288.

Do Ba, K., Nguyen, H. L., Nguyen, H. N., & Rubinfeld, R. (2011). Sublinear time algorithms for earth mover's distance. *Theory of Computing Systems*, *48*(2), 428–442.

Dudley, R. M. (1969). The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, *40*(1), 40–50.

Dvurechensky, P., Gasnikov, A., & Kroshnin, A. (2018). Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning* (pp. 1367–1376).

Evans, S. N., & Matsen, F. A. (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 569–592.

Feydy, J., Charlier, B., Vialard, F.-X., & Peyré, G. (2017). Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 291–299).

Flamary, R., & Courty, N. (2017). *POT: Python optimal transport library*. Retrieved from https://pythonot.github.io/

Flamary, R., Cuturi, M., Courty, N., & Rakotomamonjy, A. (2018). Wasserstein discriminant analysis. *Machine Learning*, *107*(12), 1923–1945.

Fournier, N., & Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, *162*(3–4), 707–738.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*(376), 817–823.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *100*(9), 881–890.

Genevay, A., Cuturi, M., Peyré, G., & Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems* (pp. 3440–3448).

Gittens, A., & Mahoney, M. W. (2016). Revisiting the Nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, *17*(1), 3977–4041.

Glasserman, P. (2013). *Monte Carlo methods in financial engineering*. Springer Science & Business Media.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Gordaliza, P., Del Barrio, E., Fabrice, G., & Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning* (pp. 2357–2365).

Grave, E., Joulin, A., & Berthet, Q. (2019). Unsupervised alignment of embeddings with Wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1880–1890).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems* (pp. 5769–5779).

Hashimoto, T., Gifford, D., & Jaakkola, T. (2016). Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning* (pp. 2417–2426).

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. CRC Press.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, *13*, 435–475.

Hütter, J.-C., & Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, *49*(2), 1166–1194. https://doi.org/10.1214/20-AOS1997

Jagarlapudi, S. N., & Jawanpuria, P. K. (2020). Statistical optimal transport posed as learning kernel embedding. *Advances in Neural Information Processing Systems* 33.

Janati, H., Cuturi, M., & Gramfort, A. (2020). Spatio-temporal alignments: Optimal transport through space and time. In *International Conference on Artificial Intelligence and Statistics* (pp. 1695–1704).

Kantorovich, L. (1942). On translation of mass (in Russian), cr. In *Doklady of the Academy of Sciences of the USSR* (Vol. 37, pp. 199–201).

Klatt, M., Tameling, C., & Munk, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, *2*(2), 419–443.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., & Gustavo, K. (2019). Generalized sliced Wasserstein distances. *Advances in Neural Information Processing Systems*.

Kolouri, S., Pope, P. E., Martin, C. E., & Rohde, G. K. (2018). Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Kroshnin, A., Spokoiny, V., & Suvorikova, A. (2021). Statistical inference for Bures–Wasserstein barycenters. *The Annals of Applied Probability*, *31*(3), 1264–1298.

Lavenant, H., Claici, S., Chien, E., & Solomon, J. (2018). Dynamical optimal transport on discrete surfaces. *ACM Transactions on Graphics (TOG)*, *37*(6), 1–16.

Le, T., Yamada, M., Fukumizu, K., & Cuturi, M. (2019). Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, *44*(3), 907–927.

Lei, J., et al. (2020). Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, *26*(1), 767–798.

Lei, N., Su, K., Cui, L., Yau, S.-T., & Gu, X. D. (2019). A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, *68*, 1–21.

Lemieux, C. (2009). *Monte Carlo and quasi-Monte Carlo sampling*. Springer.

Leobacher, G., & Pillichshammer, F. (2014). *Introduction to quasi-Monte Carlo integration and applications*. Springer.

Levina, E., & Bickel, P. (2001). The earth mover's distance is the mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (Vol. 2, pp. 251–256).

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with r*. Chapman and Hall/CRC.

Li, B., & Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, *102*(479), 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*(414), 316–327.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, *87*(420), 1025–1039.

Li, T., & Meng, C. (2021). Modern subsampling methods for large-scale least squares regression. *International Journal of Cyber-Physical Systems (IJCPS)*, *2*(2), 1–28.

Lim, S., Park, H., Lee, S.-E., Chang, S., Sim, B., & Ye, J. C. (2020). Cyclegan with a blur kernel for deconvolution microscopy: Optimal transport geometry. *IEEE Transactions on Computational Imaging*, *6*, 1127–1138.

Lin, T., Fan, C., Ho, N., Cuturi, M., & Jordan, M. I. (2020). Projection robust Wasserstein distance and Riemannian optimization. *Advances in Neural Information Processing Systems*, 33.

Lin, T., Ho, N., Chen, X., Cuturi, M., & Jordan, M. I. (2020). Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. *Advances in Neural Information Processing Systems*, 33.

Lin, T., Ho, N., & Jordan, M. I. (2019). On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning* (pp. 3982–3991).

Lin, T., Zheng, Z., Chen, E., Cuturi, M., & Jordan, M. I. (2021). On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics* (pp. 262–270).

Liu, W., Ke, Y., Liu, J., & Li, R. (2022). Model-free feature screening and FDR control with knockoff features. *Journal of the American Statistical Association*, *117*(537), 428–443.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., & Stöter, F.-R. (2019). Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning* (pp. 4104–4113).

McGregor, A., & Stubbs, D. (2013). Sketching earth-mover distance on graph metrics. In *Approximation, randomization, and combinatorial optimization. Algorithms and techniques* (pp. 274–286). Springer.

Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W., & Ma, P. (2019). Large-scale optimal transport map estimation using projection pursuit. In *Advances in Neural Information Processing Systems* (pp. 8116–8127).

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., & Ma, P. (2020). Lowcon: A design-based subsampling approach in a misspecified linear model. *Journal of Computational and Graphical Statistics*, *30*, 1–15.

Meng, C., Yu, J., Zhang, J., Ma, P., & Zhong, W. (2020). Sufficient dimension reduction for classification using principal optimal transport direction. *Advances in Neural Information Processing Systems*, 33.

Meng, C., Yu, J., Chen, Y., Zhong, W., & Ma, P. (2022). Smoothing splines approximation using Hilbert curve basis selection. *Journal of Computational and Graphical Statistics*, (just-accepted), 1–26.

Meng, C., Zhang, X., Zhang, J., Zhong, W., & Ma, P. (2020). More efficient approximation of smoothing splines via space-filling basis selection. *Biometrika*, *107*(3), 723–735.

Montavon, G., Müller, K.-R., & Cuturi, M. (2016). Wasserstein training of restricted Boltzmann machines. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 3718–3726).

Musco, C., & Musco, C. (2017). Recursive sampling for the Nyström method. In *Advances in Neural Information Processing Systems* (pp. 3833–3845).

Muzellec, B., & Cuturi, M. (2019). Subspace detours: Building transport plans that are optimal on subspace projections. *Advances in Neural Information Processing Systems* (pp. 6914–6925).

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., & Simsekli, U. (2020). Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33.

Nadjahi, K., Durmus, A., Simsekli, U., & Badeau, R. (2019). Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*.

Nguyen, K., Ho, N., Pham, T., & Bui, H. (2020). Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*.

Owen, A. B. (2003). Quasi-Monte Carlo sampling. Monte Carlo Ray Tracing: Siggraph, 1, 69–88.

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, *6*, 405–431.

Paty, F.-P., & Cuturi, M. (2019). Subspace robust Wasserstein distances. In *International Conference on Machine Learning* (pp. 5072–5081).

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, *11*(5–6), 355–607.

Pitie, F., Kokaram, A. C., & Dahyot, R. (2005). N-dimensional probability density function transfer and its application to color transfer. In *ICCV 2005. Tenth IEEE International Conference on Computer Vision* (Vol. 2, pp. 1434–1439).

Pitié, F., Kokaram, A. C., & Dahyot, R. (2007). Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, *107*(1–2), 123–137.

Rabin, J., Peyré, G., Delon, J., & Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision* (pp. 435–446).

Ramdas, A., Trillos, N. G., & Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, *19*(2), 47.

Rigollet, P., & Weed, J. (2019). Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, *8*(4), 691–717.

Rolet, A., Cuturi, M., & Peyré, G. (2016). Fast dictionary learning with a smoothed Wasserstein loss. In *Artificial Intelligence and Statistics* (pp. 630–638).

Rowland, M., Hron, J., Tang, Y., Choromanski, K., Sarlos, T., & Weller, A. (2019). Orthogonal estimation of Wasserstein distances. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 186–195).

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*, 149–171.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., & Lander, E. S. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, *176*(4), 928–943.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., & Starck, J.-L. (2018). Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, *11*(1), 643–678.

Seguy, V., & Cuturi, M. (2015). Principal geodesic analysis for probability measures under the optimal transport metric. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* (Vol. 2; pp. 3312–3320).

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., & Blondel, M. (2018). Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations. ICLR 2018* (pp. 1–15).

Singh, S. P., Hug, A., Dieuleveut, A., & Jaggi, M. (2020). Context mover's distance & barycenters: Optimal transport of contexts for building representations. In *International Conference on Artificial Intelligence and Statistics* (pp. 3437–3449).

Solomon, J., Peyré, G., Kim, V. G., & Sra, S. (2016). Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, *35*(4), 1–13.

Staib, M., Claici, S., Solomon, J. M., & Jegelka, S. (2017). Parallel streaming Wasserstein barycenters. *Advances in Neural Information Processing Systems*, *30*, 2647–2658.

Tameling, C., & Munk, A. (2018). Computational strategies for statistical inference based on empirical optimal transport. In *2018 IEEE Data Science Workshop (DSW)* (pp. 175–179).

Taylor, J., & Tibshirani, R. (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, *46*(1), 41–61.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600–620.

Tolstikhin, I., Bousquet, O., Gelly, S., & Schölkopf, B. (2018). Wasserstein auto-encoders. In *6th International Conference on Learning Representations*.

Tong, A., Huang, J., Wolf, G., Van Dijk, D., & Krishnaswamy, S. (2020). Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International Conference on Machine Learning* (pp. 9526–9536).

Vayer, T., Flamary, R., Tavenard, R., Chapel, L., & Courty, N. (2019). Sliced Gromov-Wasserstein. In *Advances in Neural Information Processing Systems*.

Villani, C. (2008). *Optimal transport: Old and new*. Springer Science & Business Media.

Wang, S., & Zhang, Z. (2013). Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *The Journal of Machine Learning Research*, *14*(1), 2729–2769.

Wang, W., Xu, H., Wang, G., Wang, W., & Carin, L. (2021). Zero-shot recognition via optimal transport. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3471–3481).

Wang, Z., Zhou, D., Yang, M., Zhang, Y., Rao, C., & Wu, H. (2020). Robust document distance with Wasserstein-Fisher-Rao metric. In *Asian Conference on Machine Learning* (pp. 721–736).

Weed, J., & Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, *25*(4A), 2620–2648.

Weed, J., & Berthet, Q. (2019). Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory* (pp. 3118–3119).

Williams, C. K., & Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems* (pp. 682–688).

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.

Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P., & Gool, L. V. (2019). Sliced Wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3713–3722).

Xu, H. (2020). Gromov-Wasserstein factorization models for graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 6478–6485).

Xu, H., Liu, J., Luo, D., & Carin, L. (2022). Representing graphs via Gromov-Wasserstein factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (pp. 1).

Xu, H., Luo, D., & Carin, L. (2019). Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in Neural Information Processing Systems*, 32.

Xu, H., Luo, D., Carin, L., & Zha, H. (2021). Learning graphons via structured gromovwasserstein barycenters. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 10505–10513).

Xu, H., Luo, D., Zha, H., & Duke, L. C. (2019). Gromov-Wasserstein learning for graph matching and node embedding. In *International Conference on Machine Learning* (pp. 6932–6941).

Xu, H., Wang, W., Liu, W., & Carin, L. (2018). Distilled Wasserstein learning for word embedding and topic modeling. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 1723–1732).

Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., & Solomon, J. M. (2019). Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems*, *32*, 1601–1611.

Zemel, Y., Panaretos, V. M., et al. (2019). Fréchet means and procrustes analysis in Wasserstein space. *Bernoulli*, *25*(2), 932–976.

Zhang, J., Zhong, W., & Ma, P. (2021). A review on modern computational optimal transport methods with applications in biomedical research. *Modern Statistical Methods for Health Research*, 279–300.

Zhou, Z.-H. (2009). Ensemble learning. *Encyclopedia of Biometrics* (Vol. 1, pp. 270–273).

How to cite this article: Zhang, J., Ma, P., Zhong, W., & Meng, C. (2023). Projection-based techniques for high-dimensional optimal transport problems. *WIREs Computational Statistics*, *15*(2), e1587. https://doi.org/10.1002/wics.1587