Approximate Discrete Optimal Transport Plan with Auxiliary Measure Method

Dongsheng An[1], Na Lei*[2], and Xianfeng Gu[1]

[1] Stony Brook University
[2] Dalian University of Technology
{doan, gu}@cs.stonybrook.edu, nalei@dlut.edu.cn

Abstract. Optimal transport (OT) between two measures plays an essential role in many fields, ranging from economy, biology to machine learning and artificial intelligence. Conventional discrete OT problem can be solved using linear programming (LP). Unfortunately, due to the large scale and the intrinsic non-linearity, achieving discrete OT plan with adequate accuracy and efficiency is challenging. Generally speaking, the OT plan is highly sparse. This work proposes an auxiliary measure method to use the semi-discrete OT maps to estimate the sparsity of the discrete OT plan with squared Euclidean cost. Although obtaining the accurate semi-discrete OT maps is difficult, we can find the sparsity information through computing the approximate semi-discrete OT maps by convex optimization. The sparsity information can be further incorporated into the downstream LP optimization to greatly reduce the computational complexity and improve the accuracy. We also give a theoretic error bound between the estimated transport plan and the OT plan in terms of Wasserstein distance. Experiments on both synthetic data and color transfer tasks demonstrate the accuracy and efficiency of the proposed method.

Keywords: Optimal transport, Convex optimization, Linear Programming, Auxiliary measure

1 Introduction

Optimal transport (OT) is a powerful tool to compute the Wasserstein distance between probability measures, which are widely used to model various natural phenomena, including those observed in economics [13], optics [15], biology [28], differential equations [17] and other domains. Recently, OT has been successfully applied in the areas of machine learning, such as parameter estimation in Bayesian nonparametric models [25], computer vision [3,10,32], natural language processing [21,34] etc. In these applications, the complex probability measures are approximated by Dirac measures supported on their samples. To compute Wasserstein distances among Dirac measures, we have to solve the discrete OT

* Corresponding author

problems. Unfortunately, solving large scale discrete OT problem with high accuracy still remains a great challenge. To tackle this problem, we propose a novel method to improve the accuracy by utilizing the sparsity of discrete OT plan .

Semi-discrete OT Problem The origin of the optimal transport problem can be traced back to 1781, when Monge asked if there existed an OT map between two measures with the given cost function. Depending on the cost function and the measures, the OT map may not exist. In 1950's, Kantorovich relaxed the OT map to OT plan, and showed the existence and the uniqueness of the plan under mild conditions [33]. In 1980's, Brenier [6] discovered that when the density of the source measure is absolutely continuous and the cost function is the squared Euclidean distance, the OT map is given by the gradient of a convex function, the so-called Brenier potential.

Recently, the equivalence between the Brenier potential and Alexandrov's convex polytope has been rigours proved in [16], both of them can be obtained by solving the non-linear Monge-Ampère equation. This connection leads to a practical algorithm to solve the semi-discrete OT problem using convex geometry. According to Thm. 2 in this paper, the Brenier potential can be represented as the upper envelope of a set of hyperplanes, and its projection induces a power diagram of the source domain, which gives the semi-discrete OT map. Moreover, the power diagram can be estimated efficiently using Monte Carlo based method in high dimensional space [2].

Discrete OT Problem In this work, we focus on computing the OT plan between two discrete measures. Suppose the source and target discrete distributions are represented by $\nu_1 = \sum_{i=1}^{m} \nu_i^1 \delta(x - x_i)$ and $\nu_2 = \sum_{j=1}^{n} \nu_j^2 \delta(y - y_j)$, respectively. The transport plan is denoted as $\pi : \nu_1 \to \nu_2$, and $\pi = \{\pi_{ij} | \sum_i \pi_{ij} = \nu_j^2, \sum_j \pi_{ij} = \nu_i^1, \pi_{ij} \geq 0\}$, where π_{ij} represents the total mass transported from x_i to y_j. For the Kantorovich problem (Eqn. (5)), there are mn unknowns in total and $m+n$ constraints. We can solve it using the conventional linear programming (LP) method, whose time complexity is $O(n^{2.5})$ with Vaidya's algorithm [8]. For large scale problems, this is prohibitively high.

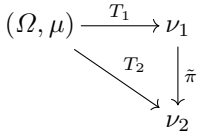
The Proposed Method In this paper, to compute the optimal transport plan between two discrete measures, we propose the auxiliary measure method. Basically, we construct an auxiliary measure μ with absolutely continuous density function defined on a convex domain Ω. Then we compute two semi-discrete OT maps $T_k : \mu \to \nu_k$, $k = 1, 2$. Each T_k induces a cell decomposition (power diagram) of Ω:

$$\Omega = \bigcup_{i=1}^{m} W_i^1 = \bigcup_{j=1}^{n} W_j^2, T_1 : W_i^1 \to x_i, T_2 : W_j^2 \to y_j.$$

The overlap of the two cell decomposition induces a refined cell decomposition:

$$\Omega = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} W_i^1 \cap W_j^2 = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} W_{ij}$$

where $W_{ij} := W_i^1 \cap W_j^2$. If we treat $T_1 : W_i^1 \to x_i$ as invertible, its inverse will be a set-valued map $T_1^{-1} : x_i \to W_i^1$, then the following diagram commutes,



The composition $\tilde{\pi} = T_2 \circ T_1^{-1}$ is a transport plan from ν_1 to ν_2, where $\tilde{\pi}_{ij} = \mu(W_{ij})$. Algorithms for finding a transport plan $\tilde{\pi}$ need to solve two semi-discrete OT problems with $m+n$ unknowns in total, which is much simpler than the LP method with mn unknowns. Both T_1 and T_2 are OT maps, but $\tilde{\pi}$ may not be optimal. Even so, $\tilde{\pi}$ is a transport plan with explicit sparsity. Namely, if W_{ij} is empty, then $\tilde{\pi}_{ij}$ is 0. Suppose the OT plan is $\hat{\pi} : \nu_1 \to \nu_2$, we can use the sparsity of $\tilde{\pi}$ to predict the sparsity of $\hat{\pi}$. Thus, by carefully choosing the auxiliary measure μ, we can make $\tilde{\pi}$ a good approximation of $\hat{\pi}$, and $\tilde{\pi}$ tells which $\hat{\pi}_{ij}$'s are zeros beforehand.

However, computing the accurate semi-discrete OT map for general OT problems is difficult [31]. In our settings, we only need the overlap information of $\{W_i^1\}$ and $\{W_j^2\}$, namely if $W_{ij} = \emptyset$ or not. To achieve this, there is no need to accurately compute the semi-discrete OT map. With the SDOT algorithm [2], we can obtain good estimations of the semi-discrete OT maps T_1, T_2, and thus get a coarse approximation of $\{W_{ij}\}$. Then by extending the coarse cell decomposition with nearest neighbour, we finally obtain the sparsity information of $\tilde{\pi}$, or equivalently $\{W_{ij}\}$. This greatly improves the efficiency of finding $\hat{\pi}$.

Contribution The contribution of the paper includes: **(i)** We propose an auxiliary measure method to solve the discrete OT problem by computing two approximate semi-discrete OT maps with $O(m+n)$ unknowns in total. With the auxiliary measure, we can greatly reduce the storage complexity of the discrete OT problem. **(ii)** The sparsity of the transport plan obtained by the auxiliary measure is used to estimate the sparsity of the discrete OT plan. The sparsity information is incorporated into the downstream LP to reduce the computational complexity and improve the accuracy of the computed OT cost. **(iii)** We give a theoretic error bound for the estimated transport plan and the OT plan in terms of Wasserstein distance. Experiments demonstrate the accuracy and efficiency of the proposed auxiliary measure method.

2 Related Work

OT plays an important role in various kinds of fields, and there is huge of research in this area. For detailed overview, we refer the readers to [26].

The semi-discrete OT problem computes the OT map between continuous and Dirac measures. Kitagawa et al. [19] use the damped Newton's method to solve such a problem. Genevay et al. [14] propose a semi-dual approach to solve the OT problems under discrete, semi-discrete or continuous settings. However, this method does not give an explicit form of the transport map. Arjovsky et al. [3] propose an approach that specializes to 1-Wasserstein distance, where the Lipschitz constraints are replaced by weight clipping at each iteration. This restricts the approximation accuracy of Wasserstein distance. By approximating the Alexandrov potential with DNN, Seguy et al. [30] solve a relaxed OT problem, and the resulting OT map can be obtained. However, their approximation using DNN is not globally convex and thus is not guaranteed to achieve global

optimum. Earlier, Gu et al. [16] propose to minimize a convex energy through the connection between the OT problem and convex geometry. In [22] the authors link the convex geometry viewed optimal transport with Kantorovich duality by Legendre dual theory. Recently, An et al. [2] extend the method to solve high dimensional semi-discrete OT problems by Monte Carlo Sampling.

When both the source and target measures are discrete, the OT problem can be treated as a standard LP task. To extend the problem into large dataset, Cuturi [11] adds an entropic regularizer into the prime OT problem. As a result, the regularized problem can be quickly solved with the Sinkhorn algorithm. Later, other entropy regularization based methods are proposed [1,12,23,9]. The problem of the Sinkhorn based methods is that they lose the sparse information of the OT plan. To solve this problem, Blondel et al. [4] incorporate structural information directly into the OT problem and keep the sparsity of the solution. However, the result, which is only an approximation of the OT plan, is not a transport plan. Schmitzer [29] then proposes a coarse-to-fine scheme to find the sparse plan for the entropy regularized problem.

Another genre to approximate the Wasserstein distance is the sliced Wasserstein distance [5], which projects the high-dimensional distribution into infinitely many one-dimensional spaces and then computes the average of the Wasserstein distance between these one-dimensional distributions. Then Kolouri et.al [20] generalize the sliced Wasserstein distance by the generalization of the Radon Transform. By selecting the most informative projection directions, Meng et.al [24] proposed the projection pursuit Monge map, which accelerates the computation of the original sliced optimal transport problem. But this kind of methods cannot give the OT plan.

3 Optimal Transport Theory

In this section, we will introduce some basic concepts and theorems in classic OT theory, focusing on the Brenier's approach and its generalization to the discrete settings. The details can be found in Villani's book [33].

Optimal Transport Problem Suppose X, Y are both subsets of d-dimensional Euclidean space $\mathbb{R}^d$, μ and ν are two probability measures defined on X and Y, respectively, with equal total measure $\mu(X) = \nu(Y)$.

Definition 1 (Measure-Preserving Map). *A map $T : X \to Y$ is* measure preserving *if, for any measurable set $B \subset Y$, the set $T^{-1}(B)$ is μ-measurable and $\mu(T^{-1}(B)) = \nu(B)$. The measure-preserving condition is denoted as $T_{\#}\mu = \nu$.*

Given the cost function $c(x, y) : X \times Y \to \mathbb{R}_{\geq 0}$, which indicates the cost of moving each unit mass from x to y, the total *transport cost* of the map $T : X \to Y$ is defined to be $\int_X c(x, T(x))d\mu(x)$.

The Monge's OT problem aims to find the measure-preserving map that minimizes the total transport cost.

Problem 1. [Monge Problem] Given the cost function $c : X \times Y \to \mathbb{R}_{\geq 0}$, find the measure preserving map $T : X \to Y$ that minimizes the total transport cost

$$(MP)\mathcal{M}_c(\mu,\nu) := \min_{T_{\#}\mu=\nu} \int_X c(x, T(x))d\mu(x). \tag{1}$$

The solution to the Monge's problem is called the *optimal transport map*, whose total transport cost is called the *optimal transport cost* between μ and ν, denoted as $\mathcal{M}_c(\mu,\nu)$.

Kantorovich's Approach Depending on the cost functions and the measures, the OT map between (X,μ) and (Y,ν) may not exist. Kantorovich relaxed the OT maps to OT plans, and defined the joint probability measure $\pi : X \times Y \to \mathbb{R}_{\geq 0}$, such that the marginal probability of π equals to μ and ν, respectively. Formally, let the projection maps be $\rho_x(x,y) = x$, $\rho_y(x,y) = y$, then we define

$$\mathit{\Pi}(\mu,\nu) := \{\pi : X \times Y \to \mathbb{R}_{\geq 0} : (\rho_x)_{\#}\pi = \mu, (\rho_y)_{\#}\pi = \nu\}$$

Problem 2 (Kantorovich Problem). Given the cost function $c : X \times Y \to \mathbb{R}_{\geq 0}$, find the joint probability measure $\pi : X \to Y$ that minimizes the total transport cost

$$(KP)\mathcal{M}_c(\mu,\nu) = \min_{\pi\in\mathit{\Pi}(\mu,\nu)} \int_{X\times Y} c(x,y)d\pi(x,y). \tag{2}$$

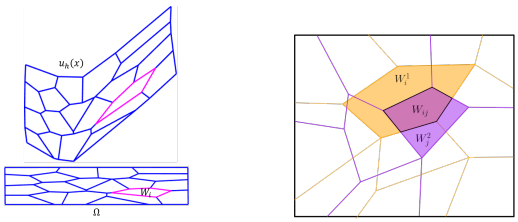
Brenier's Approach For the quadratic Euclidean distance cost, the existence, uniqueness and the intrinsic structure of the OT map were proven by Brenier [7].

Theorem 1 (Brenier Theorem). *Suppose X and Y are the subsets of the Euclidean space $\mathbb{R}^d$ and the transport cost is given by the quadratic Euclidean distance $c(x,y) = \|x-y\|^2$. Furthermore, μ is absolutely continuous, both μ and ν have finite second order moments. Then there exists a convex function $u : X \to \mathbb{R}$, the so-called the Briener potential, and its gradient map ∇u gives the solution to the Monge's problem. The Brenier potential is unique up to adding a constant, hence the optimal transport map is unique.*

Semi-discrete OT Problem Suppose the source measure μ is absolutely continuous and defined on a convex domain $\mathit{\Omega} \subset \mathbb{R}^d$, the target measure is a Dirac measure $\nu = \sum_{i=1}^n \nu_i\delta(y-y_i), i \in [n]$ and $y_i \in \mathbb{R}^d$. Also, we assume $\mu(\mathit{\Omega}) = \sum_{i=1}^n \nu_i$. *The semi-discrete OT map* is the measure-preserving map that minimizes the transport cost, $T^* := \arg\min_{T_{\#}\mu=\nu} \int_{\mathit{\Omega}} c(x,T(x))d\mu(x)$.

When the cost function is set to be the quadratic Euclidean distance $c(x,y) = \|x-y\|^2$, the Brenier potential can be expressed as $u_h(x) = \max_i\{\langle x, y_i\rangle + h_i,\ \forall i \in [n]\}$. The induced OT map pushing forward μ to ν is $T^* : W_i \to y_i$, where $W_i = \{x|\ \langle x, y_i\rangle + h_i \geq \langle x, y_k\rangle + h_k,\ \forall k \in [n]\}$.

Under the semi-discrete OT map $T^* : \mathit{\Omega} \to Y$, a cell decomposition (also a power diagram) is induced $\mathit{\Omega} = \bigcup_{i=1}^n W_i$, such that every x in the cell W_i is mapped to the target y_i, $T : x \in W_i \mapsto y_i$, and $\mu(W_i) = \nu_i$. As shown in Fig.



(a) (b)

Fig. 1. (a) Brenier potential and the corresponding power diagram. Each cell W_i is mapped to the corresponding y_i, and $\mu(W_i) = \nu_i$. (b) The illustration of the sparsity of $\hat{\pi}$. $\{W_i^1\}$ and $\{W_j^2\}$ are two cell decomposition induced by the semi-discrete OT maps from μ to ν_1 and ν_2. The refined cell decomposition $\{W_{ij}\}$ with $W_{ij} = W_i^1 \cap W_j^2$ not only gives the solution of $\tilde{\pi}$, but also gives a good approximation of the sparsity of the OT plan between ν_1 and ν_2.

1(a), the cell W_i is mapped to the corresponding y_i, which corresponds to the hyperplane $\langle x, y_i\rangle + h_i$. The total cost of T is given by $\int_\Omega c(x, T(x))d\mu(x) = \sum_{i=1}^n \int_{W_i} c(x, y_i)d\mu(x)$.

The following gives the generalization of the Brenier theorem to compute the semi-discrete OT map [16].

Theorem 2. *Let μ be a probability measure defined on a compact convex domain Ω in $\mathbb{R}^d$, $\nu = \sum_{i=1}^n \nu_i \delta(y - y_i)$ with $y_i \in \mathbb{R}^d$. If $\sum_{i=1}^n \nu_i = \mu(\Omega)$ and $c(x, y) = \|x - y\|^2$, then there exists $h = (h_1, h_2, \ldots, h_n) \in \mathbb{R}^n$, unique up to adding a constant $(k, k, \ldots, k)$, so that $w_i(h) = \nu_i \ \forall i \in [n]$, where $w_i(h) = \mu(W_i(h))$ with $W_i(h) = \{x |\ \langle x, y_i\rangle + h_i \geq \langle x, y_k\rangle + h_k,\ \forall k \in [n]\}$. The vector h is the unique minimum of the convex energy*

$$E(h) = \int_0^h \sum_{i=1}^n w_i(\eta)d\eta_i - \sum_{i=1}^n h_i\nu_i, \tag{3}$$

defined on an open convex set $\mathcal{H} = \{h \in \mathbb{R}^n : \sum_{i=1}^n h_i = 0\}$. Furthermore, if we define $u_h(x) = \max_i\{\langle x, y_i\rangle + h_i,\ \forall i \in [n]\}$, the map $\nabla u_h(x) : W_i(h) \to y_i\ \forall i \in [n]$ minimizes $\int_\Omega \|x - T(x)\|^2 d\mu(x)$ among all measure preserving maps $T_\#\mu = \nu$.

Now, the gradient of the above energy is given by:

$$\nabla E(h) = (w_1(h) - \nu_1, w_2(h) - \nu_2, \ldots, w_n(h) - \nu_n)^T \tag{4}$$

Discrete OT Problem Given both the source measure $\nu_1 = \sum_{i=1}^m \nu_i^1 \delta(x - x_i)$ and the target measure $\nu_2 = \sum_{j=1}^n \nu_j^2 \delta(y - y_j)$, with the supports defined by $X = \{x_i\}_{i=1}^m$ and $Y = \{y_j\}_{j=1}^n$, there may not exist an OT map $T : X \to Y$, but the OT plan always exists, which is the solution of the Kantorovich problem:

$$\begin{aligned} \mathcal{M}_c(\nu_1, \nu_2) &= \min_{\pi \geq 0} \sum_{i=1}^m \sum_{j=1}^n c_{ij}\pi_{ij} \\ s.t. \quad \sum_{i=1}^m \pi_{ij} &= \nu_j^2, \quad \sum_{j=1}^n \pi_{ij} = \nu_i^1 \end{aligned} \tag{5}$$

where $c_{ij} = c(x_i, y_j)$ and $\pi \in R^{m \times n}$ is the transport plan. And the problem of Eqn. (5) can be solved by classical LP.

Sparse Discrete OT Problem It requires mn unknown variables of π_{ij} to solve the discrete OT problem with LP. But in general cases, the discrete optimal transport plans are highly sparse. For example, if the OT problem is an assignment problem, that is $m = n$ and $\nu_i^1 = \nu_j^2 = 1/n \ \forall \ i, j \in [n]$, then the OT plan degenerates to an OT map, and among the n^2 π_{ij}'s, only n of them are non-zeros. If the π_{ij}s that are zeros can be determined beforehand, we will only need to consider the non-zero ones during the optimization, and this will greatly improve the computational accuracy and efficiency of the Kantorovich problem of Eqn. (5).

In the following, we give some theoretical analysis to estimate the approximation error bound for the auxiliary measure method. Suppose the discrete measures ν_1 and ν_2 are given, both of them are defined in a Euclidean space $\mathbb{R}^d$. All the probability measures defined in $\mathbb{R}^d$ form an infinite dimensional metric space $\mathcal{P}(\mathbb{R}^d)$, with the Wasserstein distance $\mathcal{W}_c$ as the metric, where c is the squared Euclidean distance. Then there is a unique geodesic γ in $\mathcal{P}(\mathbb{R}^d)$ connecting ν_1 and ν_2. Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be the auxiliary measure, the closest point on γ to μ is $\mu^* := \arg\min_{\nu \in \gamma} \mathcal{W}_c(\mu, \nu)$, and the distance from μ to the geodesic is $d = \mathcal{W}_c(\mu, \mu^*)$.

Theorem 3. *Given $\nu_1, \nu_2 \in \mathcal{P}(\mathbb{R}^d)$, the auxiliary measure μ, $T_k : \mu \to \nu_k, k = 1, 2$ are the OT maps. Suppose the distance from μ to the geodesic connecting ν_1 and ν_2 is d, then $T_2 \circ T_1^{-1} : \nu_1 \to \nu_2$ is measure preserving and its transport cost $\mathcal{C}$ is bounded by*

$$\mathcal{W}_c(\nu_1, \nu_2) \leq \mathcal{C}^{\frac{1}{2}}(T_2 \circ T_1^{-1}) \leq \mathcal{W}_c(\nu_1, \nu_2) + 2d \tag{6}$$

The proof of the theorem can be found in the supplement. From the inequality of Eqn. (6), it is obvious that the quality of the approximate transport map $T_2 \circ T_1^{-1}$ is determined by the Wasserstein distance $d = \mathcal{W}_c(\mu, \mu^*)$. If μ is on the geodesic, then $\tilde{\pi} = T_2 \circ T_1^{-1}$ is the desired OT plan between ν_1 and ν_2. If d is relatively small, then the approximated transport plan is close to the OT plan, therefore the sparsity of $\tilde{\pi}$ is similar to that of the OT plan. In practice, we use the sparsity of the approximate plan $\tilde{\pi}$ as the constraints to compute the OT plan between ν_1 and ν_2 and obtain $\hat{\pi}$. It can be seen that

$$\mathcal{W}_c^2(\nu_1, \nu_2) \leq \mathcal{C}(\hat{\pi}) \leq \mathcal{C}(T_2 \circ T_1^{-1})$$

therefore $\hat{\pi}$ is a better approximation than $T_2 \circ T_1^{-1}$, but with the same sparsity.

Assume $\{W_i^1\}$ and $\{W_j^2\}$ are the cell decomposition induced by T_1 and T_2, respectively, the newly refined cell decomposition $\{W_{ij} | W_{ij} = W_i^1 \cap W_j^2\}$ (shown in Fig. 1(b)) gives $\tilde{\pi}_{i,j} = \mu(W_{ij})$. With the cost function between ν_1 and ν_2 being $c_{ij} = \|x_i - y_j\|^2$, define $\Phi = \{(i, j) | W_i^1 \cap W_j^2 \neq \emptyset\}$, Kantorovich problem

of Eqn. (5) can be approximated by:

$$\begin{aligned} \mathcal{M}_c(\nu_1, \nu_2) &= \min_{\pi \geq 0} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \pi_{ij} \\ s.t. \quad \sum_{i=1}^{m} \pi_{ij} &= \nu_j^2, \quad \sum_{j=1}^{n} \pi_{ij} = \nu_i^1 \\ \pi_{ij} &= 0 \quad \forall (i,j) \notin \Phi \end{aligned} \tag{7}$$

4 Computational Algorithms

This section explains the computational algorithms for the auxiliary measure method in detail.

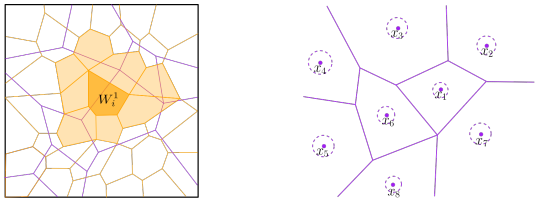
4.1 Semi-discrete OT Algorithm

Based on Thm. 2, finding the semi-discrete OT map from the given absolutely continuous measure μ to the discrete measure $\nu = \sum_{i=1}^{n} \nu_i \delta(x - x_i)$ is equivalent to optimizing the convex energy in Eqn. (3) with respect to the height vector h. The optimization can be carried out using the gradient descend method. In the gradient formula of Eqn. (4), we need to estimate the μ-volume $w_i(h)$ for each cell $W_i(h)$ with the Monte Carlo method proposed in [2]: N random samples $\{z_j\}_{j=1}^{n}$ are drawn from μ, then the μ-volume of the cell $W_i(h)$ is estimated by $\hat{w}_i(h) = \#\{z_j \mid z_j \in W_i(h)\}/N$. Given a random sample z_j, let $i = \arg\max_k \{\langle z_j, x_k \rangle + h_k, k \in [n]\}$, then $z_j \in W_i(h)$. When N goes to infinity, $\hat{w}_i(h)$ converges to $w_i(h)$. Hence the gradient of the energy can be approximated by $\nabla E \approx (\hat{w}_i(h) - \nu_i)^T$. Once the gradient is estimated, we can use the Adam algorithm [18] to minimize the energy.

When N is small, the estimated μ-volume is coarse but the computation is fast; when N is large, the estimated μ-volume is accurate but the computation is slow. To balance the efficiency and accuracy, we first use a small N to coarsely estimate h, when the energy $E(h)$ stops decreasing, we increase N to improve the accuracy of the estimation. The predefined total μ-volume distortion θ gives the stop condition, namely the algorithm stops when $\|\nabla E(h)\|_1 \leq \theta$. The sampling of z_js is independent of each other and the cell location estimation for each z_j can be paralleled, therefore the whole algorithm can be accelerated by GPUs. The algorithm is called SDOT and its details can be found in Alg. S1 of the supplement.

4.2 Discrete OT Plan with Auxiliary Measure

With the cell decompositions induced by the approximate OT maps computed by the SDOT algorithm, we firstly use a sparse matrix S to represent the overlap information. Then S is extended by nearest neighbours. Finally, we give two strategies for the choice of the auxiliary measure μ.



(a) (b)

Fig. 2. (a) The orange cells and the purple cells represent the power diagrams induced by the computed semi-discrete OT maps T_1, T_2 from μ to ν_1 and ν_2 with μ-volume accuracy θ. When computing the overlap information of W_i^1, we take both W_i^1 (the dark orange cell) and its neighbours (the light orange cells) into consideration. (b) The cell decomposition of $\nu_1 = \sum_{i=1}^m \nu_i^1 \delta(x - x_i)$ when $\mu = \sum_{i=1}^m \nu_i^1 N(x_i, \sigma I)$ with small σ. The cell W_i^1 that is mapped to x_i by the semi-discrete OT map T_1 covers x_i.

Estimate the Sparsity Matrix Given two discrete measures $\nu_1 = \sum_{i=1}^m \nu_i^1 \delta(x - x_i)$ and $\nu_2 = \sum_{j=1}^n \nu_j^2 \delta(y - y_j)$, and an auxiliary measure μ defined on a convex support Ω, we use the SDOT algorithm with μ-volume distortion parameter θ to compute the semi-discrete OT maps $T_k : \mu \to \nu_k$, $k = 1, 2$. Each map induces a cell decomposition $\{W_i^1\}$ and $\{W_j^2\}$.

We use an $m \times n$ matrix $S = (s_{ij})$ to represent the overlapping relation among the cells of the two power diagrams, and call the matrix as the sparsity matrix. The sparsity matrix is defined as

$$s_{ij} = \begin{cases} 1 \; W_i^1 \cap W_j^2 \neq \emptyset \\ 0 \; W_i^1 \cap W_j^2 = \emptyset \end{cases}$$

The sparsity matrix can be estimated by random sampling. We randomly sample $z_k \sim \mu$, then compute the cells containing z_k. If both W_i^1 and W_j^2 contain z_k, then we set s_{ij} to be 1. The procedure keeps running until the sparsity matrix S converges to the steady state. The algorithm for computing the sparsity matrix is given in Alg. S2 of the supplement.

Note that the above $\{W_i^1\}$ and $\{W_j^2\}$, which are computed by the SDOT algorithm under the μ-volume distortion parameter θ, are just approximations of the groundtruth power diagrams induced by the groundtruth semi-discrete OT maps. To make S better represent the sparse information of the groundtruth OT plan, we then not only use the computed $\{W_i^1\}$ and $\{W_j^2\}$ to compute S, the neighborhoods of each cell is also used to extend S. As shown in Fig. 2(a), to find the cells in $\{W_j^2\}$ that overlapping with the real $\hat{W}_i^1$ of the groundtruth semi-discrete OT map, we not only use the cell W_i^1 (the dark orange cell) to compute S, but also use the cells around W_i^1 (the light orange cells) to extend S. Based on the property of the semi-discrete OT map, the cells around W_i^1 corresponds to the neighbours of x_i in X. Therefore, we can use neighbours of x_i to update S.

Specifically, we extend S so that it includes the overlap information of the neighbours of each W_i^1 and W_j^2. For each sample x_i of ν_1, we find the k nearest neighbours of it, namely $x_{i1}, x_{i2}, \ldots, x_{ik}$. Then the rows of $i1, i2, \ldots, ik$ of S

are added to the ith row of S. Thus, the ith row of S includes the overlap information of both W_i^1 and its neighbour cells. Similarly, for each y_j, the k nearest neighbours of it are also found, marked as $y_{j1}, y_{j2}, \ldots, y_{jk}$. Then columns $j1, j2, \ldots, jk$ of S are added to its jth column. By replacing Φ with the new sparse matrix S, which represents the sparsity of the OT plan, the problem of Eqn. (7) can be solved effectively through LP.

Auxiliary Measure μ In theory, the auxiliary measure μ should locate at the geodesic from the source measure ν_1 to the target measure ν_2, which will make the bound tight in Thm. 3 and the computed transport plan be the OT plan. However, given two general distributions, it is hard to compute the geodesic between them without computing the OT plan first. Thus, in practice we make the distance d from the auxiliary measure μ to the geodesic between ν_1 and ν_2 small enough, then the computed transport cost should approximate the OT cost well according to Thm. 3. To achieve this, we can utilize the information inherited in the source distribution ν_1. If we can find a continuous μ that is close to ν_1, namely $\mathcal{W}_c(\mu, \nu_1)$ is small, that we can deduce that the distance d from μ to the geodesic between ν_1 and ν_2 is smaller than $\mathcal{W}_c(\mu, \nu_1)$ accordingly.

Strategy 1: if we know the continuous distribution $\hat{\nu}_1$ where ν_1 is sampled from, it is reasonable to set μ to be $\hat{\nu}_1$. In such a situation, $\mathcal{W}_c(\mu, \nu_1)$ should be reasonably small (See Alg. S3 in the supplement).

Strategy 2: alternatively, we can set μ to be a Gaussian mixture model based on ν_1, i.e. $\mu = \sum_{i=1}^m \nu_i^1 N(x_i, \sigma I_d)$, where $\sigma \ll \min d(x_i, x_k)\ \forall\ i, k \in [m]$ and $i \neq k$, and I_d represents the d-dimensional identity matrix. In such case, we have the following proposition (proof in the supplement):

Proposition 1. *Given* $\mu = \sum_{i=1}^m \nu_i^1 N(x_i, \sigma I_d)$ *and* $\nu_1 = \sum_{i=1}^m \nu_i^1 \delta(x - x_i)$, *then we have* $\mathcal{W}_c(\mu, \nu_1) \leq \sigma$ *under the quadratic Euclidean cost. Moreover, if* σ *is small enough, then the cell* W_i *of the cell decomposition induced by the semi-discrete OT map from* μ *to* ν_1 *should cover* x_i *itself.*

Then Eqn. (6) can be written as:

$$\mathcal{W}_c(\nu_1, \nu_2) \leq \mathcal{C}^{\frac{1}{2}}(T_2 \circ T_1^{-1}) \leq \mathcal{W}_c(\nu_1, \nu_2) + 2\sigma \tag{8}$$

Fig. 2(b) illustrates the relationship of the cell decomposition $\{W_i\}$ and $\{x_i\}$. Then we only need to compute the semi-discrete OT map $T_2 : \mu \to \nu_2$. The sparse matrix S can be estimated as follows: firstly we set s_{ij} to be 1 if and only if $x_i \in W_j^2$, then the sparse matrix S is extended with the neighbourhood information. This method (see Alg. S4 in the supplement) is more applicable and makes the computation much faster.

5 Experiments

This section reports our experimental results. All of our experiments are conducted on Intel Core i7-9800X CPU with 32GB RAM and NVIDIA GeForce RTX 2080 Ti GPU. We investigate the influence of different parameters, including the auxiliary measure μ, the μ-volume distortion parameter θ for the SDOT algorithm, and the number k of the nearest neighbours to extend the sparsity

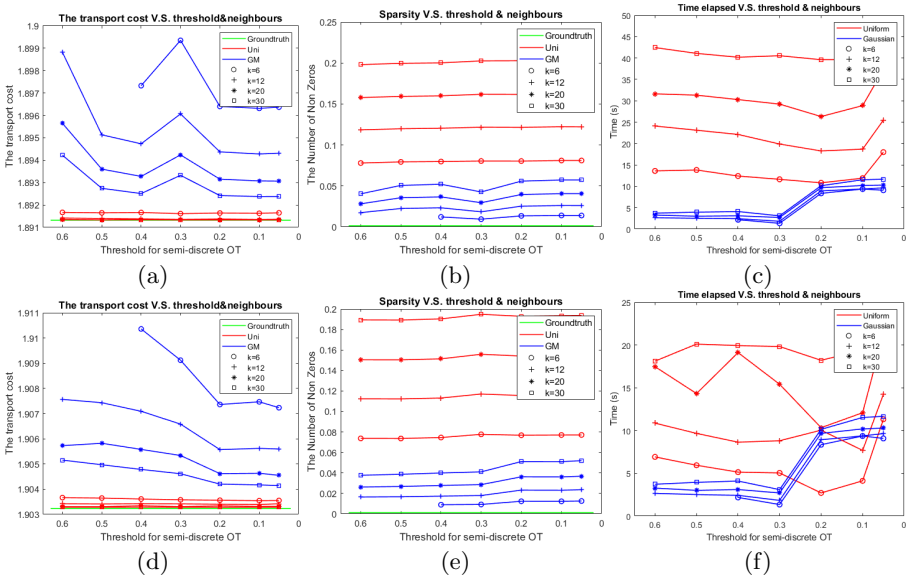


Fig. 3. The performances of the proposed algorithm on synthetic data with different parameters θ, μ and k.

matrix S. We first test our algorithm on synthetic tasks, and then apply it to the color transfer problem. Experimental results demonstrate that the proposed method outperforms the state-of-the-arts.

5.1 Performance on Synthetic dataset

We test the proposed method on two synthetic tasks with different parameters.

Two Tasks We set $\nu_1 = \sum_{i=1}^{m} \nu_i^1 \delta(x - x_i)$, where x_is are randomly sampled from the d-dimensional uniform distribution $[0,1]^d$; $\nu_2 = \sum_{j=1}^{n} \nu_j^2 \delta(y - y_j)$, where y_j's are sampled from d-dimensional Gaussian distribution $N(0, I_d)$. We conduct two tasks with $d = 5$: **(i)** compute the OT plan from ν_1 to ν_2, where $m = 1000$, $n = 2000$, the weights $\nu_i^1, i \in [m]$ and $\nu_j^2, j \in [n]$ are randomly sampled from the uniform distribution and then normalized by $\nu_i^1 = \nu_i^1 / \sum_{k=1}^{m} \nu_k^1$ and $\nu_j^2 = \nu_j^2 / \sum_{k=1}^{n} \nu_k^2$; and **(ii)** the assignment problem with $m = n = 1500$ and $\nu_i^1 = \nu_j^2 = 1/n \ \forall \ i, j \in [n]$.

Choice of parameters We use different parameters for the testing, including **(i)** the μ-volume accuracy threshold θ for computing the semi-discrete OT map between μ and ν_1, ν_2; **(ii)** the number k of the nearest neighbors to extend the sparsity matrix S; and **(iii)** the auxiliary measure μ, one is the 5-dimensional uniform distribution where ν_1 is sampled from; the other is the Gaussian mixture distribution with $\sigma = 0.1 \min_{i \neq j} d(x_i, x_j)$.

Comparison Results We compare the computational results obtained with different parameters using three indicators: **(i)** The transport cost $\mathcal{C}$ of the computed transport plan; **(ii)** The sparsity, represented by $|S|/mn$, where $|S|$ is the number of nonzero entries in the sparsity matrix S; and **(iii)** The running time of the whole pipeline.

Fig. 3 summarizes the comparison results. Fig. 3(a-c) show the statistics of the first task and Fig. 3(d-f) show results for the second task. In Fig. 3(a-f), the green curves correspond to the results of the real OT plan computed by LP. The red curves show the results computed with μ being the uniform distribution, and the blue curves are the results of μ as Gaussian mixture distribution. Different blue (or red) curves are with different ks (number of neighbors).

In Fig. 3(a), the horizontal axis represents the threshold θ; the vertical axis is the computed transport cost. Since the OT cost between ν_1 and ν_2 is independent of θ, the green curve is a horizontal line. It can be observed that the transport cost decreases when θ decreases; the cost decreases when k increases; and the costs for uniform μ, where ν_1 is sampled from, are smaller than those for Gaussian mixture μ. In Fig. 3(b), the horizontal axis represents θ, the vertical axis represents the sparsity. It is easy to see that the real OT plan is with the minimal sparsity; the Gaussian mixture μ induces better sparsity than the uniformly distributed μ; and the sparsity decreases when k increases. In Fig. 3(c), the horizontal axis is θ, the vertical axis is the running time. The green curve is not shown, because the LP method is far slower than our method. It is obvious that the method using Gaussian mixture μ is faster than that using uniform μ; when k decreases, the computation is faster. Fig. 3(d-f) show the statistics for the second task, i.e., the assignment problem. The comparison results are similar to those obtained from the first task.

In summary, comparing with the Gaussian mixture auxiliary measure μ constructed through the source measure ν_1, the known μ where ν_1 is sampled from, gives more accurate transport plan (with less transport cost), but less sparsity and slower computation speed.

5.2 Comparison with state-of-the-art techniques

We compare our algorithm with the state-of-the-art methods, including the Sinkhorn method [11], SOT [4], and PPMM [24].

To demonstrate that our method can compute accurate transport plans in different dimensions, we choose the dimension parameter d to be 2, 5 and 20. The nearest neighbor parameter k is set to be 10, the μ-volume accuracy threshold θ is 0.3. For Gaussian-mixture auxiliary measure μ, we set $\sigma = 0.1 \min_{i \neq j} d(x_i, x_j)$.

Fig. 4 shows the comparisons. The horizontal axis shows the sizes of the data sets, represented as $m \times n \times d$, meaning ν_1 has m points in $\mathbb{R}^d$ and ν_2 n points in $\mathbb{R}^d$. The vertical axis illustrates *the difference between the computed transport cost and the OT cost*, both using the squared Euclidean distance as the cost function. The green circles are the OT cost obtained by LP. The red crosses, yellow triangles, blue stars, black squares and Cyan diamonds denote the results obtain by our method with uniform μ where ν_1 is sampled from, our method

(a) Source (b) Target (c) Sinkhorn (d) SOT (e) Ours.

Fig. 5. Comparison of the results on color transfer tasks. (Zoom in/out for better visualization.)

with Gaussian mixture μ, Sinkhorn, SOT with L-BFGS solver for the smoothed semi-dual formula and PPMM, respectively.

From Fig. 4, we can see that our proposed method with auxiliary measures always obtains the minimal discrepancy in terms of OT cost. SOT method gives accurate result in general, but sometimes it leads to invalid transport plan (TP), as shown by the negative result in the second test. Similarly, the sixth result of PPMM method is negative. Furthermore, the figure shows that the results of the Sinkhorn method tends to become inaccurate when d is large.

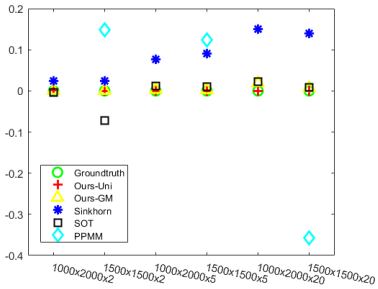


Fig. 4. Comparison among our proposed method and others.

5.3 Application to Color transfer

Given a color image, its color distribution can be represented by the histogram in the RGB color space. Assume there are m colors represented by $x_1, x_2, \ldots, x_m$, each $x_i = (r_i, g_i, b_i)$ is a point in $\mathbb{R}^3$, and the corresponding normalized frequencies are $\nu_1, \nu_2, \ldots, \nu_m$, then the image color distribution is represented by a discrete distribution in $\mathbb{R}^3$ [27]: $\nu = \sum_{i=1}^{m} \nu_i \delta(x - x_i)$. Given two images, the source image color distribution is $\nu_1 = \sum_{i=1}^{m} \nu_i^1 \delta(x - x_i)$ and the target is $\nu_2 = \sum_{j=1}^{n} \nu_j^2 \delta(y - y_j)$. We can find the OT plan $\pi : \nu_1 \to \nu_2$. For each color x_i, we apply the barycentric interpolation to get the mapping $T(x_i) = \frac{\sum_{j=1}^{n} \pi_{ij} y_j}{\sum_{j=1}^{n} \pi_{ij}}$. By replacing the color x_i of the source image with $T(x_i)$, we obtain a new image with content coming from the source and color distribution from the target.

Here we use four different images. The number of samples in the color space $\mathbb{R}^3$ for each input image is 13892 for 'autumn', 18103 for 'comunion', 17820 for 'graffiti' and 15129 for 'rainbow-bridge'. Fig. 5 shows the color transfer results from 'graffiti' to 'rainbow-bridge' using our proposed method and other state-of-the-art methods (please see the supplement for more experiments). Fig. 5(a) shows the source image and the second column the target image. Fig. 5(c) shows

	Sinkhorn			SOT			Ours		
	Cost	TP	Sparse	Cost	TP	Sparse	Cost	TP	Sparse
autumn → comunion	88.7535	√	×	69.5143	×	√	87.4280	√	√
graffiti → rainbow	84.5362	√	×	74.0894	×	√	84.3785	√	√
autumn → graffiti	131.0952	√	×	86.3683	×	√	129.6989	√	√
autumn → rainbow	83.1778	√	×	55.2500	×	√	81.9912	√	√
comunion → graffiti	70.7658	√	×	41.9410	×	√	70.1804	√	√
comunion → rainbow	39.7300	√	×	27.0912	×	√	39.4653	√	√

Table 1. The comparison between our method, Sinkhorn [11] and SOT [4] on the color transfer tasks.

the result of the Sinkhorn algorithm [11], which is blurry due to the dense transport plan. Fig. 5(d) illustrates the results of SOT [4]. Fig. 5(e) shows the result of our method, which have consistent color distribution with the target image, and is much sharper than those generated by Sinkhorn. This shows our method obtains more accurate transport plan with higher sparsity.

We also estimate the OT cost among the color distributions of the 4 input images, using the Sinkhorn method [11], SOT [4] and our method. The results are reported in Table 1. In the table, 'TP' represents 'valid transport plan'. From the table, we can find that both the results of the Sinkhorn algorithm and the proposed method are valid transport plans, and our method outperforms Sinkhorn both in the estimated OT cost and the sparsity. Though the solutions of SOT are sparse, they are not even valid transport plans. In conclusion, the proposed method gives more accurate OT plan with higher sparsity.

6 Conclusions

This work proposes an auxiliary measure method using semi-discrete OT maps to estimate the discrete OT plan by reducing the number of unknowns from $O(mn)$ to $O(m+n)$. The sparsity information of the transport plan obtained by the auxiliary measure is used to estimate the sparsity of the discrete OT plan. And the sparsity of the OT plan is incorporated into the downstream LP optimization to greatly reduce the computational complexity of the discrete Kantorovich problem and improve the accuracy. We also give a theoretic error bound for the estimated transport plan and the OT plan in terms of Wasserstein distance. Experiments on synthetic data and color transfer of real images demonstrate the accuracy and efficiency our method. In the future, we will explore to find much better auxiliary measures to further improve the accuracy of the method.

Acknowledgement Lei was supported by the National Natural Science Foundation of China No. 61936002 and the National Key R&D Program of China 2021YFA1003003. Gu is partially supported by NSF 2115095, NSF 1762287, NIH 92025 and NIH R01LM012434.

References

1. Altschuler, J., Niles-Weed, J., Rigollet, P.: Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In: Advances in Neural Information Processing Systems 30 (2017)
2. An, D., Guo, Y., Lei, N., Luo, Z., Yau, S.T., Gu, X.: Ae-ot: A new generative model based on extended semi-discrete optimal transport. In: International Conference on Learning Representations (2020)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML. pp. 214–223 (2017)
4. Blondel, M., Seguy, V., Rolet, A.: Smooth and sparse optimal transport. In: Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. pp. 880–889 (2018)
5. Bonneel, N., Rabin, J., Peyre, G., Pfister, H.: Sliced and radon wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision (2014)
6. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math. **44**(4), 375–417 (1991)
7. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics **44**(4), 375–417 (1991)
8. Bubeck, S.: Convex Optimization: Algorithms and Complexity, vol. 8. Foundations and Trends in Machine Learning (2015)
9. Chakrabarty, D., Khanna, S.: Better and simpler error analysis of the sinkhorn-knopp algorithm for matrix scaling. Mathematical Programming **188**(1), 395–407 (2021)
10. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(9), 1853–1865 (2017)
11. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transportation distances. In: International Conference on Neural Information Processing Systems. vol. 26, pp. 2292–2300 (2013)
12. Dvurechensky, P., Gasnikov, A., Kroshnin, A.: Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In: International Conference on Machine Learning. pp. 1367–1376 (2018)
13. Galichon, A.: Optimal Transport Methods in Economics. Princeton University Press (2016)
14. Genevay, A., Cuturi, M., Peyré, G., Bach, F.: Stochastic optimization for large-scale optimal transport. In: Advances in Neural Information Processing Systems. pp. 3440–3448 (2016)
15. Glimm, T., Oliker, V.: Optical design of single reflector systems and the Monge–Kantorovich mass transfer problem. Journal of Mathematical Sciences **117**(3), 4096–4108 (Sep 2003)
16. Gu, D.X., Luo, F., Sun, J., Yau, S.T.: Variational principles for minkowski type problems, discrete optimal transport, and discrete monge-ampère equations. Asian Journal of Mathematics (2016)
17. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the fokker–planck equation. SIAM Journal on Mathematical Analysis **29**(1), 1–17 (1998)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kitagawa, J., Mérigot, Q., Thibert, B.: Convergence of a newton algorithm for semi-discrete optimal transport. Journal of the European Mathematical Society **21**(9), 2603–2651 (2019)

20. Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., Rohde, G.: Generalized sliced wasserstein distances. In: Advances in Neural Information Processing Systems 32 (2019)
21. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 957–966 (2015)
22. Lei, N., Su, K., Cui, L., Yau, S.T., Gu, D.X.: A geometric view of optimal transportation and generative mode. Computer Aided Geometric Design **68**, 1–21 (2019)
23. Lin, T., Ho, N., Jordan, M.I.: On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In: International Conference on Machine Learning. pp. 3982–3991 (2019)
24. Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W., Ma, P.: Large-scale optimal transport map estimation using projection pursuit. In: Advances in Neural Information Processing Systems 32 (2019)
25. Nguyen, X.: Convergence of latent mixing measures in finite and infinite mixture models. Ann. Statist **41**, 370–400 (2013)
26. Peyré, G., Cuturi, M.: Computational Optimal Transport. https://arxiv.org/abs/1803.00567 (2018)
27. Pitie, F., Kokaram, A.C., Dahyot, R.: Automated colour grading using colour distribution transfer. Computer Vision and Image Understanding (2007)
28. 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.: Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell **176**(4), 928–943 (2019)
29. Schmitzer, B.: Stabilized sparse scaling algorithms for entropy regularized transport problems. SIAM Journal on Scientific Computing **41**(3), A1443–A1481 (2019)
30. Seguy, V., Damodaran, B.B., Flamary, R., Courty, N., Rolet, A., Blondel, M.: Large-scale optimal transport and mapping estimation. Stat **1050**, 26 (2018)
31. Taskesen, B., Shafieezadeh-Abadeh, S., Kuhn, D.: Semi-discrete optimal transport: Hardness, regularization and numerical solution. arXiv preprint arXiv:2103.06263 (2021)
32. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: ICLR (2018)
33. Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2008)
34. Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., Solomon, J.M.: Hierarchical optimal transport for document representation. In: Advances in Neural Information Processing Systems 32 (2019)