Low-rank Optimal Transport: Approximation, Statistics and Debiasing

Anonymous Author(s)
Affiliation
Address
email

Abstract

The matching principles behind optimal transport (OT) play an increasingly important role in machine learning, a trend which can be observed when OT is used to disambiguate datasets in applications (e.g. single-cell genomics) or used to improve more complex methods (e.g. balanced attention in transformers or self-supervised learning). To scale to more challenging problems, there is a growing consensus that OT requires solvers that can operate on millions, not thousands, of points. The low-rank optimal transport (LOT) approach advocated in Scetbon et al. [2021] holds several promises in that regard, and was shown to complement more established entropic regularization approaches, being able to insert itself in more complex pipelines, such as quadratic OT. LOT restricts the search for low-cost couplings to those that have a low-nonnegative rank, yielding linear time algorithms in cases of interest. However, these promises can only be fulfilled if the LOT approach is seen as a legitimate contender to entropic regularization when compared on properties of interest, where the scorecard typically includes theoretical properties (statistical bounds, relation to other methods) or practical aspects (debiasing, hyperparameter tuning, initialization). We target each of these areas in this paper in order to cement the impact of low-rank approaches in computational OT.

1 Introduction

Optimal transport (OT) is used across data-science to put in correspondence different sets of observations. These observations may come directly from datasets, or, in more advanced applications, depict intermediate layered representations of data. OT theory provides a single grammar to describe and solve increasingly complex matching problems (linear, quadratic, regularized, unbalanced, etc...), making it gain a stake in various areas of science such as as single-cell biology Schiebinger et al. [2019], Yang et al. [2020], Demetci et al. [2020], imaging Schmitz et al. [2018], Heitz et al. [2020], Zheng et al. [2020] or neuroscience Janati et al. [2020], Koundal et al. [2020].

Regularized approaches to OT. Solving OT problems at scale poses, however, formidable challenges. The most obvious among them is computational: the Kantorovich [1942] problem on discrete measures of size n is a linear program that requires $O(n^3 \log n)$ operations to be solved. A second and equally important challenge lies in the estimation of OT in high-dimensional settings, since it suffers from the curse-of-dimensionality Fournier and Guillin [2015]. The advent of regularized approaches, such as entropic regularization [Cuturi, 2013], has pushed these boundaries thanks for faster algorithms [Chizat et al., 2020, Clason et al., 2021] and improved statistical aspects [Genevay et al., 2018a]. Despite these clear strengths, regularized OT solvers remain, however, costly as they typically scale quadratically in the number of observations.

Scaling up using low-rank couplings. While it is always intuitively possible to reduce the size of measures (e.g. using k-means) prior to solving an OT between them, a promising line of work

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

proposes to combine both [Forrow et al., 2019, Scetbon et al., 2021, 2022]. Conceptually, these low-rank approaches solve simultaneously both an optimal clustering/aggregation strategy with the computation of an effective transport. This intuition rests on an explicit factorization of couplings into two sub-couplings. This has several computational benefits, since its computational cost becomes linear in n if the ground cost matrix seeded to the OT problem has itself a low-rank. While these computational improvements, mostly demonstrated empirically, hold several promises, the theoretical properties of these methods are not yet well established. This stands in stark contrast to the Sinkhorn approach, which is comparatively much better understood.

Our Contributions. The goal of this paper is to advance our knowledge, understanding and practical ability to leverage low-rank factorizations in OT. This paper provides five contributions, targeting theoretical and practical properties of LOT: *(i)* We derive the rate of convergence of the low-rank OT to the true OT with respect to the non-nnegative rank parameter. *(ii)* We provide an upper-bound of the (statistical) sample complexity when estimating LOT using the plug-in estimator, and attain a parametric rate $\mathcal{O}(\sqrt{1/n})$ that is independent of the dimension. *(iii)* We introduce a debiased version of LOT: as the Sinkhorn divergence [Feydy et al., 2018], we show that it can interpolate between the maximum mean discrepancy [Gretton et al., 2012] and OT, it is nonnegative and it metrizes the weak convergence. *(iv)* We exhibit links between the bias induced by the low-rank factorization and clustering methods. *(v)* We propose solutions to practical issues such as the tuning the step-length or the choice of the initialization when applying the algorithm proposed in [Scetbon et al., 2021].

Notations. We consider $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ two nonempty compact Polish spaces and we denote $\mathcal{M}_1^+(\mathcal{X})$ (resp. $\mathcal{M}_1^+(\mathcal{Y})$) the space of positive Radon probability measures on $\mathcal{X}$ (resp. $\mathcal{Y}$). For all $n \geq 1$, we denote Δ_n the probability simplex of size n and Δ_n^* the subset of Δ_n of positive histograms. We write $\mathbf{1}_n \triangleq (1, \dots, 1)^T \in \mathbb{R}^n$ and we denote similarly $\|\cdot\|_2$ the Euclidean norm and the Euclidean distance induced by this norm depending on the context.

2 Background on Low-rank Optimal Transport

Let $\mu \in \mathcal{M}_1^+(\mathcal{X})$, $\nu \in \mathcal{M}_1^+(\mathcal{Y})$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a nonnegative and continuous function. The Kantorovitch formulation of optimal transport between μ and ν is defined by

$$\mathrm{OT}_c(\mu, \nu) \triangleq \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) \tag{1}$$

where the feasible set is the set of distributions over the product space $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν:

$$\Pi(\mu, \nu) \triangleq \left\{\pi \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y}) \text{ s.t. } P_{1\#\pi} = \mu, P_{2\#\pi} = \nu\right\} ,$$

with $P_{1\#\pi}$ (resp. $P_{2\#\pi}$), the pushforward probability measure of π using the projection maps $P_1(x, y) = x$ (resp. $P_2(x, y) = y$). When there exists an optimal coupling solution of (1) supported on a graph of a function, we call such function a Monge map. In the discrete setting, one can reformulate the optimal transport problem as a linear program over the space of nonnegative matrices satisfying the marginal constraints. More precisely, let a and b be respectively elements of Δ_n^* and Δ_m^* and let also $\mathbf{X} \triangleq \{x_1, \dots, x_n\}$ and $\mathbf{Y} \triangleq \{y_1, \dots, y_m\}$ be respectively two subsets of $\mathcal{X}$ and $\mathcal{Y}$. By denoting $\mu_{a,\mathbf{X}} \triangleq \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu_{b,\mathbf{Y}} \triangleq \sum_{j=1}^m b_j \delta_{y_j}$ the two discrete distributions associated and writing $C \triangleq [c(x_i, y_j)]_{i,j}$, the discrete optimal transport problem can be formulated as

$$\mathrm{OT}_c(\mu_{a,\mathbf{X}}, \nu_{b,\mathbf{Y}}) = \min_{P \in \Pi_{a,b}} \langle C, P\rangle \quad \text{where } \Pi_{a,b} \triangleq \{P \in \mathbb{R}_+^{n\times m} \text{ s.t. } P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\} . \tag{2}$$

Scetbon et al. [2021] propose to constrain the discrete optimal transport problem to couplings that have a low-nonnegative rank:

Definition 1. *Given $M \in \mathbb{R}_+^{n\times m}$, the nonnegative rank of M is defined by:* $\mathrm{rk}_+(M) \triangleq \min\{q | M = \sum_{i=1}^q R_i, \forall i, \mathrm{rk}(R_i) = 1, R_i \geq 0\}$.

Note that for any $M \in \mathbb{R}_+^{n\times m}$, we always have that $\mathrm{rk}_+(M) \leq \min(n, m)$. For $r \geq 1$, we consider the set of couplings satisfying marginal constaints with nonnegative-rank of at most r as $\Pi_{a,b}(r) \triangleq \{P \in \Pi_{a,b}, \ \mathrm{rk}_+(P) \leq r\}$. The discrete Low-rank Optimal Transport (LOT) problem is defined by:

$$\mathrm{LOT}_{r,c}(\mu_{a,\mathbf{X}}, \nu_{b,\mathbf{Y}}) \triangleq \min_{P \in \Pi_{a,b}(r)} \langle C, P\rangle . \tag{3}$$

To solve this problem, Scetbon et al. [2021] show that Problem (3) is equivalent to

$$\min_{(Q,R,g)\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle \tag{4}$$

where $\mathcal{C}_1(a,b,r) \triangleq \left\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times(\mathbb{R}_+^*)^r \text{ s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b\right\}$ and $\mathcal{C}_2(r) \triangleq \left\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{r} \text{ s.t. } Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g\right\}$. They propose to solve it using a mirror descent scheme and prove the non-asymptotic stationary convergence of their algorithm. While Scetbon et al. [2021] only focus on the discrete setting, we consider here its extension for arbitrary probability measures. Following [Forrow et al., 2019], we define the set of rank-r couplings satisfying marginal constraints by:

$$\Pi_r(\mu,\nu) \triangleq \{\pi\in\Pi(\mu,\nu) : \exists (\mu_i)_{i=1}^r\in\mathcal{M}_1^+(\mathcal{X})^r,\ (\nu_i)_{i=1}^r\in\mathcal{M}_1^+(\mathcal{Y})^r,\ \lambda\in\Delta_r^* \text{ s.t. } \pi = \sum_{i=1}^r \lambda_i\mu_i\otimes\nu_i\}\ .$$

This more general definition of LOT between $\mu\in\mathcal{M}_1^+(\mathcal{X})$ and $\nu\in\mathcal{M}_1^+(\mathcal{Y})$ reads:

$$\mathrm{LOT}_{r,c}(\mu,\nu) \triangleq \inf_{\pi\in\Pi_r(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y)\ . \tag{5}$$

Note that this definition of $\mathrm{LOT}_{r,c}$ is consistent as it coincides with the one defined in (3) on discrete probability measures. Observe also that $\Pi_r(\mu,\nu)$ is compact for the weak topology and therefore the infimum in (5) is attained. See Appendix A for more details.

3 Approximation Error of LOT to original OT as a function of rank

Our goal in this section is to obtain a control of the error induced by the low-rank constraint when trying to approximate the true OT cost. We provide first a control of the approximation error in the discrete setting. The proof is given in Appendix B.1.

Proposition 1. *Let $n,m\geq 2$, $\mathbf{X}\triangleq\{x_1,\dots,x_n\}\subset\mathcal{X}$, $\mathbf{Y}\triangleq\{y_1,\dots,y_m\}\subset\mathcal{Y}$ and $a\in\Delta_n^*$ and $b\in\Delta_m^*$. Then for $2\leq r\leq\min(n,m)$, we have that*

$$|\mathrm{LOT}_{r,c}(\mu_{a,\mathbf{X}},\nu_{b,\mathbf{Y}}) - \mathrm{OT}_c(\mu_{a,\mathbf{X}},\nu_{b,\mathbf{Y}})| \leq \|C\|_\infty\, ln(\min(n,m)/(r-1))$$

Remark 1. *Note that this result improves the control obtained in [Liu et al., 2021], where they obtain that $|\mathrm{LOT}_{r,c}(\mu_{a,\mathbf{X}},\nu_{b,\mathbf{Y}}) - \mathrm{OT}_c(\mu_{a,\mathbf{X}},\nu_{b,\mathbf{Y}})| \lesssim \|C\|_\infty\sqrt{nm}(\min(n,m)-r)$ as we have for any $z,z'\geq 1$, $|\ln(z)-\ln(z')|\leq|z-z'|$.*

It is in fact possible to obtain another control of the approximation error by partitioning the space where the measures are supported. For that purpose let us introduce the notion of entropy numbers.

Definition 2. *Let $(\mathcal{Z},d)$ a metric space, $\mathcal{W}\subset\mathcal{Z}$ and $k\geq 1$ an integer. Then by denoting $B_{\mathcal{Z}}(z,\varepsilon)\triangleq\{y\in\mathcal{Z}:\ d(z,y)\leq\varepsilon\}$, we define the k-th (dyadic) entropy number of $\mathcal{W}$ as*

$$\mathcal{N}_k(\mathcal{W},d)\triangleq\inf\{\varepsilon \text{ s.t. } \exists\, z_1,\dots,z_{2^k}\in\mathcal{Z}:\ \ \mathcal{W}\subset\cup_{i=1}^{2^k}B_{\mathcal{Z}}(z_i,\varepsilon)\}$$

For example, any compact set $\mathcal{W}$ of $\mathbb{R}^d$ admits finite entropy numbers, and by denoting $R\triangleq\sup_{w\in\mathcal{W}}\|w\|_2$, we have $\mathcal{N}_k(\mathcal{W},\|\cdot\|_2)\leq 4R/2^{k/d}$. We obtain next a control the approximation error of $\mathrm{LOT}_{r,c}$ to the true OT cost using entropy numbers (see proof in Appendix B.2).

Proposition 2. *Let $\mu\in\mathcal{M}_1^+(\mathcal{X})$, $\nu\in\mathcal{M}_1^+(\mathcal{Y})$ and assume that c is L-Lipschitz w.r.t. x and y. Then for any $r\geq 1$, we have*

$$|\mathrm{LOT}_{r,c}(\mu,\nu) - \mathrm{OT}_c(\mu,\nu)| \leq 2L\max(\mathcal{N}_{\lfloor\log_2(\lfloor\sqrt{r}\rfloor)\rfloor}(\mathcal{X},d_{\mathcal{X}}),\mathcal{N}_{\lfloor\log_2(\lfloor\sqrt{r}\rfloor)\rfloor}(\mathcal{Y},d_{\mathcal{Y}}))$$

This results in the following bound for the p-Wasserstein distance for any $p\geq 1$ on $\mathbb{R}^d$.

Corollary 1. *Let $d\geq 1$, $p\geq 1$, $\mathcal{X}$ a compact subspace of $\mathbb{R}^d$ and $\mu,\nu\in\mathcal{M}_1^+(\mathcal{X})$. By denoting $R\triangleq\sup_{x\in\mathcal{X}}\|x\|_2$, we obtain that for any $r\geq 1$,*

$$|\mathrm{LOT}_{r,\|\cdot\|_2^p}(\mu,\nu) - \mathrm{OT}_{\|\cdot\|_2^p}(\mu,\nu)| \leq 4dp\frac{(8R^2)^p}{r^{p/2d}}\ .$$

As per the Proof of Proposition 2 we can provide a tighter control, assuming a Monge map exists.

Corollary 2. *Under the same assumptions of Proposition 2 and by assuming in addition that there exists a Monge map solving $\mathrm{OT}_c(\mu,\nu)$, we obtain that for any $r \geq 1$,*

$$|\mathrm{LOT}_{r,c}(\mu,\nu) - \mathrm{OT}_c(\mu,\nu)| \leq L\mathcal{N}_{\lfloor \log_2(r) \rfloor}(\mathcal{Y}, d_{\mathcal{Y}})$$

When $\mathcal{X} = \mathcal{Y}$ are a subspaces of $\mathbb{R}^d$, a sufficient condition for a Monge map to exists is that either μ or ν is absolutely continuous with respect to the Lebesgue measure and that c is of the form $h(x-y)$ where $h : \mathcal{X} \to \mathbb{R}_+$ is a strictly convex function [Santambrogio, 2015, Theorem 1.17]. Therefore if μ is absolutely continuous with respect to the Lebesgue measure, we obtain that for any $r, p \geq 1$

$$|\mathrm{LOT}_{r,\|\cdot\|_2^p}(\mu,\nu) - \mathrm{OT}_{\|\cdot\|_2^p}(\mu,\nu)| \leq 2dp\frac{(8R^2)^p}{r^{p/d}} \ .$$

4 Sample Complexity of LOT

We now focus on the statistical performance of the plug-in estimator for LOT. In the following we assume that $\mathcal{X} = \mathcal{Y}$ for simplicity. Given $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$, we denote the empirical measures associated $\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where $(X_i, Y_i)_{i=1}^n$ are sampled independently from $\mu \otimes \nu$. We consider the plug-in estimator defined as $\mathrm{LOT}_{r,c}(\hat{\mu}_n, \hat{\nu}_n)$, and we aim at quantifying the rate at which it converges towards the true low-rank optimal transport cost $\mathrm{LOT}_{r,c}(\mu,\nu)$. Before doing so, in the next Proposition we show that this estimator is consistent on compact spaces. The proof is given in Appendix B.3.

Proposition 3. *Let $r \geq 1$ and $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$, then* $\mathrm{LOT}_{r,c}(\hat{\mu}_n, \hat{\nu}_n) \xrightarrow[n\to+\infty]{} \mathrm{LOT}_{r,c}(\mu,\nu)$ *a.s.*

Next we aim at obtaining the convergence rates of our plug-in estimator. In the following Proposition, we obtain a non-asymptotic upper-bound of the statistical error. See Appendix B.4 for the proof.

Proposition 4. *Let $r \geq 1$ and $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$. Then, there exists a constant K_r such that for any $\delta > 0$ and $n \geq 1$, we have, with a probability of at least $1 - 2\delta$, that*

$$\mathrm{LOT}_{r,c}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{LOT}_{r,c}(\mu,\nu) \leq 11\|c\|_\infty\sqrt{\frac{r}{n}} + K_r\|c\|_\infty\left[\sqrt{\frac{\log(40/\delta)}{n}} + \frac{\sqrt{r}\log(40/\delta)}{n}\right]$$

This result shows that the estimation of $\mathrm{LOT}_{r,c}$ is independent of the dimension and can be performed on general compact metric spaces. Therefore, $\mathrm{LOT}_{r,c}$ presents a clear statistical benefit compared to unregularized OT which suffers from the curse of dimensionality [Dudley, 1969]. In addition, LOT compares favorably to known results on entropic optimal transport. The rate of entropy regularized OT does not depend on the ambient dimension with respect to n, but carries an exponential dependence in dimension with respect to the regularization parameter ε [Mena and Niles-Weed, 2019]. By contrast, the term associated with the nonnegative rank r has no direct dependence on dimension.

Our next aim is to obtain an explicit rate with respect to r and n. In Proposition 4, we cannot control explicitly K_r in the general setting. Indeed, in our proof, we obtain that $K_r \triangleq 14/\min_i \lambda_i^*$ where $(\lambda_i^*)_{i=1}^r \in \Delta_r^*$ are the weights involved in the decomposition of one optimal solution of the true $\mathrm{LOT}_{r,c}(\mu,\nu)$. Therefore the control of K_r requires additional assumptions on the optimal solutions of $\mathrm{LOT}_{r,c}(\mu,\nu)$. In the following Proposition, we obtain an explicit upper-bound of the statistical error with respect to r and n in the asymptotic regime.

Proposition 5. *Let $r \geq 1$, $\delta > 0$ and $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$. Then there exists a constant $N_{r,\delta}$ such that if $n \geq N_{r,\delta}$ then with a probability of at least $1 - 2\delta$, we have*

$$\mathrm{LOT}_{r,c}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{LOT}_{r,c}(\mu,\nu) \leq 11\|c\|_\infty\sqrt{\frac{r}{n}} + 77\|c\|_\infty\sqrt{\frac{\log(40/\delta)}{n}} \ .$$

Note that one cannot recover the result obtained in Proposition 5 from the one obtained in Proposition in 4 as we have that $K_r \geq 14r \xrightarrow[r\to+\infty]{} +\infty$. In order to prove the above result, we use an extension of the McDiarmid's inequality when differences are bounded with high probability [Kutin, 2002]. See proof in Appendix B.5 for more details.

5 Debiased Formulation of LOT

We introduce here the debiased formulation of $\mathrm{LOT}_{r,c}$ and show that it is able to distinguish two distributions, metrize the convergence in law and can be used as a new objective in order to learn distributions. We focus next on the debiasing terms involving measures with themselves $\mathrm{LOT}_{r,c}(\mu,\mu)$ in this new divergence, and show that they can be interpreted as defining a new clustering method generalizing k-means for any geometry.

5.1 On the Proprieties of the Debiased Low-rank Optimal Transport

When it comes to learn (or generate) a distribution in ML applications given samples, it is crucial to consider a divergence that is able to distinguish between two distributions and metrize the convergence in law. In general, $\mathrm{LOT}_{r,c}(\mu,\mu) \neq 0$ and the minimum of $\mathrm{LOT}_{r,c}(\nu,\mu)$ with respect to ν will not necessarily recover μ. In order to alleviate this issue we propose a debiased version of $\mathrm{LOT}_{r,c}$ defined for any $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$ as

$$\mathrm{DLOT}_{r,c}(\mu,\nu) \triangleq \mathrm{LOT}_{r,c}(\mu,\nu) - \frac{1}{2}[\mathrm{LOT}_{r,c}(\mu,\mu) + \mathrm{LOT}_{r,c}(\nu,\nu)] \ .$$

Note that $\mathrm{DLOT}_{r,c}(\nu,\nu) = 0$. In the next Proposition, we show that, as the Sinkhorn divergence [Genevay et al., 2018b, Feydy et al., 2018], $\mathrm{DLOT}_{r,c}$ interpolates between the Maximum Mean Discrepancy (MMD) and OT. See proof in Appendix B.6.

Proposition 6. *Let $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$. Let us assume that c is symmetric, then we have*

$$\mathrm{DLOT}_{1,c}(\mu,\nu) = \frac{1}{2}\int_{\mathcal{X}^2} -c(x,y)d[\mu-\nu]\otimes d[\mu-\nu](x,y) \ .$$

If in addition we assume the c is Lipschitz w.r.t to x and y, then we have

$$\mathrm{DLOT}_{r,c}(\mu,\nu) \xrightarrow[r\to+\infty]{} \mathrm{OT}_c(\mu,\nu) \ .$$

Next, we aim at showing some useful properties of the debiased low-rank OT for machine learning applications. For that purpose, let us first recall some definitions.

Definition 3. *We say that the cost $c : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ is a semimetric on $\mathcal{X}$ if for all $x,x' \in \mathcal{X}$, $c(x,x') = c(x',x)$ and $c(x,x') = 0$ if and only if $x = x'$. In addition we say that c has a negative type if $\forall n \geq 2$, $x_1,\dots,x_n \in \mathcal{X}$ and $\alpha_1,\dots,\alpha_n \in \mathbb{R}$ such that $\sum_{i=1}^n \alpha_i = 0$, $\sum_{i,j=1}^n \alpha_i\alpha_j c(x_i,x_j) \leq 0$. We say also that c has a strong negative type if for all $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$, $\mu \neq \nu \implies \int_{\mathcal{X}^2} c(x,y)d[\mu-\nu]\otimes[\mu-\nu] < 0$.*

Note that if c has a strong negative type, then c has a negative type too. For example, all Euclidean spaces and even separable Hilbert spaces endowed with the metric induced by their inner products have strong negative type. Also, on $\mathbb{R}^d$, the squared Euclidean distance has a negative type [Sejdinovic et al., 2013].

We can now provide stronger geometric guarantees for $\mathrm{DLOT}_{r,c}$. In the next Proposition, we show that for a large class of cost functions, $\mathrm{DLOT}_{r,c}$ is nonnegative, able to distinguish two distributions, and metrizes the convergence in law. The proof is given in Appendix B.7.

Proposition 7. *Let $r \geq 1$, and let us assume that c is a semimetric of negative type. Then for all $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$, we have that*

$$\mathit{DLOT}_r(\mu,\nu) \geq 0 \ .$$

In addition, if c has strong negative type then we have also that

$$\begin{aligned} \mathrm{DLOT}_{r,c}(\mu,\nu) = 0 &\iff \mu = \nu \ \textit{and} \\ \mu_n \to \mu &\iff \mathrm{DLOT}_{r,c}(\mu_n,\mu) \to 0 \ . \end{aligned}$$

where the convergence of the sequence of probability measures considered is the convergence in law.

Observe that when c has strong negative type, $\nu \to \mathrm{DLOT}_{r,c}(\nu,\mu) \geq 0$ and it admits a unique global minimizer at $\nu = \mu$. Therefore, $\mathrm{DLOT}_{r,c}$ has desirable properties to be used as a loss.

5.2 Low-Rank Transport Bias and Clustering

We turn next to the debiasing terms appearing in DLOT and exhibit links between LOT and clustering methods. Indeed, in the discrete setting, the low-rank bias of a probability measure μ defined as $\mathrm{LOT}_{k,c}(\mu,\mu)$ can be seen as a generalized version of the k-means method for any geometry. In the next Proposition we obtain a new formulation of $\mathrm{LOT}_{k,c}(\mu,\mu)$ viewed as a general clustering method on arbitrary metric space. See proof in Appendix B.8.

Proposition 8. *Let $n \geq k \geq 1$, $\mathbf{X} \triangleq \{x_1,\dots,x_n\} \subset \mathcal{X}$ and $a \in \Delta_n^*$. If c is a semimetric of negative type, then by denoting $C = (c(x_i,x_j))_{i,j}$, we have that*

$$\mathrm{LOT}_{k,c}(\mu_{a,\mathbf{X}},\mu_{a,\mathbf{X}}) = \min_{Q} \langle C, Q\mathit{Diag}(1/Q^T\mathbf{1}_n)Q^T\rangle \;\; s.t. \;\; Q \in \mathbb{R}_+^{n\times k} \;\;,\;\; Q\mathbf{1}_k = a \;. \tag{6}$$

Let us now explain in more details the link between (15) and k-means. When $\mathcal{X}$ is a subspace of $\mathbb{R}^d$, c is the squared Euclidean distance and $a = \mathbf{1}_n$, we recover exactly the k-means algorithm.

Corollary 3. *Let $n \geq k \geq 1$ and $\mathbf{X} \triangleq \{x_1,\dots,x_n\} \subset \mathbb{R}^d$. We have that*

$$\mathrm{LOT}_{k,\|\cdot\|_2^2}(\mu_{\mathbf{1}_n,\mathbf{X}},\mu_{a,\mathbf{X}}) = 2 \min_{Q,z_1,\dots,z_k} \sum_{i=1}^{n}\sum_{q=1}^{k} Q_{i,q}\|x_i - z_q\|_2^2 \;\; s.t. \;\; Q \in \{0,1\}^{n\times k},\;\; Q\mathbf{1}_k = \mathbf{1}_n \;.$$

In the general setting, solving $\mathrm{LOT}_{k,c}(\mu_{a,\mathbf{X}},\mu_{a,\mathbf{X}})$ for a given geometry c, and a prescribed histogram a offers a new clustering method where the assignment of the points to the clusters is determined by the matrix Q^* solution of (15).

6 Computing LOT: Adaptive Stepsizes and Better Initializations

We target in this section practical issues that arises when using [Scetbon et al., 2021, Algo.3] to solve (4). Scetbon et al. [2021] propose to apply a mirror descent scheme with respect to the Kullback-Leibler divergence which boils down to solve at each iteration $k \geq 0$ the following convex problem:

$$(Q_{k+1},R_{k+1},g_{k+1}) \triangleq \operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)} \mathrm{KL}(\boldsymbol{\zeta},\boldsymbol{\xi}_k) \tag{7}$$

where $(Q_0,R_0,g_0) \in \mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)$, $\boldsymbol{\xi}_k \triangleq (\xi_k^{(1)},\xi_k^{(2)},\xi_k^{(3)})$, $\xi_k^{(1)} \triangleq Q_k \odot \exp(-\gamma_k C R_k \,\mathrm{diag}(1/g_k))$, $\xi_k^{(2)} \triangleq R_k \odot \exp(-\gamma_k C^T Q_k \,\mathrm{diag}(1/g_k))$, $\xi_k^{(3)} \triangleq g_k \odot \exp(\gamma_k\omega_k/g_k^2)$ with $[\omega_k]_i \triangleq [Q_k^T C R_k]_{i,i}$ for all $i \in \{1,\dots,r\}$, $\mathrm{KL}(\mathbf{w},\mathbf{r}) \triangleq \sum_i w_i \log(w_i/r_i)$ and $(\gamma_k)_{k\geq 0}$ is a sequence of positive step sizes. In the general setting, each iteration of their algorithm requires $\mathcal{O}(nmr)$ operations and when the ground cost matrix C admits a low-rank factorization of the form $C = AB^T$ where $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{m\times d}$ with $d \ll \min(n,m)$, then the total complexity per iteration becomes linear $\mathcal{O}((n+m)rd)$. In the following we investigate two practical aspects of the algorithm: the choice of the step sizes and the initialization.

Adaptive choice of γ_k. Scetbon et al. [2021] show experimentally that the choice of $(\gamma_k)_{k\geq 0}$ does not impact the solution obtained upon convergence, but rather the speed at which it is attained. Indeed the larger γ_k is, the faster the algorithm will converge. As a result, their algorithm simply relies on a fixed γ schedule. However, the range of admissible γ depends on the problem considered and it may vary from one problem to another. It may be of particular interest for practitioners to have a generic range of admissible values for γ independently of the considered problem, in order to alleviate parameter tuning issues. We propose to consider instead an adaptive choice of $(\gamma_k)_{k\geq 0}$ along iterations. D'Orazio et al. [2021], Bayandina et al. [2018] have proposed adaptive mirror descent schemes where, at each iteration, the step-size is normalized by the squared dual-norm of the gradient. Applying such a strategy in our case amounts to consider at each iteration

$$\gamma_k = \frac{\gamma}{\| \left(CR\,\mathrm{diag}(1/g), C^TQ\,\mathrm{diag}(1/g), -\mathcal{D}(Q^TRC)/g^2\right)\|_\infty^2} \;.$$

where the initial $\gamma > 0$ is fixed. We recommend to set such as global $\gamma \in [1,10]$, and observe that this range works whatever the problem considered.

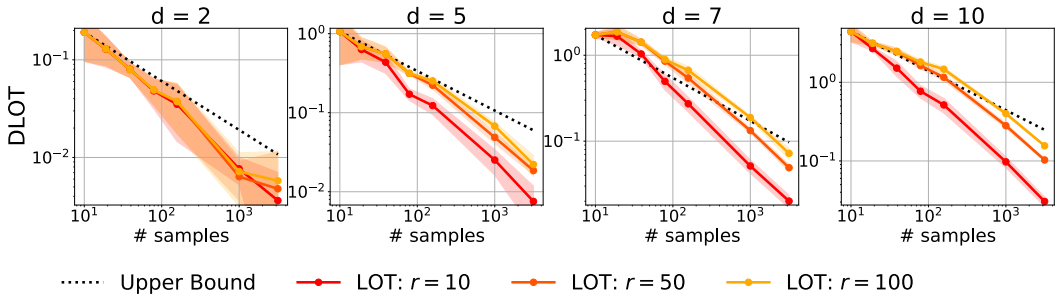

Figure 1: In this experiment, we consider a mixture of 10 anisotropic Gaussians and we plot the value of $\mathrm{DLOT}_{r,c}$ between two independent empirical measures associated to this mixture when varying the number of samples n and the dimension d for multiple ranks r. The ground cost considered is the squared Euclidean distance. Note that $\mathrm{LOT}_r(\mu,\mu) \neq 0$ and therefore we use $\mathrm{DLOT}_{r,c}(\mu,\mu)$ instead to evaluate the rates. Each point has been obtained by repeating 10 times the experiment. We compare the empirical rates obtained with the theoretical one derived in Proposition 4. We observe that our theoretical results match the empirical ones and, as expected, the rates do not depend on d.

On the choice of the initialization. As $\mathrm{LOT}_{r,c}$ (4) is a non-convex optimization problem, the question of choosing an efficient initialization arises in practice. Scetbon et al. [2021] show experimentally that the convergence of the algorithm does not depend on the initalization chosen if no stopping criterion is used. Indeed, their experimental findings support that only well behaved local minimas are attractive. However, in practice one needs to use a stopping criterion in order to terminate the algorithm. We do observe in many instances that using trivial initializers may result in spurious local minima, which trigger the stopping criterion early on and prevent the algorithm to reach a good solution. Based on various experimentations, we propose to consider a novel initialization of the algorithm. Our initialization aims at being close to a well-behaved local minimum by clustering the input measures. When the measures are supported on Euclidean space, we propose to find r centroids $(z_i)_{i=1}^r$ of one of the two input discrete probability measures using k-means and to solve the following convex barycenter problem:

$$\min_{Q,R} \langle C_{X,Z}, Q\rangle + \langle C_{Y,Z}, R\rangle - \varepsilon H(Q) - \varepsilon H(R) \;\text{ s.t. }\; Q\mathbf{1}_n = a,\;\; R\mathbf{1}_n = b,\;\; Q^T\mathbf{1}_r = R^T\mathbf{1}_r \quad (8)$$

where $C_{X,Z} = (c(x_i, z_j))_{i,j}$, $C_{Y,Z} = (c(y_i, z_j))_{i,j}$, and $H(P) = -\sum_{i,j} P_{i,j}(\log(P_{i,j} - 1)$. In practice we fix $\varepsilon = 1/10$ and we then initialize $\mathrm{LOT}_{r,c}$ using (Q, R) solution of (8) and $g \triangleq Q^T\mathbf{1}_r (= R^T\mathbf{1}_r)$. Note that (Q, R, g) is an admissible initialization and finding the centroids as well as solving (8) requires $\mathcal{O}((n+m)r)$ algebraic operations. Therefore such initialization does not change the total complexity of the algorithm.

In the general (non-Euclidean) case, we propose to initialize the algorithm by applying our generalized k-means approach defined in (15) on each input measure where we fix the common marginal to be $g = \mathbf{1}_r/r$. More precisely, by denoting $C_{X,X} = (c(x_i, x_j))_{i,j}$ and $C_{Y,Y} = (c(y_i, y_j))_{i,j}$, we initialize the algorithm by solving:

$$\begin{aligned} &Q \in \operatorname*{argmin}_{Q} \langle C_{X,X}, Q\mathrm{Diag}(1/Q^T\mathbf{1}_n)Q^T\rangle \;\text{ s.t. }\; Q \in \mathbb{R}_+^{n\times k}\;,\; Q\mathbf{1}_k = a,\;\; Q^T\mathbf{1}_n = \mathbf{1}_r/r \;\text{ and} \\ &R \in \operatorname*{argmin}_{R} \langle C_{Y,Y}, R\mathrm{Diag}(1/R^T\mathbf{1}_m)R^T\rangle \;\text{ s.t. }\; R \in \mathbb{R}_+^{m\times k}\;,\; R\mathbf{1}_k = b,\;\; R^T\mathbf{1}_n = \mathbf{1}_r/r\;. \end{aligned} \quad (9)$$

Note that again the (Q, R, g) obtained is an admissible initialization and the complexity of solving (9) is of the same order as solving (4), thus the total complexity of the algorithm remains the same.

7 Experiments

In this section, we illustrate experimentally our theoretical findings and show how our initialization provide practical improvements. For that purpose we consider 3 synthetic problems and one real world dataset to: *(i)* provide illustrations on the statistical rates of $\mathrm{LOT}_{r,c}$, *(ii)* exhibit the gradient flow of the debiased formulation $\mathrm{DLOT}_{r,c}$, *(iii)* use the clustering method induced by $\mathrm{LOT}_{r,c}$, and *(iv)* show the effect of the initialization. All experiments were run on a MacBook Pro 2019 laptop.

Statistical rates. We aim at showing the statistical rates of the plug-in estimator of $\mathrm{LOT}_{r,c}$. As $\mathrm{LOT}_{r,c}(\mu,\mu) \neq 0$ and as we do not have access to the this value given samples from μ, we consider

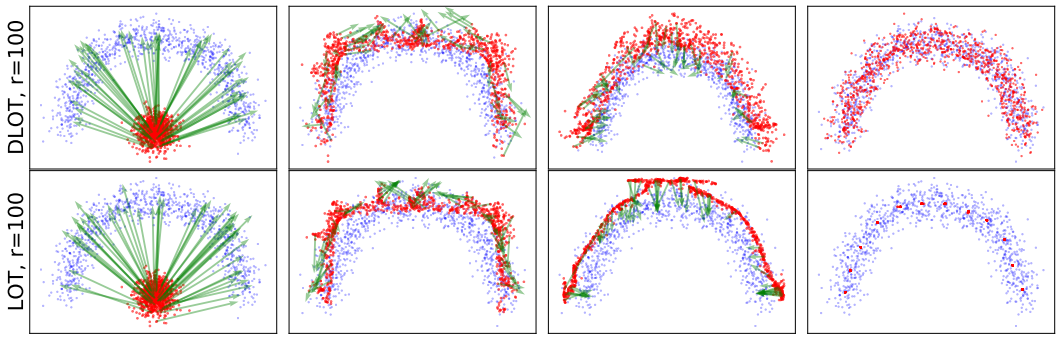

Figure 2: We compare the gradient flows $(\mu_t)_{t\geq 0}$ (in red) starting from a Gaussian distribution, μ_0, to a moon shape distribution (in blue), ν, in 2D when minimizing either $L(\mu) \triangleq \text{DLOT}_{r,c}(\mu,\nu)$ or $L(\mu) \triangleq \text{LOT}_{r,c}(\mu,\nu)$. The ground cost is the squared Euclidean distance and we fix $r = 100$. We consider 1000 samples from each distribution and and we plot the evolution of the probability measure obtained along the iterations of a gradient descent scheme. We also display in green the vector field in the descent direction. We show that the debiased version allows to recover the target distribution while $\text{LOT}_{r,c}$ is learning a biased version with a low-rank structure.

instead the debiased version of the low-rank optimal transport, $\text{DLOT}_{r,c}$. In figure 1, we show that the empiricial rates match the theoretical bound obtained in Proposition 4. In particular, we show that that these rates does not depend on the dimension of the ground space.

Gradient Flows using DLOT. We illustrate here a practical use of DLOT for ML application. In figure 2, we consider $Y_1, \dots, Y_n$ independent samples from a moon shape distribution in 2D, and by denoting $\hat{\nu}_n$ the empirical measure associated, we show the iterates obtained by a gradient descent scheme on the following optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times 2}} \text{DLOT}_{r,c}(\mu_{\mathbf{1}_n/n,\mathbf{X}}, \hat{\nu}_n) \ .$$

We initialize the algorithm using $n = 1000$ samples drawn from a Gaussian distribution. We show that the gradient flow of our debiased version is able to recover the target distribution. We also compare it with the gradient flow of the biased version (LOT) and show that it fails to reproduce the target distribution as it is learning a biased one with a low-rank structure.

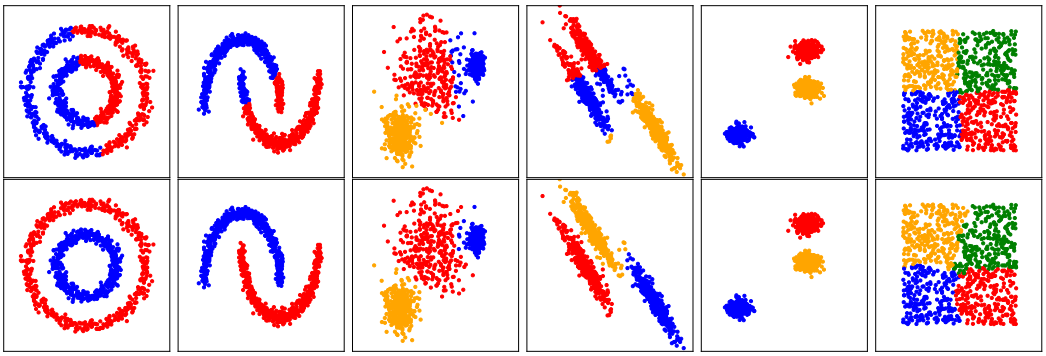

Figure 3: In this experiment, we draw 1000 samples from multiple distributions from the python package scikit-learn [Pedregosa et al., 2011] and we apply the method proposed in (15) for two different costs: in the top row we consider the squared Euclidean distance while in the bottom row, we consider the shortest path distance on the graph associated with the ground cost $c(x,y) = 1 - k(x,y)$ where k is a Gaussian kernel. In the two first problem (starting from the left), we fix $r = 2$, in the next three problem we fix $r = 3$ and in the last one we fix $r = 4$. We observe that the flexibility of our method allows to recover the clustering for a well chosen ground cost.

Application to Clustering. In this experiment we show some applications of the clustering method induced by $\text{LOT}_{r,c}$. In figure 3, we consider 6 datasets with different structure and we aim at recovering the clusters using (15) for some well chosen costs. We compare the clusters obtained

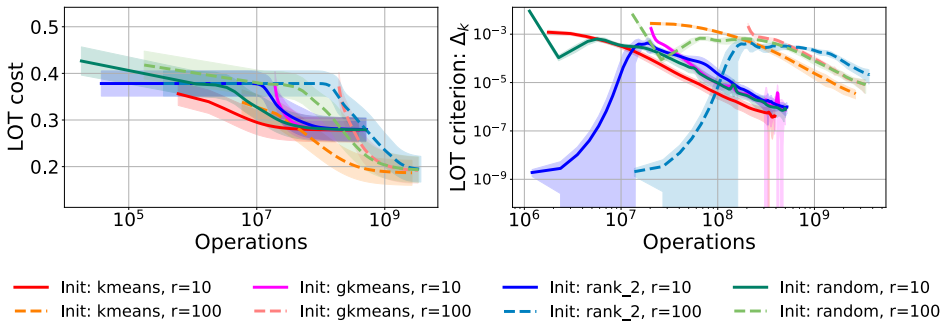

Figure 4: In this experiment, we consider the Newsgroup20 dataset [Pedregosa et al., 2011] constituted of texts and we embed them into distributions in 50D using the same pre-processing steps as in [Cuturi et al., 2022]. We compare different initialization when applying the algorithm of [Scetbon et al., 2021] to compare random texts viewed as distributions for multiple choices of rank r. The ground cost considered in the squared Euclidean distance. We repeat the experiments 50 times by sampling randomly multiple problems of similar size ($\simeq$ 250 samples). Note that we normalize the cost matrix by its maximum value in order to have comparable LOT cost. We consider 4 different initialization: the one using k-means algorithm (8), the one using the generalized k-means (9), the rank-2 initialization [Scetbon et al., 2021] and a random initialization where Q, R and g are drawn from Gaussians. We compare both the cost value and the criterion value (Δ_k) along the iterations of the MD scheme. First we observe that whatever the initialization considered, the algorithm converges toward the same value. In addition, we observe that both k-means and general k-means are able to initialize well the algorithm by avoiding bad local minima at initialization while the two other initialization are close to spurious local minima at initialization.

when considering either the squared Euclidean cost (which amounts at applying the k-means) and the shortest-path distance on the data viewed as a graph. We show that our method is able to recover the clusters on these settings for well chosen costs and therefore the proposed algorithm in Scetbon et al. [2021] can be seen as a new alternative in order to clusterize data.

Effect of the Initialization. Our goal here is to show the effect of the initialization. In figure 4, we display the evolution of the cost as well as the value of the stopping criterion along the iterations of the MD scheme solving (4) when considering different initialization. Recall that the stopping criterion introduced in [Scetbon et al., 2021] is defined for all $k \geq 1$ by

$$\Delta_k \triangleq \frac{1}{\gamma_k^2}(\mathrm{KL}((Q_k, R_k, g_k), (Q_{k-1}, R_{k-1}, g_{k-1})) + \mathrm{KL}((Q_{k-1}, R_{k-1}, g_{k-1}), (Q_k, R_k, g_k))),$$

where $((Q_k, R_k, g_k))_{k\geq 0}$ is the sequence solution of (7). We show that whatever the initialization chosen, the algorithm manages to converge to an efficient solution if no stopping criterion is used. However, the choice of the initialization may impact the termination of the algorithm as some initialization might be too close to some spurious local minima. We show also that the initialization we propose in (8) and (9) are sufficiently far away from bad local minima and allow the algorithm to converge directly toward the desired solution.

8 Conclusion

We assembled in this work theoretical and practical arguments to support low-rank factorizations for OT. We have presented two controls: one concerning the approximation error to the true optimal transport and another concerning the statistical rates of the plug-in estimator. The latter is showed to be independent of the dimension, which is of particular interest when studying OT in ML settings. We have motivated further the use of LOT as a loss by introducing its debiased version and showed that it possesses desirable properties: positivity and metrization of the convergence in law. We have also presented the links between the bias induced by such regularization and clustering methods, and studied empirically the effects of hyperparameters involved in the practical estimation of LOT. The strong theoretical foundations provided in this paper motivate further studies of the empirical behaviour of LOT estimator, notably on finding suitable local minima and on improvements on the convergence of the MD scheme using other adaptive choices for step sizes.

References

Anastasia Bayandina, Pavel Dvurechensky, Alexander Gasnikov, Fedor Stonyakin, and Alexander Titov. Mirror descent and convex optimization problems with non-smooth inequality constraints. In *Large-scale and distributed optimization*, pages 181–213. Springer, 2018.

Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33, 2020.

Christian Clason, Dirk A. Lorenz, Hinrich Mahler, and Benedikt Wirth. Entropic regularization of continuous optimal transport problems. *Journal of Mathematical Analysis and Applications*, 494 (1):124432, 2021. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2020.124432.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, 2020.

Ryan D'Orazio, Nicolas Loizou, Issam Laradji, and Ioannis Mitliagkas. Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize. *arXiv preprint arXiv:2110.15412*, 2021.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR, 2019.

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

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018a.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 09–11 Apr 2018b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Matthieu Heitz, Nicolas Bonneel, David Coeurjolly, Marco Cuturi, and Gabriel Peyré. Ground metric learning on graphs. *Journal of Mathematical Imaging and Vision*, pages 1–19, 2020.

Hicham Janati, Thomas Bazeille, Bertrand Thirion, Marco Cuturi, and Alexandre Gramfort. Multi-subject meg/eeg source imaging with sparse multi-task regression. *NeuroImage*, page 116847, 2020.

Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

Sunil Koundal, Rena Elkin, Saad Nadeem, Yuechuan Xue, Stefan Constantinou, Simon Sanggaard, Xiaodan Liu, Brittany Monte, Feng Xu, William Van Nostrand, et al. Optimal mass transport with lagrangian workflow reveals advective and diffusion driven solute transport in the glymphatic system. *Scientific reports*, 10(1):1–18, 2020.

Samuel Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. *Dept. Comput. Sci., Univ. Chicago, Chicago, IL, USA, Tech. Rep. TR-2002-04*, 2002.

Weijie Liu, Chao Zhang, Nenggan Zheng, and Hui Qian. Approximating optimal transport via low-rank and sparse factorization. *arXiv preprint arXiv:2111.06546*, 2021.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization, 2021.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. *ICML*, 2022.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

Karren Dai Yang, Karthik Damodaran, Saradha Venkatachalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

Xinye Zheng, Jianbo Ye, James Z Wang, and Jia Li. Scott: Shape-location combined tracking with optimal transport. *SIAM Journal on Mathematics of Data Science*, 2(2):284–308, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 3, 4, 5, 6.
   (b) Did you describe the limitations of your work? [Yes] See Section 8.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 3, 4, 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 7.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 7.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 7.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]