Low-Rank Sinkhorn Factorization

Meyer Scetbon [1] **Marco Cuturi** [1 2] **Gabriel Peyré** [3 4]

Abstract

Several recent applications of optimal transport (OT) theory to machine learning have relied on regularization, notably entropy and the Sinkhorn algorithm. Because matrix-vector products are pervasive in the Sinkhorn algorithm, several works have proposed to *approximate* kernel matrices appearing in its iterations using low-rank factors. Another route lies instead in imposing low-nonnegative rank constraints on the feasible set of couplings considered in OT problems, with no approximations on cost nor kernel matrices. This route was first explored by Forrow et al. (2018), who proposed an algorithm tailored for the squared Euclidean ground cost, using a proxy objective that can be solved through the machinery of regularized 2-Wasserstein barycenters. Building on this, we introduce in this work a generic approach that aims at solving, in full generality, the OT problem under low-nonnegative rank constraints with arbitrary costs. Our algorithm relies on an explicit factorization of low-rank couplings as a product of *sub-coupling* factors linked by a common marginal; similar to an NMF approach, we alternatively updates these factors. We prove the non-asymptotic stationary convergence of this algorithm and illustrate its efficiency on benchmark experiments.

1. Introduction

By providing a simple and comprehensive framework to compare probability distributions, optimal transport (OT) theory has inspired many developments in machine learning (Peyré & Cuturi, 2019). A flurry of works have recently connected it to other trending topics, such as normalizing flows or convex neural networks (Makkuva et al., 2020; Korotin et al., 2021; Tong et al., 2020), while the scope of its applications has now reached several fields of science such as single-cell biology (Schiebinger et al., 2019; Yang et al., 2020), imaging (Schmitz et al., 2018; Heitz et al., 2020) or neuroscience (Janati et al., 2020; Koundal et al., 2020).

Challenges when computing OT. Solving optimal transport problems at scale poses, however, formidable challenges. The most obvious among them is computational: Instantiating the Kantorovich (1942) problem on discrete measures of size n can be solved with a linear program (LP) of complexity $O(n^3 \log n)$. A second and equally important challenge lies in the statistical performance of using that LP to estimate OT between densities: the LP solution between i.i.d samples converges very slowly to that between densities (Fournier & Guillin, 2015). It is now increasingly clear that regularizing OT in some way or another is the only way to mitigate these two issues (Genevay et al., 2018; Chizat et al., 2020; Clason et al., 2021). A popular approach consists in penalizing the OT problem with a strongly convex function of the coupling (Cuturi, 2013; Dessein et al., 2018). We explore in this work an alternative, and more direct approach to add regularity: we restrict, instead, the set of feasible couplings to have a small nonnegative rank.

Low-Rank Kernel Factorization. Low-rank factorizations are not new to regularized OT. They have been used to speed-up the resolution of entropy regularized OT with the Sinkhorn algorithm, pending some approximations: Given a data-dependent $n \times m$ cost matrix C, the Sinkhorn iterations consist in matrix-vector products of the form Kv or $K^T u$ where $K \triangleq \exp(-C/\varepsilon)$ and u, v are n, m- vectors. Altschuler et al. (2018) and Altschuler & Boix-Adsera (2020) have proposed to approximate the kernel K with a product of thin rank r matrices, $\widetilde{K} = AB^T$. Naturally, the ability to approximate K with a low-rank $\widetilde{K}$ degrades as ε decreases, making this approach valid only for sufficiently large ε. Thanks to this approximation, however, each Sinkhorn iteration is linear in n or m, and the coupling outputted by the Sinkorn algorithm is of the form $\widetilde{P} = CD^T$ where $C = \mathrm{diag}(u)A$, $D = \mathrm{diag}(v)B$. This approximation results therefore in a *low-rank* solution that is not, however, rigorously optimal for the original problem as defined by K but rather that defined by $\widetilde{K}$. However the solution obtained with $\widetilde{K}$ can be arbitrary close to the true solution by increasing the rank considered. Similarly, Scetbon & Cuturi (2020)

[1]CREST, ENSAE [2]Google Brain [3]Ecole Normale Supérieure, PSL University [4]CNRS. Correspondence to: Meyer Scetbon <meyer.scetbon@ensae.fr>, Marco Cuturi <cuturi@google.com>, Gabriel Peyré <gabriel.peyre@ens.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

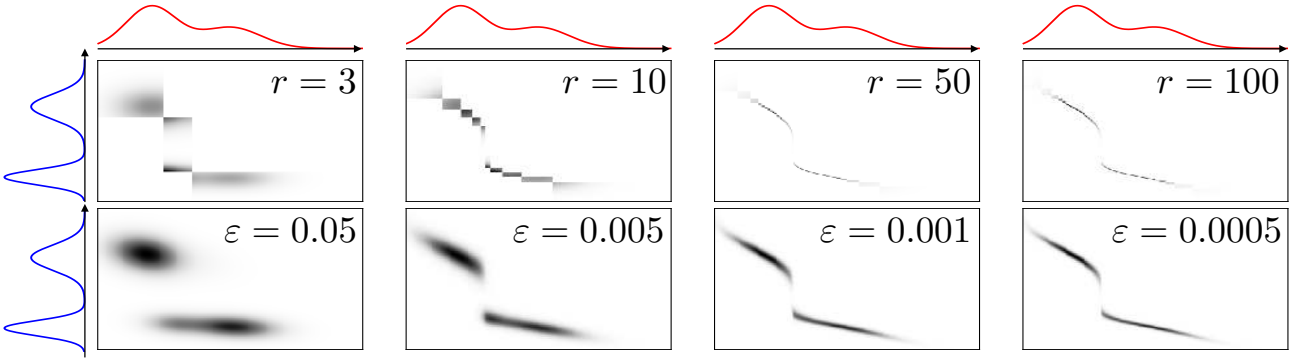

Figure 1. Two Gaussian mixture densities evaluated on $n = 200$ and $m = 220$ sized grids in 1D, displayed as blue/red curves. Between them, $n \times m$ optimal coupling matrices obtained by our proposed low-rank OT method for varying rank constraint values r (in increasing order, top row) and the Sinkhorn algorithm, for various ε (in decreasing order, bottom row). The ground cost is the 1.5-norm.

consider instead *nonnegative low-rank* approximations for K of the form $\widetilde{K} = QR^T$ where $Q, R > 0$.

Low-Nonnegative Rank Couplings. To our knowledge, only Forrow et al. (2018) have used low rank considerations for couplings, rather than costs or kernels. Their work studies the case where the ground cost is the squared Euclidean distance. They obtain for that cost a proxy for rank-constrained OT problems using 2-Wasserstein barycenters (Agueh & Carlier, 2011). Their algorithm blends those in (Cuturi & Doucet, 2014; Benamou et al., 2015) and results in an intuitive mass transfer plan that goes through a small number of r points, where r is the coupling's nonnegative rank.

Our Contributions. In this work, we tackle directly the low-rank problem formulated by (Forrow et al., 2018) but make no assumption on the cost matrix; we address instead the low-rank OT problem in its full generality. We consider couplings $P = Q\,\mathrm{diag}(1/g)R^T$ decomposed as the product of two sub-couplings Q, R, with common right marginal g, and left-marginal given by those of P on each side. Each of these sub-couplings minimizes a transport cost that involves the original cost matrix C and the other sub-coupling. We handle this problem by optimizing jointly on Q, R and g using a mirror-descent approach. We prove the non-asymptotic stationary convergence of this approach. In addition, we show that the time complexity of our algorithm can become linear when exploiting low rank assumptions on the *cost* (not the kernel) involved in the OT problem.

Differences with previous work. Our approach borrows ideas from (Forrow et al., 2018) but is generic as it applies to all ground costs. Our approach constrains the non-negative rank of the coupling solution P by construction, rather than relying on a low rank approximation $\widetilde{K}$ for kernel $K = e^{-C/\varepsilon}$. This is a crucial point, because the ability to approximate K with a low rank $\widetilde{K}$ significantly degrades as ε decreases. By contrast, our approach applies to all ranks, small and large. Interestingly, we also show that a low-rank assumption on the cost matrix (not on the kernel) can also be leveraged, providing therefore a "best of both worlds" scenario in which both the *coupling*'s and the *cost*'s (not the kernel) low rank properties can be enforced and exploited. Finally, a useful parallel can be drawn between our approach and that of the vanilla Sinkhorn algorithm, in the sense that they propose different regularization schemes. Indeed, the (discrete) path of solutions obtained by our algorithm when varying r between 1 and $\min(n, m)$ can be seen as an alternative to the entropic regularization path. Both paths contain at their extremes the original OT solution (maximal rank and minimal entropy) and the product of marginals (minimal rank and maximal entropy), as illustrated in Fig. 1.

2. Discrete Optimal Transport

OT as a linear program. Let a and b be two histograms in Δ_n, Δ_m, the probability simplices of respective size n, m. Assuming $a > 0$ and $b > 0$, set $X \triangleq (x_1, \dots, x_n)$ and $Y \triangleq (y_1, \dots, y_m)$ two families of points taken each within arbitrary sets, and define discrete distributions $\mu \triangleq \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu \triangleq \sum_{j=1}^m b_j \delta_{y_j}$. The set of couplings with marginals a, b is:

$$\Pi_{a,b} \triangleq \{P \in \mathbb{R}_+^{n \times m} \text{ s.t. } P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\}\,.$$

Given a cost function c defined on pairs of points in X, Y and writing $C \triangleq [c(x_i, y_j)]_{i,j}$ its associated matrix, the optimal transport (OT) problem can be written as follows:

$$\mathrm{OT}(\mu, \nu) \triangleq \min_{P \in \Pi_{a,b}} \langle C, P \rangle\,. \tag{1}$$

Entropic regularization. Several works have shown recently (Genevay et al., 2018; Chizat et al., 2020) that when X and Y are sampled from a continuous space, it is preferable to regularize (1) using, for instance, an entropic regularizer (Cuturi, 2013) to achieve both better computational

and statistical efficiency,

$$\mathrm{OT}_\varepsilon(\mu,\nu) \triangleq \min_{P\in\Pi_{a,b}} \langle C, P\rangle - \varepsilon H(P)\,. \tag{2}$$

where $\varepsilon \geq 0$ and H is the Shannon entropy defined as $H(P) \triangleq -\sum_{ij} P_{ij}(\log P_{ij} - 1)$. If ε goes to 0, one recovers the classical OT problem and for any $\varepsilon > 0$, Eq. (2) becomes ε-strongly convex on $\Pi_{a,b}$ and admits a unique solution P_ε, of the form

$$\exists u_\varepsilon \in \mathbb{R}^n_+, v_\varepsilon \in \mathbb{R}^m_+ \text{ s.t. } P_\varepsilon = \mathrm{diag}(u_\varepsilon) K \mathrm{diag}(v_\varepsilon) \tag{3}$$

where $K \triangleq \exp(-C/\varepsilon)$. Cuturi (2013) shows that the scaling vectors u_ε and v_ε can be obtained efficiently thanks to the Sinkhorn algorithm (see Alg. 1, where $\odot$ and $/$ denote entry-wise operation). Each iteration can be performed in $\mathcal{O}(nm)$ algebraic operations as it involves only matrix-vector products. The number of Sinkhorn iterations needed to converge to a precision δ (monitored by the difference between the column-sum of $\mathrm{diag}(u)\mathbf{K}\mathrm{diag}(v)$ and b) is controlled by the scale of elements in C relative to ε (Franklin & Lorenz, 1989). That convergence deteriorates with smaller ε, as studied in more detail by (Altschuler et al., 2017; Dvurechensky et al., 2018).

Algorithm 1 Sinkhorn(K, a, b, δ)

Inputs: K, a, b, δ, u
repeat
  $v \leftarrow b/K^T u,\ u \leftarrow a/Kv$
until $\|u \odot Kv - a\|_1 + \|v \odot K^T u - b\|_1 < \delta$;
Result: u, v

Mirror descent and ε schedule. A possible interpretation of the entropic regularization in the OT problem is that it can be seen as the k_ε-th update of a Mirror Descent (MD) algorithm applied to the objective (1) where $k_\varepsilon \geq 1$ depends on ε and the gradient steps used in the MD. Several works have proposed such links between a gradual decrease in ε to obtain a better approximation of the unregularized OT problem (Schmitzer, 2019; Lin et al., 2019; Xie et al., 2020). More precisely, the MD algorithm associated to the Kullback–Leibler divergence (KL) applied to the objective (1) makes for all $k \geq 0$ the following update:

$$Q^{k+1} \triangleq \operatorname*{argmin}_{Q\in\Pi_{a,b}} \langle C, Q\rangle + \frac{1}{\gamma_k}\mathrm{KL}(Q, Q_k) \tag{4}$$

where $(\gamma_k)_{k\geq 0}$ is a sequence of positive real numbers, $Q_0 \in \Pi_{a,b}$ is an initial point and KL is the Kullback–Leibler divergence defined asw. If $Q_0 \triangleq ab^T$, then one obtains that for all $k \geq 0$, updating the coupling according to Eq. (4) is the same as solving

$$Q^{k+1} \triangleq \operatorname*{argmin}_{Q\in\Pi_{a,b}} \langle C, Q\rangle - \varepsilon_k H(Q)$$

where $\varepsilon_k \triangleq (\sum_{j=0}^k \gamma_j)^{-1}$. Therefore the MD algorithm applied to (1) produces the sequence $(P_{\varepsilon_k})_{k\geq 0}$ of optimal couplings according to the objective (2). We show next that this viewpoint can be applied when one adds also some structures to the couplings considered in the OT problem (1), leading to a new regularized approach.

3. Nonnegative Factorization of the Optimal Coupling

Here we aim at regularizing the OT problem by decomposing the couplings involved into a product of two low-rank couplings. We introduce the associated non-convex problem and develop a mirror-descent algorithm which operates by solving a succession of convex programs.

3.1. Low-Rank and Factored Couplings

We introduce low rank couplings and explain how they can be parameterized as factored couplings.

Definition 1. *Given $M \in \mathbb{R}^{n\times m}$, the nonnegative rank of M is the smallest number of nonnegative rank-one matrices into which the matrix can be decomposed additively:*

$$\mathrm{rk}_+(M) \triangleq \min\left\{ q | M = \sum_{i=1}^q R_i, \forall i, \mathrm{rk}(R_i) = 1, R_i \geq 0 \right\}.$$

Let $r \geq 1$, and let us denote

$$\Pi_{a,b}(r) \triangleq \{P \in \Pi_{a,b}, \mathrm{rk}_+(P) \leq r\}.$$

From Definition 1, one has

$$\Pi_{a,b}(r) = \Big\{ \sum_{i=1}^r g_i q_i r_i^T \text{ s.t. } \forall\, i\ q_i \in \Delta_n,\ r_i \in \Delta_m,\ g \in \Delta_r,\ \sum_{i=1}^r g_i q_i = a \text{ and } \sum_{i=1}^r g_i r_i = b \Big\}$$

from which we deduce directly that $\Pi_{a,b}(r)$ is compact. Moreover for $g \in \Delta_r^* \triangleq \{h \in \Delta_r \text{ s.t. } \forall i\ h_i > 0\}$, we write

$$\Pi_{a,g,b} \triangleq \Big\{ P \in \mathbb{R}^{n\times m}_+, P = Q\,\mathrm{diag}(1/g) R^T,\ Q \in \Pi_{a,g},\ \text{and } R \in \Pi_{b,g} \Big\}.$$

Note that $\Pi_{a,g,b}$ is compact and a subset of $\Pi_{a,b}(r)$ since for all $P \in \Pi_{a,g,b}$, $P \in \Pi_{a,b}$ and one has $\mathrm{rk}(P) \leq \mathrm{rk}_+(P) \leq r$. Moreover, for any $P \in \Pi_{a,b}$ such that $\mathrm{rk}_+(P) \leq r$, there exists $g \in \Delta_r^*$, $Q \in \Pi_{a,g}$ and $R \in \Pi_{b,g}$ such that $P = Q\,\mathrm{diag}(1/g) R^T$ (Cohen & Rothblum, 1993). Therefore

$$\bigcup_{g\in\Delta_r^*} \Pi_{a,g,b} = \Pi_{a,b}(r). \tag{5}$$

We exploit next this identity to build an efficient algorithm in order to solve the optimal transport problem under low nonnegative rank constraints.

3.2. The Low-rank OT Problem (LOT)

The problem of interest in this work is:

$$\mathrm{LOT}_r(\mu,\nu) \triangleq \min_{P\in\Pi_{a,b}(r)} \langle C, P\rangle. \tag{6}$$

Here the minimum is always attained as $\Pi_{a,b}(r)$ is compact and the objective is continuous. Thanks to (5), problem (6) is equivalent to

$$\min_{(Q,R,g)\in\mathcal{C}(a,b,r)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle \tag{7}$$

where $\mathcal{C}(a,b,r) \triangleq \mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)$, with

$$\mathcal{C}_1(a,b,r) \triangleq \Big\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times(\mathbb{R}_+^*)^r \\ \text{s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b\Big\}$$

$$\mathcal{C}_2(r) \triangleq \Big\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{r} \\ \text{s.t. } Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g\Big\}.$$

In the following, we also consider regularized version of the problem (7) by adding an entropic term to the objective which leads for all $\varepsilon\geq 0$ to the following problem

$$\mathrm{LOT}_{r,\varepsilon}(\mu,\nu) \triangleq \inf_{(Q,R,g)\in\mathcal{C}(a,b,r)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle \\ -\varepsilon H((Q,R,g)). \tag{8}$$

Here the entropy of (Q,R,g) is to be understood as that of the values of the three respective entropies evaluated for each term. We will see that adding an entropic term to the objective allows to stabilize the MD scheme employed to solve (6). For all $\varepsilon\geq 0$, the objective function defined in (8) is lower semi-continuous, and admits therefore a minimum in $\overline{\mathcal{C}_1(a,b,r)}\cap\mathcal{C}_2(r)$ where $\overline{\mathcal{C}_1(a,b,r)}$ is the closure of $\mathcal{C}_1(a,b,r)$. However, the existence of a solution for problem (8) requires more care, as shown in the following proposition.

Proposition 1. *If $\varepsilon = 0$ then the infimum of (8) is always attained. If $\varepsilon > 0$, then if $r = 1$, the infimum of (8) is attained and for $r\geq 2$, problem (8) admits a minimum if $LOT_{r,\varepsilon}(\mu,\nu) < LOT_{r-1,\varepsilon}(\mu,\nu)$.*

Stabilized Formulation using Lower Bounds In order to ensure stability of the mirror descent method, and enable its theoretical analysis, we introduce a lower bound α on the weight vector g.

Let us assume in the following that we consider (r,ε) satisfying the conditions of Proposition 1. In particular if $\varepsilon = 0$, r can be arbitrarily chosen and we recover the problem defined in (6). Under this assumption, there exists $(Q_\varepsilon^*, R_\varepsilon^*, g_\varepsilon^*)\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)$ solution of Eq. (8) from which follows the existence of $\frac{1}{r}\geq\alpha^*>0$, such that $g_\varepsilon^*\geq\alpha^*$ coordinate-wise. Let us now define for any $\frac{1}{r}\geq\alpha>0$, the following set

$$\mathcal{C}_1(a,b,r,\alpha) \triangleq \Big\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{r} \\ \text{s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b,\ g\geq\alpha\Big\}.$$

Then if α is sufficiently small (i.e. $\alpha\leq\alpha^*$) we have that the problem (8) is equivalent to

$$\mathrm{LOT}_{r,\varepsilon,\alpha}(\mu,\nu) = \min_{(Q,R,g)\in\mathcal{C}(a,b,r,\alpha)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle \\ -\varepsilon H((Q,R,g)), \tag{9}$$

where $\mathcal{C}(a,b,r,\alpha) \triangleq \mathcal{C}_1(a,b,r,\alpha)\cap\mathcal{C}_2(r)$. Note that for any $\frac{1}{r}\geq\alpha>0$, the set of constraints is not empty, compact and the minimum always exists.

3.3. Mirror Descent Optimization Scheme

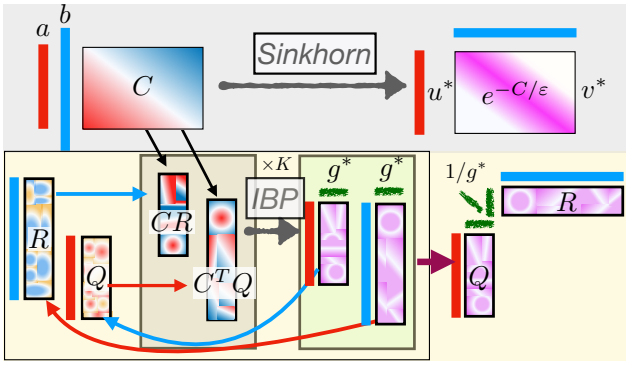

Figure 2. Comparison of the Sinkhorn algorithm with our proposed procedure.

Mirror descent outer loop. We propose to use a Mirror Descent scheme with a KL divergence to solve Eq. (9). It leads, for all $k\geq 0$, to the following updates which necessitate the solution of a convex problem at each step

$$(Q_{k+1}, R_{k+1}, g_{k+1}) \triangleq \operatorname*{argmin}_{\zeta\in\mathcal{C}(a,b,r,\alpha)} \mathrm{KL}(\zeta, \boldsymbol{\xi}_k) \tag{10}$$

where $(Q_0,R_0,g_0)\in\mathcal{C}(a,b,r,\alpha)$ is an initial point such that $Q_0>0$ and $R_0>0$, $\boldsymbol{\xi}_k \triangleq (\xi_k^{(1)},\xi_k^{(2)},\xi_k^{(3)})$, $\xi_k^{(1)} \triangleq \exp(-\gamma_k C R_k\,\mathrm{diag}(1/g_k) - (\gamma_k\varepsilon-1)\log(Q_k))$, $\xi_k^{(2)} \triangleq \exp(-\gamma_k C^T Q_k\,\mathrm{diag}(1/g_k) - (\gamma_k\varepsilon-1)\log(R_k))$, $\xi_k^{(3)} \triangleq \exp(\gamma_k\omega_k/g_k^2 - (\gamma_k\varepsilon-1)\log(g_k))$ with $[\omega_k]_i \triangleq [Q_k^T C R_k]_{i,i}$ for all $i\in\{1,\dots,r\}$ and $(\gamma_k)_{k\geq 0}$ is a sequence of positive step sizes. Note that for all $k\geq 0$, (Q_k,R_k,g_k) live in $(\mathbb{R}_+^*)^{n\times r}\times(\mathbb{R}_+^*)^{m\times r}\times(\mathbb{R}_+^*)^r$, and therefore $\boldsymbol{\xi}_k$ is well defined and lives also in $(\mathbb{R}_+^*)^{n\times r}\times(\mathbb{R}_+^*)^{m\times r}\times(\mathbb{R}_+^*)^r$.

Dykstra's inner loop. In order to solve Eq. (10), we use the Dykstra's Algorithm (Dykstra, 1983). Given a closed convex set $\mathcal{C} \subset \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r$, we denote for all $\boldsymbol{\xi} \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{m\times r} \times (\mathbb{R}_+^*)^r$ the projection according to the Kullback-Leibler divergence as

$$\mathcal{P}_{\mathcal{C}}^{\mathrm{KL}}(\boldsymbol{\xi}) \triangleq \operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}).$$

Starting from $\boldsymbol{\zeta}_0 \triangleq \boldsymbol{\xi}$ and $\boldsymbol{q}_0 = \boldsymbol{q}_{-1} = (\mathbf{1}, \mathbf{1}, \mathbf{1}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r$, the Dykstra's Algorithm consists in computing for all $j \geq 0$,

$$\begin{aligned}
\boldsymbol{\zeta}_{2j+1} &= \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_1(a,b,r,\alpha)}(\boldsymbol{\zeta}_{2j} \odot \boldsymbol{q}_{2j-1}) \\
\boldsymbol{q}_{2j+1} &= \boldsymbol{q}_{2j-1} \odot \frac{\boldsymbol{\zeta}_{2j}}{\boldsymbol{\zeta}_{2j+1}} \\
\boldsymbol{\zeta}_{2j+2} &= \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_2(r)}(\boldsymbol{\zeta}_{2j+1} \odot \boldsymbol{q}_{2j}) \\
\boldsymbol{q}_{2j+2} &= \boldsymbol{q}_{2j} \odot \frac{\boldsymbol{\zeta}_{2j+1}}{\boldsymbol{\zeta}_{2j+2}}.
\end{aligned}$$

As $\mathcal{C}_1(a, b, r, \alpha)$ and $\mathcal{C}_2(r)$ are closed convex subspaces and $\boldsymbol{\xi} \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{m\times r} \times (\mathbb{R}_+^*)^r$, one can show that $(\boldsymbol{\zeta}_j)_{j\geq 0}$ converges towards the unique solution of Eq. (10), (Bauschke & Lewis, 2000). The following propositions detail how to compute the relevant projections involved in the Dykstra's Algorithm.

Proposition 2. *For* $\tilde{\boldsymbol{\xi}} \triangleq (\tilde{Q}, \tilde{R}, \tilde{g}) \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^r$*, one has, denoting* $\hat{g} \triangleq \max(\tilde{g}, \alpha)$

$$\mathcal{P}^{KL}_{\mathcal{C}_1(a,b,r,\alpha)}(\tilde{\boldsymbol{\xi}}) = \left(\mathrm{diag}\left(\frac{a}{\tilde{Q}\mathbf{1}_r}\right)\tilde{Q}, \mathrm{diag}\left(\frac{b}{\tilde{R}\mathbf{1}_r}\right)\tilde{R}, \hat{g}\right).$$

Let us now show the solution of the projection on $\mathcal{C}_2(r)$.

Proposition 3. *For* $\tilde{\boldsymbol{\xi}} \triangleq (\tilde{Q}, \tilde{R}, \tilde{g}) \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^r$*, the projection* $(Q, R, g) = \mathcal{P}^{KL}_{\mathcal{C}_2(r)}(\tilde{\boldsymbol{\xi}})$ *satisfies*

$$\begin{aligned}
Q &= \tilde{Q}\,\mathrm{diag}(g/\tilde{Q}^T\mathbf{1}_n), \quad R = \tilde{R}\,\mathrm{diag}(g/\tilde{R}^T\mathbf{1}_m) \\
g &= (\tilde{g} \odot \tilde{Q}^T\mathbf{1}_n \odot \tilde{R}^T\mathbf{1}_m)^{1/3}.
\end{aligned}$$

Efficient computation of the updates. The projection obtained in Proposition 2, 3 lead to simple updates of the couplings. Indeed, starting with $\boldsymbol{\zeta}_0 \triangleq \boldsymbol{\xi} = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})$ the Dysktra's Algorithm applied to our problem (10) needs only to compute scaling vectors as presented in Alg. 2. More precisely, the Dykstra's Algorithm produces the iterates $(\boldsymbol{\zeta}_j)_{j\geq 0}$ which satisfy for all $j \geq 0$ $\boldsymbol{\zeta}_j = (Q_j, R_j, g_j)$ where

$$\begin{aligned}
Q_j &= \mathrm{diag}(u_j^1)\xi^{(1)}\,\mathrm{diag}(v_j^1) \\
R_j &= \mathrm{diag}(u_j^2)\xi^{(2)}\,\mathrm{diag}(v_j^2)
\end{aligned}$$

for the sequences $(u_j^i, v_j^i)_{j\geq 0}$ initialized as, $u_0^i \triangleq \mathbf{1}_n$, $v_0^i \triangleq \mathbf{1}_m$ for all $i \in \{1, 2\}$, $q_{0,1}^{(3)} = q_{0,2}^{(3)} = q_0^{(1)} = q_0^{(2)} = \mathbf{1}_r$ and computed with the iterations

$$\begin{aligned}
u_{n+1}^{k,i} &= \frac{p_i}{\xi_k^i v_n^{k,i}} \\
\tilde{g}_{n+1} &= \max(\alpha, g_n \odot q_{n,1}^{(3)}),\ q_{n+1,1}^{(3)} = (g_n \odot q_{n,1}^{(3)})/\tilde{g}_{n+1} \\
g_{n+1} &= (\tilde{g}_{n+1} \odot q_{n,2}^{(3)})^{1/3} \prod_{i=1}^{2} (v_n^{k,i} \odot q_n^{(i)} \odot (\xi_k^i)^T u_n^{k,i})^{1/3} \\
v_{n+1}^{k,i} &= \frac{g_{n+1}}{(\xi_k^i)^T u_n^{k,i}} \\
q_{n+1}^{(i)} &= (v_n^{k,i} \odot q_n^{(i)})/v_{n+1}^{k,i},\ q_{n+1,2}^{(3)} = (\tilde{g}_{n+1} \odot q_{n,2}^{(3)})/g_{n+1}
\end{aligned}$$

We have denoted $p_1 \triangleq a$ and $p_2 \triangleq b$ to simplify the notations.

Algorithm 2 LR-Dykstra$((\xi^{(i)})_{1\leq i\leq 3}, p_1, p_2, \alpha, \delta)$

Inputs: $\xi^{(1)}, \xi^{(2)}, \tilde{g} \triangleq \xi^{(3)}, p_1, p_2, \alpha, \delta, q_1^{(3)} = q_2^{(3)} = \mathbf{1}_r, \forall i \in \{1, 2\}$, $\tilde{v}^{(i)} = \mathbf{1}_r, q^{(i)} = \mathbf{1}_r$

repeat

$u^{(i)} \leftarrow p_i/\xi^{(i)}\tilde{v}^{(i)}\ \forall i \in \{1, 2\}$,
$g \leftarrow \max(\alpha, \tilde{g} \odot q_1^{(3)}),\ q_1^{(3)} \leftarrow (\tilde{g} \odot q_1^{(3)})/g,\ \tilde{g} \leftarrow g$,
$g \leftarrow (\tilde{g} \odot q_2^{(3)})^{1/3} \prod_{i=1}^{2} (v^{(i)} \odot q^{(i)} \odot (\xi^{(i)})^T u^{(i)})^{1/3}$,
$v^{(i)} \leftarrow g/(\xi^{(i)})^T u^{(i)}\ \forall i \in \{1, 2\}$,
$q^{(i)} \leftarrow (\tilde{v}^{(i)} \odot q^{(i)})/v^{(i)}\ \forall i \in \{1, 2\},\ q_2^{(3)} \leftarrow (\tilde{g} \odot q_2^{(3)})/g$,
$\tilde{v}^{(i)} \leftarrow v^{(i)}\ \forall i \in \{1, 2\},\ \tilde{g} \leftarrow g$

until $\sum_{i=1}^{2} \|u^{(i)} \odot \xi^{(i)} v^{(i)} - p_i\|_1 < \delta$;
$Q \leftarrow \mathrm{diag}(u^{(1)})\xi^{(1)}\,\mathrm{diag}(v^{(1)})$
$R \leftarrow \mathrm{diag}(u^{(2)})\xi^{(2)}\,\mathrm{diag}(v^{(2)})$
Result: Q, R, g

Let us now introduce the proposed MD algorithm applied to (9). By denoting $\mathcal{D}(\cdot)$ the operator extracting the diagonal of a square matrix we obtain Alg. 3. See also Figure 2 for an illustration of the proposed algorithm.

Algorithm 3 LOT$(C, a, b, r, \alpha, \delta)$

Inputs: $C, a, b, (\gamma_k)_{k\geq 0}, Q, R, g, \alpha, \delta$

for $k = 1, \ldots$ **do**

$\xi^{(1)} \leftarrow \exp(-\gamma_k C R\,\mathrm{diag}(1/g) - (\gamma_k\varepsilon - 1)\log(Q))$,
$\xi^{(2)} \leftarrow \exp(-\gamma_k C^T Q\,\mathrm{diag}(1/g) - (\gamma_k\varepsilon - 1)\log(R))$,
$\omega \leftarrow \mathcal{D}(Q^T C R)$,
$\xi^{(3)} \leftarrow \exp(\gamma_k\omega/g^2 - (\gamma_k\varepsilon - 1)\log(g))$,
$Q, R, g \leftarrow$ LR-Dykstra$((\xi^{(i)})_{1\leq i\leq 3}, a, b, \alpha, \delta)$ (Alg. 2)

end

Result: $\langle C, Q\,\mathrm{diag}(1/g)R^T\rangle$

Computational Cost. Note that $(\xi^{(i)})_{1\leq i\leq 3}$ considered in Alg. 3 live in $\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^r$ and therefore given those matrices, each iteration of Alg. 2 requires $\mathcal{O}((n+m)r)$ algebraic operations, since it involves only matrix/vector multiplications of the form $\xi^{(i)}v_i$ and $(\xi^{(i)})^Tu_i$. However without any assumption on the cost matrix C, computing $(\xi^{(i)})_{1\leq i\leq 3}$ requires $\mathcal{O}(nmr)$ algebraic operations since CR and C^TQ must be evaluated. We show in §3.5 how to reduce the quadratic cost of computing $(\xi^{(i)})_{1\leq i\leq 3}$ to a linear cost with respect to the number of samples if one assumes that the considered *cost* matrix can be factored, either exactly (ensured with a squared Euclidean distance cost) or approximately if that cost is a distance. Writing N the number of iterations of the MD scheme and T the number of iterations considered in Algorithm 2 at each step of the MD, we end up with a total computational cost of $\mathcal{O}(NT(n+m)r+Nnmr)$.

Remark 1. *Note that our algorithm can be applied in the specific case where $\varepsilon=0$ in order to solve Eq. (6). Moreover, our algorithm can be applied for an arbitrary choice of the cost function. For example in Figure 5, we run our algorithm on graphs where the distance considered in the shortest-path distance.*

3.4. Convergence of the Mirror Descent

Even if the objective (9) is not convex in (Q,R,g), we obtain the non-asymptotic stationary convergence of the MD algorithm in this setting. For that purpose we introduce a stronger convergence criterion than the one presented in (Ghadimi et al., 2013) to obtain non-asymptotic stationary convergence of the MD scheme. Indeed let F_ε be the objective function of the problem (9) defined on $\mathcal{C}(a,b,r,\alpha)$ and let us denotes for any $\gamma>0$ and $\boldsymbol{\xi}\in\mathcal{C}(a,b,r,\alpha)$

$$\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)\triangleq\operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}(a,b,r,\alpha)}\{\langle\nabla F_\varepsilon(\boldsymbol{\xi}),\boldsymbol{\zeta}\rangle+\frac{1}{\gamma}KL(\boldsymbol{\zeta},\boldsymbol{\xi})\}.$$

Then the criteron used in (Ghadimi et al., 2013) to show the stationary convergence of the MD scheme is defined as the square norm of the following vector:

$$P_{\mathcal{C}(a,b,r,\alpha)}(\boldsymbol{\xi},\gamma)\triangleq\frac{1}{\gamma}(\boldsymbol{\xi}-\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)).$$

This vector can be seen as a generalized projected gradient of F_ε at $\boldsymbol{\xi}$. Indeed if $X=\mathbb{R}^d$ and by replacing the *prox-function* $\mathrm{KL}(u,x)$ by $\frac{1}{2}\|u-x\|_2^2$, we would have $P_X(x,\gamma)=\nabla F_\varepsilon(x)$. Here we consider instead the following criterion to establish convergence:

$$\Delta_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)\triangleq\frac{1}{\gamma^2}(\mathrm{KL}(\boldsymbol{\xi},\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma))+\mathrm{KL}(\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma),\boldsymbol{\xi})).$$

Such criterion is in fact stronger than the one used in (Ghadimi et al., 2013) as we have

$$\begin{aligned}\Delta_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)&=\frac{1}{\gamma^2}(\langle\nabla h(\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma))-\nabla h(\boldsymbol{\xi}),\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)-\boldsymbol{\xi}\rangle\\&\geq\frac{1}{2\gamma^2}\|\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)-\boldsymbol{\xi}\|_1^2\\&=\frac{1}{2}\|P_{\mathcal{C}(a,b,r,\alpha)}(\boldsymbol{\xi},\gamma)\|_1^2\end{aligned}$$

where h denotes the minus entropy function and the last inequality comes from the strong convexity of h on $\mathcal{C}(a,b,r,\alpha)$.

For any $\frac{1}{r}\geq\alpha>0$, we show in the following proposition the non-asymptotic stationary convergence of the MD scheme applied to the problem (9). To prove this result, we show that for any $\varepsilon\geq 0$, the objective is smooth relatively to the negative entropy function (Bauschke et al., 2017) and we extend the proof of (Ghadimi et al., 2013) to this case.

Proposition 4. *Let $\varepsilon\geq 0$, $\frac{1}{r}\geq\alpha>0$ and $N\geq 1$. By denoting*

$$L_{\varepsilon,\alpha}\triangleq\sqrt{3\left(2\frac{\|C\|_2^2}{\alpha^4}+\left(\varepsilon+\frac{2\|C\|_2}{\alpha^3}\right)^2\right)}$$

and by considering a constant stepsize in the MD scheme (10) such that for all $k=1,\dots,N$ $\gamma_k=\frac{1}{2L_{\varepsilon,\alpha}}$, we obtain that

$$\min_{1\leq k\leq N}\Delta_{\varepsilon,\alpha}((Q_k,R_k,g_k),\gamma_k)\leq\frac{4L_{\varepsilon,\alpha}D_0}{N}.$$

where $D_0\triangleq F_\varepsilon(Q_0,R_0,g_0)-\mathrm{LOT}_{r,\varepsilon,\alpha}$ is the distance of the initial value to the optimal one.

Thanks to Proposition 4, for α sufficiently small (i.e. $\alpha\leq\alpha^*$), we have $\mathrm{LOT}_{r,\varepsilon,\alpha}=\mathrm{LOT}_{r,\varepsilon}$ and therefore we obtain a stationary point of (8). In particular, if $\varepsilon=0$, the proposed algorithm converges towards a stationary point of (6).

Remark 2. *We also propose an algorithm to directly solve (8). The main difference is that the updates of the MD can be solved using the Iterative Bregman Projections (IBP) Algorithm. See Appendix F for more details.*

Remark 3. *For all $\varepsilon\geq 0$, the MD scheme implies that each iteration k of our proposed algorithm outputs $(Q_k,R_k,g_k)\in\mathcal{C}_1(a,b,r,\alpha)\cap\mathcal{C}_2(r)$, and therefore the matrix obtained a each iteration $P_k^{LOT}=Q_k\,\mathrm{diag}(1/g_k)R_k^T$ is a coupling which sastifies the marginal constraints while in the Sinkhorn algorithm, the matrix defined at each iteration by $P_k^{Sin}=\mathrm{diag}(u_k)K\,\mathrm{diag}(v_k)$ becomes a coupling which satisfies the marginal constraints only at convergence.*

In the following section, we aim at accelerating our method in order to obtain a linear time algorithm to solve (8).

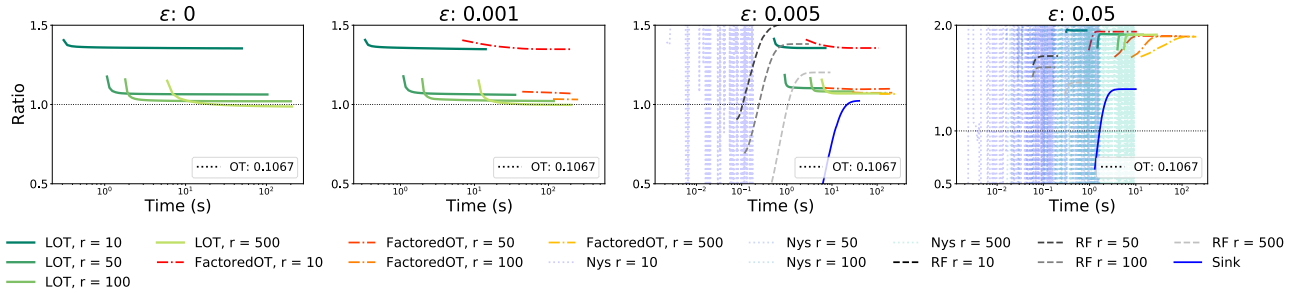

Figure 3. In this experiment, we consider two Gaussian distributions evaluated on $n = m = 5000$ in 2D. The first one has a mean of $(1, 1)^T$ and identity covariance matrix I_2 while the other has 0 mean and covariance $0.1 \times I_2$. The ground cost is the squared Euclidean distance. Note that for this cost, an exact low-rank factorization of the cost is available, and therefore all low-rank methods, including ours, have a linear time complexity. *Left:* we show that when $\varepsilon = 0$ our method is able to quickly obtain the exact OT by forcing the nonnegative rank of the coupling to be relatively small compared to the number of samples. Note that in this setting, all the other methods cannot be applied. *Middle left, middle right:* In these plots, we show that our method can obtain high accuracy for either estimate the true OT or its regularized version with order of magnitude faster than the other low-rank methods for any rank r. Moreover, our methods outperforms **Sin** in these regimes of small regularizations. Note that **Sin** does not converge for $\varepsilon = 0.001$ as we do not consider its stabilized version using log-sum-exp function but rather its classical version which is less costly to compute. *Right:* Here we change the scale of the y-axis of the plot. We see that the regime of the entropic regularizations for the Sinkhorn algorithm and our method differs. Indeed, the Sinkhorn algorithm has a larger range of ε such that it provides an efficient approximation of the OT, whereas **LOT** is regularizing *twice*, namely with respect to both rank *and* entropy.

3.5. Linear time approximation of the Low-Rank Optimal Transport

Here we aim at obtaining the optimal solution of Eq. (8) in linear time with respect to the number of samples. For that purpose let us introduce our main assumption on the cost matrix C.

Assumption 1. *Assume that C admits a low-rank factorization, that is there exists $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{m\times d}$ such that $C = AB^T$.*

From the Assumption 1 one can in fact accelerate the computation in the iterations of the proposed Alg. (3) and obtain a linear time algorithm with respect to the number of samples. Indeed recall that given $\boldsymbol{\xi} = (\xi^{(i)})_{1\leq i\leq 3}$, each iteration of the Dykstra's Alg. (2) can be performed in linear time. Moreover, thanks to Assumption 1, the computation of $\boldsymbol{\xi}$, which requires to compute both CR and C^TQ can be performed in $\mathcal{O}((n+m)dr)$ algebraic operations and thus Alg. (3) requires only a linear number of algebraic operations with respect to the number of samples at each iteration.

Let us now justify why the Assumption 1 of a low-rank factorization for the cost matrix is well suited in the problem of computing the Optimal Transport.

Squared Euclidean Metric. In the specific case where C is a Square Euclidean distance matrix, it admits a low-rank decomposition. Indeed let $X \triangleq [x_1, \dots, x_n] \in \mathbb{R}^{d\times n}$, let $Y \triangleq [y_1, \dots, y_m] \in \mathbb{R}^{d\times m}$ and let $D \triangleq (\|x_i - y_j\|_2^2)_{i,j}$. Then by denoting $p = [\|x_1\|_2^2, \dots, \|x_n\|_2^2]^T \in \mathbb{R}^n$ and $q = [\|y_1\|_2^2, \dots, \|y_m\|_2^2]^T \in \mathbb{R}^m$ we can rewrite D as the following:

$$D = p\mathbf{1}_m^T + \mathbf{1}_n q^T - 2X^TY.$$

Therefore by denoting $A = [p, \mathbf{1}_n, -2X^T] \in \mathbb{R}^{n\times(d+2)}$ and $B = [\mathbf{1}_m, q, Y^T] \in \mathbb{R}^{n\times(d+2)}$ we obtain that

$$D = AB^T.$$

General Case: Distance Matrix. In the following we denote a distance matrix $D \in \mathbb{R}^{n\times m}$, any matrix such that there exists a metric space $(\mathcal{X}, d)$, $\{x_i\}_{i=1}^n \in \mathcal{X}^n$ and $\{y_j\}_{j=1}^m \in \mathcal{X}^m$ which satisfy for all i, j, $D_{i,j} = d(x_i, y_j)$. In fact it is always possible to obtain a low-rank approximation of a distance matrix in linear time. In (Bakshi & Woodruff, 2018; Indyk et al., 2019), the authors proposed an algorithm such that for any distance matrix $D \in \mathbb{R}^{n\times m}$ and $\gamma > 0$ it outputs matrices $M \in \mathbb{R}^{n\times d}$, $N \in \mathbb{R}^{m\times d}$ in $\mathcal{O}((m+n)\text{poly}(\frac{d}{\gamma}))$ algebraic operations such that with probability at least 0.99 we have

$$\|D - MN^T\|_F^2 \leq \|D - D_d\|_F^2 + \gamma\|D\|_F^2$$

where D_d denotes the best rank-d approximation to D. Therefore one can always obtain a low-rank factorization of a distance matrix in linear time with respect to the number of samples. See Appendix D for more details.

4. Numerical Results

4.1. Comparison with other regularization schemes

We consider three synthetic problems in which we study the time-accuracy trade-off as well as the couplings obtained,

by comparing our method with other low-rank methods, as well as Sinkhorn's algorithm. More precisely, we compare our proposed method, **LOT**, with the factored Optimal Transport (Forrow et al., 2018), **FactoredOT**, the Nystrom-based method (Altschuler et al., 2018), **Nys**, the random features-based method (Scetbon & Cuturi, 2020), **RF** and the Sinkhorn algorithm (Cuturi, 2013), **Sin**. For **LOT**, we set the lower bound on g to $\alpha = 10^{-5}$.

Time-accuracy Tradeoff We consider two problems where the ground cost involved in the OT problem is either the *squared Euclidean* distance or the *Euclidean* distance. In the first one, we consider measures supported on $n = 5000$ points in $\mathbb{R}^2$, while the second we consider $n = 10000$ samples in $\mathbb{R}^2$. The method proposed by (Forrow et al., 2018) can only be used with the squared Euclidean distance (2-Wasserstein) while ours works for any cost. For all the low-ranks methods, we vary the ranks between 10 and 500. For all the randomized methods, we consider the mean over 10 runs to estimate the OT.

In Fig. 3, 4 we plot the ratio w.r.t. the (non-regularized) optimal transport cost defined as $\mathrm{R} := \langle C, \widetilde{P}\rangle / \langle C, P^*\rangle$ where $\widetilde{P}$ is the coupling obtained by the method considered and P^* is the ground truth (we ensure this optimal cost is large enough to avoid spurious divisions by 0). We present the time-accuracy tradeoffs of the methods for different regularizations ε and ranks r. We show that our method provides consistently a better approximation of the OT while being much faster than the other low-rank methods for various targeted rank values r. We also show that our method is able to approximate arbitrarily well the OT and so faster than the Sinkhorn algorithm thanks to the low-rank constraints. We compare the methods in the same setting but we increase the dimensionality of the problems considered and we observe similar results. See Appendix G for more details.

Remark 4. *Adding an entropic regularization in our objective allows to stabilize the MD scheme and therefore obtain faster convergence. Indeed if $\varepsilon > 0$, then the number of iterations required to solve each iteration of the MD scheme (10) by Algorithm (2) is monitored by ε given a certain precision δ while in the case where $\varepsilon = 0$, the number of iterations required for Algorithm 2 to reach the precision δ increases as the number of iterations in the MD scheme increases.*

Comparison of the Couplings Seeking to take a deeper look at the phenomenon highlighted in Fig. 1, we study differences in the regularization paths of **LOT** and **Sin**. We consider distributions supported on graphs of $n = 1000$ nodes, endowed with the shortest path distance (Bondy et al., 1976). We consider **LOT** with *no* entropic regularization (i.e. $\varepsilon = 0$ in Eq. (9)) against **Sin** for various pairs of regularizers. Results are displayed in Fig. 5, where the discrete path of regularizations parameterized by the rank r of **LOT** is compared with that obtained by **Sin** when varying ε. The gaps in couplings (in ℓ_1) between the two methods are displayed. Both methods are able to approximate arbitrarily well the OT but offer two different paths to interpolate from the independent coupling ab^T of rank 1 to the optimal one. More precisely, we see that the range of ε for which the entropic OT provides an efficient approximation of the true coupling is very localized, while the rank r needed for **LOT** to obtain such approximation is wider. Moreover, we see that the decay of the ratio of **LOT** with respect to r is faster than the decay of **Sin** w.r.t. ε.

Remark 5. *A comparative advantage of using the low-rank parameterization of OT over the Sinkhorn approach lies in the simple bounds that r admits, between 1 and n, and the fact that r encodes directly, through an integer, a direct property of the resulting coupling. In that sense, the same value r can be used across experiments that compare measures of various sizes and supports. By contrast, selecting a suitable regularization strength ε in the Sinkhorn algorithm is usually challenging, as the parameter is continuous and its magnitude depends directly on the cost matrix values, making a common choice across experiments difficult.*

Real World Application In Figure 6 we consider the single-cell trajectory inference problem (Schiebinger et al., 2019) where the goal is to infer the ancestors of some specific cells (iPSCs) from temporal snapshots sampled several times a day for a period of 18 days. We apply the exact same pre-processing suggested by (Schiebinger et al., 2019), and we obtain that our proposed method is able to recover a similar path as the one obtained by the Sinkhorn algorithm.

4.2. On the non-convexity of LOT

As our problem (6) is non-convex, we investigate the effect of the initialization as well as the choice of the gradient step γ in the proposed MD scheme. In addition, we consider a specific situation where the optimal coupling solution of (1) admits a nonnegative low-rank to see if our method is able to recover the global minimum in such situation. In the following experiments we set $\varepsilon = 0$ and the lower bound on g to $\alpha = 10^{-5}$.

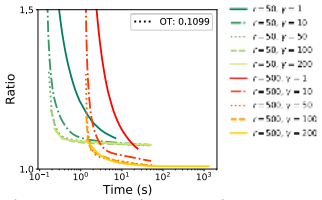

Figure 7. In this experiment, we consider the same situation as in Figure 3 with $n = m = 1000$ varying γ for $r = 50$ or 500.

Effect of γ In Figure 7, we plot the ratio on the same experiment presented in Figure 3 when varing γ. We show that our algorithm is robust to the choice of γ as it manages to converge for a large range of γ. Moreover if the rank is large

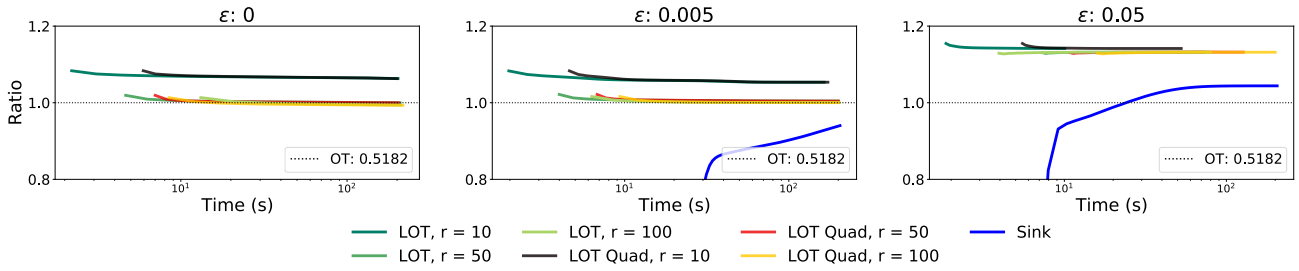

Figure 4. Here we consider two Gaussian mixture densities sampled with $n = m = 10000$ points in 2D (See Appendix G for more details). The ground cost is the Euclidean distance. As this cost is a distance, we can apply our linear version of the algorithm and we denote **LOT Quad** to refer to its quadratic counterpart. We see that **LOT** and **LOT Quad** provide similar results while **LOT** is faster. All kernel-based methods (**Nys**, **RF**) fail to converge in this setting. As in Fig. 3, we see that our method is able to approximate faster than **Sin** the true OT thanks to the low-rank constraint.

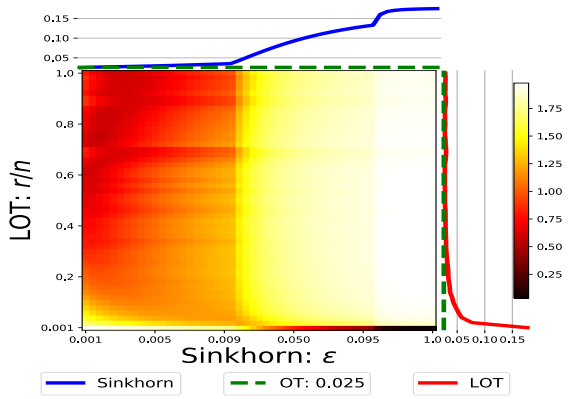

Figure 5. We illustrate in this plot the gaps between the couplings reached by **Sin** and **LOT** for varying regularization strengths. Measures were sampled on a complete graph obtained by sampling $2n = 2000$ points from a 2-D standard normal distribution, the edge weights set to their squared Euclidean distances. The supports are obtained by randomly splitting the nodes of the graphs into two subsets of same size. We vary the entropic regularization ε and the nonnegative rank r. We consider ε in log-scale ranging from 0.001 to 1 and r ranging from 1 to 1000, represented as a fraction of n. The blue (resp. red) curve stands for **Sin** (resp. **LOT**). We plot the ℓ_1 distance between their respective couplings.

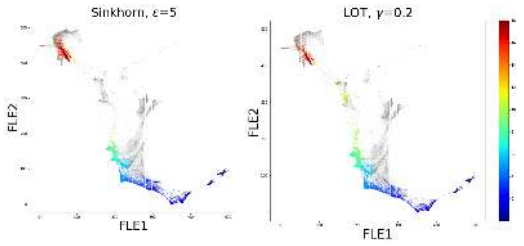

Figure 6. Here we compare the paths recovered by both the Sinkhorn algorithm with $\varepsilon = 5$, and our method with $\gamma = 1/\varepsilon$ and $r500$. Each sub-optimal transport problem between two temporal snapchots contains $n \simeq 5000$ cells.

enough, our method is able to find the optimal solution of the true OT problem (1).

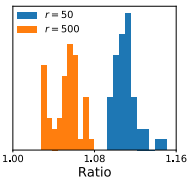

Figure 8. Same setting as in Figure 7.

Effect of the Initialization In Figure 8 we plot the ratios to LP solution of LOT costs, with 50 random initializations (Gaussian entries for Q, R, rescaled to have left/right marginals a and b). We show that our method is robust to the choice of the initialization. We also design an OT problem where the ground truth OT solution of Eq. (1) has low nonnegative rank. Indeed, by fixing $z_1, \dots, z_r \in \mathbb{R}^d$ anchors and by defining the cost $c(x, y) = \min_{k \in 1,..,r} \|x - z_k\| + \|z_k - y\|$, we show that the true optimal coupling has a low nonnegative rank r. Our algorithm recovers consistently the OT coupling for multiple random initializations. See Appendix H for more details.

Conclusion We proposed a new approach to regularize the OT problem by restricting solutions to have a small non-negative rank. Our algorithm leverages both low-rank constraints and entropic smoothing. Our method can leverage the factorization of the ground cost (and *not* that of the kernel usually associated to Sinkhorn) to propose a linear time complexity alternative to solve OT problems.

Acknowledgements This work was supported by a "Chaire d'excellence de l'IDEX Paris Saclay", by the European Research Council (ERC project NORIA) and by the French government under management of ANR as part of the "Investissements d'avenir" program (ANR19-P3IA-0001, PRAIRIE 3IA Institute).

References

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *arXiv preprint arXiv:1705.09634*, 2017.

Altschuler, J., Bach, F., Rudi, A., and Niles-Weed, J. Massively scalable sinkhorn distances via the nyström method, 2018.

Altschuler, J. M. and Boix-Adsera, E. Polynomial-time algorithms for multimarginal optimal transport problems with structure, 2020.

Bakshi, A. and Woodruff, D. P. Sublinear time low-rank approximation of distance matrices, 2018.

Bauschke, H. H. and Lewis, A. S. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

Bauschke, H. H., Bolte, J., and Teboulle, M. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Bondy, J. A., Murty, U. S. R., et al. *Graph theory with applications*, volume 290. Macmillan London, 1976.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3): 200–217, 1967.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33, 2020.

Clason, C., Lorenz, D. A., Mahler, H., and Wirth, B. 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.

Cohen, J. E. and Rothblum, U. G. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149 – 168, 1993. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(93)90224-C. URL http://www.sciencedirect.com/science/article/pii/002437959390224C.

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

Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *Proceedings of ICML*, volume 32, pp. 685–693, 2014.

Dessein, A., Papadakis, N., and Rouas, J.-L. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1367–1376. PMLR, 10–15 Jul 2018.

Dykstra, R. L. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. Statistical optimal transport via factored couplings, 2018.

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

Franklin, J. and Lorenz, J. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114: 717–735, 1989.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, 2013.

Heitz, M., Bonneel, N., Coeurjolly, D., Cuturi, M., and Peyré, G. Ground metric learning on graphs. *Journal of Mathematical Imaging and Vision*, pp. 1–19, 2020.

Indyk, P., Vakilian, A., Wagner, T., and Woodruff, D. Sample-optimal low-rank approximation of distance matrices, 2019.

Janati, H., Bazeille, T., Thirion, B., Cuturi, M., and Gramfort, A. Multi-subject meg/eeg source imaging with sparse multi-task regression. *NeuroImage*, pp. 116847, 2020.

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

Korotin, A., Li, L., Solomon, J., and Burnaev, E. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3tFAs5E-Pe.

Koundal, S., Elkin, R., Nadeem, S., Xue, Y., Constantinou, S., Sanggaard, S., Liu, X., Monte, B., Xu, F., Van Nostrand, W., 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.

Lin, T., Ho, N., and Jordan, M. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3982–3991. PMLR, 09–15 Jun 2019.

Lu, H., Freund, R. M., and Nesterov, Y. Relatively-smooth convex optimization by first-order methods, and applications, 2017.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681. PMLR, 13–18 Jul 2020.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019. ISSN 1935-8245.

Scetbon, M. and Cuturi, M. Linear time sinkhorn divergences using positive features, 2020.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

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

Schmitzer, B. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

Tong, A., Huang, J., Wolf, G., Van Dijk, D., and Krishnaswamy, S. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9526–9536. PMLR, 13–18 Jul 2020.

Xie, Y., Wang, X., Wang, R., and Zha, H. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in Artificial Intelligence*, pp. 433–453. PMLR, 2020.

Yang, K. D., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A. C., Shivashankar, G., and Uhler, C. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

Zhang, K. S., Peyré, G., Fadili, J., and Pereyra, M. Wasserstein control of mirror langevin monte carlo, 2020.