Supplementary material

In Sec. A, we introduce some important notions linked to the mirror-descent scheme. We also prove in this section a general result which states the non-asymptotic stationary convergence of the mirror-descent according to a specific criterion introcuded in this work. In Sec. B, we detail the computation of the Dykstra's algorithm 2 for which we have obtained a simple expression of the updates of the couplings. In Sec. C, we provides all the proofs of the Propositions introduced in this work in the main text. In Sec D, we detail the algorithm presented in (Indyk et al., 2019). In Sec. E, F, we give two variants of our algorithm when either the marginal g is fixed or when no lower bound is provided on the coordinates of g. In Sec. G, we provides more experiment to illustrate our method.

A. Mirror Descent Algorithm

Let $\mathcal{X}$ a closed convex subset in a Euclidean space $\mathbb{R}^q$, $f : \mathcal{X} \to \mathbb{R}$ continuously differentiable and let us consider the following problem

$$\min_{x \in \mathcal{X}} f(x). \tag{11}$$

Given a convex function $h : \mathcal{X} \to \mathbb{R}$ continuously differentiable, one can define the *Bregman Divergence* associated to h as

$$D_h(x, z) := h(x) - h(z) - \langle \nabla h(z), x - z \rangle.$$

To solve Eq. (11), one can employ the mirror-descent (MD) algorithm. Given an initial point $x_0 \in \mathcal{X}$ and a sequence of positive step-size $(\gamma_k)_{k \geq 0}$, the mirror-descent scheme associated to the *prox-function* D_h computes

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \langle \nabla f(x_k), x \rangle + \frac{1}{\gamma_k} D_h(x, x_k).$$

In the following, we need to introduce two notions of relative strong convexity and relative smoothness in order to prove non-asymptotic stationary convergence of the MD scheme.

Definition (Relative smoothness.)**.** *Let $L > 0$ and f continuously differentiable on $\mathcal{X}$. f is said to be L-smooth relatively to h if*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x)$$

Definition (Relative strong convexity)**.** *Let $\alpha > 0$ and f continuously differentiable on $\mathcal{X}$. f is said to be α-strongly convex relatively to h if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \alpha D_h(y, x) \; \forall \, x, y \in \mathcal{X}$$

Note that h is always 1-strongly convex relatively to h. Let us now prove a general result to show non-asymptotic stationary convergence of the MD scheme. For that purpose, we introduce for all $k \geq 0$ the following criterion to establish convergence:

$$\Delta_k \triangleq \frac{1}{\gamma_k^2} (D_h(x_k, x_{k+1}) + D_h(x_{k+1}, x_k)).$$

Proposition 5. *Let $N \geq 1$, f continuously differentiable on $\mathcal{X}$ which is L-smooth relatively to h. By considering for all $k = 1, \ldots, N$, $\gamma_k = 1/2L$, and by denoting $D_0 = f(x_0) - \min_{x \in \mathcal{X}} f(x)$, we have*

$$\min_{0 \leq k \leq N-1} \Delta_k \leq \frac{4 L D_0}{N}.$$

Proof. Let $k \geq 0$, then by L-smoothness of f, we have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + L D_h(x_{k+1}, x_k),$$

and by optimality of x_{k+1}, we have for all $x \in \mathcal{X}$,

$$\langle \nabla f(x_k) + \frac{1}{\gamma_k} [\nabla h(x_{k+1}) - \nabla h(x_k)], x - x_{k+1} \rangle \geq 0,$$

which implies, by taking $x = x_k$, that

$$\begin{aligned}\langle \nabla f(x_k), x_k - x_{k+1}\rangle &\geq \frac{1}{\gamma_k}[-\langle \nabla h(x_{k+1}), x_k - x_{k+1}\rangle - \langle \nabla h(x_k), x_{k+1} - x_k\rangle] \\ &\geq \frac{1}{\gamma_k}[D_h(x_k, x_{k+1}) + D_h(x_{k+1}, x_k)].\end{aligned}$$

Then we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{\gamma_k}[D_h(x_k, x_{k+1}) + D_h(x_{k+1}, x_k)] + LD_h(x_{k+1}, x_k) + LD_h(x_k, x_{k+1})$$

where the last term is added by positivity of $D_h(\cdot, \cdot)$ (as h is supposed to be convex on $\mathcal{X}$). Finally we obtain that

$$\left(\sum_{k=0}^{N-1} \gamma_k(1 - \gamma_k L)\Delta_k\right) \leq f(x_0) - f(x_N) \leq D_0,$$

and as soon as $\gamma_k < \frac{1}{L}$, we have

$$\min_{0\leq k\leq N-1} \Delta_k \leq \frac{D_0}{\left(\sum_{k=0}^{N-1} \gamma_k(1 - \gamma_k L)\right)}.$$

Then by taking $\gamma_k = \frac{1}{2L}$, the result follows. □

In this paper, we consider h to be the negative entropy function defined on Δ_q^* as

$$h(x) = \sum_{i=1}^{q} x_i \log(x_i). \tag{12}$$

Therefore the *prox-function* associated is just the Kullback–Leibler divergence (KL) defined as,

$$\mathrm{KL}(x, z) = \sum_{i=1}^{q} x_i \log(x_i/z_i).$$

Moreover if $\mathcal{X} \subset \prod_{i=1}^{p} \Delta_{q_i}^*$ for $p \geq 1$, we consider instead

$$h((x^{(1)}, \ldots, x^{(p)})) := \sum_{i=1}^{p}\sum_{j=1}^{q_i} x_j^{(i)} \log(x_j^{(i)})$$

where the associated *prox-function* is

$$D_h((x^{(1)}, \ldots, x^{(p)}), (z^{(1)}, \ldots, z^{(p)})) = \sum_{i=1}^{p} \mathrm{KL}(x^{(i)}, z^{(i)}).$$

B. The Dykstra's Algorithm

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 2 applied to our problem consists in computing for all $j \geq 0$,

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

In fact these operations can be simplified to simple matrix/vector multiplications. 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

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

C. Proofs

C.1. Proof of Proposition 1

Proof. The case when $\varepsilon = 0$ is clear. Assume now that $\varepsilon > 0$. When $r = 1$, note that $\mathcal{C}_1(a, b, r) \cap \mathcal{C}_2(r)$ is closed as $g = 1$ and bounded, therefore and by continuity of the objective the mininum exists. Let $r \geq 2$. First remarks that we always have $\mathrm{LOT}_{r,\varepsilon}(\mu, \nu) \leq \mathrm{LOT}_{r-1,\varepsilon}(\mu, \nu)$. Let us assume that (8) does not admits a minimum. Because the objective F_ε is a lower semi-continuous function on $\overline{\mathcal{C}_1(a, b, r)} \cap \mathcal{C}_2(r)$, and by compacity of $\overline{\mathcal{C}_1(a, b, r)} \cap \mathcal{C}_2(r)$, the objective function admits a minimum $(Q, R, g) \in \overline{\mathcal{C}_1(a, b, r)} \cap \mathcal{C}_2(r)$ and we have $\mathrm{LOT}_{r,\varepsilon}(\mu, \nu) = F_\varepsilon(Q, R, g)$. But as the minimum is not attained on $\mathcal{C}_1(a, b, r) \cap \mathcal{C}_2(r)$, it means that there exists at least one coordinate $i \in \{1, \ldots, r\}$ such that $g_i = 0$. Then because the constraints, Q and R both admit a column which is the null vector. By deleting these coordinates in Q, R, g, we obtain that $\mathrm{LOT}_{r,\varepsilon}(\mu, \nu) = \mathrm{LOT}_{r-1,\varepsilon}(\mu, \nu)$. □

C.2. Proof of Proposition 2

Proof. The first oder conditions of the projection gives that there exists $(\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_+^r$ such that

$$\begin{aligned}
\log(Q/\tilde{Q}) + \lambda_1 \mathbf{1}^T &= 0 \\
\log(R/\tilde{R}) + \lambda_2 \mathbf{1}^T &= 0 \\
\log(g/\tilde{g}) + \lambda_3 &= 0
\end{aligned}$$

Moreover the conditions $Q\mathbf{1} = a$, $R\mathbf{1} = b$ and $g \geq \alpha$ imply that

$$\begin{aligned} Q &= \text{Diag}(a/\tilde{Q}\mathbf{1})\tilde{Q} \\ R &= \text{Diag}(b/\tilde{R}\mathbf{1})\tilde{R} \\ g &= \max(\alpha, \tilde{g}). \end{aligned}$$

□

C.3. Proof of Proposition 3

Proof. The first order conditions of the projection states that there exists $(\lambda_1, \lambda_2) \in \mathbb{R}^r \times \mathbb{R}^r$ such that

$$\begin{aligned} \log(Q/\tilde{Q}) + \mathbf{1}_n \lambda_1^T &= 0 \\ \log(R/\tilde{R}) + \mathbf{1}_m \lambda_2^T &= 0 \\ \log(g/\tilde{g}) - (\lambda_1 + \lambda_2) &= 0 \end{aligned}$$

Moreover the conditions $Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g$ imply that

$$\begin{aligned} Q &= \tilde{Q}\text{Diag}(g/\tilde{Q}^T\mathbf{1}_n) \\ R &= \tilde{R}\text{Diag}(g/\tilde{R}^T\mathbf{1}_m) \\ g^3 &= \tilde{g} \odot \tilde{Q}^T\mathbf{1}_n \odot \tilde{R}^T\mathbf{1}_m \end{aligned}$$

from which the result follows. □

C.4. Proof of Proposition 4

Proof. To show the result, we just need to show that

$$F_\varepsilon : (Q, R, g) \in \mathcal{C}(a, b, r, \alpha) \to \langle C, Q\,\text{diag}(1/g)R^T \rangle - \varepsilon H(Q, R, g)$$

is smooth relatively to

$$H(Q, R, g) := \sum_{i,j} Q_{i,j} \log(Q_{i,j}) + \sum_{i,j} R_{i,j} \log(R_{i,j}) + \sum_j g_j \log(g_j),$$

then by applying Proposition 5, the result will follow. Let us now show that F_ε is $L_{\varepsilon,\alpha}$-smooth. To do so, it is enough to show that (Lu et al., 2017; Zhang et al., 2020)

$$\|\nabla F_\varepsilon(Q_1, R_1, g_1) - \nabla F_\varepsilon(Q_2, R_2, g_2)\|_2 \leq L_{\varepsilon,\alpha} \|H(Q_1, R_1, g_1) - H(Q_2, R_2, g_2)\|_2.$$

We first have that

$$\nabla F_\varepsilon(Q, R, g) = \big(CR\,\text{diag}(1/g) + \varepsilon(\log Q + \mathbf{1}), C^TQ\,\text{diag}(1/g) + \varepsilon(\log R + \mathbf{1}), -\mathcal{D}(Q^TRC)/g^2 + \varepsilon(\log g + 1)\big)$$

Now we have,

$$\begin{aligned} \|\nabla F_\varepsilon(Q_1) - \nabla F_\varepsilon(Q_2)\|_2^2 &\leq \|CR_1\,\text{diag}(1/g_1) - CR_2\,\text{diag}(1/g_2)\|_2^2 + \varepsilon^2 \|\log Q_1 - \log Q_2\|_2^2 \\ &\quad + 2\varepsilon \|\log Q_1 - \log Q_2\|_2 \|CR_1\,\text{diag}(1/g_1) - CR_2\,\text{diag}(1/g_2)\|_2 \\ &\leq \|C\|_2^2 \|(R_1 - R_2)\,\text{diag}(1/g_1) + (\text{diag}(1/g_1) - \text{diag}(1/g_2))R_2\|_2^2 + \varepsilon^2 \|\log Q_1 - \log Q_2\|_2^2 \\ &\quad + 2\varepsilon \|\log Q_1 - \log Q_2\|_2 \|CR_1\,\text{diag}(1/g_1) - CR_2\,\text{diag}(1/g_2)\|_2 \\ &\leq \|C\|_2^2 \left[\frac{\|R_1 - R_2\|_2^2}{\alpha^2} + \|1/g_1 - 1/g_2\|_2^2 + \frac{\|R_1 - R_2\| \|1/g_1 - 1/g_2\|_2}{\alpha} \right] + \varepsilon^2 \|\log Q_1 - \log Q_2\|_2^2 \\ &\quad + 2\varepsilon \|\log Q_1 - \log Q_2\|_2 \|CR_1\,\text{diag}(1/g_1) - CR_2\,\text{diag}(1/g_2)\|_2. \end{aligned}$$

As $Q \to H(Q)$ is 1-strongly convex w.r.t to the ℓ_2-norm on $\Delta_{n\times r}$, we have

$$\begin{aligned}\|Q_1 - Q_2\|_2^2 &\leq \langle \log Q_1 - \log Q_2, Q_1 - Q_2\rangle \\ &\leq \|\log Q_1 - \log Q_2\|_2 \|Q_1 - Q_2\|_2\end{aligned}$$

from which follows that

$$\|Q_1 - Q_2\|_2 \leq \log Q_1 - \log Q_2\|_2.$$

Moreover we have

$$\|1/g_1 - 1/g_2\|_2 \leq \frac{\|g_1 - g_2\|_2}{\alpha^2} \leq \|\frac{\|\log g_1 - \log g_2\|_2}{\alpha^2}$$

Therefore we obtain that

$$\|\nabla F_\varepsilon(Q_1) - \nabla F_\varepsilon(Q_2)\|_2^2 \leq \left(\frac{\|C\|_2}{\alpha}\|\log R_1 - \log R_2\|_2 + \frac{\|C\|_2}{\alpha^2}\|\log g_1 - \log g_2\|_2 + \varepsilon\|\log Q_1 - \log Q_2\|_2\right)^2.$$

An analogue proof leads to

$$\|\nabla F_\varepsilon(R_1) - \nabla F_\varepsilon(R_2)\|_2^2 \leq \left(\frac{\|C\|_2}{\alpha}\|\log Q_1 - \log Q_2\|_2 + \frac{\|C\|_2}{\alpha^2}\|\log g_1 - \log g_2\|_2 + \varepsilon\|\log R_1 - \log R_2\|_2\right)^2.$$

Let us now consider smoothness of F_ε w.r.t g,

$$\begin{aligned}\|\nabla F_\varepsilon(g_1) - \nabla F_\varepsilon(g_2)\|_2^2 \leq &\left\|\frac{\mathcal{D}(Q_1^T C R_1)}{g_1^2} - \frac{\mathcal{D}(Q_2^T C R_2)}{g_2^2}\right\|_2^2 + \varepsilon^2\|\log g_1 - \log g_2\|_2^2 \\ &+ 2\varepsilon\left\|\frac{\mathcal{D}(Q_1^T C R_1)}{g_1^2} - \frac{\mathcal{D}(Q_2^T C R_2)}{g_2^2}\right\|_2 \|\log g_1 - \log g_2\|_2.\end{aligned}$$

but we have that

$$\begin{aligned}\left\|\frac{\mathcal{D}(Q_1^T C R_1)}{g_1^2} - \frac{\mathcal{D}(Q_2^T C R_2)}{g_2^2}\right\|_2^2 &\leq \|(1/g_1^2 - 1/g_2^2)\operatorname{diag}(Q_1^T C R_1)\|_2^2 + \|\mathcal{D}(Q_1^T C R_1) - \mathcal{D}(Q_2^T C R_2)/g_2^2\|_2^2 \\ &\quad + 2\|(1/g_1^2 - 1/g_2^2)\operatorname{diag}(Q_1^T C R_1)\|_2 \|\mathcal{D}(Q_1^T C R_1) - \mathcal{D}(Q_2^T C R_2)/g_2^2\|_2 \\ &\leq \left(\frac{2\|C\|_2}{\alpha^3}\|\log g_1 - \log g_2\|_2 + \frac{\|C\|_2}{\alpha^2}\left[\|Q_1 - Q_2\|_2^2 + \|R_1 - R_2\|_2\right]\right)^2.\end{aligned}$$

Therefore we obtain that

$$\|\nabla F_\varepsilon(g_1) - \nabla F_\varepsilon(g_2)\|_2^2 \leq \left(\left(\varepsilon + \frac{2\|C\|_2}{\alpha^3}\right)\|\log g_1 - \log g_2\|_2 + \frac{\|C\|}{\alpha^2}\|Q_1 - Q_2\|_2 + + \frac{\|C\|}{\alpha^2}\|R_1 - R_2\|_2\right)^2$$

Finally we obtain that

$$\begin{aligned}\|\nabla F_\varepsilon(Q_1, R_1, g_1) - \nabla F_\varepsilon(Q_2, R_2, g_2)\|_2^2 \leq &\, 3\left(\frac{\|C\|_2^2}{\alpha^2} + \frac{\|C\|_2^2}{\alpha^4} + \varepsilon^2\right)\left[\|\log Q_1 - \log Q_2\|_2^2 + \|\log R_1 - \log R_2\|_2^2\right] \\ &+ 3\left(\frac{2\|C\|_2^2}{\alpha^4} + \left(\varepsilon + \frac{2\|C\|_2}{\alpha^3}\right)^2\right)\|\log g_1 - \log g_2\|_2^2\end{aligned}$$

Thus we obtain that

$$\|\nabla F_\varepsilon(Q_1, R_1, g_1) - \nabla F_\varepsilon(Q_2, R_2, g_2)\|_2 \leq L_{\varepsilon,\alpha}\|\nabla H(Q_1, R_1, g_1) - \nabla H(Q_2, R_2, g_2)\|_2$$

and the result follows. □

D. Low-Rank Factorization of Distance Matrix

In this section we present the algorithm used to perform a low-rank approximation of a distance matrix (Bakshi & Woodruff, 2018; Indyk et al., 2019). Given a metric space $(\mathcal{X}, d)$, $X = \{x_i\}_{i=1}^n \in \mathcal{X}^n$ and $Y = \{y_j\}_{j=1}^m \in \mathcal{X}^m$ we aim at obtaining a low-rank approximation of the distance matrix $D = (d(x_i, y_j))_{i,j}$ with a precision $\gamma > 0$. Let us now present the algorithm considered where we have denoted $t = \lfloor r/\gamma \rfloor$.

Algorithm 4 LR-Distance(X, Y, r, γ)

Inputs: X, Y, r, γ
Choose $i^* \in \{1, \dots, n\}$, and $j^* \{1, \dots, m\}$ uniformly at random.
For $i = 1, \dots, n$, $p_i \leftarrow d(x_i, y_j^*)^2 + d(x_i^*, y_j^*)^2 + \frac{1}{m}\sum_{j=1}^m d(x_i^*, y_j)^2$.
Independently choose $i^{(1)}, \dots, i^{(t)}$ according $(p_1, \dots, p_n)$.
$X^{(t)} \leftarrow [x_{i^{(1)}}, \dots, x_{i^{(t)}}]$, $P^{(t)} \leftarrow [\sqrt{t p_{i^{(1)}}}, \dots, \sqrt{t p_{i^{(t)}}}]$, $S \leftarrow d(X^{(t)}, Y)/P^{(t)}$
Denote $S = [S^{(1)}, \dots, S^{(m)}]$,
For $j = 1, \dots, m$, $q_j \leftarrow \|S^{(j)}\|_2^2 / \|S\|_F^2$
Independently choose $j^{(1)}, \dots, j^{(t)}$ according $(q_1, \dots, q_m)$.
$S^{(t)} \leftarrow [S^{j^{(1)}}, \dots, S^{j^{(t)}}]$, $Q^{(t)} \leftarrow [\sqrt{t q_{j^{(1)}}}, \dots, \sqrt{t q_{j^{(t)}}}]$, $W \leftarrow S^{(t)}/Q^{(t)}$
$U_1, D_1, V_1 \leftarrow \text{SVD}(W)$ (decreasing order of singular values).
$N \leftarrow [U_1(1), \dots, U_1^{(r)}]$, $N \leftarrow S^T N / \|W^T N\|_F$
Choose $j^{(1)}, \dots, j^{(t)}$ uniformly at random in $\{1, \dots, m\}$.
$Y^{(t)} \leftarrow [y_{j^{(1)}}, \dots, y_{j^{(t)}}]$, $D^{(t)} \leftarrow d(X, Y^{(t)})/\sqrt{t}$.
$U_2, D_2, V_2 = \text{SVD}(N^T N)$, $U_2 \leftarrow U_2/D_2$, $N^{(t)} \leftarrow [(N^T)^{(j^{(1)})}, \dots, (N^T)^{(j^{(t)})}]$, $B \leftarrow U_2^T N^{(t)}/\sqrt{t}$, $A \leftarrow (BB^T)^{-1}$.
$Z \leftarrow AB(D^{(t)})^T$, $M \leftarrow Z^T U_2^T$
Result: M, N

E. Positive low-rank factorization with fixed marginal

Let $g \in \Delta_r^*$, and let us for now consider the following problem

$$\text{LOT}_{r,g}(\mu, \nu) := \min_{P \in \Pi_{a,g,b}} \langle C, P \rangle. \tag{13}$$

By definition of $\Pi_{a,g,b}$, this problem can be formulated as follows:

$$\text{LOT}_{r,g}(\mu, \nu) = \min_{\substack{Q \in \Pi_{a,g} \\ R \in \Pi_{b,g}}} \langle C, Q \text{Diag}(1/g) R^T \rangle. \tag{14}$$

As in the classical OT problem, one can extend the above objective and consider for any $\varepsilon \geq 0$ an entropic version of the problem defined as

$$\text{LOT}_{r,g,\varepsilon}(\mu, \nu) := \min_{\substack{Q \in \Pi_{a,g} \\ R \in \Pi_{b,g}}} \langle C, Q \text{Diag}(1/g) R^T \rangle - \varepsilon H((Q, R)) \tag{15}$$

Note that for any $\varepsilon \geq 0$, the minimum always exists as the objective is continuous and $\Pi_{a,g,b}$ is compact. Moreover we clearly have that $\text{LOT}_{r,g,0}(\mu, \nu) = \text{LOT}_{r,g}(\mu, \nu)$. Applying a MD method to the objective (14) leads for all $k \geq 0$ to the following updates

$$Q_{k+1} := \underset{Q \in \Pi_{a,g}}{\text{argmin}} \langle C_k^{(1)}, Q \rangle - \frac{1}{\gamma_k} H(Q)$$
$$R_{k+1} := \underset{R \in \Pi_{a,g}}{\text{argmin}} \langle C_k^{(2)} R \rangle - \frac{1}{\gamma_k} H(R)$$

where, $(Q_0, R_0) \in \Pi_{a,g} \times \Pi_{b,g}$ is an initial point, $C_k^{(1)} := C R_k \text{Diag}(1/g) + (\varepsilon - \frac{1}{\gamma_k}) \log(Q_k)$, $C_k^{(2)} := C^T Q_k \text{Diag}(1/g) + (\varepsilon - \frac{1}{\gamma_k}) \log(R_k)$ and γ_k is a sequence of positive real numbers. Therefore a MD method bowls down to solve at each iteration two regularized OT problems which can be done efficiently using the Sinkhorn algorithm (1).

Convergence of the Mirror Descent. Even if the objective (14) is not convex in (Q, R), one can obtain the non-asymptotic stationary convergence of the MD algorithm in this setting.

Let f_ε be the objective function of the problem (15) defined on $X := \Pi_{a,g} \times \Pi_{b,g}$ and let us denotes for any $\gamma > 0$ and $x \in X$

$$\mathcal{G}_\varepsilon(x, \gamma) := \operatorname*{argmin}_{u \in X} \{\langle \nabla f_\varepsilon(x), u\rangle + \frac{1}{\gamma} KL(u, x)\}.$$

Let us now define the following criterion to establish convergence:

$$\Delta_\varepsilon(x, \gamma) := \frac{1}{\gamma^2}(KL(x, \mathcal{G}_\varepsilon(x, \gamma)) + KL(\mathcal{G}_\varepsilon(x, \gamma), x)).$$

To show the non-asymptotic stationary convergence, we show that for any $\varepsilon \geq 0$, the objective is smooth relative to the entropy function (Bauschke et al., 2017) and we extend the proof of (Ghadimi et al., 2013) to this case.

Proposition. *Let $\varepsilon \geq 0$ and $N \geq 1$. By denoting $L_\varepsilon := \sqrt{2(\|C\|_2^2 \|Diag(1/g)\|_2^2 + \varepsilon^2)}$ and by considering a constant stepsize in the MD scheme such that for all $k = 1, \ldots, N$ $\gamma_k = \frac{1}{L_\varepsilon}$, we obtain that*

$$\min_{1 \leq k \leq N} \Delta_\varepsilon((Q_k, R_k), \gamma_k) \leq \frac{2 L_\varepsilon D_0}{N}.$$

where $D_0 := f_\varepsilon(Q_0, R_0) - LOT_{r,g,\varepsilon}$ is the distance of the initial value to the optimal one.

Proof. A similar proof of the one given for Proposition 4 gives that f_ε is L_ε-smooth relatively to H. □

Let us now introduce our first algorithm (5) to compute a positive low-rank factorization of the optimal coupling. Here we consider the case where $g := \mathbf{1}_r/r$. Before introducing our algorithm it is worth noting that a trivial initialization may lead to a trivial fixed point in the MD updates. Indeed if one initialize $Q := ag^T$ and $R := bg^T$, then $CR\text{Diag}(1/g) = Ca\mathbf{1}^T$ and $C^TQ\text{Diag}(1/g) = C^Tb\mathbf{1}^T$ and therefore (Q, R) is a fixed point of the MD. To avoid this, we initialize our algorithm in the following way: let $\lambda := \min_{i,j,k}(a_i, b_j, g_k)/2$, $a_1 \in \Delta_n^* \backslash \{a\}$, $a_2 := (a - \lambda a_1)/(1 - \lambda)$, $b_1 \in \Delta_n^* \backslash \{b\}$, $b_2 := (b - \lambda b_1)/(1 - \lambda)$, $g_1 \in \Delta_r^* \backslash \{g\}$ and $g_2 := (g - \lambda g_1)/(1 - \lambda)$. We can now define our initialization as $Q := \lambda a_1 g_1^T + (1 - \lambda) a_2 g_2^T$, $R := \lambda b_1 g_1^T + (1 - \lambda) b_2 g_2^T$.

Algorithm 5 LOT-F(C, a, b, δ)

Inputs: $C, a, b, \delta, Q, R, g, \gamma, \delta_S$
repeat
  $Q_{\text{old}} \leftarrow Q,\ R_{\text{old}} \leftarrow R$
  $C^{(1)} \leftarrow CR\text{Diag}(1/g) - \frac{1}{\gamma}\log(Q)$,
  $C^{(2)} \leftarrow C^TQ\text{Diag}(1/g) - \frac{1}{\gamma}\log(R)$,
  $K^{(1)} \leftarrow \exp(-\gamma C^{(1)})$,
  $K^{(2)} \leftarrow \exp(-\gamma C^{(2)})$,
  $u, v \leftarrow \text{Sinkhorn}(K^{(1)}, a, g, \delta_S)$ (Algorithm (1)),
  $Q \leftarrow \text{Diag}(u)K^{(1)}\text{Diag}(v)$,
  $u, v \leftarrow \text{Sinkhorn}(K^{(2)}, a, g, \delta_S)$ (Algorithm (1)),
  $R \leftarrow \text{Diag}(u)K^{(2)}\text{Diag}(v)$
until $\Delta((Q, R), \gamma) < \delta$;
Result: Q, R

Computational Cost. Note that the kernels $(K^{(i)})_{1 \leq i \leq 2}$ considered in algorithm (5) live in $\mathbb{R}_+^{n \times r} \times \mathbb{R}_+^{m \times r}$ and therefore each iteration of both Sinkhorn algorithms can be computed either in $\mathcal{O}(nr)$ or in $\mathcal{O}(mr)$ algebraic operations as it involves only matrix/vector multiplications of the form $K^{(i)}v$ and $(K^{(i)})^Tu$. However without any assumption on the cost matrix C, computing $(K^{(i)})_{1 \leq i \leq 2}$ costs $\mathcal{O}(nmr)$ algebraic operations as it requires to compute both CR and C^TQ at each iteration. Thanks to assumption 1, such multiplications can be performed in $\mathcal{O}((n+m)dr)$ algebraic operations and thus algorithm (5) requires only a linear number of algebraic operations with respect to the number of samples at each iteration.

In the following, we will see that if we do not fix the marginal, the problem can also be solved efficiently as each iteration of the MD algorithm can be seen as a wasserstein barycenter problem.

F. A Positive low-rank factorization with free marginal

Applying a MD method to the objective (8) leads, for all $k \geq 0$, to the following updates

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\boldsymbol{\zeta} \in \mathcal{C}_1(a,b,r) \cap \mathcal{C}_2(r)} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_k) \tag{16}$$

where $(Q_0, R_0, g_0) \in \mathcal{C}_1(a,b,r) \cap \mathcal{C}_2(r)$ is an initial point, $\boldsymbol{\xi}_k := (\xi_k^{(1)}, \xi_k^{(2)}, \xi_k^{(3)})$, $\xi_k^{(1)} := \exp(-\gamma_k C R_k \mathrm{Diag}(1/g_k)_k - (\gamma_k \varepsilon - 1)\log(Q_k))$, $\xi_k^{(2)} := \exp(-\gamma_k C^T Q_k \mathrm{Diag}(1/g_k) - (\gamma_k\varepsilon - 1)\log(R_k))$, $\xi_k^{(3)} := \exp(\gamma_k \omega_k / g_k^2 - (\gamma_k \varepsilon - 1)\log(g_k))$ with $[\omega_k]_i := [Q_k^T C R_k]_{i,i}$ for all $i \in \{1, \dots, r\}$ and $(\gamma_k)_{k \geq 0}$ is a sequence of positive real numbers.

Eq. (16) is well defined. Indeed as the kernels $(\xi_k^{(i)})$ are matrices with positive coefficients, the infimum is attained in $\mathcal{C}_1(a,b,r) \cap \mathcal{C}_2(r)$ and the problem admits a unique solution. Moreover solving Eq. (16) bowls down to solve

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\boldsymbol{\zeta} \in \overline{\mathcal{C}_1(a,b,r)} \cap \mathcal{C}_2(r)} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_k) \tag{17}$$

In order to solve Eq. (17), we consider the Iterative Bregman Projections (IBP) algorithm. Starting from $\boldsymbol{\zeta}_0^{(k)} := \boldsymbol{\xi}_k$, the IBP algorithm consists in computing for all $j \geq 0$,

$$\boldsymbol{\zeta}_{2j+1}^{(k)} = \mathcal{P}^{\mathrm{KL}}_{\overline{\mathcal{C}_1(a,b,r)}}(\boldsymbol{\zeta}_{2j}^{(k)})$$
$$\boldsymbol{\zeta}_{2j+2}^{(k)} = \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_2(r)}(\boldsymbol{\zeta}_{2j+1}^{(k)}).$$

As $\overline{\mathcal{C}_1(a,b,r)}$ and $\mathcal{C}_2(r)$ are affine subspaces (note that nonnegativity constraints are already in the definition of the objective) one can show that $\boldsymbol{\zeta}_j^{(k)}$ converges towards the unique solution of Eq. (17), (Bregman, 1967). Remarks that the projection on $\overline{\mathcal{C}_1(a,b,r)}$ can be computed very easily as one has for any $\tilde{\boldsymbol{\xi}} := (\tilde{Q}, \tilde{R}, \tilde{g}) \in \mathbb{R}_+^{n \times r} \times \mathbb{R}_+^{n \times r} \times \mathbb{R}_+^r$,

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

and the solution of the projection on $\mathcal{C}_2(r)$ is already given in Proposition 3.

Efficient computation of the updates. For all $k \geq 0$, starting with $\boldsymbol{\zeta}_0^{(k)} := \boldsymbol{\xi}_k$ the IBP algorithm leads to a simple algorithm (6) which computes only scaling vectors. More precisely, the IBP algorithm produces the iterates $(\boldsymbol{\zeta}_n^{(k)})_{n \geq 0}$ which satisfy for all $n \geq 0$ $\boldsymbol{\zeta}_n^{(k)} = (Q_n^{(k)}, R_n^{(k)}, g_n^{(k)})$ where

$$Q_n^{(k)} = \mathrm{Diag}(u_n^{k,1}) \xi_k^1 \mathrm{Diag}(v_n^{k,1})$$
$$R_n^{(k)} = \mathrm{Diag}(u_n^{k,2}) \xi_k^2 \mathrm{Diag}(v_n^{k,2})$$

for the sequences $(u_n^{k,i}, v_n^{k,i})$ initialized as $v_0^{k,i} := \mathbf{1}$ for all $i \in \{1, 2\}$ and computed with the iterations

$$u_n^{k,i} = \frac{p_i}{\xi_k^i v_n^{k,i}}$$
$$g_{n+1}^{(k)} = (g_n^{(k)})^{1/3} \prod_{i=1}^{2} (v_n^{k,i} \odot (\xi_k^i)^T u_n^{k,i})^{1/3}$$
$$v_{n+1}^{k,i} = \frac{g_{n+1}^{(k)}}{(\xi_k^i)^T u_n^{k,i}}$$

where we have denoted $p_1 := a$ and $p_2 := b$ to simplify the notations.

Algorithm 6 LR-IBP$((\xi^{(i)})_{1\leq i\leq 3}, p_1, p_2, \delta)$

Inputs: $\xi^{(1)}, \xi^{(2)}, g := \xi^{(3)}, p_1, p_2, \delta, v^{(i)}$
repeat
  $u^{(i)} \leftarrow p_i / \xi^{(i)} v^{(i)} \; \forall i \in \{1,2\}$,
  $g \leftarrow (g)^{1/3} \prod_{i=1}^{2} (v^{(i)} \odot (\xi^{(i)})^T u^{(i)})^{1/3}$,
  $v^{(i)} \leftarrow g / (\xi^{(i)})^T u^{(i)} \; \forall i \in \{1,2\}$
until $\sum_{i=1}^{2} \|u^{(i)} \odot \xi^{(i)} v^{(i)} - p_i\|_1 < \delta$;
$Q \leftarrow \text{Diag}(u^{(1)}) \xi_k^{(1)} \text{Diag}(v^{(1)})$
$R \leftarrow \text{Diag}(u^{(2)}) \xi_k^{(2)} \text{Diag}(v^{(2)})$
Result: Q, R, g

Let us now introduce the proposed MD algorithm applied to (7). By denoting $\mathcal{D}(\cdot)$ the operator extracting the diagonal of a square matrix we obtain the following algorithm (7) to solve Eq. (6). We initialize our algorithm with the exact same procedure as in algorithm (5).

Algorithm 7 LOT(C, a, b, r, δ)

Inputs: $C, a, b, (\gamma_k)_{k\geq 0}, Q, R, g, \delta$
for $k = 1, \ldots$ **do**
  $\xi^{(1)} \leftarrow \exp(-\gamma_k C R \text{Diag}(1/g) - (\gamma_k \varepsilon - 1)\log(Q))$,
  $\xi^{(2)} \leftarrow \exp(-\gamma_k C^T Q \text{Diag}(1/g) - (\gamma_k \varepsilon - 1)\log(R))$,
  $\omega \leftarrow \mathcal{D}(Q^T C R)$, $\xi^{(3)} \leftarrow \exp(\gamma_k \omega / g^2 - (\gamma_k \varepsilon - 1)\log(g))$,
  $Q, R, g \leftarrow$ LR-IBP$((\xi^{(i)})_{1\leq i\leq 3}, a, b, \delta)$ (Algorithm (6))
end
Result: $\langle C, Q \text{Diag}(1/g) R^T \rangle$

Computational Cost. Note that $(\xi^{(i)})_{1\leq i\leq 3}$ considered in algorithm (7) lives in $\mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \times \mathbb{R}_+^r$ and therefore each iteration of algorithm (6) can be computed in $\mathcal{O}((n+m)r)$ algebraic operations as it involves only matrix/vector multiplications of the form $\xi^{(i)} v_i$ and $(\xi^{(i)})^T u_i$. However without any assumption on the cost matrix C, computing $(\xi^{(i)})_{1\leq i\leq 3}$ costs $\mathcal{O}(nmr)$ algebraic operations as it requires to compute both CR and $C^T Q$ at each iteration. Thanks to assumption 1, such multiplications can be performed in $\mathcal{O}((n+m)dr)$ algebraic operations and thus algorithm (7) requires only a linear number of algebraic operations with respect to the number of samples at each iterations.

G. Addiational Experiments

In Fig. 4, we compare two Gaussian mixture densities sampled with $n = m = 10000$ points in 2D. The two densities considered are

$$f_X(x) = \frac{1}{3}\frac{\exp\left((x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\sqrt{2\pi|\Sigma|}} + \frac{1}{3}\frac{\exp\left((x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)}{\sqrt{2\pi|\Sigma|}} + \frac{1}{3}\frac{\exp\left((x-\mu_3)^T\Sigma^{-1}(x-\mu_3)\right)}{\sqrt{2\pi|\Sigma|}}$$
$$f_Y(x) = \frac{1}{2}\frac{\exp\left((x-\nu_1)^T\Sigma^{-1}(x-\nu_1)\right)}{\sqrt{2\pi|\Sigma|}} + \frac{1}{2}\frac{\exp\left((x-\nu_2)^T\Sigma^{-1}(x-\nu_2)\right)}{\sqrt{2\pi|\Sigma|}}$$

where

$$\mu_1 = [0,0], \quad \mu_2 = [0,1], \quad \mu_3 = [1,1], \quad \nu_1 = [0.5, 0.5], \quad \nu_2 = [-0.5, 0.5], \quad \Sigma = 0.05 \times \text{Id}_2.$$

We show in Fig. 9 a plot of the two distributions considered. In Fig. 10, we consider the exact same setting as the one presented in Fig. 4 but we increase the dimension of the problem. More precisely we consider two Gaussian mixture

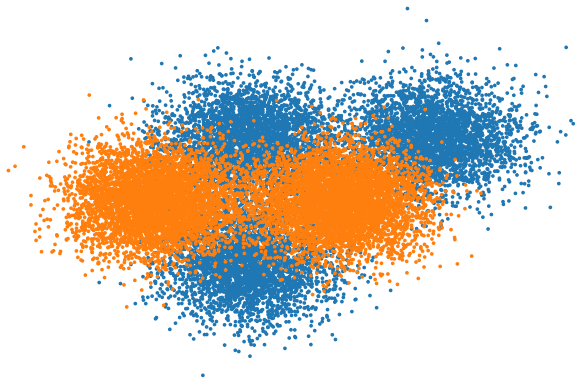

Figure 9. Plot of the Gaussian mixtures considered in Fig. 4.

densities samples with $n = m = 10000$ points in 10D where

$$\mu_1 = [0, \dots, 0],\ \mu_2 = [0, 1, 0, \dots, 0],\ \mu_3 = [1, 1, 0, \dots, 0],$$
$$\nu_1 = [0.5, 0.5, 0, \dots, 0],\ \nu_2 = [-0.5, 0.5, 0, \dots, 0],$$
$$\Sigma = 0.05 \times \mathrm{Id}_{10}.$$

Similarly as in Fig. 4, we observe that **LOT** and **LOT Quad** provide similar results while **LOT** is faster. All kernel-based methods fail to converge in this setting. Moreover we see that for small regularizations ε, our method is able to approximate faster than **Sin** the true OT thanks to the low-rank constraint. Note also that we observe again a difference between the two entropic regularizations of the **Sin** objective and **LOT** objective. Indeed the range of ε where **Sin** provides an efficient approximation of the true OT is larger than the one of **LOT**. Indeed recall that for **LOT**, we regularize *twice* as we constraint the nonnegative rank of the couplings and we add an entropic term to regularize the objective.

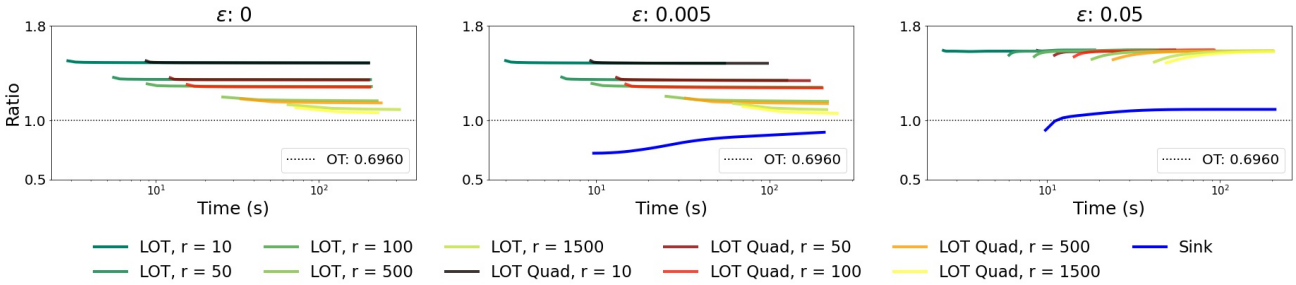

Figure 10. Comparison of the time-accuracy tradeoff for different methods for estimating the OT or its regularized version between two mixture of gaussians in 10D.

In Fig. 3, we compare the time-accuracy tradeoff for different methods on a synthetic problem where we aim at estimating either the OT or its regularized version between two gaussians in 2D. Here we consider the exact same setting but we increase the dimension of the problem: $d = 10$. As in Fig. 3, our proposed method obtains an efficient approximation of the OT or its regularized version for all rank r faster than other low-rank methods in the regime of small ε. We also see that for all low-rank methods, a rank of $r = 500$ is not enough in this setting to obtain the exact OT, but as the rank increases, the approximation gets better.

H. Tight solution

Let $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_m)$, $a, b \in \Sigma_n, \Sigma_m$ be probability weights, and $Z = (z_1, \dots, z_k)$ be points in a set endowed with a cost c. We consider the network problem from sources X to target Y passing through Z. This

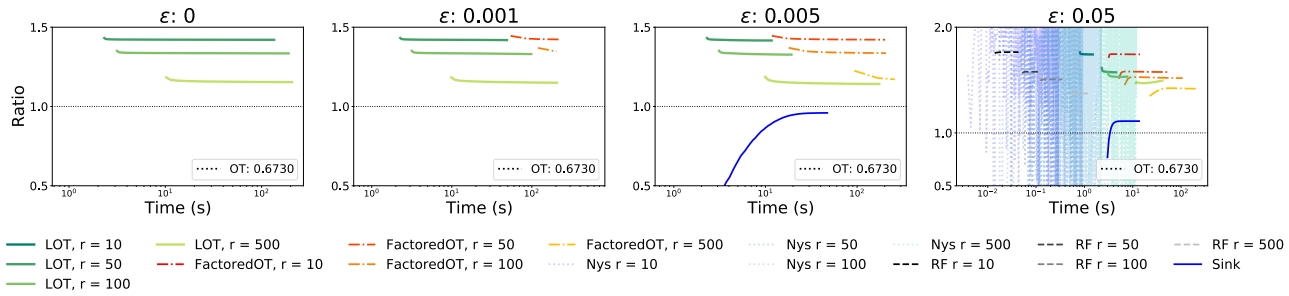

Figure 11. In this experiment, we consider two Gaussian distributions evaluated on $n = m = 5000$ in 10D. The first one has a mean of $(1, \dots, 1)^T \in \mathbb{R}^{10}$ and identity covariance matrix I_{10} while the other has 0 mean and covariance $0.1 \times I_{10}$. The ground cost is the squared Euclidean distance.

is equivalent to solving the regular $n \times m$ OT problem with cost matrix $C_{ij} = \min_k c(x_i, z_k) + c(z_k, y_j)$. We write $k_{ij} = \operatorname{argmin}_k c(x_i, z_k) + c(z_k, y_j)$, $D = [c(x_i, z_k)]_{ik}$ and $D' = [c(z_k, y_j)]_{kj}$.

Lemma 1. *Let $P^\star$ be an optimal solution for the problem $\min_{P \in U(a,b)} \langle P, C \rangle$. Write*

$$g^*_k = \sum_{i,j} P_{ij} 1_{k=k_{ij}}, U^*_{ik} = \sum_j P_{ij} 1_{k=k_{ij}}, V^*_{kj} = \sum_i P_{ij} 1_{k=k_{ij}}$$

Then matrices $U^* \in U(a, g^*), V^* \in U(g^*, b)$ and are respectively optimal for the OT problems with costs D and D' respectively. Additionally, $\langle P^\star, C \rangle = \langle U^*, D \rangle + \langle V^*, D' \rangle$.

Proof. It is easy to check that $U^* \in U(a, g^*), V^* \in U(g^*, b)$ and that we have:

$$\langle P^\star, C \rangle = \langle U^*, D \rangle + \langle V^*, D' \rangle$$

Moreover let $U \in U(a, g^*), V \in U(g^*, b)$, then we have

$$\begin{aligned}
\langle P^\star, C \rangle \leq \langle C, UD(1/g^*)V \rangle &= \sum_k \frac{1}{g^*_k} \sum_{ij} C_{ij} U_{ik} V_{kj} \\
&= \sum_k \frac{1}{g^*_k} \sum_{i,j} \min_{k'} (D_{ik'} + D_{k'j}) U_{ik} V_{kj} \\
&\leq \sum_k \frac{1}{g^*_k} \sum_{i,j} (D_{ik} + D_{kj}) U_{ik} V_{kj} \\
&\leq \sum_k \frac{1}{g^*_k} \sum_{i,j} D_{ik} U_{ik} V_{kj} + \sum_{i,j} D_{kj} U_{ik} V_{kj} \\
&\leq \sum_k \sum_i D_{ik} U_{ik} + \sum_j D_{kj} V_{kj} \\
&= \langle U, D \rangle + \langle V, D' \rangle
\end{aligned}$$

Therefore for any $U \in U(a, g), V \in U(g, b)$ we have

$$\langle U^*, D \rangle + \langle V^*, D' \rangle \leq \langle U, D \rangle + \langle V, D' \rangle$$

from which follows the optimality of U^* and V^*. □

Proposition 6. *$U^* D(1/g^*) V^*$ is optimal for the OT problem between X and Y with costs C.*

Proof. Obviously $U^* D(1/g^*) V^*$ has the right marginals. Moreover from the computation obtained in the proof of Lemma 1, we have

$$\langle C, U^* D(1/g^*) V^* \rangle \leq \langle U^*, D \rangle + \langle V^*, D' \rangle = \langle P^\star, C \rangle$$

from which follows the optimality of $U^* D(1/g^*) V^*$. □

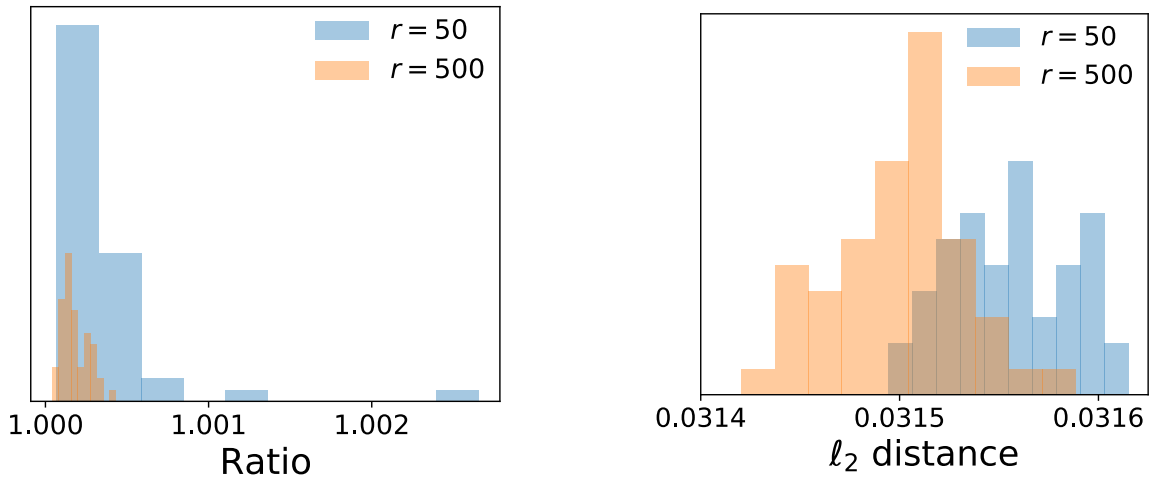

Figure 12. Here we consider the same setting as in Figure 10 where the cost functions is defined as $c(x, y) = \min_{k \in 1,..,r} \|x - z_k\| + \|z_k - y\|$ and $z_1, \dots, z_r \in \mathbb{R}^{10}$ are fixed anchors.

In the following experiment we aim at showing that our method is able to recover the exact true solution of Eq. (1) when the optimal coupling admits a low nonnegative rank. Moreover we show that our algorithm is robust to the choice of the initialization. Indeed in Figure 12, we plot both the histograms of the ratios to the LP solution of LOT costs and the ℓ_2 distance between the true optimal coupling and the coupling obtained by our algorithm for multiple random initializations. We show that our method is able to recover consistenly the true optimal coupling.