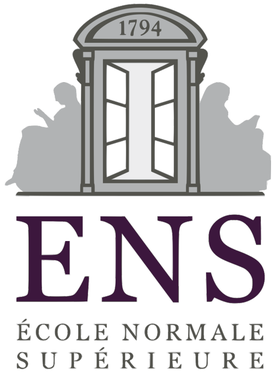
Low-Rank Sinkhorn Factorization

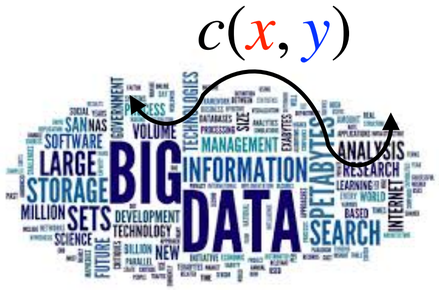
M. Scetbon M. Cuturi G. Peyré

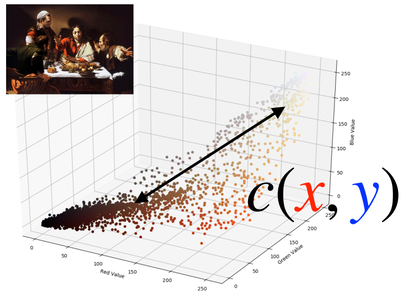


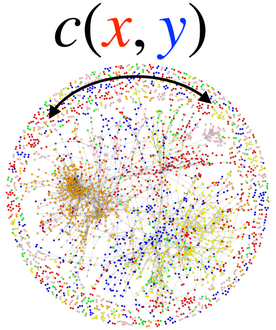
Thirty-eighth International Conference on Machine Learning

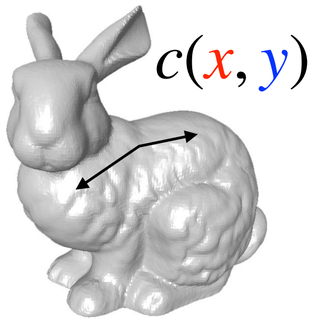
Optimal Transport: Comparing Distributions

Some Examples of Probability Measures:



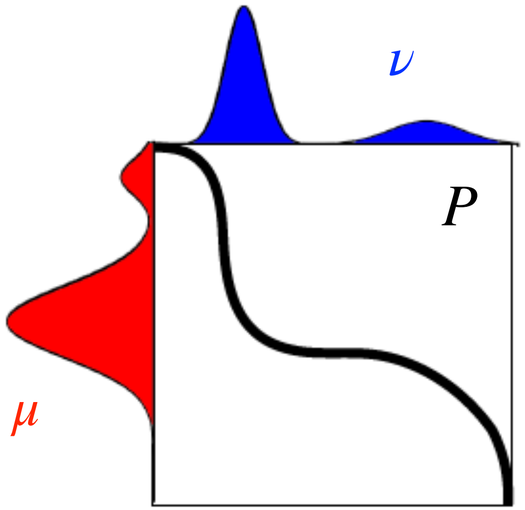






Discrete Optimal Transport:

- Discrete Distributions: $\mu = \sum_{i=1}^{n} a_i \, \delta_{x_i} \in \mathscr{M}_1^+(\mathscr{X}) \quad , \quad \nu = \sum_{j=1}^{m} b_j \, \delta_{y_j} \in \mathscr{M}_1^+(\mathscr{Y})$
- Set of Couplings: $\Pi_{a,b} = \left\{ P \in \mathbb{R}_+^{n \times m} \text{ s.t. } P\mathbf{1}_m = a, P^T\mathbf{1}_n = b \right\}$
- Cost Matrix: $\forall i, j \; C_{i,j} = c(x_i, y_j)$ where $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$



$$W_c(\mu, \nu) = \min_{P \in \Pi_{a,b}} \langle P, C \rangle$$

Costly to compute ⟶ Linear Program: $\mathcal{O}(n^3 \log(n))$ complexity

Entropic Regularization

Shannon Entropy: $\mathrm{H}(P) = -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1)$

Regularized Optimal Transport:

$$W_{c,\varepsilon}(\mu, \nu) := \min_{P \in \mathbb{R}_+^{n \times m}} \langle P, C \rangle - \varepsilon \mathrm{H}(P)$$
$$P\mathbf{1}_m = a, P^T\mathbf{1}_n = b$$

Sinkhorn Algorithm:

- Kernel Matrix: $K := \exp(-C/\varepsilon)$
- Until convergence, at each iteration compute : $v \leftarrow \frac{b}{K^T u}$, $u \leftarrow \frac{a}{Kv}$
- Output: $P_\varepsilon^* = \mathrm{Diag}(u) K \mathrm{Diag}(v)$

Computing $K^T u$ and Kv requires $\mathcal{O}(nm)$ operations

Quadratic time algorithm

Low-Rank Approximation of the Kernel:

Replace K in the Sinkhorn iterations by $\widetilde{K} = AB^T$ where $(A, B) \in (\mathbb{R}_*^+)^{n \times r} \times (\mathbb{R}_*^+)^{m \times r}$

Linear time algorithm

$\widetilde{P_\varepsilon^*} = \mathrm{Diag}(u) A B^T \mathrm{Diag}(v)$ admits a low rank factorization

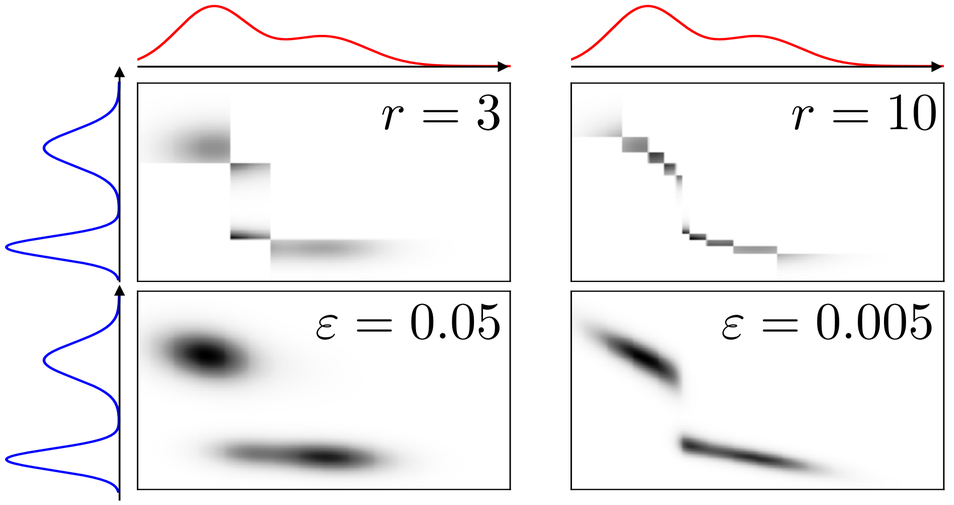
⟶ Computing $BA^T u$ and $AB^T v$ requires $\mathcal{O}(nr)$ algebraic operations

In this work, we propose *instead* to directly constraint the coupling to admit a low-NN rank

Low-Rank Optimal Transport

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

Low-NN Rank Couplings: $\Pi_{a,b}(r) := \left\{ P \in \Pi_{a,b} \text{ s.t. } \mathrm{rk}_+(P) \leq r \right\}$



Definition of Low-rank Optimal Transport

$$\mathrm{LOT}_r(\mu, \nu) := \min_{P \in \Pi_{a,b}(r)} \langle P, C \rangle$$

Characterization of Low-NN Rank Couplings:

$$\Pi_{a,b}(r) = \{ P \in \mathbb{R}_+^{n \times m} \,|\, P = Q \mathrm{Diag}(1/g) R^T,\ Q \in \Pi_{a,g}, R \in \Pi_{b,g},\ g > 0 \text{ and } g \in \Delta_r \}$$

Reparametrization of LOT

$$\mathrm{LOT}_r(\mu,\nu) = \min_{(Q,R,g)\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)} \langle C, Q\mathrm{Diag}(1/g)R^T\rangle$$

where
$$\begin{cases} \mathcal{C}_1(a,b,r) := \left\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times(\mathbb{R}_+^*)^r \text{ s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b\right\} \\ \mathcal{C}_2(r) := \left\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times(\mathbb{R}_+)^r \text{ s.t. } Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g\right\} \end{cases}$$

Entropic Regularization of LOT

$$\mathrm{LOT}_{r,\varepsilon}(\mu,\nu) := \inf_{(Q,R,g)\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)} \langle C, Q\mathrm{Diag}(1/g)R^T\rangle - \varepsilon H((Q,R,g))$$

Proposition

If $\epsilon = 0$ then the infimum is attained. If $\epsilon > 0$, then if $r = 1$, the infimum is attained and for $r \geq 2$, the problem admits a minimum if $\mathrm{LOT}_{r,\varepsilon}(\mu,\nu) < \mathrm{LOT}_{r-1,\varepsilon}(\mu,\nu)$.

Mirror Descent Scheme

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \min_{(\zeta_1,\zeta_2,\zeta_3)\in\mathscr{C}_1(a,b,r)\cap\mathscr{C}_2(r)} \mathrm{KL}((\zeta_1,\zeta_2,\zeta_3),(K_k^{(1)},K_k^{(2)},K_k^{(3)}))$$

Quadratic time algorithm

where $(Q_0, R_0, g_0) \in \mathscr{C}_1(a,b,r)\cap\mathscr{C}_2(r)$, $K_k^{(1)} := \exp(-\gamma_k C R_k \mathrm{Diag}(1/g_k) - (\gamma_k\varepsilon - 1)\log(Q_k))$,

$K_k^{(2)} := \exp(-\gamma_k C^T Q_k \mathrm{Diag}(1/g_k) - (\gamma_k\varepsilon - 1)\log(R_k))$, $K_k^{(3)} := \exp(\gamma_k\omega_k/g_k^2 - (\gamma_k\varepsilon - 1)\log(g_k))$, $[\omega_k]_i := [Q_k^T C R_k]_{i,i}$, $\gamma_k > 0$

Remarks:

- Each iteration of the MD is a convex problem which can be solved efficiently using the IBP algorithm
- Given $(K_k^{(i)})_{i=1}^3$, each iteration of the IBP algorithm requires $\mathcal{O}((n+m)r)$ algebraic operations
- Computing $(K_k^{(i)})_{i=1}^2$ requires $\mathcal{O}(nmr)$ algebraic operations

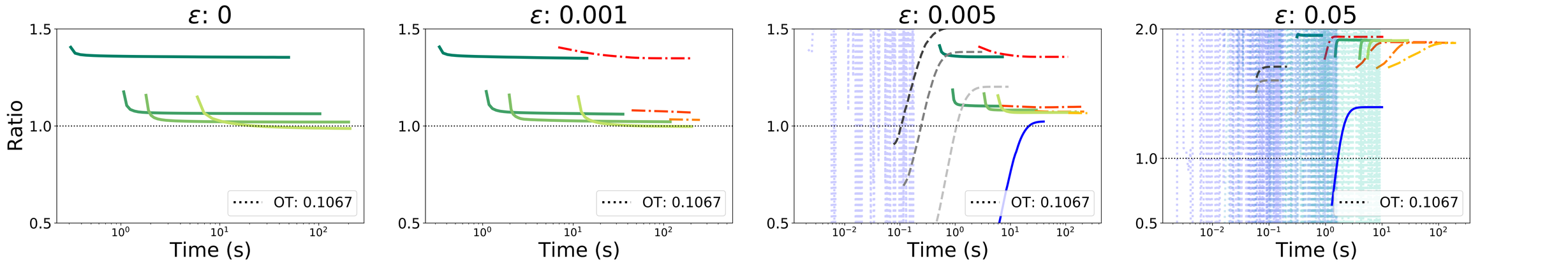
Linear time algorithm

Linear Time Approximation

If $C \simeq AB^T$ where $(A,B) \in \mathbb{R}^{n\times d}\times\mathbb{R}^{m\times d}$ and $d \ll \min(n,m)$ ⟶ $(K_k^{(i)})_{i=1}^2$ can be performed in $\mathcal{O}((n+m)dr)$

Example: The squared Euclidean distance, or more generally any distance matrix.

Experiments



Thank you

Advantages of the proposed method

→ Faster to compute than Sinkhorn.

→ Its parametrization, which is the rank r, encodes directly a property of the resulting coupling.