Low-Rank Sinkhorn Factorization

Meyer SCETBON, Marco CUTURI, Gabriel PEYRÉ

Overview

Problem: How to obtain an efficient procedure to compute with high accuracy the Optimal Transport?

Contributions:

- We introduce *LOT* (**L**ow-rank **O**ptimal **T**ransport), a new regularization scheme which aims at solving the optimal transport problem under the constraint that the couplings have a *low-nonnegative rank*.
- We propose a new *parametrization* of the problem where the couplings can be decomposed as the product of *two couplings* with common right marginal.
- We derive a simple algorithm using a *mirror-descent* approach and prove the non-asymptotic *stationary convergence*.
- We show that our approach can become *linear in time* with respect to the number of samples by exploiting *low-rank* assumptions on the *cost* **(not the kernel)**.
- We show *experimentally* the efficiency of our approach.

Sinkhorn Algorithm

Discrete Distributions: $\mu = \sum_{i=1}^{n} a_i\, \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j\, \delta_{y_j}$

Cost matrix: $\forall i,j\ C_{i,j} = c(x_i, y_j)$

Definition of Entropic Optimal Transport

Shannon entropy

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

Sinkhorn algorithm

Compute the kernel matrix: $K = \exp(-C/\varepsilon)$

Each iteration compute : $v \leftarrow \dfrac{b}{K^T u}$, $u \leftarrow \dfrac{a}{Kv}$

Output: $P_\varepsilon^* = \mathrm{Diag}(u) K \mathrm{Diag}(v)$

Remark:

Computing $K^T u$ and Kv requires $\mathcal{O}(n^2)$ algebraic operations

Low-rank Kernel

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

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

⟶ The entries of $\widetilde{K} = AB^T$ are positive and Sinkhorn converges

⚠ The smaller ϵ, the more difficult the approximation of K

Remark:

Sinkhorn outputs $\widetilde{P_\epsilon^*} = \mathrm{Diag}(u)\widetilde{K}\mathrm{Diag}(v)$ with with $\widetilde{K} = AB^T$

$\widetilde{P_\epsilon^*}$ admits a Low-NN rank factorization

Low-Rank Optimal Transport

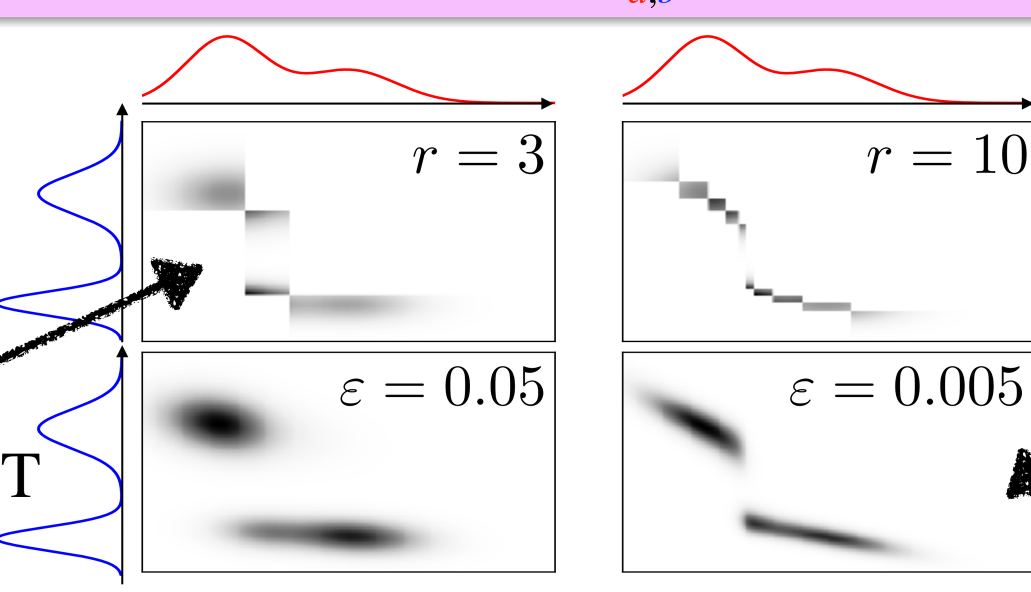
NN-rank: $\mathrm{rk}_+(M) := \min \left\{ q \,|\, M = \sum_{i=1}^{q} R_i, \forall i, \mathrm{rk}(R_i) = 1, R_i \geq 0 \right\}$

Low-NN rank couplings:

$$\Pi_{a,b}(r) := \{P \in \mathbb{R}_+^{n\times m} \text{ s.t. } P\mathbf{1}_m = a,\ P^T\mathbf{1}_n = b \text{ and } \mathrm{rk}_+(P) \leq r\}$$

Definition of Low-rank Optimal Transport

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



Entropic OT

Low-Rank OT

Reparametrization of LOT

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

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

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

Mirror-Descent Scheme

Each update compute:

$$K_k^{(1)} := \exp(-\gamma_k C R_k \mathrm{Diag}(1/g_k) + \log(Q_k))$$

$$K_k^{(2)} := \exp(-\gamma_k C^T Q_k \mathrm{Diag}(1/g_k) + \log(R_k))$$

$$K_k^{(3)} := \exp(\gamma_k \omega_k / g_k^2 + \log(g_k)) \quad \text{where} \quad [\omega_k]_i := [Q_k^T C R_k]_{i,i}$$

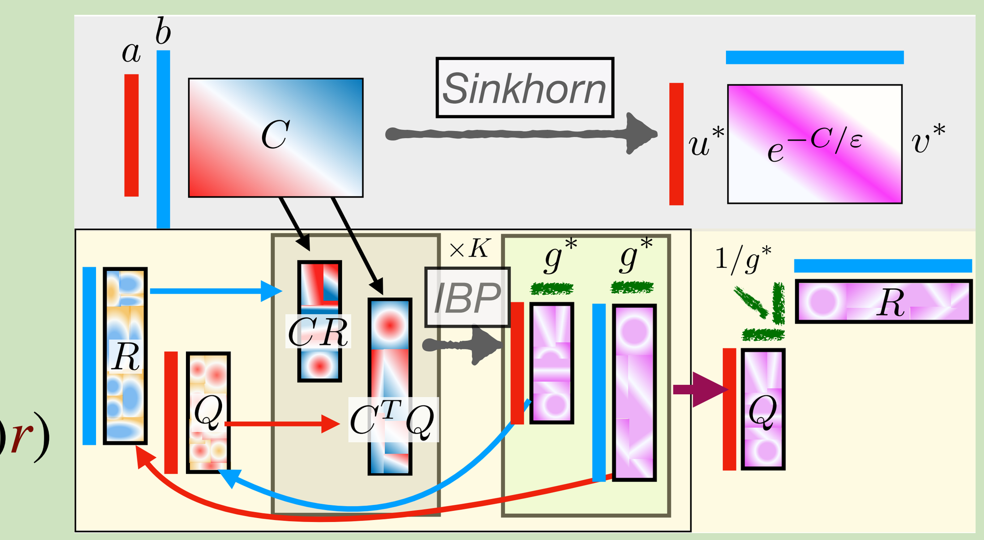
And solve:

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

Can be solved efficiently using the IBP algorithm

Remarks:

- Computing $(K_k^{(i)})_{i=1}^2$ requires $\mathcal{O}(nmr)$ operations
- Solving the barycenter problem requires $\mathcal{O}((n+m)r)$ operations.



Linear Time Approximation

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

Examples: the squared Euclidean distance or any distance matrix

Experiments

ℓ_1 distance between the couplings

Advantage of the method:

- Faster to compute than Sinkhorn with better accuracy
- Its parametrization, which is the rank r, encodes directly a property on the coupling