Screening Sinkhorn Algorithm for Regularized Optimal Transport

Mokhtar Z. Alaya, Maxime Bérar, Gilles Gasso, Alain Rakotomamonjy

LITIS EA4108, University of Rouen Normandy

Objectives

- Introduce a novel strategy for efficiently approximating the optimal transport (OT) distance between two discrete measures.
- Propose the SCREENKHORN algorithm: solve a smaller Sinkhorn problem while ensuring approximation.
- Illustrate the efficiency of SCREENKHORN on complex tasks such as domain adaptation with regularized OT.

Introduction

- OT is a method for comparing probability distributions with the ability to incorporate spatial information.

(courtesy of M. Cuturi)

- Given $\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{\mu}_i \delta_{\boldsymbol{x}_i}$ and $\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{\nu}_j \delta_{\boldsymbol{x}_j}$ two discrete probability distributions and a nonnegative cost matrix $\boldsymbol{C} = (\boldsymbol{C}_{ij}) \in \mathbb{R}_+^{n \times m}$, the OT (Wasserstein) distance writes as

$$\mathcal{S}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P} \in \boldsymbol{\Pi}(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle,$$

where

$$\boldsymbol{\Pi}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m}, \boldsymbol{P}\mathbf{1}_m = \boldsymbol{\mu}, \boldsymbol{P}^\top \mathbf{1}_n = \boldsymbol{\nu}\}.$$

.

- **Sinkhorn Divergence:** The entropic regularization of OT distances relies on the addition of a penalty term:

$$\mathcal{S}_\eta(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P} \in \boldsymbol{\Pi}(\boldsymbol{\mu}, \boldsymbol{\nu})} \{\langle \boldsymbol{C}, \boldsymbol{P} \rangle - \eta H(\boldsymbol{P})\},$$

where $H(\boldsymbol{P}) = -\sum_{i,j} \boldsymbol{P}_{ij} \log(\boldsymbol{P}_{ij})$ is the negative entropy and $\eta > 0$ is a regularization parameter.

Dual of Sinkhorn Divergence

- The dual of Sinkhorn divergence is given by

$$\mathcal{S}_\eta^{\mathsf{d}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{v} \in \mathbb{R}^m} \{\Psi(\boldsymbol{u}, \boldsymbol{v})\},$$

where $\Psi(\boldsymbol{u}, \boldsymbol{v}) := \mathbf{1}_n^\top B(\boldsymbol{u}, \boldsymbol{v}) \mathbf{1}_m - \langle \boldsymbol{u}, \boldsymbol{\mu} \rangle - \langle \boldsymbol{v}, \boldsymbol{\nu} \rangle$, $B(\boldsymbol{u}, \boldsymbol{v}) := \mathrm{diag}(e^{\boldsymbol{u}})\, \boldsymbol{K}\, \mathrm{diag}(e^{\boldsymbol{v}})$, $\boldsymbol{K} := e^{-\boldsymbol{C}/\eta}$.

- The optimal solution $\boldsymbol{P}^\star$ of the primal problem $\mathcal{S}_\eta^{\mathsf{d}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ takes the form $\boldsymbol{P}^\star = \mathrm{diag}(e^{\boldsymbol{u}^\star})\, \boldsymbol{K}\, \mathrm{diag}(e^{\boldsymbol{v}^\star})$, where $(\boldsymbol{u}^\star, \boldsymbol{v}^\star) = \mathrm{argmin}_{\boldsymbol{u}, \boldsymbol{v}} \{\Psi(\boldsymbol{u}, \boldsymbol{v})\}$.

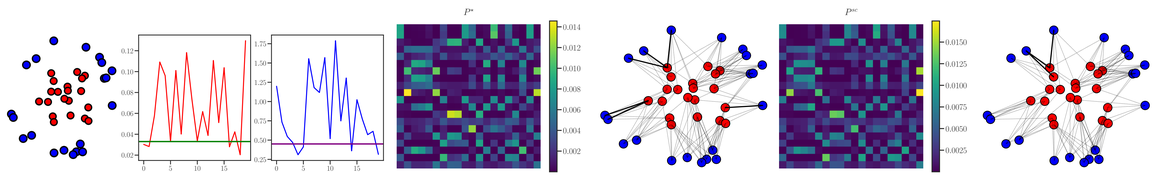
Static Screening Test

- OT plans present a large number of neglectable values. This favorites the using of *static screening test* like in supervised learning (Lasso).
- Based on this idea, we define a so-called **approximate dual of Sinkhorn divergence**

$$\mathcal{S}_\eta^{\mathsf{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{u} \in \mathcal{C}_{\frac{\varepsilon}{\kappa}}^n, \boldsymbol{v} \in \mathcal{C}_{\varepsilon\kappa}^m} \{\Psi_\kappa(\boldsymbol{u}, \boldsymbol{v})\},$$

$\Psi_\kappa(\boldsymbol{u}, \boldsymbol{v}) := \mathbf{1}_n^\top B(\boldsymbol{u}, \boldsymbol{v}) \mathbf{1}_m - \langle \kappa \boldsymbol{u}, \boldsymbol{\mu} \rangle - \langle \frac{\boldsymbol{v}}{\kappa}, \boldsymbol{\nu} \rangle$, and $\mathcal{C}_\alpha^r = \{w \in \mathbb{R}^r : e^{w_i} \geq \alpha\}$, for $\alpha > 0$.

Toy Example



Screening Preprocessing

Let $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ be an optimal solution of $\mathcal{S}_\eta^{\mathsf{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ then $e^{\boldsymbol{u}_i^*} = \frac{\varepsilon}{\kappa}$ and $e^{\boldsymbol{v}_j^*} = \varepsilon\kappa$ for all $i \in I_{\varepsilon,\kappa}^{\boldsymbol{C}}$ and $j \in J_{\varepsilon,\kappa}^{\boldsymbol{C}}$ where

$$I_{\varepsilon,\kappa} = \Big\{i = 1, \ldots, n : \boldsymbol{\mu}_i \geq \frac{\varepsilon^2}{\kappa} r_i(\boldsymbol{K})\Big\} \text{ and } J_{\varepsilon,\kappa} = \Big\{j = 1, \ldots, m : \boldsymbol{\nu}_j \geq \kappa\varepsilon^2 c_j(\boldsymbol{K})\Big\}.$$

Screening with a Fixed Number Budget of Points

- Let $\boldsymbol{\xi} = \boldsymbol{\mu} \oslash r(\boldsymbol{K}) \in \mathbb{R}^n$ and $\boldsymbol{\zeta} = \boldsymbol{\nu} \oslash c(\boldsymbol{K}) \in \mathbb{R}^m$ sorted in descending order. To keep only n_b-budget and m_b-budget of points, the parameters κ and ε satisfy $\frac{\varepsilon^2}{\kappa} = \boldsymbol{\xi}_{n_b}$ and $\varepsilon^2\kappa = \boldsymbol{\zeta}_{m_b}$.
- We restrict the constraints feasibility to the *screened domain* defined by $\mathcal{U}^{\mathsf{sc}} \cap \mathcal{V}^{\mathsf{sc}}$, where $\mathcal{U}^{\mathsf{sc}} = \{\boldsymbol{u} \in \mathbb{R}^{n_b} : e^{\boldsymbol{u}_i} \geq \frac{\varepsilon}{\kappa}\}$ and $\mathcal{V}^{\mathsf{sc}} = \{\boldsymbol{v} \in \mathbb{R}^{m_b} : e^{\boldsymbol{v}_j} \geq \varepsilon\kappa\}$, then we derive the **screened dual of Sinkhorn divergence problem** as

$$\mathcal{S}_\eta^{\mathsf{scd}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{u} \in \mathcal{U}_{\mathsf{sc}}, \boldsymbol{v} \in \mathcal{V}_{\mathsf{sc}}} \{\Psi_{\varepsilon,\kappa}(\boldsymbol{u}, \boldsymbol{v})\}.$$

Screenkhorn Algorithm

SCREENKHORN($\boldsymbol{C}, \eta, \boldsymbol{\mu}, \boldsymbol{\nu}, n_b, m_b$)

1. $\boldsymbol{K} \leftarrow e^{-\boldsymbol{C}/\eta}$;
2. $\boldsymbol{\xi} \leftarrow \mathrm{sort}(\boldsymbol{\mu} \oslash r(\boldsymbol{K}))$, $\boldsymbol{\zeta} \leftarrow \mathrm{sort}(\boldsymbol{\nu} \oslash c(\boldsymbol{K}))$; //(decreasing order)
3. $\varepsilon \leftarrow (\boldsymbol{\xi}_{n_b}\boldsymbol{\zeta}_{m_b})^{1/4}$, $\kappa \leftarrow \sqrt{\boldsymbol{\zeta}_{m_b}/\boldsymbol{\xi}_{n_b}}$;
4. $I_{\varepsilon,\kappa} \leftarrow \{i = 1, \ldots, n : \boldsymbol{\mu}_i \geq \varepsilon^2\kappa^{-1} r_i(\boldsymbol{K})\}$;
5. $J_{\varepsilon,\kappa} \leftarrow \{j = 1, \ldots, m : \boldsymbol{\nu}_j \geq \varepsilon^2\kappa c_j(\boldsymbol{K})\}$;
6. $\boldsymbol{K}_{\min} = \min_{I_{\varepsilon,\kappa}, J_{\varepsilon,\kappa}} \boldsymbol{K}_{ij}$;
7. $\underline{\boldsymbol{\mu}} \leftarrow \min_{i \in I_{\varepsilon,\kappa}} \boldsymbol{\mu}_i$, $\bar{\boldsymbol{\mu}} \leftarrow \max_{i \in I_{\varepsilon,\kappa}} \boldsymbol{\mu}_i$; $\underline{\boldsymbol{\nu}} \leftarrow \min_{j \in J_{\varepsilon,\kappa}} \boldsymbol{\nu}_i$, $\bar{\boldsymbol{\nu}} \leftarrow \max_{j \in J_{\varepsilon,\kappa}} \boldsymbol{\nu}_i$;
8. $\underline{\boldsymbol{u}} \leftarrow \log\big(\frac{\varepsilon}{\kappa} \vee \frac{\underline{\boldsymbol{\mu}}}{\varepsilon(m-m_b)+\varepsilon\vee\frac{\kappa}{m\varepsilon}\boldsymbol{K}_{\min}^{-}m_b}\big)$, $\bar{\boldsymbol{u}} \leftarrow \log\big(\frac{\varepsilon}{\kappa} \vee \frac{\bar{\boldsymbol{\mu}}}{m\varepsilon\,\boldsymbol{K}_{\min}}\big)$;
9. $\underline{\boldsymbol{v}} \leftarrow \log\big(\varepsilon\kappa \vee \frac{\underline{\boldsymbol{\nu}}}{\varepsilon(n-n_b)+\varepsilon\vee\frac{\kappa\underline{\boldsymbol{\mu}}}{m\varepsilon}\boldsymbol{K}_{\min}^{-}n_b}\big)$, $\bar{\boldsymbol{v}} \leftarrow \log\big(\varepsilon\kappa \vee \frac{\bar{\boldsymbol{\nu}}}{n\varepsilon\,\boldsymbol{K}_{\min}}\big)$;
10. $\widetilde{\boldsymbol{u}} \leftarrow \mathrm{stack}(\bar{\boldsymbol{u}}\mathbf{1}_{n_b}, \bar{\boldsymbol{v}}\mathbf{1}_{m_b})$, $\underline{\widetilde{\boldsymbol{u}}} \leftarrow \mathrm{stack}(\underline{\boldsymbol{u}}\mathbf{1}_{n_b}, \underline{\boldsymbol{v}}\mathbf{1}_{m_b})$;

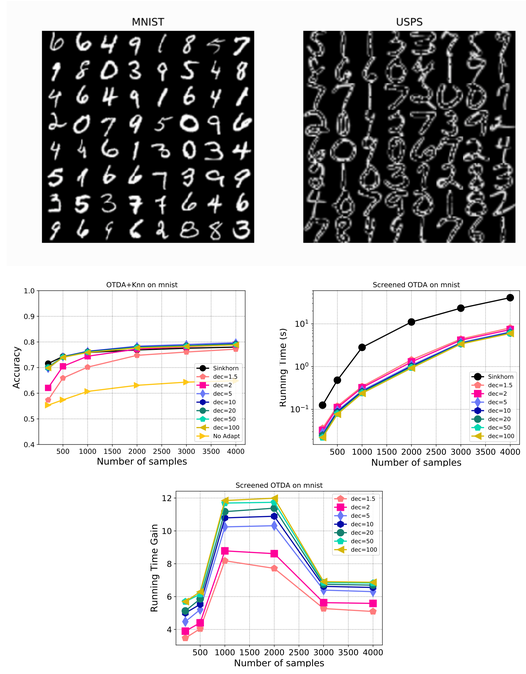
Step 1: Screening

11. $\boldsymbol{u}^{(0)} \leftarrow \log(\varepsilon\kappa^{-1})\mathbf{1}_{n_b}$, $\boldsymbol{v}^{(0)} \leftarrow \log(\varepsilon\kappa)\mathbf{1}_{m_b}$;
12. $\boldsymbol{\theta}^{(0)} \leftarrow \mathrm{stack}(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)})$;
13. $\boldsymbol{\theta} \leftarrow$ L-BFGS-B$(\boldsymbol{\theta}^{(0)}, \underline{\widetilde{\boldsymbol{u}}}, \widetilde{\boldsymbol{u}})$;
14. $\boldsymbol{\theta}_u \leftarrow (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{n_b})^\top$;
15. $\boldsymbol{\theta}_v \leftarrow (\boldsymbol{\theta}_{n_b+1}, \ldots, \boldsymbol{\theta}_{n_b+m_b})^\top$;

Step 2: L-BFGS-B

16. $\boldsymbol{u}_i^{\mathsf{sc}} \leftarrow (\boldsymbol{\theta}_u)_i$ if $i \in I_{\varepsilon,\kappa}$ and $\boldsymbol{u}_i^{\mathsf{sc}} \leftarrow \log(\varepsilon\kappa^{-1})$ if $i \in I_{\varepsilon,\kappa}^{\boldsymbol{C}}$;
17. $\boldsymbol{v}_j^{\mathsf{sc}} \leftarrow (\boldsymbol{\theta}_v)_j$ if $j \in J_{\varepsilon,\kappa}$ and $\boldsymbol{v}_j^{\mathsf{sc}} \leftarrow \log(\varepsilon\kappa)$ if $j \in J_{\varepsilon,\kappa}^{\boldsymbol{C}}$;
18. return $B(\boldsymbol{u}^{\mathsf{sc}}, \boldsymbol{v}^{\mathsf{sc}})$.

OT Domain Adaptation (OTDA): MNIST to USPS



Key References

1. M. Cuturi. Sinkhorn distances: Light speed computation of optimal transport, NIPS 2013.
2. R. Flamary and N. Courty. POT: Python optimal transport library, 2017.

Acknowledgments

This work was supported by grants from the Normandie Projet GRR-DAISI, European funding FEDER DAISI and OATMIL ANR-17-CE23-0012 Project of the French National Research Agency (ANR).