Screening Sinkhorn Algorithm for Regularized Optimal Transport

Mokhtar Z. Alaya

UNION EUROPÉENNE
Fonds européen de développement régional

litis

RÉGION NORMANDIE

UNIVERSITÉ DE ROUEN NORMANDIE

Réunion GdR ISIS/MIA, 9 Juillet 2019

Joint work with ...

Maxime Bérar

LITIS, Université de Rouen Normandie

Gilles Gasso

LITIS, INSA Rouen

Alain Rakotomamonjy

LITIS, Université de Rouen Normandie

& Criteo AI Lab, Criteo Paris

Outline

UNION EUROPÉENNE
Fonds européen de développement régional

RÉGION NORMANDIE

Regularized discrete OT framework: Kantorovich's formula

- We consider two discrete probability measures:
  $\boldsymbol{\mu} = \sum_{i=1}^{n} \mu_i \delta_{\boldsymbol{x}_i} \in \boldsymbol{\Sigma}_n$ and $\boldsymbol{\nu} = \sum_{j=1}^{m} \nu_j \delta_{\boldsymbol{x}_j} \in \boldsymbol{\Sigma}_m$.
- We denote their probabilistic couplings set as
  $\boldsymbol{\Pi}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m}, \boldsymbol{P}\mathbf{1}_m = \boldsymbol{\mu}, \boldsymbol{P}^\top \mathbf{1}_n = \boldsymbol{\nu}\}$.
- Cost matrix: $\boldsymbol{C} = (\boldsymbol{C}_{ij}) \in \mathbb{R}_+^{n\times m}$, (e.g., $\boldsymbol{C}_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$).
- Computing OT between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ amounts to solving a linear problem

Kantorovich [1942]

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

Regularized discrete OT framework: Sinkhorn divergence

- Linear programming problem that requires generally super $\mathcal{O}(n^3)$ arithmetic operations [Pele and Werman, 2009].
- Entropic regularization of OT distances relies on the addition of a penalty term as follows

Sinkhorn divergence, Cuturi [2013]

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

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

Regularized discrete OT framework: Dual of $\mathcal{S}_\eta(\mu, \nu)$

- **Dual of Sinkhorn divergence** is given by

Dual of Sinkhorn divergence

$$\mathcal{S}_\eta^{\mathrm{d}}(\mu, \nu) = \min_{u \in \mathbb{R}^n, v \in \mathbb{R}^m} \left\{ \Psi(u, v) := \mathbf{1}_n^\top B(u, v) \mathbf{1}_m - \langle u, \mu \rangle - \langle v, \nu \rangle \right\}.$$

- $B(u, v) := \mathrm{diag}(e^{u})\, K \mathrm{diag}(e^{v})$ and $K := e^{-C/\eta}$ (Gibbs kernel).
- The primal optimal solution $P^\star$ takes the form

Optimal transportation plan

$$P^\star = \mathrm{diag}(e^{u^\star})\, K \mathrm{diag}(e^{v^\star}), \text{ where } (u^\star, v^\star) = \operatorname*{argmin}_{u,v} \{\Psi(u, v)\}.$$

UNION EUROPÉENNE
Fonds européen de développement régional

RÉGION NORMANDIE

- $\boldsymbol{P}^\star$ can be solved efficiently by Sinkhorn iterations (near-$\mathcal{O}(n^2)$) complexity [Altschuler et al., 2017])

Algorithm 1: SINKHORN($\boldsymbol{C}, \mu, \nu$)

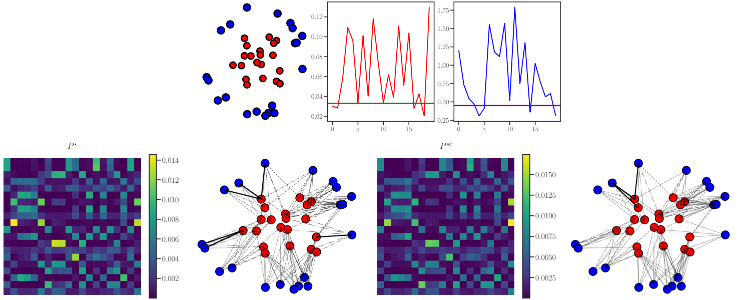
1. $\boldsymbol{u}^{(0)} \leftarrow \mathbf{1}_n/n$, $\boldsymbol{v}^{(0)} \leftarrow \mathbf{1}/m$;
2. $\boldsymbol{K} \leftarrow e^{-\boldsymbol{C}/\eta}$;
3. **for** $k = 1, 2, 3, \ldots$ **do**
   $\boldsymbol{u}^{(k)} \leftarrow \mu \oslash \boldsymbol{K}\boldsymbol{v}^{(k-1)}$; $\boldsymbol{v}^{(k)} \leftarrow \nu \oslash \boldsymbol{K}^\top \boldsymbol{u}^{(k-1)}$;
4. **return** $diag(\boldsymbol{u}^{(k)})\, \boldsymbol{K} diag(\boldsymbol{v}^{(k)})$.

- POT [Flamary and Courty, 2017]

```
from ot import sinkhorn
P_star = sinkhorn(mu, nu, C, eta)
```

Screened dual of Sinkhorn divergence: Motivation

- OT plan presents a large number of neglectable values (sparse) [Blondel et al., 2018].
- **Static screening test** in Lasso [Ghaoui et al., 2010].
- We define the convex set $\mathcal{C}_\alpha^r = \{w \in \mathbb{R}^r : e^{w_i} \geq \alpha\}$, for $\alpha > 0$.



- Identify these indices and fixed at the thresholds before solving the problem. $\rightarrow$ Reduce the scale of the optim. procedure.

Static screening test: Approximate dual of $\mathcal{S}_\eta(\mu, \nu)$

- Based on this idea, we define a so-called **approximate dual of Sinkhorn divergence**

Approximate dual of Sinkhorn divergence

$$\mathcal{S}_\eta^{\mathrm{ad}}(\mu, \nu) = \min_{u \in \mathcal{C}^n_{\frac{\varepsilon}{\kappa}}, v \in \mathcal{C}^m_{\varepsilon\kappa}} \Big\{ \Psi_\kappa(u, v) := \mathbf{1}_n^\top B(u, v) \mathbf{1}_m - \langle \kappa u, \mu \rangle - \langle \frac{v}{\kappa}, \nu \rangle \Big\}.$$

- This is a simply dual Sinkhorn with lower-bounded variables, where the bounds are $\alpha_u = \frac{\varepsilon}{\kappa}$ and $\alpha_v = \varepsilon\kappa$ with $\varepsilon > 0$ and $\kappa > 0$ being fixed numeric constants.

- The κ-parameter plays a role of **scaling factor**

  $\rightarrow$ closed order of the potential components e^{u} and e^{v}.
- The ε-parameter acts like a **threshold** for these components.
- The static screening test aims at locating two subsets of indices (I, J) in $\{1, \ldots, n\} \times \{1, \ldots, m\}$ satisfying:

Static screening test $\mathcal{T}(I, J)$

$$(u, v) \in \mathcal{C}^n_{\alpha_u} \times \mathcal{C}^m_{\alpha_v} \equiv \begin{cases} e^{u_i} > \alpha_u \text{ and } e^{v_j} > \alpha_v, \forall (i,j) \in I \times J \\ e^{u_{i'}} = \alpha_u \text{ and } e^{v_{j'}} = \alpha_v, \forall (i',j') \in I^{\complement} \times J^{\complement} \end{cases}$$

UNION EUROPÉENNE
Fonds européen de développement régional

RÉGION NORMANDIE

Proposition [A., Bérar, Gasso, Rakotomamonjy (2019)]

Let $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ be an optimal solution of $\mathcal{S}_\eta^{\text{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Define $I_{\varepsilon,\kappa} = \{i = 1, \ldots, n : \boldsymbol{\mu}_i \geq \frac{\varepsilon^2}{\kappa} r_i(\,\boldsymbol{K})\}$ and $J_{\varepsilon,\kappa} = \{j = 1, \ldots, m : \boldsymbol{\nu}_j \geq \kappa\varepsilon^2 c_j(\,\boldsymbol{K})\}$. Then one has $e^{\boldsymbol{u}_i^*} = \frac{\varepsilon}{\kappa}$ and $e^{\boldsymbol{v}_j^*} = \varepsilon\kappa$ for all $i \in I^\complement_{\varepsilon,\kappa}$ and $j \in J^\complement_{\varepsilon,\kappa}$.

- The parameters ε and κ are difficult to interpret, we exhibit their relations with a **fixed number budget of points** from the supports of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Screening with a fixed number budget of points

- We denote by $n_b \in \{1, \dots, n\}$ and $m_b \in \{1, \dots, m\}$ the number of points that are going to be optimized in $\mathcal{S}_\eta^{\mathrm{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$.
- Let $\boldsymbol{\xi} \in \mathbb{R}^n$ and $\boldsymbol{\zeta} \in \mathbb{R}^m$ to be the ordered decreasing vectors of $\boldsymbol{\mu} \oslash r(\boldsymbol{K})$ and $\boldsymbol{\nu} \oslash c(\boldsymbol{K})$ respectively.
- 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 $\varepsilon^2 \kappa = \boldsymbol{\zeta}_{m_b}$. Then

$$\varepsilon = (\boldsymbol{\xi}_{n_b} \boldsymbol{\zeta}_{m_b})^{1/4} \text{ and } \kappa = \sqrt{\frac{\boldsymbol{\zeta}_{m_b}}{\boldsymbol{\xi}_{n_b}}}.$$

- This guarantees that

$$|I_{\varepsilon,\kappa}| = n_b \text{ and } |J_{\varepsilon,\kappa}| = m_b.$$

Screening with a fixed number budget of points

- Any solution $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ of $\mathcal{S}_\eta^{\text{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ satisfies $\mathcal{T}(I_{\varepsilon,\kappa}, J_{\varepsilon,\kappa})$ with $\alpha_{\boldsymbol{u}^*} = \frac{\varepsilon}{\kappa}$ and $\alpha_{\boldsymbol{v}^*} = \varepsilon\kappa$.
- We can restrict the variables in $\mathcal{S}_\eta^{\text{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ to variables in $I_{\varepsilon,\kappa}$ and $J_{\varepsilon,\kappa}$.
- This boils down to restricting the constraints feasibility $\mathcal{C}^n_{\frac{\varepsilon}{\kappa}} \cap \mathcal{C}^m_{\varepsilon\kappa}$ to the **screened domain** defined by $\mathcal{U}^{\text{sc}} \cap \mathcal{V}^{\text{sc}}$, where

$$\mathcal{U}^{\text{sc}} = \{\boldsymbol{u} \in \mathbb{R}^{n_b} : e^{\boldsymbol{u}_i} \geq \frac{\varepsilon}{\kappa}\} \text{ and } \mathcal{V}^{\text{sc}} = \{\boldsymbol{v} \in \mathbb{R}^{m_b} : e^{\boldsymbol{v}_j} \geq \varepsilon\kappa\}.$$

UNION EUROPÉENNE Fonds européen de développement régional

RÉGION NORMANDIE

Screening with a fixed number budget of points

- By replacing in $\mathcal{S}_\eta^{\text{ad}}(\boldsymbol{\mu}, \boldsymbol{\nu})$, the variables belonging to $(I^{\complement}_{\varepsilon,\kappa} \times J^{\complement}_{\varepsilon,\kappa})$ by $\frac{\varepsilon}{\kappa}$ and $\varepsilon\kappa$, we derive the **screened dual of Sinkhorn divergence problem** as

Screened dual of Sinkhorn divergence

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

where

$$\Psi_{\varepsilon,\kappa}(\boldsymbol{u}, \boldsymbol{v}) := (e^{\boldsymbol{u}_{I_{\varepsilon,\kappa}}})^\top \boldsymbol{K}_{(I_{\varepsilon,\kappa}, J_{\varepsilon,\kappa})} e^{\boldsymbol{v}_{J_{\varepsilon,\kappa}}} + \varepsilon\kappa (e^{\boldsymbol{u}_{I_{\varepsilon,\kappa}}})^\top \boldsymbol{K}_{(I_{\varepsilon,\kappa}, J^{\complement}_{\varepsilon,\kappa})} \mathbf{1}_{m_b}$$
$$+ \frac{\varepsilon}{\kappa} \mathbf{1}_{n_b}^\top \boldsymbol{K}_{(I^{\complement}_{\varepsilon,\kappa}, J_{\varepsilon,\kappa})} e^{\boldsymbol{v}_{J_{\varepsilon,\kappa}}} - \kappa \boldsymbol{\mu}_{I_{\varepsilon,\kappa}}^\top \boldsymbol{u}_{I_{\varepsilon,\kappa}} - \kappa^{-1} \boldsymbol{\nu}_{J_{\varepsilon,\kappa}}^\top \boldsymbol{v}_{J_{\varepsilon,\kappa}} + \Xi$$

with $\Xi = \varepsilon^2 \sum_{i\in I^{\complement}_{\varepsilon,\kappa}, j\in J^{\complement}_{\varepsilon,\kappa}} \boldsymbol{K}_{ij} - \kappa \log(\varepsilon\kappa^{-1}) \sum_{i\in I^{\complement}_{\varepsilon,\kappa}} \boldsymbol{\mu}_i - \kappa^{-1} \log(\varepsilon\kappa) \sum_{j\in J^{\complement}_{\varepsilon,\kappa}} \boldsymbol{\nu}_j$.

L-BFGS-B: Box constraints on $(\boldsymbol{u}^{\mathrm{sc}}, \boldsymbol{v}^{\mathrm{sc}})$

- $\mathcal{S}_\eta^{\mathrm{scd}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ uses only the restricted parts $\boldsymbol{K}_{(I_{\varepsilon,\kappa}, J_{\varepsilon,\kappa})}$, $\boldsymbol{K}_{(I_{\varepsilon,\kappa}, J^{\complement}_{\varepsilon,\kappa})}$, and $\boldsymbol{K}_{(I^{\complement}_{\varepsilon,\kappa}, J_{\varepsilon,\kappa})}$ of the Gibbs kernel $\boldsymbol{K}$ for calculating the objective function $\Psi_{\varepsilon,\kappa}$.

Proposition [A., Bérar, Gasso, Rakotomamonjy (2019)]

Let $(\boldsymbol{u}^{\mathrm{sc}}, \boldsymbol{v}^{\mathrm{sc}})$ be an optimal pair solution of the screened dual $\mathcal{S}_\eta^{\mathrm{scd}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\boldsymbol{K}_{\min} = \min_{i \in I_{\varepsilon,\kappa}, j \in J_{\varepsilon,\kappa}} \boldsymbol{K}_{ij}$. Then, one has

$$\frac{\varepsilon}{\kappa} \vee \frac{\min_{i \in I_{\varepsilon,\kappa}} \boldsymbol{\mu}_i}{\varepsilon(m - m_b) + \varepsilon \vee \frac{\max_{j \in J_{\varepsilon,\kappa}} \boldsymbol{\nu}_j}{n\varepsilon\kappa \boldsymbol{K}_{\min}} m_b} \leq e^{\boldsymbol{u}_i^{\mathrm{sc}}} \leq \frac{\varepsilon}{\kappa} \vee \frac{\max_{i \in I_{\varepsilon,\kappa}} \boldsymbol{\mu}_i}{m\varepsilon \boldsymbol{K}_{\min}},$$

$$\varepsilon\kappa \vee \frac{\min_{j \in J_{\varepsilon,\kappa}} \boldsymbol{\nu}_j}{\varepsilon(n - n_b) + \varepsilon \vee \frac{\kappa \max_{i \in I_{\varepsilon,\kappa}} \boldsymbol{\mu}_i}{m\varepsilon \boldsymbol{K}_{\min}} n_b} \leq e^{\boldsymbol{v}_j^{\mathrm{sc}}} \leq \varepsilon\kappa \vee \frac{\max_{j \in J_{\varepsilon,\kappa}} \boldsymbol{\nu}_j}{n\varepsilon \boldsymbol{K}_{\min}}$$

for all $i \in I_{\varepsilon,\kappa}$ and $j \in J_{\varepsilon,\kappa}$.

Algorithm 2: $\textsc{Screenkhorn}(\boldsymbol{C}, \eta, \boldsymbol{\mu}, \boldsymbol{\nu}, n_b, m_b)$

Step 1: Initialization

$\boldsymbol{K} \leftarrow e^{-\boldsymbol{C}/\eta}$;

$\boldsymbol{\xi} \leftarrow \texttt{sort}(\boldsymbol{\mu} \oslash r(\boldsymbol{K}))$, $\boldsymbol{\zeta} \leftarrow \texttt{sort}(\boldsymbol{\nu} \oslash c(\boldsymbol{K}))$; //(decreasing order)

$\varepsilon \leftarrow (\boldsymbol{\xi}_{n_b}\boldsymbol{\zeta}_{m_b})^{1/4}$, $\kappa \leftarrow \sqrt{\boldsymbol{\zeta}_{m_b}/\boldsymbol{\xi}_{n_b}}$;

$I_{\varepsilon,\kappa} \leftarrow \{i = 1, \dots, n : \boldsymbol{\mu}_i \geq \varepsilon^2\kappa^{-1} r_i(\boldsymbol{K})\}$;
$J_{\varepsilon,\kappa} \leftarrow \{j = 1, \dots, m : \boldsymbol{\nu}_j \geq \varepsilon^2\kappa c_j(\boldsymbol{K})\}$, $\boldsymbol{K}_{\min} = \min_{I_{\varepsilon,\kappa}, J_{\varepsilon,\kappa}} \boldsymbol{K}_{ij}$;

$\underline{\boldsymbol{\mu}} \leftarrow \min_{i \in I_{\varepsilon,\kappa}} \boldsymbol{\mu}_i$, $\bar{\boldsymbol{\mu}} \leftarrow \max_{i \in I_{\varepsilon,\kappa}} \boldsymbol{\mu}_i$; $\underline{\boldsymbol{\nu}} \leftarrow \min_{j \in J_{\varepsilon,\kappa}} \boldsymbol{\mu}_i$, $\bar{\boldsymbol{\nu}} \leftarrow \max_{j \in J_{\varepsilon,\kappa}} \boldsymbol{\mu}_i$;

$\underline{\boldsymbol{u}} \leftarrow \log\left(\frac{\varepsilon}{\kappa} \vee \frac{\underline{\boldsymbol{\mu}}}{\varepsilon(m-m_b)+\varepsilon \vee \frac{\bar{\boldsymbol{\nu}}}{n\varepsilon\kappa \boldsymbol{K}_{\min}} m_b}\right)$, $\bar{\boldsymbol{u}} \leftarrow \log\left(\frac{\varepsilon}{\kappa} \vee \frac{\bar{\boldsymbol{\mu}}}{m\varepsilon \boldsymbol{K}_{\min}}\right)$;

$\underline{\boldsymbol{v}} \leftarrow \log\left(\varepsilon\kappa \vee \frac{\underline{\boldsymbol{\nu}}}{\varepsilon(n-n_b)+\varepsilon \vee \frac{\kappa\bar{\boldsymbol{\mu}}}{m\varepsilon \boldsymbol{K}_{\min}} n_b}\right)$, $\bar{\boldsymbol{v}} \leftarrow \log\left(\varepsilon\kappa \vee \frac{\bar{\boldsymbol{\nu}}}{n\varepsilon \boldsymbol{K}_{\min}}\right)$;

$\bar{\boldsymbol{\theta}} \leftarrow \texttt{stack}(\bar{\boldsymbol{u}}\mathbf{1}_{n_b}, \bar{\boldsymbol{v}}\mathbf{1}_{m_b})$, $\underline{\boldsymbol{\theta}} \leftarrow \texttt{stack}(\underline{\boldsymbol{u}}\mathbf{1}_{n_b}, \underline{\boldsymbol{v}}\mathbf{1}_{m_b})$;

Step 2: L-BFGS-B

$\boldsymbol{u}^{(0)} \leftarrow \log(\varepsilon\kappa^{-1})\mathbf{1}_{n_b}$, $\boldsymbol{v}^{(0)} \leftarrow \log(\varepsilon\kappa)\mathbf{1}_{m_b}$, $\boldsymbol{\theta}^{(0)} \leftarrow \texttt{stack}(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)})$;

$\boldsymbol{\theta} \leftarrow \text{L-BFGS-B}(\boldsymbol{\theta}^{(0)}, \underline{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}})$;

$\boldsymbol{\theta}_u \leftarrow (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n_b})^\top$, $\boldsymbol{\theta}_v \leftarrow (\boldsymbol{\theta}_{n_b+1}, \dots, \boldsymbol{\theta}_{n_b+m_b})^\top$;

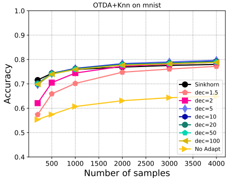
$\boldsymbol{u}^{\mathrm{sc}}_i \leftarrow (\boldsymbol{\theta}_u)_i$ if $i \in I_{\varepsilon,\kappa}$ and $\boldsymbol{u}^{\mathrm{sc}}_i \leftarrow \log(\varepsilon\kappa^{-1})$ if $i \in I^{\complement}_{\varepsilon,\kappa}$;

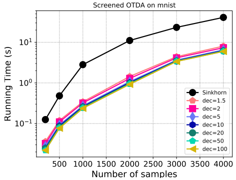
$\boldsymbol{v}^{\mathrm{sc}}_j \leftarrow (\boldsymbol{\theta}_v)_j$ if $j \in J_{\varepsilon,\kappa}$ and $\boldsymbol{v}^{\mathrm{sc}}_j \leftarrow \log(\varepsilon\kappa)$ if $j \in J^{\complement}_{\varepsilon,\kappa}$;

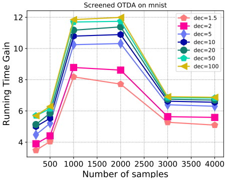
return $B(\boldsymbol{u}^{sc}, \boldsymbol{v}^{sc})$.

Integrating SCREENKHORN into machine learning pipelines

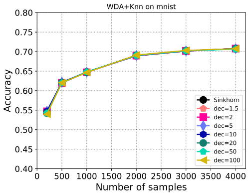
Optimal Transport Domain Adaptation (OTDA) [Courty et al., 2017]: MNIST to USPS data.

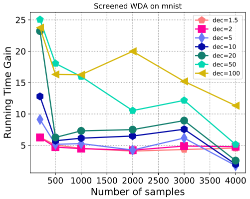






Wasserstein Discriminant Analysis (WDA) [Flamary et al., 2018]: MNIST data.





Take Home Message

- We introudce a novel approach for approximating the Sinkhorn divergence based on a screening strategy with a carefully analyzing its optimality conditions.
- Integrated in some complex machine learning pipelines, our SCREENKHORN algorithm achieves strong gain in efficiency while not compromising on accuracy.

References

Alaya, M. Z., M. Bérar, G. Gasso, and A. Rakotomamonjy (2019). Screening sinkhorn algorithm for regularized optimal transport.

Altschuler, J., J. Weed, and P. Rigollet (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 1964–1974. Curran Associates, Inc.

Blondel, M., V. Seguy, and A. Rolet (2018, 09–11 Apr). Smooth and sparse optimal transport. In A. Storkey and F. Perez-Cruz (Eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Volume 84 of *Proceedings of Machine Learning Research*, Playa Blanca, Lanzarote, Canary Islands, pp. 880–889. PMLR.

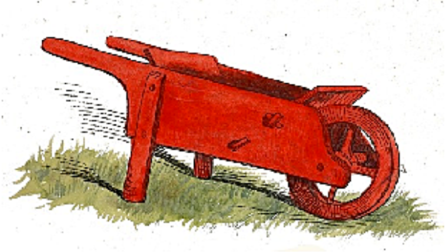
Courty, N., R. Flamary, D. Tuia, and A. Rakotomamonjy (2017). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence 39*(9), 1853–1865.

Flamary, R. and N. Courty (2017). Pot python optimal transport library.

Flamary, R., M. Cuturi, N. Courty, and A. Rakotomamonjy (2018). Wasserstein discriminant analysis. *Machine Learning 107*(12), 1923–1945.

Ghaoui, L. E., V. Viallon, and T. Rabbani (2010). Safe feature elimination in sparse supervised learning. *CoRR abs/1009.4219*.

Pele, O. and M. Werman (2009). Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467.



Thank You!

[Credit image: P. Lemberger]