Linear Time Sinkhorn Divergence using Positive Features

M. Scetbon

M. Cuturi

ENSAE IP PARIS

Google Brain

NEURAL INFORMATION PROCESSING SYSTEMS

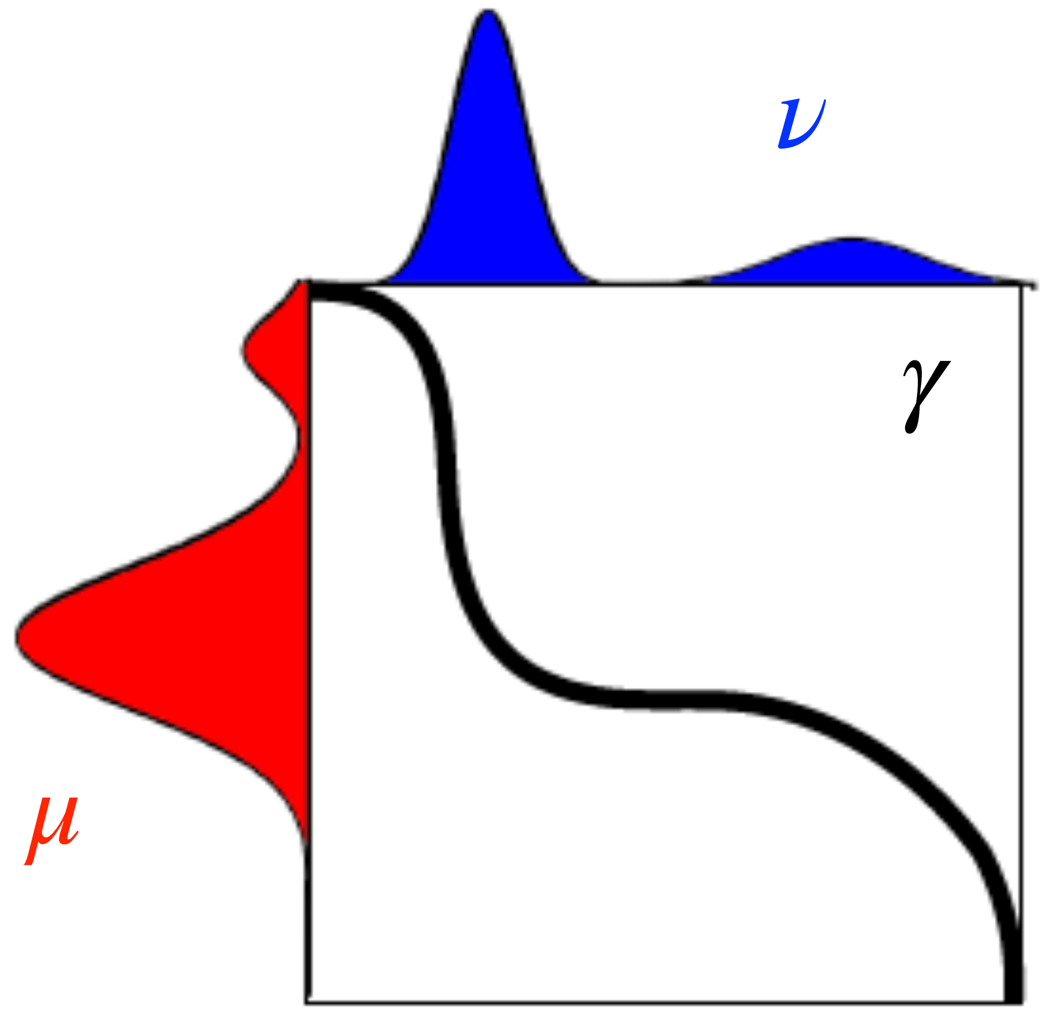
Thirty-fourth Conference on Neural Information Processing Systems

OPTIMAL TRANSPORT

Distributions: $\mu \in \mathcal{M}_1^+(\mathcal{X})$ and $\nu \in \mathcal{M}_1^+(\mathcal{Y})$

Couplings: $\Pi(\mu, \nu) := \left\{ \gamma \text{ s.t. } \Pi_{1\sharp}\gamma = \mu \, , \, \Pi_{2\sharp}\gamma = \nu \right\}$

Cost function: $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$



Optimal coupling[1]

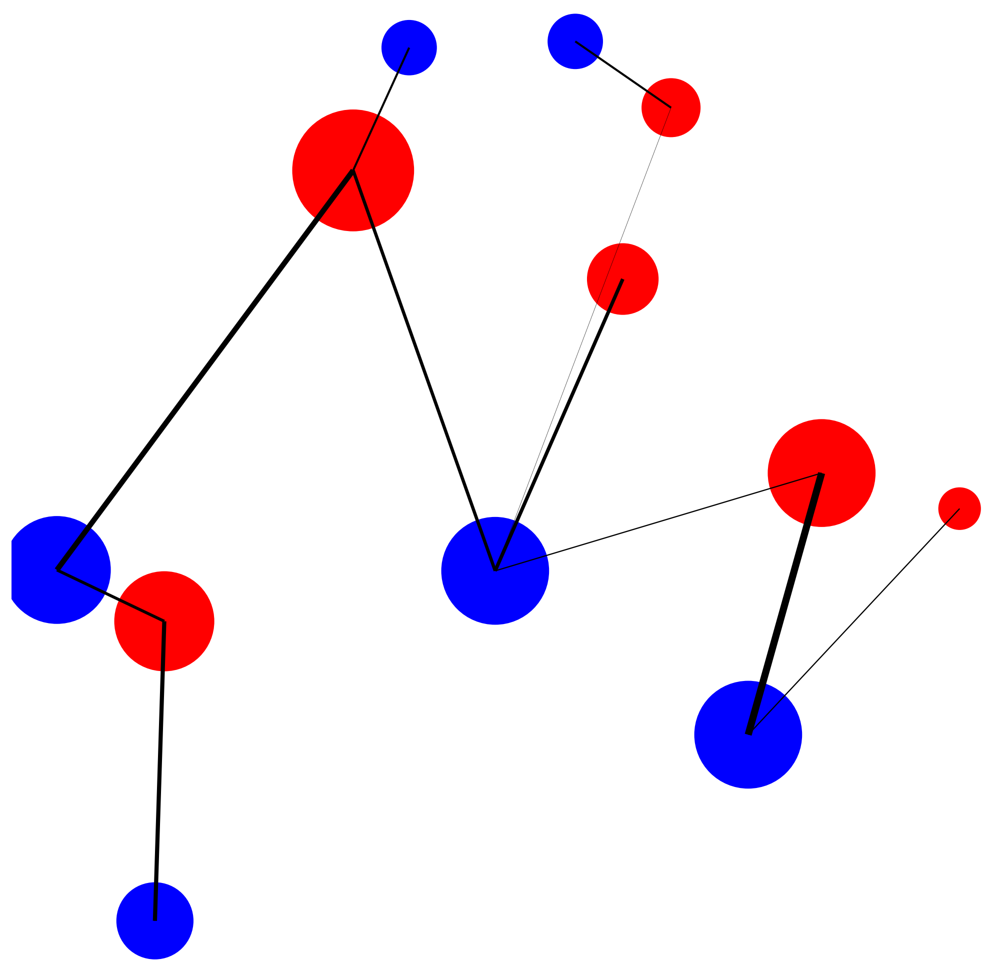
Definition of Optimal Transport

$$W_c(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y)$$

1: Graph was taken from Computational Optimal Transport book, written by Gabriel Peyré and Marco Cuturi

How to compute the OT in practice ?

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



Discrete OT: $W_c(\mu, \nu) = \min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}} \langle P, C \rangle$ where $\forall i, j \; C_{i,j} = c(x_i, y_j)$

Main issues

- Costly to compute ⟶ LP: $\mathcal{O}(n^3 \log(n))$ complexity
- Not differentiable with respect to the measures
- Suffers from the curse of dimensionality

Entropic Regularization

Relative Entropy: $\mathrm{KL}(\gamma || \pi) = \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\gamma}{d\pi}(x,y)\right) d\gamma(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} (d\pi(x,y) - d\gamma(x,y))$

Definition of the Regularized OT

$$W_{c,\varepsilon}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) + \varepsilon \mathrm{KL}(\gamma || \mu\otimes\nu)$$

Approximation of OT $\lim_{\varepsilon\to 0} W_{c,\varepsilon}(\mu,\nu) \to W_c(\mu,\nu)$

Advantages

- It is differentiable with respect to the measures
- It does not suffer from the curse of dimension
- It is faster to compute $\longrightarrow$ Sinkhorn algorithm: $\mathcal{O}(n^2)$ per iteration

Sinkhorn Algorithm

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

Until convergence, at each iteration compute : $v \leftarrow \dfrac{b}{K^T u}$, $u \leftarrow \dfrac{a}{Kv}$

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

⚠ The Sinkhorn algorithm converges iff all the entries of K are positive

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

→ Cannot be applied for large scale problems

Positive Low-rank Factorization of the Kernel

- Random version to approximate the ROT for usual cost functions
- Constructive and differentiable method to learn an adapted kernel

Positive Random Features

Kernel of the form: $k(x,y) = \int_{u \in \mathcal{U}} \varphi(x,u)^T \varphi(y,u) d\rho(u)$ where $\forall\, x, u \in \mathcal{X} \times \mathcal{U},\ \varphi(x,u) \in (\mathbb{R}_*^+)^p$

Positive low-rank Factorization: $k_\theta(x,y) = \langle \varphi_\theta(x), \varphi_\theta(y) \rangle$ where $\begin{cases} \varphi_\theta(x) = \dfrac{1}{\sqrt{r}} \left(\varphi(x,u_1), \ldots, \varphi(x,u_r)\right) \in (\mathbb{R}_*^+)^{p \times r} \\ \theta = (u_1, \ldots, u_r) \in \mathcal{U}^r \text{ and } u_i \sim \rho \text{ i.i.d} \end{cases}$

Example: RBF Kernel $$e^{-\frac{\|x-y\|_2^2}{\varepsilon}} = \left(\frac{4}{\pi\varepsilon}\right)^{d/2} \int_{u \in \mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right) \exp\left(-2\varepsilon^{-1}\|y-u\|_2^2\right) du$$

Positive Low-rank Factorization of the Kernel

Approximation of ROT:

$$W_{c_\theta,\varepsilon}(\mu,\nu) = \min_{\substack{P \in \mathbb{R}_+^{n\times m} \\ P\mathbf{1}_m = a,\, P^T\mathbf{1}_n = b}} \varepsilon \mathrm{KL}(P\,||\,K_\theta)$$

where $K_\theta = \xi^T\zeta$, $\xi = [\varphi_\theta(x_1), \ldots, \varphi_\theta(x_n)] \in (\mathbb{R}_*^+)^{r\times n}$, $\zeta = [\varphi_\theta(y_1), \ldots, \varphi_\theta(y_m)] \in (\mathbb{R}_*^+)^{r\times m}$

Remarks:

- Computing $K_\theta^T u$ and $K_\theta v$ requires $\mathcal{O}(nr)$ algebraic operations
- All the entries of K_θ are positive $\longrightarrow$ the Sinkhorn algorithm converges
- $W_{c_\theta,\varepsilon} \simeq W_{c,\varepsilon}$ where $c(x,y) = -\varepsilon \log(k(x,y))$

Example: $k(x,y) = e^{-\frac{\|x-y\|^2}{\varepsilon}}$ and therefore $c(x,y) = \|x-y\|^2$

Positive Low-rank Factorization of the Kernel

Theorem

Let where $\psi = \sup_{x,y,u} |\varphi(x,u)^T\varphi(y,u)/k(x,y)|$, then with a probability $1-\tau$, the Sinkhorn Algorithm with inputs K_θ, a and b output a δ-approximation of the ROT distance in $\tilde{\mathcal{O}}\left(\frac{n}{\varepsilon\delta^3}\|\mathbf{C}\|_\infty^4\psi^2\log\left(\frac{n}{\tau}\right)\right)$ algebraic operations.

Constructive Positive Features: Differentiability

Proposition

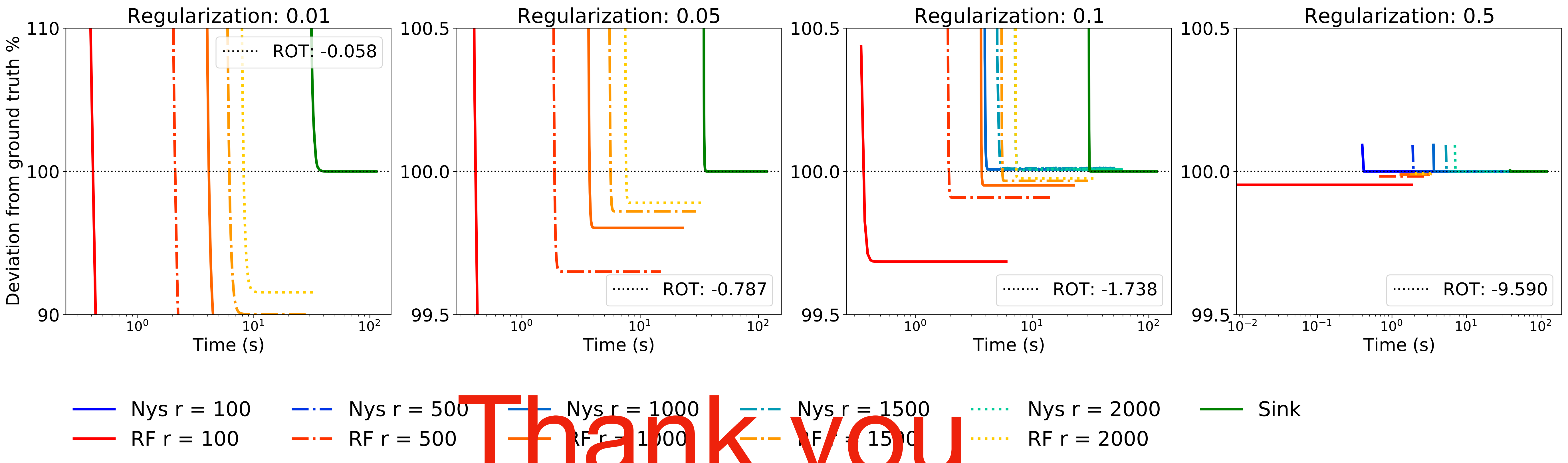
Let $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$, $\mu(\mathbf{X}) = \sum_{i=1}^{n} a_i\delta_{x_i}$, $\nu = \sum_{j=1}^{m} b_j\,\delta_{y_j}$ and $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^r \to \varphi_\theta(x) \in (\mathbb{R}_+^*)^r$ a differentiable map.

Denote $k_\theta(x,y) = \langle\varphi_\theta(x), \varphi_\theta(y)\rangle$. Then $\theta \to W_{c_\theta,\varepsilon}(\mu(X),\nu)$ and $\mathbf{X} \to W_{c_\theta,\varepsilon}(\mu(X),\nu)$ are differentiable.

$\longrightarrow$ Learn an adapted kernel/cost function to compare two distributions via OT

Experiments

- Efficiency vs. Approximation trade-off using positive features



Thank you

- Using positive features to learn adversarial kernels in GANs