Linear Time Sinkhorn Divergence using Positive Features

Meyer SCETBON (ENSAE, IP Paris), Marco CUTURI (ENSAE, IP Paris & Google Brain)

Overview

Problem: How to compute in linear time the Optimal Transport (OT) ?

Contributions:

- We obtain a *random approximation* of the OT in linear time for usual cost functions.
- We propose a *constructive and differentiable* method to learn an adapted cost to compute the OT in linear time.

Discrete Optimal Transport

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

Cost function: $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, Cost matrix: $\forall i,j \; C_{i,j} = c(x_i, y_j)$

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

Main issues:

- *Costly to compute* $\longrightarrow$ LP: $\mathcal{O}(n^3 \log(n))$ *complexity*
- $W_c(\mu,\nu)$ *is not differentiable with respect to the measures*

Entropic Regularization

Discrete Regularized OT:

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

Sinkhorn Algorithm

Until convergence, 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*
- $W_{c,\varepsilon}(\mu,\nu)$ *is differentiable with respect to the measures*

Positive Low Rank Factorization

A first Idea: Approximate $K \simeq \xi^T \zeta$ where $\xi, \zeta \in \mathbb{R}^{r\times n}$

$\longrightarrow$ $\widetilde{K}^T u$ *and* $\widetilde{K} v$ *requires only* $\mathcal{O}(nr)$ *algebraic operations*

▲ *Sinkhorn converges iff all the entries of* K *are positive.*

A low rank approximation of $K = \exp(-C/\varepsilon)$ for a given C does not ensure the convergence of Sinkhorn.

Choose the Kernel instead of the Cost:

$$k(x,y) = \int_{u \in \mathcal{U}} \varphi(x,u)\varphi(y,u)d\rho(u) \quad \text{where} \quad \varphi(x,u) \in \mathbb{R}_*^+$$

Associated cost: $c(x,y) = -\varepsilon \log(k(x,y))$

Example: RBF Kernel

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

$$c(x,y) = \|x-y\|_2^2$$

Random Approximation

$\theta = (u_1, \ldots, u_r) \in \mathcal{U}^r$, $u_i \sim \rho$ i.i.d

$$\varphi_\theta(x) = \frac{1}{\sqrt{r}} \left(\varphi(x,u_1), \ldots, \varphi(x,u_r)\right) \in (\mathbb{R}_*^+)^r$$

$$k_\theta(x,y) = \langle \varphi_\theta(x), \varphi_\theta(y) \rangle$$

Approximation of the Discrete Regularized OT:

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

where $K_\theta = \xi^T \zeta$

$\xi = [\varphi_\theta(x_1), \ldots, \varphi_\theta(x_n)] \in (\mathbb{R}_*^+)^{r\times n}$

$\zeta = [\varphi_\theta(y_1), \ldots, \varphi_\theta(y_m)] \in (\mathbb{R}_*^+)^{r\times m}$

Theorem

With probability $1-\tau$, the Sinkhorn Algorithm with inputs K_θ, a and b output a δ-approximation of $W_{c,\varepsilon}(\mu,\nu)$ in

$$\tilde{\mathcal{O}}\left(\frac{n}{\varepsilon\delta^3}\|\mathbf{C}\|_\infty^4 \log\left(\frac{n}{\tau}\right)\right)$$

Constructive Method: Differentiability

Learn an adapted cost to compute the regularized OT:

Let $\psi : (x,\theta) \in \mathbb{R}^d \times \mathbb{R}^r \to \varphi_\theta(x) \in (\mathbb{R}_+^*)^r$ a differentiable map

Denote $k_\theta(x,y) = \langle \varphi_\theta(x), \varphi_\theta(y) \rangle$, $c_\theta(x,y) = -\varepsilon \log k_\theta(x,y)$

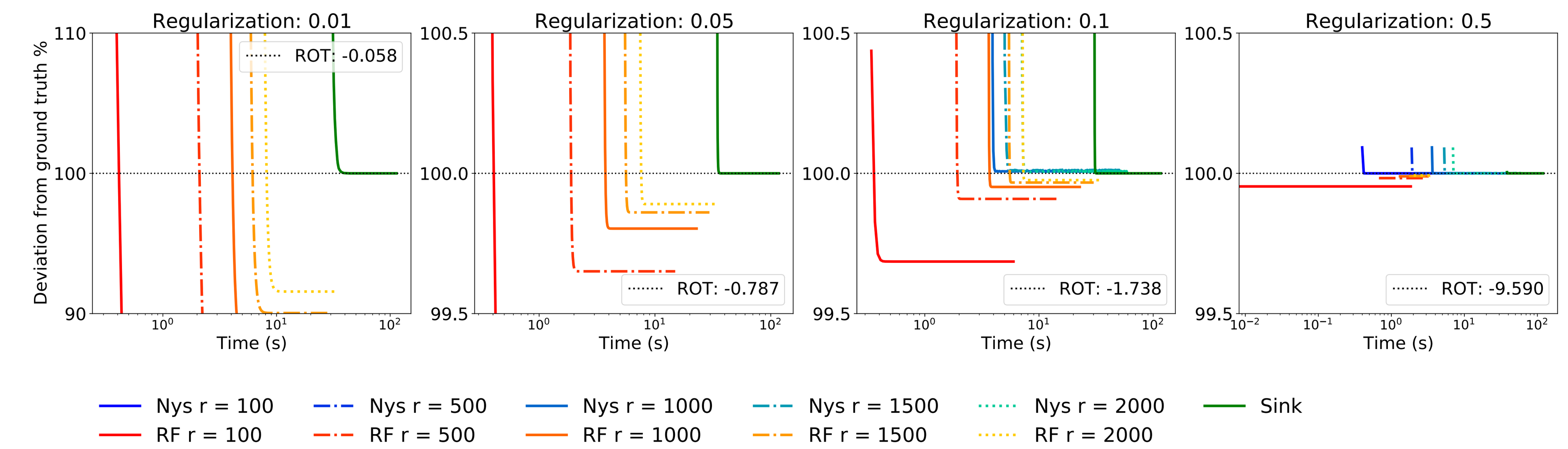
Proposition

Let $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$ and $\mu(\mathbf{X}) = \sum_{i=1}^{n} a_i \, \delta_{x_i}$. Then

$\theta \to W_{c_\theta,\varepsilon}(\mu(X),\nu)$ and $\mathbf{X} \to W_{c_\theta,\varepsilon}(\mu(X),\nu)$ are differentiable.

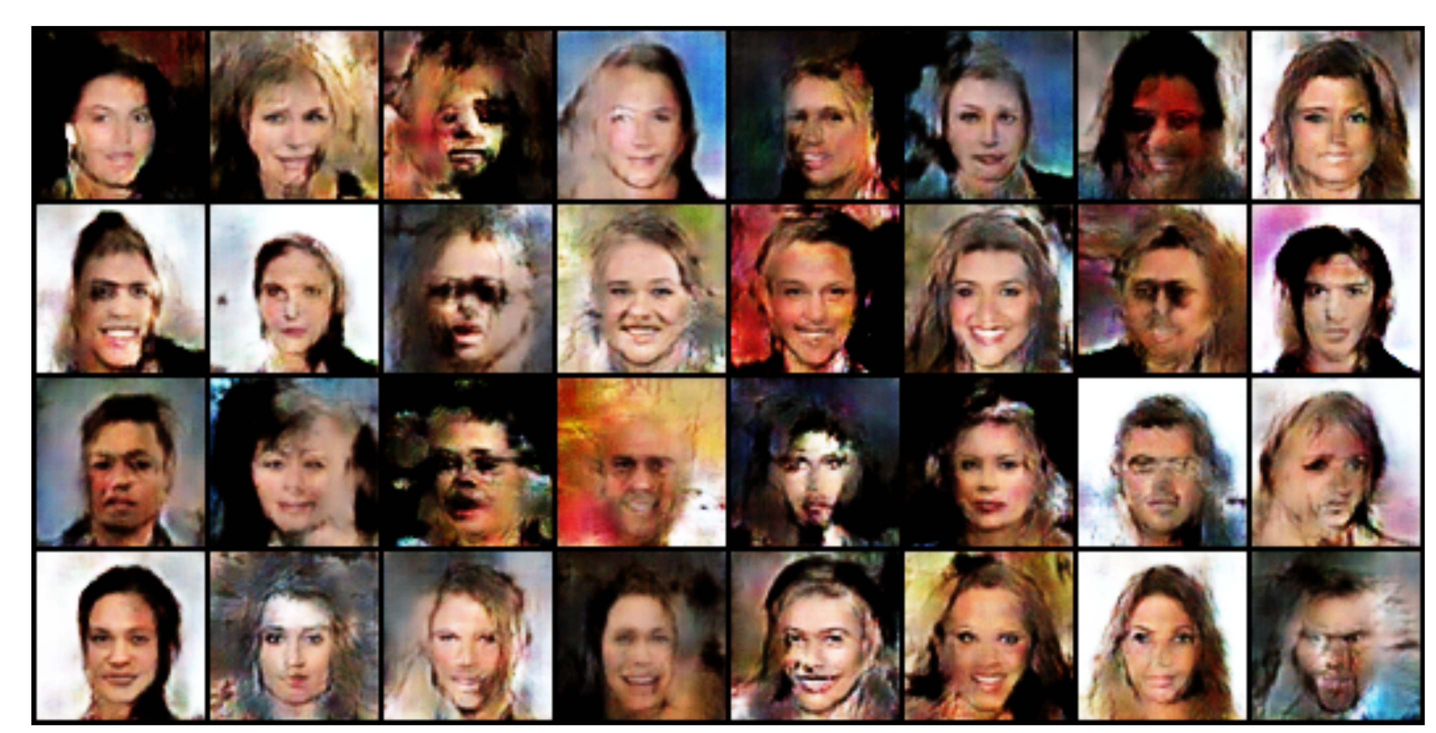
Experiments

- **Efficiency vs. Approximation trade-off using positive features**



Here the two samples are drawn from two gaussians and $n = m = 20000$

- **Using positive features to learn adversarial costs in GANs**



Images generated by a learned generative model trained by optimizing $W_{c_\theta,\varepsilon}$