On the Complexity of Approximating Wasserstein Barycenters

Alexey Kroshnin, Darina Dvinskikh,
Pavel Dvurechensky, Alexander Gasnikov,
Nazarii Tupitsa, César A. Uribe

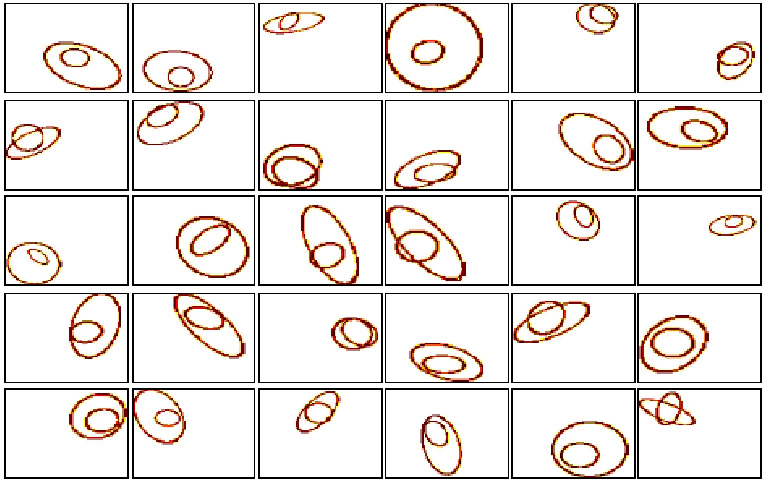
International Conference on Machine Learning 2019

Wasserstein barycenter

$$\hat{\nu} = \arg\min_{\nu \in \mathcal{P}_2(\Omega)} \sum_{i=1}^{m} \mathcal{W}(\mu_i, \nu),$$

where $\mathcal{W}(\mu, \nu)$ is the Wasserstein distance between measures μ and ν on Ω.

WB is efficient in machine learning problems with geometric data, e.g. template image reconstruction from random sample:



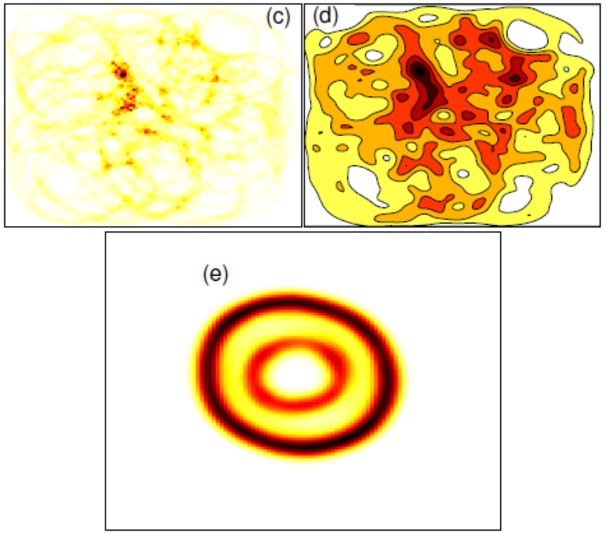


Figure: Images from [Cuturi & Doucet, 2014]

Motivation

We consider a set of discrete measures $p_1, \ldots, p_m \in S_n(1)$.

Main question: How much work is it needed to find their barycenter $\hat{q}$ with accuracy ε?

$$\frac{1}{m}\sum_{l=1}^{m}\mathcal{W}(p_l, \hat{q}) - \min_{q \in S_n(1)} \frac{1}{m}\sum_{l=1}^{m}\mathcal{W}(p_l, q) \leq \varepsilon$$

Beyond that challenges are:

- Fine discrete approximation for continuous ν and $\mu_i \Rightarrow$ large n ,
- Large amount of data $\Rightarrow$ large m ,
- Data produced and stored distributedly (e.g. produced by a network of sensors).

Background

Following [Cuturi & Doucet, 2014], we use entropic regularization.

$$\min_{q\in S_n(1)} \frac{1}{m}\sum_{l=1}^{m} \mathcal{W}_\gamma(p_l, q) = \min_{\substack{q\in S_n(1),\\ \pi_l\in\Pi(p_l,q),\ l=1,\dots,m}} \frac{1}{m}\sum_{l=1}^{m}\left\{\langle \pi_l, C_l\rangle + \gamma H(\pi_l)\right\}, \quad (1)$$

- $H(\pi) = \sum_{i,j=1}^{n} \pi_{ij}\left(\ln \pi_{ij} - 1\right) = \langle \pi, \ln \pi - \mathbf{1}\mathbf{1}^\top\rangle$.
- $\Pi(p,q) = \{\pi \in \mathbb{R}_+^{n\times n} : \pi\mathbf{1} = p, \pi^\top\mathbf{1} = q\}$.
- C_{ij} — transport cost from point z_i to y_j of the supports.

Cost of finding $\mathcal{W}_0(p,q)$

- Sinkhorn's algorithm $O\left(\frac{n^2}{\varepsilon^2}\right)$, [Altschuler, Weed, Rigollet, NeurIPS'17; Dvurechensky, Gasnikov, Kroshnin, ICML'18]
- Accelerated Gradient Descent $O\left(\min\left\{\frac{n^{2.5}}{\varepsilon}, \frac{n^2}{\varepsilon^2}\right\}\right)$, [Dvurechensky, Gasnikov, Kroshnin, ICML'18; Lin, Ho, Jordan, ICML'19]

Background

Algorithms for barycenter

$$\min_{q \in S_n(1)} \frac{1}{m} \sum_{l=1}^{m} \mathcal{W}_\gamma(p_l, q) = \min_{\substack{q \in S_n(1), \\ \pi_l \in \Pi(p_l, q),\ l=1,\dots,m}} \frac{1}{m} \sum_{l=1}^{m} \left\{ \langle \pi_l, C_l \rangle + \gamma H(\pi_l) \right\}.$$

- Sinkhorn + Gradient Descent [Cuturi, Doucet, NeurIPS'13]
- Iterative Bregman Projections [Benamou et al., SIAM J Sci Comp'15]
- (Accelerated) Gradient Descent [Cuturi, Peyre, SIAM J Im Sci'16; Dvurechensky et al, NeurIPS'18; Uribe et al., CDC'18].
- Stochastic Gradient Descent [Staib et al., NeurIPS'17; Claici, Chen, Solomon, ICML'18]

Question of complexity was open.

Contributions

- Prove that to find an ε approximation of the γ-regularized WB
  - Iterative Bregman Projections (IBP) needs $\frac{1}{\gamma\varepsilon}$ iterations;
  - Accelerated Gradient descent (AGD) needs $\sqrt{\frac{n}{\gamma\varepsilon}}$ iterations.
- Setting $\gamma = \Theta\left(\varepsilon/\ln n\right)$ allows to find an ε-approximation for the non-regularized WB with arithmetic operations complexity
  - $\widetilde{O}\left(\frac{mn^2}{\varepsilon^2}\right)$ for IBP ,
  - $\widetilde{O}\left(\frac{mn^{2.5}}{\varepsilon}\right)$ for AGD .
- We propose a proximal-IBP algorithm to solve the issue of instability of IBP and AGD caused by small gamma.
- We discuss scalability of the algorithms via their distributed versions.
  - IBP can be realized distributedly in a centralized architecture (master/slaves),
  - AGD can be realized in a general decentralized architecture.

Iterative Bregman Projections

$$\min_{\substack{\pi_l \mathbf{1} = p_l,\ \pi_l^\top \mathbf{1} = \pi_{l+1}^\top \mathbf{1} \\ \pi_l \in \mathbb{R}_+^{n\times n},\ l=1,\dots,m}} \frac{1}{m} \sum_{l=1}^{m} \left\{ \langle \pi_l, C_l \rangle + \gamma H(\pi_l) \right\}$$

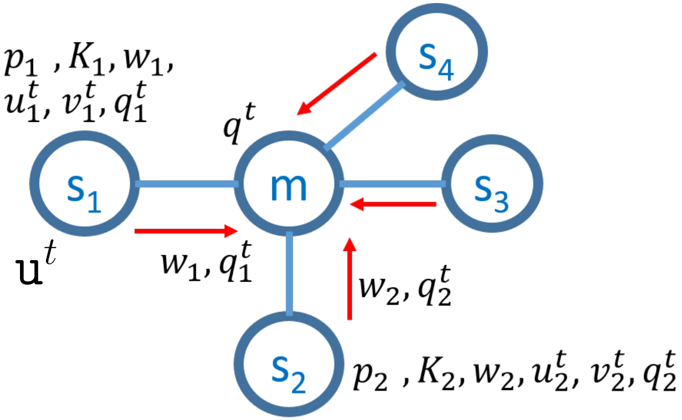
Dual problem:

$$\min_{\substack{\mathbf{u},\mathbf{v} \\ \frac{1}{m}\sum_{l=1}^m v_l = 0}} f(\mathbf{u},\mathbf{v}) := \frac{1}{m} \sum_{l=1}^{m} \left\{ \langle \mathbf{1}, B_l(u_l, v_l)\mathbf{1} \rangle - \langle u_l, p_l \rangle \right\},$$

$\mathbf{u} = [u_1, \dots, u_m]$, $\mathbf{v} = [v_1, \dots, v_m]$, $u_l, v_l \in \mathbb{R}^n$,
$B_l(u_l, v_l) := \operatorname{diag}(e^{u_l}) \exp(-C_l/\gamma) \operatorname{diag}(e^{v_l})$.

IBP is equivalent to alternating minimization for the dual problem.

- $u_l^{t+1} := \ln p_l - \ln K_l e^{v_l^t}$, $\mathbf{v}^{t+1} := \mathbf{v}^t$
- $v_l^{t+1} := \frac{1}{m} \sum_{k=1}^{m} \ln K_k^\top e^{u_k^t} - \ln K_l^\top e^{u_l^t}$, $\mathbf{u}^{t+1} := \mathbf{u}^t$



Accelerated Gradient Descent

Define symmetric p.s.d. matrix $\bar{W}$ s.t. $\mathrm{Ker}(\bar{\mathrm{W}}) = \mathrm{span}(\mathbf{1})$.
For $W := \bar{W} \otimes I_n$ and $\mathsf{q} = (q_1^\top, \ldots, q_m^\top)^\top$ it holds
$q_1 = \cdots = q_m \iff \sqrt{W}\mathsf{q} = 0$

Equivalent form of problem (1)

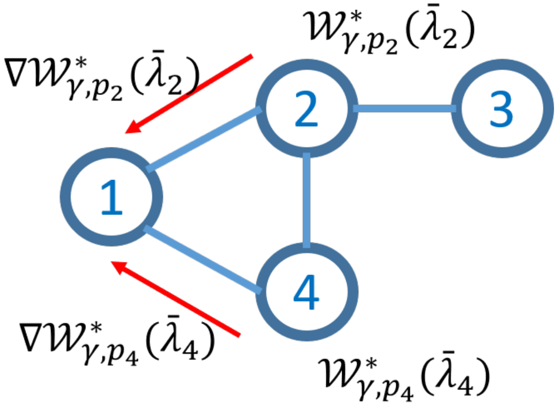
$$\max_{\substack{q_1,\ldots,q_m \in S_1(n) \\ \sqrt{W}\mathsf{q}=0}} -\frac{1}{m}\sum_{l=1}^{m} \mathcal{W}_{\gamma,p_l}(q_l).$$

Dual problem

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{mn}} \mathcal{W}_\gamma^*(\boldsymbol{\lambda}) := \frac{1}{m}\sum_{l=l}^{m} \mathcal{W}^*_{\gamma,p_l}(\overbrace{m[\sqrt{W}\boldsymbol{\lambda}]_l}^{\bar{\lambda}_l}).$$

Run (A)GD for the dual and reconstruct the primal solution

- $\bar{\lambda}_l^{k+1} = \bar{\lambda}_l^k - \frac{\alpha_{k+1}}{m}\sum_{j=1}^m W_{lj}\nabla\mathcal{W}^*_{\gamma,p_j}(\bar{\lambda}_j^{k+1})$
- $q_l^{k+1} = \frac{1}{A_{k+1}}\sum_{i=0}^{k+1}\alpha_i q_i(\bar{\lambda}_l^{k+1})$, where
  $q_l(\cdot) = \nabla\mathcal{W}^*_{\gamma,p_l}(\cdot)$



Thank you!

Welcome to poster #203,
Pacific Ballroom.