Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters

Pavel Dvurechensky, Darina Dvinskikh, Alexander Gasnikov, César A. Uribe, Angelia Nedić

Conference on Neural Information Processing Systems 2018

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:

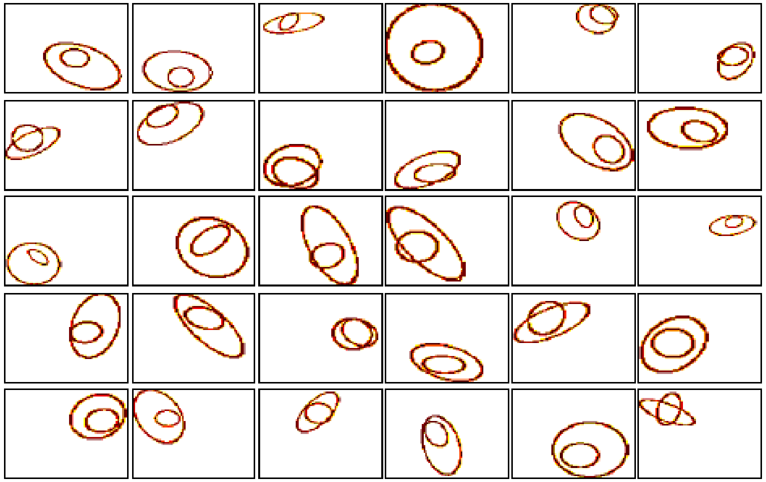
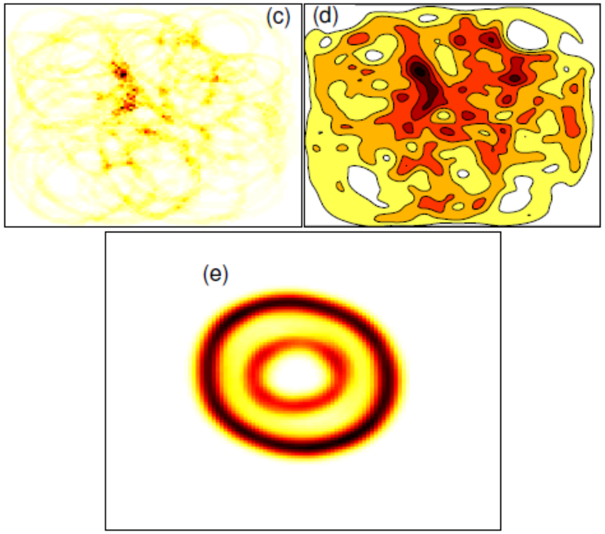


Figure: Images from [Cuturi, 2013]



Motivation

We fix the support z_i, $i = 1, ..., n$ of the barycenter: $\nu = \sum_{i=1}^{n} p_i \delta(z_i)$.

We add Entropic regularization with parameter γ.

$$\hat{p} = \arg \min_{p \in S_1(n)} \sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(p).$$

Challenges:

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

Background and contribution

PAPER	LARGE m, n	DIST. DATA	CONT. μ_i	COMPL-TY
SINKHORN-TYPE [CUTURI&DOUCET'14, BENAMOU ET AL.'15]	$\surd$	$\times$	$\times$	?
DISTRIBUTED AGD [SCAMAN ET AL.'17, URIBE ET AL.'17, LAN ET AL.'17]	$\surd$	$\surd$	$\times$	?
SGD-BASED [STAIB ET.AL.'17, CLAICI ET AL.'18]	$\surd$	$\times$	$\surd$	$1/\varepsilon^2$
THIS PAPER	$\surd$	$\surd$	$\surd$	$1/\varepsilon^2$

Mitglied der Leibniz Gemeinschaft

Contributions

- Novel Accelerated Primal-Dual Stochastic Gradient Method (APDSGD) for general class of stochastic optimization problems with linear constraints

$$(P): \quad \min_{x \in Q \subseteq E} \{f(x) : Ax = b\}, \quad (D): \quad \min_{\lambda} \left\{\langle \lambda, b\rangle + \mathbb{E}_{\xi} F^*(-A^T\lambda, \xi)\right\}.$$

  with complexity

$$O\left(\max\left\{\sqrt{\frac{L_D R_D^2}{\varepsilon}}, \frac{\sigma^2 R_D^2}{\varepsilon^2}\right\}\right)$$

  to obtain

$$f(\mathbb{E}\hat{x}) - f^* \leq \varepsilon \text{ and } \|A\mathbb{E}\hat{x} - b\|_2 \leq \varepsilon.$$

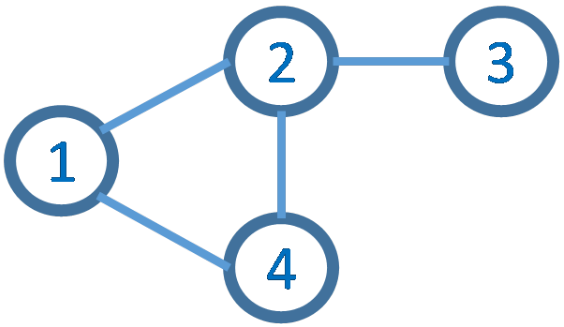
- Decentralized distributed algorithm for γ-regularized Wasserstein barycenter of a set of continuous measures stored over a network with arbitrary topology with complexity

$$O\left(mn \max\left\{\frac{1}{\sqrt{\varepsilon\gamma}}, \frac{m}{\varepsilon^2}\right\}\right) \text{ a.o.}$$

- Experimens on the MNIST digit dataset and the IXI Magnetic Resonance dataset.

Distributed optimization framework[1]

$$\min_{x\in\mathbb{R}} \sum_{i=1}^{m} f_i(x) \iff \min \sum_{i=1}^{m} f_i(x_i) \text{ s.t. } x_1 = \ldots = x_m \in \mathbb{R}.$$



Laplacian matrix

$$W = \begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ -1 & -1 & 0 & 2 \end{pmatrix}$$

$$x_1 = \ldots = x_m \iff \sqrt{W}\boldsymbol{x} = 0 \longrightarrow \max_{\boldsymbol{x}\in\mathbb{R}^m : \sqrt{W}\boldsymbol{x}=0} -\sum_{i=1}^{m} f_i(x_i).$$

Distributed reformulation through dual problem

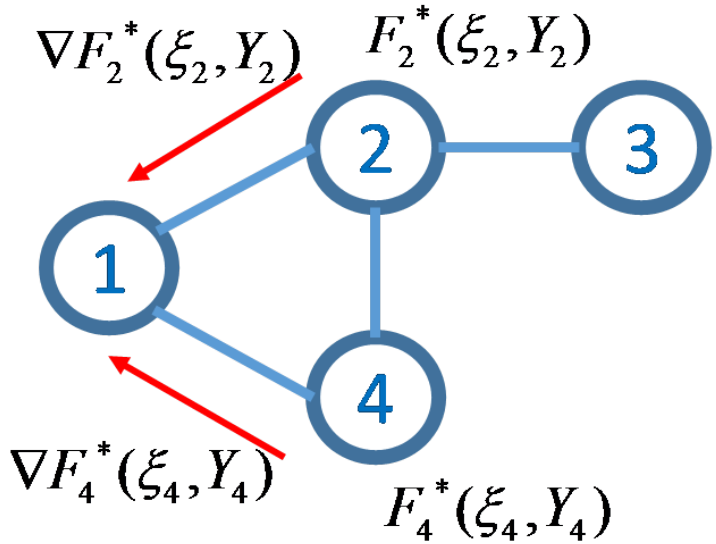
$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^m} \sum_{i=1}^{m} f_i^*\left(\left[\sqrt{W}\boldsymbol{\lambda}\right]_i\right) = \min_{\boldsymbol{\lambda}\in\mathbb{R}^m} \sum_{i=1}^{m} \mathbb{E}_{Y_i\sim\mu_i} F_i^*\left(\left[\sqrt{W}\boldsymbol{\lambda}\right]_i, Y_i\right).$$

[1] [Boyd et al.'11, Jakovetić et al.'15, Scaman et al.'17, Uribe et al.'17, Lan et al.'17]

Distributed stochastic gradient method in the dual

Change the variables $\boldsymbol{\xi} := \sqrt{W}\boldsymbol{\lambda}$.

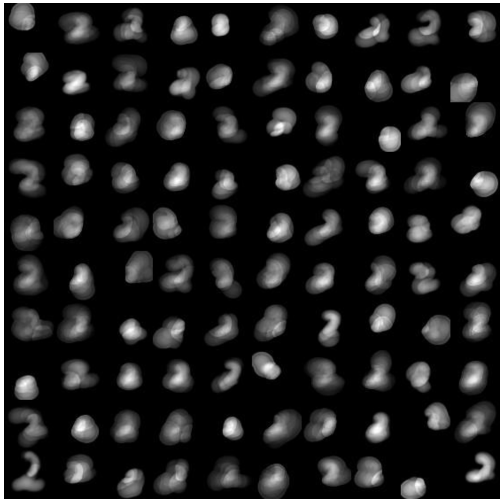
SGD step for each node i: $\xi_i^{(k+1)} = \xi_i^{(k)} - \alpha \sum_{j=1}^{m} [W]_{ij} \nabla F_j^* (\xi_j, Y_j)$.



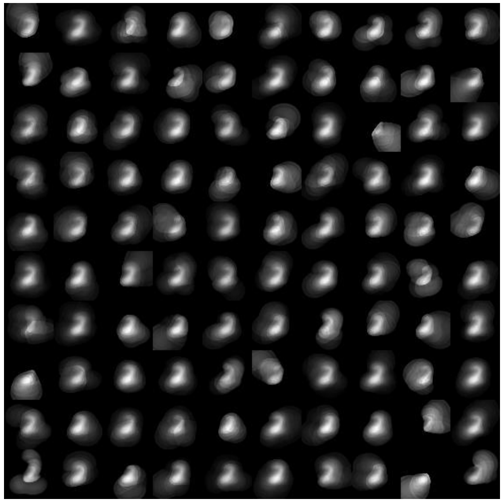
Our contribution: Acceleration and careful Primal-Dual analysis for solving the primal problem.

Experiments on MNIST dataset

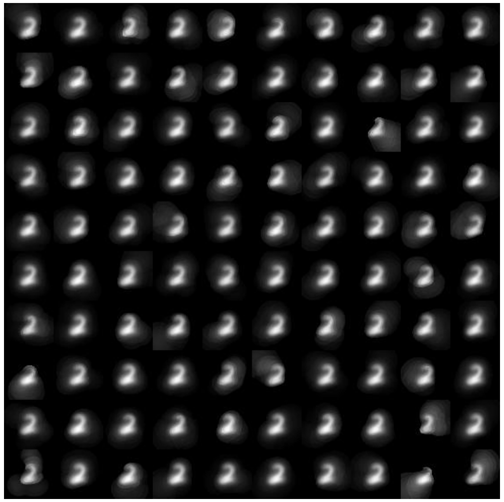
$k = 0$



$k = 10$



$k = 20$



$k = 30$

Thank you!

Welcome to poster #15,
Room 210 & 230 AB.