On the Complexity of Approximating Wasserstein Barycenters

Alexey Kroshnin, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, Nazarii Tupitsa, César A. Uribe
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 \( \mu \) and \( \nu \) on \( \Omega \).

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 $\varepsilon$?

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

Beyond that challenges are:

- Fine discrete approximation for continuous $\nu$ 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.

\[
\min_{q \in S_n(1)} \frac{1}{m} \sum_{l=1}^{m} W_\gamma(p_l, q) = \min_{q \in S_n(1), \pi_l \in \Pi(p_l, q), l=1,\ldots,m} \frac{1}{m} \sum_{l=1}^{m} \left\{ \langle \pi_l, C_l \rangle + \gamma H(\pi_l) \right\}, \tag{1}
\]

- \( H(\pi) = \sum_{i,j=1}^{n} \pi_{ij} (\ln \pi_{ij} - 1) = \langle \pi, \ln \pi - 11^T \rangle \).
- \( \Pi(p, q) = \{ \pi \in \mathbb{R}_+^{n \times n} : \pi 1 = p, \pi^T 1 = q \} \).
- \( C_{ij} \) — transport cost from point \( z_i \) to \( y_j \) of the supports.

Cost of finding \( W_0(p, q) \)

- **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

\[ \min_{q \in S_n(1)} \frac{1}{m} \sum_{l=1}^{m} W_\gamma(p_l, q) = \min_{q \in S_n(1), \pi_l \in \Pi(p_l,q), l=1,...,m} \frac{1}{m} \sum_{l=1}^{m} \left\{ \langle \pi_l, C_l \rangle + \gamma H(\pi_l) \right\}. \]

- 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 $\varepsilon$ approximation of the $\gamma$-regularized WB
  - Iterative Bregman Projections (IBP) needs $\frac{1}{\gamma \varepsilon}$ iterations;
  - Accelerated Gradient descent (AGD) needs $\sqrt{\frac{n}{\gamma \varepsilon}}$ iterations.

- Setting $\gamma = \Theta(\varepsilon / \ln n)$ allows to find an $\varepsilon$-approximation for the non-regularized WB with arithmetic operations complexity
  - $\tilde{O}\left(\frac{mn^2}{\varepsilon^2}\right)$ for IBP,
  - $\tilde{O}\left(\frac{mn^2.5}{\varepsilon}\right)$ 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

\[
\min_{\pi_l 1 = p_l, \pi_l^T 1 = \pi_{l+1}^T 1} \frac{1}{m} \sum_{l=1}^{m} \left\{ \langle \pi_l, C_l \rangle + \gamma H(\pi_l) \right\}
\]

Dual problem:

\[
\min_{u, v} \quad f(u, v) := \frac{1}{m} \sum_{l=1}^{m} \left\{ \langle 1, B_l(u_l, v_l)1 \rangle - \langle u_l, p_l \rangle \right\},
\]

\[
\frac{1}{m} \sum_{l=1}^{m} v_l = 0
\]

\[
u = [u_1, \ldots, u_m], \quad v = [v_1, \ldots, v_m], \quad u_l, v_l \in \mathbb{R}^n,
\]

\[
B_l(u_l, v_l) := \text{diag}(e^{u_l}) \exp(-C_l/\gamma) \text{diag}(e^{v_l}).
\]

IBP is equivalent to alternating minimization for the dual problem.

\[
\begin{align*}
&u_{l+1}^{t+1} := \ln p_l - \ln K_l e^{v_t^l}, \quad v^{t+1} := v^t \\
&v_{l}^{t+1} := \frac{1}{m} \sum_{k=1}^{m} \ln K_k^T e^{u_k^t} - \ln K_l^T e^{u_t^l}, \quad u^{t+1} := u^t
\end{align*}
\]
Accelerated Gradient Descent

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

Equivalent form of problem (1)

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

Dual problem

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

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

$$\bar{\lambda}_{l+1}^{k+1} = \bar{\lambda}_l^k - \frac{\alpha_{k+1}}{m} \sum_{j=1}^{m} W_{lj} \nabla_{p_j} \mathcal{W}_{\gamma,p_j}^* (\bar{\lambda}_{j}^{k+1})$$

$$q_{l}^{k+1} = \frac{1}{A_{k+1}} \sum_{i=0}^{k+1} \alpha_i q_i (\bar{\lambda}_l^{k+1}), \text{ where }$$

$$\nabla \mathcal{W}_{\gamma,p_l}^*(\cdot) = \nabla_{p_l} \mathcal{W}_{\gamma,p_l}(\cdot)$$
Thank you!

Welcome to poster #203, Pacific Ballroom.