Randomized distributed computation of Wasserstein barycenter with mini-batch

Darina Dvinskikh

MIPT
Skoltech

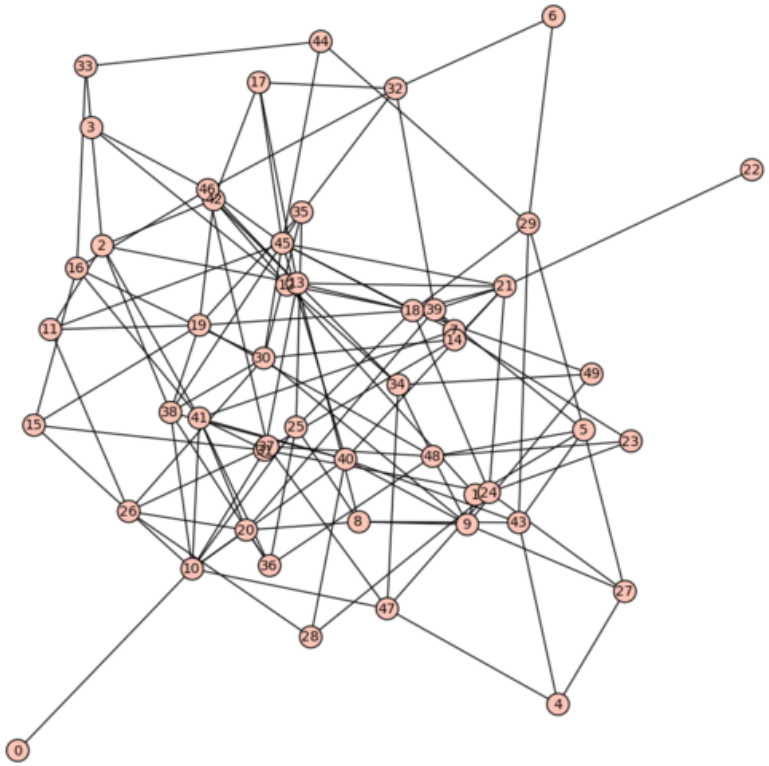
Joint work with
Gasnikov A., Dvurechensky P., Uribe C.A. and Nedić A.

April 14, 2018

Outline

Network

Connected undirected graph $\mathcal{G} = (V, E)$,
where V is a set of m nodes and E is a set of edges



$\forall\ i = 1 \dots m \quad p_i \in V_i,\ p_i \in \mathbb{R}^n$ stacked column vector $p = [p_1^T, p_2^T, \dots, p_m^T]^T$,
d_{max} and d_{min} are the maximum and the minimum degree of nodes in the network.

Network

Laplacian matrix

$$\bar{W} = D - A,$$

where D is a degree matrix, A is an adjacency matrix.

$$\bar{W}_{ij} = \begin{cases} -1, & \text{if } (i,j) \in E \\ \deg(\text{i}), & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

$p \in \mathbb{R}^n \longrightarrow$ define $W = \bar{W} \otimes I_n$, $W_{ij} = \bar{W}_{ij} \otimes I_n$,
where $\otimes$ is the Kronecker product.

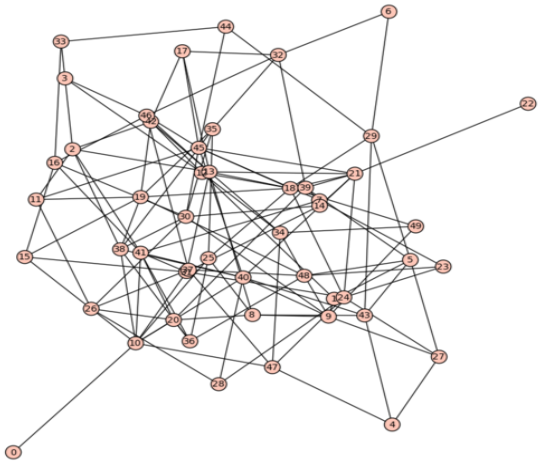
- $\bar{W}$ is singular $\Longrightarrow \lambda = 0$ is an eigenvalue and $p = [1, 1, \ldots, 1]$ is a solution $\bar{W}p = 0$.
- $dim\left(Ker(\bar{W})\right)$ is equal to the number of connected components in the graph $\mathcal{G}$

Graph is connected then

$$Wp = 0 \iff p_1 = \cdots = p_m \quad \text{and} \quad \sqrt{W}p = 0 \iff p_1 = \cdots = p_m$$

Multi-agent network system

- task is separated between agents
- agents exchange the informations with their neighbors



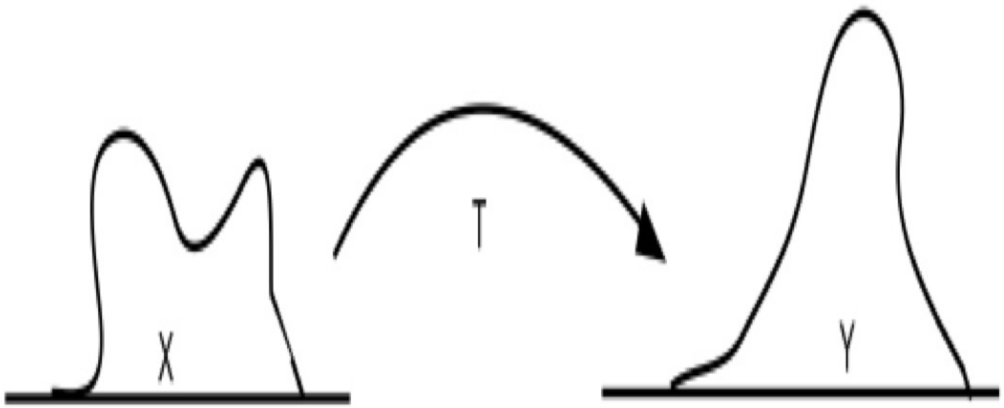
For problems of the form: $\min_x f(x) := \sum_{i=1}^{m} f_i(x)$

$$\min_{x_1=\cdots=x_m} f(x) := \sum_{i=1}^{m} f_i(x_i), \quad x = [x_1^T, x_2^T \ldots, x_m^T]$$

$$\min_{\sqrt{W}x=0} f(x) := \sum_{i=1}^{m} f_i(x_i)$$

What is optimal transport?

A geometric toolbox to compare probability measures supported on a metric space



Wasserstein metric for discrete probability measures

Wasserstein distance

$$\mathcal{W}(p,q) = \min_{X \in U(p,q)} \langle C, X \rangle,$$

where p and q are discrete probability measures,
X is a transport plan,
C is a cost matrix,
$U(p,q) = \{X \in \mathbb{R}_+^{n \times n}, \quad X\mathbf{1} = p, \; X^T\mathbf{1} = q\}$.

Regularized Wasserstein distance, $\gamma \geq 0$

$$\mathcal{W}_\gamma(p,q) = \min_{X \in U(p,q)} \langle C, X \rangle - \gamma E(X),$$

$E = -\sum_{i,j=1}^n X_{ij} \log X_{ij}$ — entropy.

Wasserstein barycenter

$$\min_{p \in S_1(n)} \sum_{k=1}^m \lambda_k \mathcal{W}_{\gamma, q_k}(p),$$

where $S_1(n)$ is a probability symplex.

Wasserstein barycenter on the network

$$\min_{p \in S_{\mathbf{1}}(n)} \sum_{i=1}^{m} \mathcal{W}_{\gamma, q_i}(p).$$

Stacked column vector $p = [p_1^T, p_2^T, \ldots, p_m^T]^T \in \mathbb{R}^{nm}$.

$$\min_{\substack{p_{\mathbf{1}}, \ldots, p_m \in S_{\mathbf{1}}(n), \\ p_{\mathbf{1}} = \cdots = p_m}} \sum_{i=1}^{m} \mathcal{W}_{\gamma, q_k}(p_i).$$

$$\min_{\substack{p_{\mathbf{1}}, \ldots, p_m \in S_{\mathbf{1}}(n), \\ \sqrt{W}p = 0}} \sum_{i=1}^{m} \mathcal{W}_{\gamma, q_i}(p_i) = - \max_{\substack{p_{\mathbf{1}}, \ldots, p_m \in S_{\mathbf{1}}(n), \\ \sqrt{W}p = 0}} - \sum_{k=1}^{m} \mathcal{W}_{\gamma, q_i}(p_i) =$$

$$= \min_{y} \max_{p_{\mathbf{1}}, \ldots, p_m \in S_{\mathbf{1}}(n)} \left\{ \langle y, \sqrt{W}p \rangle - \sum_{i=1}^{m} \mathcal{W}_{\gamma, q_i}(p_i) \right\} = \min_{y} \sum_{i=1}^{m} \mathcal{W}^*_{\gamma, q_i}([\sqrt{W}y]_i),$$

where $y = [y_1^T, y_2^T, \ldots, y_m^T]^T \in \mathbb{R}^{nm}$.

$$\min_{y} \sum_{i=1}^{m} \mathcal{W}^*_{\gamma, q_i}([\sqrt{W}y]_i).$$

Computation of the gradient

Reminder: $U(p,q) = \{X \in \mathbb{R}_+^{n\times n}, \;\; X\mathbf{1} = p, \; X^T\mathbf{1} = q\}$.

$$\mathcal{W}^*_{\gamma,q_i}([\sqrt{W}y]_i) = \max_{p_i \in S_{\mathbf{1}}(n)} \left\{ \langle [\sqrt{W}y]_i, p_i\rangle - \min_{X\in U(p_i,q_i)} \{\langle C, X\rangle - \gamma E(X)\} \right\}$$

$$= \max_{p_i \in S_{\mathbf{1}}(n)} \left\{ \langle [\sqrt{W}y]_i, p_i\rangle - \max_{\lambda_i,\mu_i} \left(\langle \lambda_i, p_i\rangle + \langle \mu_i, q_i\rangle - \gamma \sum_{t,k}^{n} \exp\left(\frac{\lambda_{i_t} - C_{tk} + \mu_{i_k}}{\gamma} - 1 \right) \right) \right\}$$

$$= \max_{p_i \in S_{\mathbf{1}}(n)} - \max_{\lambda_i,\mu_i} \left\{ \langle \lambda_i - [\sqrt{W}y]_i, p_i\rangle + \langle \mu_i, q_i\rangle - \gamma \sum_{t,k}^{n} \exp\left(\frac{\lambda_{i_t} - C_{tk} + \mu_{i_k}}{\gamma} - 1 \right) \right\} =$$

$$= /C^n = C - [\sqrt{W}y]_i\mathbf{1}^T/ = \min_{p_i \in S_{\mathbf{1}}(n)} -\mathcal{W}_{\gamma,q_i}(p_i) =$$

$$= - \min_{p_i \in S_{\mathbf{1}}(n)} \min_{X\in U(p_i,q_i)} \{\langle C^n, X\rangle - \gamma E(X)\} = - \min_{X\geq 0, X^T\mathbf{1} = q_i} \left\{ \langle C - [\sqrt{W}y]_i\mathbf{1}^T, X\rangle - \gamma E(X) \right\}$$

$$= \langle C - [\sqrt{W}y]_i\mathbf{1}^T, X^*\rangle - \gamma E(X^*)$$

$$\nabla_y \mathcal{W}^*_{\gamma,q_i}([\sqrt{W}y]_i) = [\sqrt{W}X^*\mathbf{1}]_i$$

Computation of the gradient

$$\nabla_y \mathcal{W}^*_{\gamma,q_i}([\sqrt{W}y]_i) = [\sqrt{W}X^*\mathbf{1}]_i = [\sqrt{W}p^*]_i = \sum_j^m \sqrt{W}_{ij} p_j^*,$$

where p^* is a solution of $\max_{p_1,\dots,p_m \in S_1(n)} \left\{ \langle y, \sqrt{W}p \rangle - \sum_{i=1}^m \mathcal{W}_{\gamma,q_i}(p_i) \right\}$

Reminder: $p = [p_1^T, p_2^T, \dots, p_m^T]^T \in \mathbb{R}^{nm}$ is a stacked column vector.

The solution of i^{th} transport problem is following

$$X_{tk} = q_{i_k} \frac{\exp\left(\frac{[\sqrt{W}y]_{i_t} - C_{tk}}{\gamma}\right)}{\sum_{t=1}^n \exp\left(\frac{[\sqrt{W}y]_{i_t} - C_{tk}}{\gamma}\right)}$$

Then using the change $\tilde{y} = \sqrt{W}y$

$$p_i^* = X^*\mathbf{1} = \sum_{k=1}^n q_{i_k} \frac{\exp\left(\frac{\tilde{y}_{i_t} - C_{tk}}{\gamma}\right)}{\sum_{t=1}^n \exp\left(\frac{\tilde{y}_{i_t} - C_{tk}}{\gamma}\right)} \quad \forall\ t \quad = \sum_{k=1}^n q_{i_k} \nabla f_{i_k},\ q_i \in S_1(n).$$

Idea of mini-batch

$$f(x) = \sum_{k=1}^{n} q_k f_k(x),$$

where q_k is a weight of function f_k and $\sum_{k=1}^{n} q_k = 1$ ($q \in S_1(n)$).

Stochastic approximation of $\nabla f(x)$ with mini-batch

$$\tilde{\nabla} f(x) = \frac{1}{b} \sum_{k=1}^{b} \nabla f_{\xi_k}(x), \quad \{\xi\}_{k=1}^{b} \text{ — i.i.d.}$$

Gradient unbiased condition is equal to

$$\mathbb{E}_{\{\xi_k\}_{i=1}^{b}} \left[\frac{1}{b} \sum_{i=1}^{b} \nabla f_{\xi_k}(x) \right] = \sum_{i=1}^{n} q_k \nabla f_k(x) \quad \text{and} \quad \mathbb{P}(\xi_k = i) = \begin{cases} q_1, & i = 1 \\ q_2, & i = 2 \\ \vdots & \\ q_n, & i = n \end{cases}$$

$$\mathbb{D}_{\{\xi_k\}_{k=1}^{b}} \left[\frac{1}{b} \sum_{k=1}^{b} \nabla f_{\xi_k}(x) \right] = \frac{1}{b^2} \sum_{k=1}^{b} \mathbb{D}_{\xi_k} \left[\nabla f_{\xi_k}(x) \right] = \frac{1}{b} \mathbb{D}_{\xi} \left[\nabla f_{\xi}(x) \right] \leq \frac{D}{b}.$$

Estimation of the variance

$$\mathbb{D}_\xi\left[\nabla f_\xi(x)\right] = \mathbb{E}_\xi\left\|\nabla f_\xi(x) - \sum_{i=1}^n q_k \nabla f_k(x)\right\|_2^2 = \mathbb{E}_\xi\left\|\nabla f_\xi(x)\right\|_2^2 - \left\|\sum_{k=1}^n q_k \nabla f_k(x)\right\|_2^2$$

$$= \sum_{i=1}^n q_k \|\nabla f_k(x)\|_2^2 - \left\|\sum_{k=1}^n q_k \nabla f_i(x)\right\|_2^2 \leq \sum_{k=1}^n q_k \|\nabla f_k(x)\|_2^2 \leq G^2,$$

where $\|\nabla f_k(x)\| \leq G \ \forall \ k = 1, \ldots, n.$

Estimation of smoothness parameter

$\mathcal{W}_{\gamma,q}(p)$ is γ strongly convex $\Longrightarrow$

$$\mathcal{W}^*_{\gamma,q}([\sqrt{W}y]) = \max_{p_1,\dots,p_m \in S_1(n)} \left\{ \langle y, \sqrt{W}p \rangle - \mathcal{W}_{\gamma,q}(p) \right\}$$

is L-smooth.

$$L = \frac{\|\sqrt{W}\|^2_{l^1 \to l^2}}{\gamma}$$

$$\begin{aligned} \|\sqrt{W}\|^2_{l^1 \to l^2} &= \max_{i=1,\cdots,m} \|\sqrt{W}_i\|_2^2 = \max_{i=1,\cdots,m} \sqrt{W}_i^T \sqrt{W}_i \\ &= \max_{i=1,\cdots,m} [W]_{ii} = d_{\max}, \end{aligned}$$

where $\sqrt{W}_i$ is a i^{th} column of $\sqrt{W}$.

$$L = \frac{d_{max}}{\gamma}$$

Algorithm: Similar Triangles Method

Require: λ_0 – starting point, ε – desired accuracy, N — number of iteration.

Set $k = 0$, $A_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$, $M_k = 2L$

For each agent $i \in V$

for $k = 0, \ldots, N-1$ **do**

$$\alpha_{k+1} = (1 + \sqrt{1 + 4M_k\alpha_k^2})/2M_k, \quad A_{k+1} = A_k + \alpha_{k+1}$$

$$\lambda_{k+1}^i = \frac{\alpha_{k+1}\zeta_k^i + A_k\eta_k^i}{A_{k=1}}$$

Randomly choose mini-batch $I_k \subset \{1, ..., n\}$ of size l with probability equal to weights

Compute a stochastic estimate $\tilde{p}_j^*(\lambda_{k+1}^j) = \frac{1}{b}\sum_{t \in I_k} \nabla f_t(\lambda_{k+1}^j)$

$$\zeta_{k+1}^i = \zeta_k^i - \alpha_{k+1}\sum_{j=1}^m W_{ij}\tilde{p}_j^*(\lambda_{k+1}^j)$$

$$\eta_{k+1}^i = \frac{\alpha_{k+1}\zeta_{k+1}^i + A_k\eta_k^i}{A_{k+1}}$$

$$(p_{k+1}^*)_i = \frac{\alpha_{k+1}p_i^*(\lambda_{k+1}) + A_k(\tilde{p}_k^*)_i}{A_{k+1}}$$

$(y_n^*)^i = \eta_N \ \forall \ i \in V$

Approximation error of the gradient

Use Fenchel inequality $\langle g, x\rangle \leq \frac{1}{2\zeta}\|g\|_*^2 + \frac{\zeta}{2}\|x\|^2$, estimate the error of approximation ($\zeta = L$)

$$\langle \nabla f(x) - \tilde{\nabla} f(x), y - x\rangle \leq \frac{1}{2L}\|\nabla f(x) - \tilde{\nabla} f(x)\|_2^2 + \frac{L}{2}\|y - x\|_2^2$$

Assume f is L-smooth: $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2$.

Adding the term $\langle -\tilde{\nabla} f(x), y - x\rangle$, we get

$$f(y) - \langle \tilde{\nabla} f(x), y - x\rangle \leq f(x) + \langle \nabla f(x) - \tilde{\nabla} f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2.$$

Then using Fenchel inequality, we obtain

$$f(y) \leq f(x) + \langle \tilde{\nabla} f(x), y - x\rangle + \frac{2L}{2}\|y - x\|_2^2 + \frac{1}{2L}\|\nabla f(x) - \tilde{\nabla} f(x)\|_2^2. \quad (1)$$

Inequality (1) is equivalent to $(\delta, 2L)$-oracle with $\delta = \frac{1}{2L}\|\nabla f(x) - \tilde{\nabla} f(x)\|_2^2$.

Randomized computation of gradient with mini-batch

$$\mathbb{E}_{\{\xi_k\}_{k=1}^b}\delta = O\left(\frac{\varepsilon}{N(\varepsilon)}\right) = \frac{C\varepsilon}{N(\varepsilon)}.$$

For similar triangles method $C = \frac{3}{2}$ (follows from stochastic estimation of complexity bounds of algorithm).

$$\mathbb{E}_{\{\xi_k\}_{k=1}^b}\delta = \mathbb{E}_{\{\xi_k\}_{k=1}^b}\left[\frac{1}{2L}\|\nabla f(x) - \tilde{\nabla} f(x)\|_2^2\right] \leq \frac{G^2}{2Lb}$$

Considering the number of iteration for similar triangles method

$$N(\varepsilon) = \sqrt{\frac{16LR^2}{\varepsilon}}$$

The number of function for gradient estimation

$$b = \max\left\{1, G^2\sqrt{\frac{16R^2}{9L\varepsilon^3}}\right\}$$

Overall complexity

The number of iteration is N,
arithmetical complexity of calculating the gradient is $O(n)$,
the number of gradients is b and n (respect to with and without mini-batch)

With mini-batch	Without mini-batch
$N \cdot n \cdot b$	$N \cdot n \cdot n$
$O\left(n\frac{G^2R^2}{\varepsilon^2}\right)$	$O\left(n^2\sqrt{\frac{LR^2}{\varepsilon}}\right)$

To compare

$\frac{n}{\tilde{\varepsilon}^2}$	$\frac{n^2}{\sqrt{\tilde{\varepsilon}}}$

where $\frac{G^2R}{\varepsilon} \sim 1/\tilde{\varepsilon}$ and $\frac{LR^2}{\varepsilon} \sim 1/\tilde{\varepsilon}$ are relative errors.

Comparing of the approach

For real image $\sim 1024 \times 1024$ pixels:

- The size of histogram $n = 1024^2 \sim 10^6$

$$\frac{n}{\tilde{\varepsilon}^2} \text{ VS } \frac{n^2}{\sqrt{\tilde{\varepsilon}}}$$

If $\tilde{\varepsilon} \geq 10^{-4}$ then $n/\tilde{\varepsilon}^2 \leq n^2/\sqrt{\tilde{\varepsilon}}$.

Conclusion:

If the size of image n is large and the accuracy ε is not high then mini-batch works!

Results publication

- César A. Uribe, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Angelia Nedić.
  Distributed Computation of Wasserstein Barycenters over Networks.
  arXiv preprint:1803.02933, 2018.

MNIST

3 agents on the different stage of iteration for finding the barycenters of all digits (from 0 to 9).