Decentralized Computation of Wasserstein Barycenter over Time-Varying Networks

Olga Yufereva
N. N. Krasovski Institute of Mathematics and Mechanics
olga.o.yufereva@gmail.com

Michael Persiianov
Moscow Institute of Physics and Technology
persiianov.mi@phystech.edu

Pavel Dvurechensky
Weierstrass Institute for Applied Analysis and Stochastics
pavel.dvurechensky@wias-berlin.de

Alexander Gasnikov
Moscow Institute of Physics and Technology
gasnikov@yandex.ru

Dmitry Kovalev
King Abdullah University of Science and Technology
dmitry.kovalev@kaust.edu.sa

Abstract

Inspired by recent advances in distributed algorithms for approximating Wasserstein barycenters, we propose a novel distributed algorithm for this problem. The main novelty is that we consider time-varying computational networks, which are motivated by examples when only a subset of sensors can make an observation at each time step, and yet, the goal is to average signals (e.g., satellite pictures of some area) by approximating their barycenter. We prove non-asymptotic accelerated in the sense of Nesterov convergence rates and explicitly characterize their dependence on the parameters of the network and its dynamics. Our approach is based on our novel distributed non-smooth optimization method on time-varying networks, which may be of separate interest. In the experiments, we demonstrate the efficiency of the proposed algorithm.

Introduction

In this paper we consider Wasserstein barycenter problem \cite{agueh2011entropies} (WB problem) and propose accelerated dual decentralized algorithm on time-varying network. The interest to WB problem has grown significantly in the last decade (see e.g. \cite{peyre2019computational} and references there in). In ML applications this problem is typically expensive, although it is convex \cite{dvinskikh2021decentralized}. So distributed methods are required.

At the same time in decentralized convex optimization a significant progress was also made in the last decade (see e.g. \cite{gurbunov2020fast} and references there in). In particular, for smooth strongly
convex optimization problems decentralized methods for time-varying networks were proposed (for primal oracle see [Li and Lin [2021], Kovalev et al. [2021a] and for dual one see Kovalev et al. [2021b]).

Starting with Uribe et al. [2018] it was observed that decentralized methods with dual oracle are well suited for WB problem. In the cycle of subsequent papers Dvurechenskii et al. [2018], Kroshnin et al. [2019], Dvinskikh et al. [2019], Krawtschenko et al. [2020], Dvinskikh [2021a] different decentralized accelerated (randomized) algorithms were proposed for dual WB problem. In Dvinskikh and Tlapkin [2021] it was mentioned that WB problem can be reformulated as bilinear saddle-point problem (SPP). Decentralized algorithm for this problem was also proposed Rogozin et al. [2021]. But in all these papers only the networks that do not change over time were considered.

SPP approach from Rogozin et al. [2021] significantly required that network does not change over time. In the core of the dual approach lies the dual problem for WB problem with affine-consensus constraints. This problem was solved by different variants of accelerated methods Dvinskikh [2021a]. If communication network changes over time, then affine-consensus constraints also changes over time so does the dual problem. Thus, in the dual approach we should solve the family of dual problems. Since accelerated methods have potential functions that depend on target functions it is difficult to generalize directly algorithms from Dvurechenskii et al. [2018], Kroshnin et al. [2019], Dvinskikh et al. [2019], Krawtschenko et al. [2020], Dvinskikh [2021a] for time-varying networks.

So in any case (for the dual approach and SPP approach) it is also not obvious how to generalize known decentralized algorithms for WB problem for time-varying network.

The natural idea is to use ADOM that is an optimal decentralized algorithm for smooth strongly convex unconstrained problems for time-varying networks with dual oracle Kovalev et al. [2021b]. ADOM for WB problem can be considered as projected accelerated algorithm with inexact consensus-based projection applied to specific dual reformulation of WB problem. Since WB problem

- (Smoothness) is not smooth;
- (Constraints) is not unconstrained; it has natural simplex constraint;
- (Strongly convex) is not strongly convex;

it seems impossible to apply ADOM for WB problem directly.

The main result of this paper is generalization of ADOM for general convex decentralized optimization problem (with simple constraints). Then we apply the method we have obtained for WB problem. The problem of Strong convexity can be easily solved by the regularization of the primal problem. For WB problem we use entropy-based regularization Peyré et al. [2019]. We solve the problems of Smoothness and Constraints by using special regularization of the dual problem. This regularization goes back to Devolder et al. [2012], Gasnikov et al. [2016] and by infimal convolution can be considered as Moreau–Yosida smoothing of the primal problem Rockafellar [2015], Lemaréchal and Sagastizábal [1997]. We emphasis that the proposed dual regularization (primal smoothing) was earlier investigated only for non time-varying networks Uribe et al. [2020]. For time-varying networks and problems with simple constraints the analysis is different.

**Notation**

We use bold or normal font ($\mathbf{x}$ or $x$) for different spaces $\mathbf{x} \in (\mathbb{R}^d)^m$, $x \in \mathbb{R}^d$. The $l$-th component of a vector $x \in \mathbb{R}^d$ is denoted by $[x]_l$ and $l$-th component of $x \in (\mathbb{R}^d)^m$ is denoted by $[x]_l$, which is the corresponding vector from $\mathbb{R}^d$. We also use capital and lowercase letters for function to visually indicate their domain, as we consider $\min_{x \in S} F(x) = \frac{1}{m} \sum f_i(x_i)$.

Let $1$ denote a column vector with all entries equal to 1. The $d$-dimension simplex is denoted $S_1(d)$, that means $S_1(d) := \{ p \in [0, 1]^d | p^T 1 = 1 \}$. For matrices $A$ and $B$, $A \circ B$ and $A / B$ stands for the element-wise product and division, respectively. Another product we define as follows $\langle M, X \rangle := \sum_{i=1}^d \sum_{j=1}^d M_{ij} X_{ij}$. We use $\mathbb{R}_+$ for $[0, +\infty)$.

**Limitation**

All communication networks are supposed to be non-trivial and to have same number of vertices, however vertices are allowed to be isolated. Our numerical method relies on Assumption...
restricts the conditional number of communication matrices. In definitions of Wasserstein distance, a cost (loss) matrix \( M \) is a non-negative symmetric matrix with zeros on the diagonal. In WB problem we deal with finite-supported probabilities distributions, i.e. we assume that initial data vectors belong to \( d \)-dimension simplexes; Theorem 3.3 also requires these vectors to be bounded away from zero.

1 Wasserstein Barycenter Problem

Wasserstein barycenter problem is a famous and commonly occurring optimization problem. By Wasserstein barycenter we call a minimum of a sum of Wasserstein distances for given set of probability distributions. It is known (see Cuturi and Doucet [2014]) that barycenters captures the mean structure of given data. To provide insight how local nodes’ data change and tend to global barycenter, we refer to Figure 1 that illustrates our theoretical result, Theorem 3.3, on example of a dataset of hand-written digits ‘4’ of MNIST 784 [LeCun [1998]], which barycenter resembles ‘4’ as well.

More precisely, Figure 1a presents the evolution of local ‘barycenters’ and Figure 1b compares the terminal state at one node with the ‘true’ barycenter while communication networks are Erdős–Rényi random networks and change every 5 iterations.

General theory, that begins with Wasserstein distance, can be found in Villani [2009]. We provide only necessary definitions and take into consideration only finite-supported distribution since we deal with numerical experiments.

Consider two probability distributions \( p, q \in S_1(d) \) with support on a finite set of points \( \{x_i \in \mathbb{R}^d\}_{i=1}^d \) such that \( p(x_i) = p_i \) and \( q(x_i) = q_i \). A cost (loss) matrix \( M \), which elements satisfy \( [M]_{ij} \in \mathbb{R}_+ \), represents the cost of moving a unit mass from \( i \)-th to \( j \)-th component. So \( M \) should be a non-negative symmetric matrix with zeros on the diagonal. It is often taken as the Euclidean distances matrix, i.e. \( [M]_{ij} = \|x_i - x_j\|_2^2 \). Define the set of transport plans as

\[
U(p, q) := \{ X \in \mathbb{R}_+^{d \times d} \mid X 1 = p, X^T 1 = q \}.
\]

Wasserstein distance between two probability distributions defines as the minimum

\[
\mathcal{W}(p, q) := \min_{X \in U(p, q)} \langle M, X \rangle.
\]

The uniform Wasserstein barycenter of the family of \( q_i \in S_1(d) \) is defined as the solution to the following optimization problem

\[
\frac{1}{m} \sum_{i=1}^m \mathcal{W}_{q_i}(p) \rightarrow \min_{p \in S_1(d)} (1)
\]

where \( \mathcal{W}_{q_i}(p) \) is one-argument function that equals \( \mathcal{W}(q_i, p) \). This is a convex optimization problem, that brought a number of challenges even for centralized computation (see Peyré et al. [2019]). Further we consider (1) as decentralized convex optimization problem and Theorem 3.3 states a computation method to solve it.

2 Decentralized optimization

2.1 Decentralized computation problem

Decentralized computation simulates computation on distributed individual devices. The devices are considered as nodes of an undirected connected graph called a communication network. It means that each node can perform computations based only on its local data and the data of its neighbors in communication network. Decentralized computation of an optimization problem

\[
\sum_{i=1}^n f_i(x) \rightarrow \min_{x \in S}, (2)
\]

\footnote{As the ‘true’ barycenter we take the output computed via standard method of Python Optimal Transport (POT) package Flamary et al. [2021] that computes the entropy regularized Wasserstein barycenter of distributions.}
requires numerical computation assuming that each function $f_i$ is stored on the corresponding node $i \in [m] := \{1, 2, \ldots, m\}$. Such approach brings us to an effective reformulation of the optimization problem.

Figure 1: Evolution of local data converging to Wasserstein barycenter of the hand-written digit 4 of the MNIST784 dataset for a subset of 7 nodes out of 50 over Erdős–Rényi random networks varying each 5 iterations; regularization parameters are $\gamma = 0.03$, $\rho = 0.001$
2.1.1 Consensus condition

Since each computational node carries its own local data approximation, we can substitute formally different variables \(x_i\) for the mutual argument \(x\) in (2) assuming that they belong to the so-called consensus space. It means the new variables \(x_i\) must eventually coincide with each other and with the wanted barycenter. We obtain an equivalent optimization problem in the following form:

\[
F(x) = \frac{1}{m} \sum_{i=1}^{m} f_i([x]_i) \rightarrow \min_{x \in S},
\]

where \(S = \{x = ([x]_1, \ldots, [x]_m) \in (\mathbb{S})^m \mid [x]_1 = \ldots = [x]_m\}\).

Here \(i\)-th component \([x]_i\) is a corresponding \(d\)-dimension vector.

2.1.2 Time-varying communication network

Decentralized communication between nodes is typically represented via a matrix-vector multiplication with a communication (gossip) matrix. Typically, for each time step \(k \in \{0, 1, 2, \ldots\}\) we have a corresponding communication network with its Laplacian \(\hat{W}_k\), but generally \(\hat{W}_k\) must fulfill only the following:

1. \(\hat{W}_k\) is symmetric and positive semi-definite,
2. \([\hat{W}_k]_{i,j} \neq 0\) if and only if \((i, j)\) are connected by the network,
3. \(\ker \hat{W}_k = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 = \ldots = x_n\}\).

Then let us introduce the matrix \(W_k = \hat{W}_k \otimes I_d\). This matrix can be represented as a block matrix \((W_k)_{i,j} = (\hat{W}_k)_{i,j} I_d\), where each block \((\hat{W}_k)_{i,j}\) is a \(d \times d\) matrix proportional to \(I_d\). Hence, decentralized communication of vectors \(x_1, \ldots, x_n \in \mathbb{R}^d\) stored on the nodes among neighboring nodes at time step \(k\) can be represented as a multiplication of the \(md\)-dimensional vector by matrix \(W_k\). Indeed, consider \(x = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^m\), \(y = (y_1, \ldots, y_n) \in (\mathbb{R}^d)^m\), where each \(x_i\) is stored by node \(i\), and let \(y = W_k x\). One can observe that

\[
y_i = \sum_{j=1}^{n} (\hat{W}_k)_{i,j} x_j = \sum_{j \in N_i} (\hat{W}_k)_{i,j} x_j,
\]

where \(N_i\) is the set of the neighboring nodes for the node \(i\) according to the communication network at \(k\)-th iteration. Hence, for each node \(i\), vector \(y_i\) is a linear combination of vectors \(x_j\), stored at the neighboring nodes \(j \in N_i\).

The considered Algorithm\[1\] requires also these matrices \(W_k\) to have conditional numbers bounded for all \(k \in \{0, 1, 2, \ldots\}\). Namely, it requires the following assumption.

**Assumption 1.** Let there exist constants \(0 < \lambda^+_{\min} < \lambda_{\max}\) such that

\[
\lambda^+_{\min} \leq \lambda^+_{\min}(W_k) \leq \lambda_{\max}(W_k) \leq \lambda_{\max} \quad \forall k
\]

where \(\lambda^+_{\min}(W_k)\) is the smallest positive eigenvalue of \(W_k\) and \(\lambda_{\max}(W_k)\) is the biggest one.

A condition number of the matrix \(W_k\) is given as \(\frac{\lambda_{\max}(W_k)}{\lambda_{\min}(W_k)}\) and is known to be a measure of the connectivity of a network, and appears in convergence rates of many decentralized algorithms.

2.2 ADOM and its assumptions

The state of art numerical computation method for time-varying networks is developed in Kovalev et al.\[2021b\]. It has natural restrictions on the class of suitable problems and, e.g., WB problem lies beyond the requirements of its algorithm since its domain is not \(\mathbb{R}^d\) and the function is non-smooth and is not strongly convex (until it is regularized). So we modify ADOM to solve general optimization problems with restrictions.

This subsection is to present the main objects of Kovalev et al.\[2021b\]. For the sake of consistency, original notation are slightly changed.
where functions $h_i : \mathbb{R}^d \to \mathbb{R}$ are assumed to be smooth and strongly convex. Then, its argmin is approximated in terms of the Fenchel transform $H^*$ of the function $H$ since here set $S = \mathbb{R}^d$ and problem (4) is equivalent to the following:

$$H^*(z) \to \min_{z \in L^\perp},$$

where $L^\perp = \{ z = ([z]_1, \ldots, [z]_m) \in (\mathbb{R}^d)^m \mid \sum_{i=1}^m [z]_i = 0 \}$.

**Theorem 2.1** ([Kovalev et al., 2021b Theorem 1]). Let functions $h_i : \mathbb{R}^d \to \mathbb{R}$ be $L$ smooth and $\mu$ strongly convex, $x^*$ be the argminimum of the optimization problem (4). $W_k$ be a communication network at the $k$-th iteration satisfying Assumption [4]. Set parameters $\alpha, \eta, \theta, \sigma, \tau$ of Algorithm [1] to $\alpha = \frac{1}{4L}, \eta = \frac{2L}{\lambda_{\max}} \sqrt{\mu}, \theta = \frac{\mu}{\lambda_{\max}}, \sigma = \frac{1}{\lambda_{\max}}$, and $\tau = \frac{\lambda_{\max}}{\lambda_{\max}} \sqrt{\mu}$. Then there exists $C > 0$, such that for $H^*(z) = \frac{1}{m} \sum_{i=1}^m h_i([z]_i)$

$$\|\nabla H^*(z^*_n) - x^*\|^2_2 \leq C \left( 1 - \frac{\lambda_{\min}}{\lambda_{\max}} \sqrt{\frac{\mu}{L}} \right)^n$$

Addressing details of the proof of Theorem 1 of [Kovalev et al., 2021b] we see that there is a particular choice of the constant $C$, namely

$$C = \max \left\{ \frac{2}{\mu^2},\tau(1-\tau)L \right\} = \frac{1}{\mu^2} \max \left\{ \frac{2\lambda_{\min} \sqrt{\mu}}{\tau \lambda_{\max} \sqrt{L}}, \frac{1}{2}\right\} = \frac{1}{2\mu^2}.$$

It means that the actual convergence rate is $n = O \left( \frac{\lambda_{\min}}{\lambda_{\max}} \sqrt{\frac{\mu}{L}} \ln \frac{1}{\mu^2} \right)$.

**Algorithm 1** ADOM: Accelerated Decentralized Optimization Method

1: **input:** $\nabla H^* : (\mathbb{R}^d)^m \to \mathbb{R}, z^0 \in L^\perp, m^0 \in (\mathbb{R}^d)^V, \alpha, \eta, \theta, \sigma, \tau > 0, \tau \in (0, 1)$
2: set $z^0_g = z^0$
3: for $k = 0, 1, 2, \ldots$ do
4: $z^k_g = \tau z^k + (1-\tau)z^k$
5: $\Delta^k = \sigma W_k(m^k - \eta \nabla H^*(z^k_g))$
6: $m^{k+1} = m^k - \eta \nabla H^*(z^k_g) - \Delta^k$
7: $z^{k+1}_g = z^k + \eta \alpha (z^k_g - z^k) + \Delta^k$
8: $z^{k+1}_i = z^k_i - \theta W_k \nabla H^*(z^k_g)$
9: end for

3 Main results

One of our main features is that functions of optimization problem can be defined on a convex set $S \subset \mathbb{R}^d$ instead of the whole $\mathbb{R}^d$, that is crucial for duality (5) of the consensus condition in combination with smoothness. The main result, Theorem 3.1, is stated for the following decentralized optimization problem over time-varying networks:

$$x^*_n = \arg\min_{x \in S} F^\tau(x) = \arg\min_{x \in S} \frac{1}{m} \sum_{i=1}^m f_i^\tau(x),$$

where $S = \{ x = ([x]_1, \ldots, [x]_m) \in (S)^m \subset (\mathbb{R}^d)^m \mid [x]_1 = \ldots = [x]_m \}$, where functions $f_i^\tau$ are $\gamma$ strongly convex and defined on a convex set $S$. Recall that $i$-th component $[x]_i$ is a corresponding $d$-dimension vector.

Then we see that similarly we can approximate convex functions (Corollary 3.2). We obtain an important Theorem 3.3 by applying Theorem 3.1 to the Wasserstein barycenter problem; its particularity allows to estimate parameters well. In the next section we provide numerical experiments on the example of the Wasserstein barycenter problem.
3.1 General case

**Theorem 3.1.** Let \( S \subset \mathbb{R}^d \) be a convex set, let functions \( f_i^\gamma : S \to \mathbb{R} \) of the problem (8) be \( \gamma \) strongly convex, let \( \mathcal{W}_k \) be the \( k \)-th communication network satisfying Assumption 1 for some \( \lambda_{\text{min}}, \lambda_{\text{max}} > 0 \). Algorithm 1 applied for

\[
\nabla (H^{r,\gamma})^* (z) = (\nabla h_1^\gamma ([z]_1), \ldots, \nabla h_m^\gamma ([z]_m))^\top, \\
\nabla h_i^\gamma ([z]_i) = \frac{1}{m} \nabla (f_i^\gamma)^* ([z]_i) + \frac{r}{mK} [z]_i,
\]

for \( x_{r,\gamma} = \nabla (H^{r,\gamma})^* (z_{r,\gamma}) \), and for a sufficient number of iteration \( n \) yields:

1. **consensus condition approximation:** for each \( i \) and \( j \)

\[
\| [x_{r,\gamma}]_j - [x_{r,\gamma}]_i \|_2^2 \leq C_1 \left( 1 - \frac{\lambda_{\text{min}}^+}{\lambda_{\text{max}}^+} \sqrt{\frac{r,\gamma}{1 + r,\gamma}} \right)^n \leq \varepsilon; \tag{10}
\]

2. **value approximation:**

\[
F^{\gamma} (x_{r,\gamma}) - \min_{x \in S} F^{\gamma} (x) \leq \frac{r}{2 (1 + r,\gamma)} mK^2 + C_2 \left( 1 - \frac{\lambda_{\text{min}}^+}{\lambda_{\text{max}}^+} \sqrt{\frac{r,\gamma}{1 + r,\gamma}} \right)^{n/2} \leq \varepsilon \tag{11}
\]

where \( r \leq \frac{\varepsilon}{2mK^2} \), \( C_1 = \frac{(1 + r,\gamma)^2}{2^2} \), \( C_2 = \frac{(1 + r,\gamma)\sqrt{mK}}{2^2} \), \( K \) is such that \( \| \nabla f_i^\gamma (x) \|_2 \leq K \) for all \( x \in \{ y \in S \mid \| y - \arg \min_{x \in S} F^{\gamma} (x) \|_2 \leq \varepsilon/\gamma \} \) and Algorithm’s input is \( \alpha = \frac{\varepsilon}{2} \).

\[
\eta = \frac{2\lambda_{\text{min}}^+ \sqrt{\gamma}}{\eta \lambda_{\text{max}}^+ \sqrt{r (1 + r,\gamma)}}, \quad \theta = \frac{\gamma}{\lambda_{\text{max}}^+(1 + r,\gamma)}, \quad \sigma = \frac{1}{\lambda_{\text{max}}^+}, \quad \tau = \frac{\lambda_{\text{min}}^+}{\lambda_{\text{max}}^+} \sqrt{\frac{r,\gamma}{1 + r,\gamma}}.
\]

Then the rate of the number of iterations is

\[
n = O \left( \frac{\lambda_{\text{max}}^+}{\lambda_{\text{min}}^+} \sqrt{\left( 1 + \frac{r,\gamma}{r,\gamma} \right)} \ln C_2 \varepsilon \right) = O \left( \frac{1}{\sqrt{\gamma \varepsilon}} \ln \frac{1}{\varepsilon} \right)
\]

**Corollary 3.2.** Let \( S \) be a convex set in \( \mathbb{R}^d \) and let \( f_i : S \to \mathbb{R} \) be continuous convex function for \( i = 1, \ldots, m \). Decentralized convex optimization problem \( \frac{1}{m} \sum_{i=1}^{m} f_i (x) \to \min_{x \in S} \) over time-varying communication networks can be \( \varepsilon \)-approximated numerically by Algorithm 1 with sufficiently small \( r \), sufficiently great \( n \) and

\[
\nabla h_i^\gamma ([z]_i) = \frac{1}{m} \nabla (f_i^\gamma)^* ([z]_i) + \frac{r}{2m} [z]_i
\]

for \( \gamma \) strongly convex regularizing functions \( f_i^\gamma (x) \) that satisfy

\[
0 \leq \min_{z \in S} \sum_{i=1}^{m} f_i (x) - \min_{x \in S} \sum_{i=1}^{m} f_i^\gamma (x) \leq \varepsilon/2,
\]

e.g., for \( f_i^\gamma (x) = f_i (x) + \frac{\gamma}{2} \| x \|_2^2 \).

3.2 Wasserstein barycenter problem computation

Turning to the Wasserstein barycenter problem, Assumption 2 provides us with precisely defined function \( \nabla (H^{r,\gamma})^* \) and a bound on the parameter \( K = K_\zeta \).

**Assumption 2.** Let vectors \( q_i \in S_1 (d) \subset \mathbb{R}^d \) be such that

\[
\min_{i=1, \ldots, d} |q_i|_1 = \delta > 0.
\]

**Theorem 3.3.** Let vectors of initial data \( q_i \) satisfy Assumption 2 and let \( \mathcal{W}_k \) be the \( k \)-th communication network satisfying Assumption 1 for some \( \lambda_{\text{min}}^+, \lambda_{\text{max}}^+ > 0 \). Algorithm 1 applied for

\[
\nabla h_i^\gamma (z) = \frac{r}{2m} z + \frac{m}{m} \sum_{j=1}^{m} [q]_j \frac{\exp (\frac{1}{2} ([z]_j - M_{ij}))}{\sum_{l=1}^{m} \exp (\frac{1}{2} ([z]_l - M_{ij}))},
\]

\[
\nabla (H^{r,\gamma})^* (z) = (\nabla h_1^\gamma ([z]_1), \ldots, \nabla h_m^\gamma ([z]_m))^\top,
\]

7
approximates Wasserstein barycenter problem \([1]\) with the barycenter \(p^*\) as

\[
\left| \sum_{i=1}^{m} W_{q_i}([x^n_i]_i) - \sum_{i=1}^{m} W_{q_i}([p^*]_i) \right| 
\leq 2\gamma \ln d + \frac{r}{4(1+r\gamma)} mK^2 + C \left( 1 - \frac{\lambda_{\text{min}}}{\lambda_{\text{max}}} \sqrt{\frac{r\gamma}{1+r\gamma}} \right)^{n/2} \leq \varepsilon,
\]

where it suffices to take \(\gamma = \frac{1}{8} \varepsilon \ln d\), \(K = \sqrt{\sum_{i=1}^{d} \left( 2\gamma \ln d + \inf_j \sup_i |M_{ij} - M_{il}| - \gamma \ln \delta \right)}\), \(r = \frac{\varepsilon}{4mK^2}\), \(C = \frac{(1+r\gamma)\sqrt{MK}}{2\sqrt{2\gamma}} \sqrt{\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} + \frac{(1+r\gamma)^2}{8r\gamma^2}}\); i.e. \(K = \mathcal{O}\left( \|M\|_\infty \sqrt{d} + \frac{\gamma \ln d}{\ln 3} \sqrt{d} \right)\) and number of iterations

\(n = \mathcal{O}\left( \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \sqrt{\frac{1 + r\gamma}{r\gamma}} \ln \frac{C}{\varepsilon} \right) = \mathcal{O}\left( \frac{1}{\varepsilon} \ln \frac{1}{\varepsilon} \right)\).

4 Numerical Experiments

We provide numerical experiments to demonstrate performance of the proposed method. We can fruitfully test our method on an artificial set of univariate, discrete and truncated Gaussian distributions. The reason is that we have an analytic formula for the zero-entropy Wasserstein barycenter of a set of Gaussian distributions.
For all figures of this section we generated a dataset of truncated Gaussians of size 100. The number of Gaussians (same as the number of nodes) is indicated at figures. For entropy regularized Wasserstein distance we use normalized Euclidean cost matrix and regularization parameter $\gamma = 0.01$. The method’s regularization parameter is $r = 0.001$.

One natural way to sample a random network is to independently sample each edge with a probability $p$. Such networks are called Erdős–Rényi network or $(m, p)$-Erdős–Rényi network, where $m$ is the number of nodes and $p$ is the probability of an edge.

We can see the difference in error value in two frequency-limiting cases: Figure 2a presents evolution of errors computed for constant networks of different topologies, while at Figure 2b networks change every iterations within indicated topology; the exception is complete network that cannot change. Let us notice also that star, cycle and minimum spanning tree (of $(m, 0.9)$-Erdős–Rényi) networks have $m − 1$, $m$, and $m + 1$ edges respectively in contrast to the complete network with $m(m − 1)/2$ edges and $(m, 0.5)$-Erdős–Rényi network that has $m(m − 1)/4$ edges in average.

![Graphs](image)

(a) cycle networks

(b) $(m, 0.9)$-Erdős–Rényi networks

Figure 3: Different number of iteration between network changing: errors over time

For the two ‘most efficient’ topologies that prove themselves at Figures 2a–2b we compute at Figure 3 the error evolution for different frequency of the networks varying. We indicate the lengths of epoch, i.e. number of iteration between network changing. Notice that the evolution for constant networks, computed above, matches to infinite epoch length. The number of iterations remain the same on all figures despite it is insufficient for convergence on cycle networks. Nonetheless one can see the
trends of convergence and notice that there is no monotonicity with respect to frequency of networks varying.

Funding Transparency Statement

The work of A. Gasnikov in the first part of the paper was funded by Russian Science Foundation (project 18-71-10108). The work of O. Yufereva in the second part of the paper was performed as part of research conducted in the Ural Mathematical Center with the financial support of the Ministry of Science and Higher Education of the Russian Federation (Agreement number 075-02-2022-874).

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

J. Bigot, E. Cazelles, and N. Papadakis. Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755, 2019.

M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 2015.

O. Devolder, F. Glineur, and Y. Nesterov. Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM Journal on Optimization*, 22(2):702–727, 2012.

D. Dvinskikh. *Decentralized Algorithms for Wasserstein Barycenters*. PhD thesis, Humboldt Universitaet zu Berlin (Germany), 2021a.

D. Dvinskikh. Stochastic approximation versus sample average approximation for wasserstein barycenters. *Optimization Methods and Software*, pages 1–33, 2021b.

D. Dvinskikh and D. Tiapkin. Improved complexity bounds in wasserstein barycenter problem. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1738–1746. PMLR, 2021.

D. Dvinskikh, E. Gorbunov, A. Gasnikov, P. Dvurechensky, and C. A. Uribe. On primal and dual approaches for distributed stochastic convex optimization over networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7435–7440. IEEE, 2019.

P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich. Decentralize and randomize: Faster algorithm for wasserstein barycenters. *Advances in Neural Information Processing Systems*, 31, 2018.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Roiot, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL [http://jmlr.org/papers/v22/20-451.html](http://jmlr.org/papers/v22/20-451.html).

A. V. Gasnikov, E. Gasnikova, Y. E. Nesterov, and A. Chernov. Efficient numerical methods for entropy-linear programming problems. *Computational Mathematics and Mathematical Physics*, 56(4):514–524, 2016.

E. Gorbunov, A. Rogozin, A. Beznosikov, D. Dvinskikh, and A. Gasnikov. Recent theoretical advances in decentralized distributed convex optimization. *arXiv preprint arXiv:2011.13259*, 2020.

D. Kovalev, E. Gasanov, A. Gasnikov, and P. Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems*, 34, 2021a.

D. Kovalev, E. Shulgin, P. Richtárik, A. V. Rogozin, and A. Gasnikov. Adom: accelerated decentralized optimization method for time-varying networks. In *International Conference on Machine Learning*, pages 5784–5793. PMLR, 2021b.

R. Krawtchenko, C. A. Uribe, A. Gasnikov, and P. Dvurechensky. Distributed optimization with quantization for computing wasserstein barycenters. *arXiv preprint arXiv:2010.14325*, 2020.
Moreover, its conjugate is equivalent to
\[ (h^r_{i,\gamma})^*(z_i^k) = \nabla (H^{r,\gamma})^*(z_i^k), \]
where \( \nabla (H^{r,\gamma})^*(z_i^k) \) is the output of Algorithm 1.

**A.1 Derivation \((H^{r,\gamma})^*\)**

In brief, in this subsection we define \( \frac{1}{r} \) smooth and \( \frac{1}{1+r+\gamma} \) functions \( h^r_{i,\gamma} \) such that \( \nabla (H^{r,\gamma})^* \) from (9) is the gradient of the conjugate function \((H^{r,\gamma})^*\) of \( H^{r,\gamma} = \frac{1}{m} \sum_{i=1}^{m} h^r_{i,\gamma} \). Then the consensus condition (10) becomes a corollary of Theorem 2.1 with \( L = \frac{1}{r} \) and \( \mu = \frac{\gamma}{1+r+\gamma} \).

From now on let functions \( h^r_{i,\gamma}: \mathbb{R}^d \rightarrow \mathbb{R} \) and \( H^{r,\gamma}: (\mathbb{R}^d)^m \rightarrow \mathbb{R} \) be
\[
H^{r,\gamma}(x) = \frac{1}{m} \sum_{i=1}^{m} h^r_{i,\gamma}([x]_i), \text{ where } h^r_{i,\gamma}(x) = \inf_{y \in S} \left\{ f_i^r(y) + \frac{1}{2r} \|y - x\|_2^2 \right\}. \tag{13}
\]

Define their conjugate as \((h^r_{i,\gamma})^*\) and \((H^{r,\gamma})^*\).

**Lemma A.1.** If functions \( h^r_{i,\gamma} \) and \( H^{r,\gamma} \) are defined by (13), then their Fenchel conjugate functions \((h^r_{i,\gamma})^*\) and \((H^{r,\gamma})^*: (\mathbb{R}^d)^m \rightarrow \mathbb{R}\) are
\[
(H^{r,\gamma})^*(z) = \frac{1}{m} \sum_{i=1}^{m} (h^r_{i,\gamma})^*([z]_i), \text{ where } (h^r_{i,\gamma})^*(z) = (f_i^r)^*(z) + \frac{\gamma}{2} \|z\|_2^2.
\]

Moreover, its conjugate \((H^{r,\gamma})^{**}\) coincides with \( H^{r,\gamma} \).
Moreover, since
Thus we construct a relaxation
\[ \tilde{f}_i^\gamma(x) = \begin{cases} f_i^\gamma(x) & \text{if } x \in S \\ +\infty & \text{otherwise.} \end{cases} \]  
Such \( \tilde{f}_i^\gamma \) are \( \gamma \) strongly convex as well. Moreover, substitution \( \tilde{f}_i^\gamma \) for \( f_i^\gamma \) affect neither primal \( h_i^{\gamma} \):
\[ h_i^{\gamma}(x) = \inf_{y \in S} \left\{ f_i^\gamma(y) + \frac{1}{2r} \|y - x\|^2_2 \right\} = \inf_{y \in \mathbb{R}^d} \left\{ \tilde{f}_i^\gamma(y) + \frac{1}{2r} \|y - x\|^2_2 \right\}, \]

nor \( (f_i^\gamma)^*(z) + \frac{r}{2} \|z\|^2_2 \):
\[
(f_i^\gamma)^*(z) + \frac{r}{2} \|z\|^2_2 = \max_{x \in S} \left\{ \langle z, x \rangle - f_i^\gamma(x) \right\} + \frac{r}{2} \|z\|^2_2
\]
\[ = \max_{x \in \mathbb{R}^d} \left\{ \langle z, x \rangle - \tilde{f}_i^\gamma(x) \right\} + \frac{r}{2} \|z\|^2_2 = (f_i^\gamma)^*(z) + \frac{r}{2} \|z\|^2_2. \]

For each \( i \) one can see that \( (h_i^{\gamma})^* = (f_i^\gamma)^* + \frac{r}{2} \|z\|^2_2 \) is the Fenchel conjugate of \( h_i^{\gamma} \) and vice versa. Indeed, for proper, convex and lower semicontinuous \( g_1, g_2 : \mathbb{R}^d \to \mathbb{R} \) we have \( (g_1 + g_2)^*(x) = g_1^* \square g_2^* \) and \( (g_1 \square g_2)^* = g_1^* + g_2^* \), where \( (g_1 \square g_2)(x) \) means the convolution \( \inf \{ g_1(y) + g_2(x - y) \mid y \in \mathbb{R}^d \} \).

Hence the Fenchel conjugate for the function \( H^{\gamma} \) will be
\[
\sup_{x \in (\mathbb{R}^n)^m} \left\{ \langle z, x \rangle - H^{\gamma}(x) \right\} = \sup_{x \in (\mathbb{R}^n)^m} \left\{ \sum_{i=1}^m \left( \langle [z]_i, [x]_i \rangle - \frac{1}{m} h_i^{\gamma}([x]_i) \right) \right\}
\]
\[ = \sum_{i=1}^m \sup_{[x]_i \in \mathbb{R}^d} \left\{ \langle [z]_i, [x]_i \rangle - \frac{1}{m} h_i^{\gamma}([x]_i) \right\} = \frac{1}{m} \sum_{i=1}^m (h_i^{\gamma})^*([z]_i) = (H^{\gamma})^*(z). \]

In the same way one can see that \( H^{\gamma} \) and \( (H^{\gamma})'' \) coincide.

**Remark 1.** For each \( i \) the function \( (h_i^{\gamma})^* \) from \( [13] \) is \( \left( \frac{1}{\gamma} + r \right) \) smooth and \( r \) strongly convex by definition, so we have \( h_i^{\gamma} = (h_i^{\gamma})'' \) being \( \frac{1}{r} \) smooth and \( \frac{1}{r + \gamma} \) strongly convex. In addition
\[
\nabla (h_i^{\gamma})^*(z) = \frac{1}{m} (f_i^\gamma)^*(z) + \frac{r}{2m} z
\]
as required in \([9]\). Then we can apply Algorithm 1 for \( L = r^{-1} \) smooth and \( \mu = \frac{\gamma}{r + \gamma} \) strongly convex functions \( h_i^{\gamma} \) and get the values of \( \nabla (h_i^{\gamma})^*(z) \) as output.

Thus we construct a relaxation \( H^{\gamma}(x) \to \min_{x \in \mathcal{L}} \) of the constrained convex optimization problem \( F^{\gamma}(x) \to \min_{x \in S} \).

**Corollary A.2.** Let a function \( H^{\gamma} \) be defined in \([13]\) and let \( x^*_{\gamma} = \arg \min_{x \in \mathcal{L}} H^{\gamma}(x) \). Then applying Algorithm 1 for
\[
\nabla (h_i^{\gamma})^*(z) = \frac{1}{m} (f_i^\gamma)^*(z) + \frac{r}{2m} z
\]
we get by Theorem 2.1
\[
\|x^*_{\gamma} - x^k_{\gamma}\|_2^2 \leq C \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1 + r\gamma}} \right)^k
\]
where \( x^k_{\gamma} = \nabla (H^{\gamma})^*(z^k_{\gamma}) \) and
\[
C = C_1 = \frac{(1 + r\gamma)^2}{2r^2}. \]

Moreover, since \( x^*_{\gamma} \in \mathcal{L}, \) i.e. \( [x^*_{\gamma}]_i = [x^*_{\gamma}]_j \) for all \( i \) and \( j \), the consensus condition is approximated as follows
\[
\| [x^*_{\gamma}]_i - [x^*_{\gamma}]_j \|_2^2 \leq 2C \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1 + r\gamma}} \right)^k.
\]
A.1.1 Value bounds on $H^{r, \gamma}$

Despite we defined $h^{r, \gamma}_i$ for all $\mathbb{R}^d$, some properties hold true on the initial set $S$ only.

**Lemma A.3.** Let functions $h^{r, \gamma}_i$ be defined in [13]. If $x \in S$, then for any $r > 0$, for each $i = 1, \ldots, m$ we have

$$f_i^r(x) - \frac{r}{2(1 + r\gamma)} \|\nabla f_i^r(x)\|^2 \leq h^{r, \gamma}_i(x) \leq f_i^r(x). \quad (17)$$

**Proof.** The second inequality directly follows from the definition (13). To prove the first one we recall that $f_i^r$ is $\gamma$ strongly convex and the following holds:

$$h^{r, \gamma}_i(x) = \inf_{y \in S} \{ f_i^r(y) + (2r)^{-1}\|x - y\|_2^2 \}$$

$$= \inf_{y : (x - y) \in S} \{ f_i^r(x - y) + (2r)^{-1}\|y\|_2^2 \}$$

$$\geq \inf_{y : (x - y) \in S} \{ f_i^r(x) + (\nabla f_i^r(x), y) + \gamma/2\|y\|_2^2 + (2r)^{-1}\|y\|_2^2 \}$$

$$\geq \inf_{y \in \mathbb{R}^d} \{ f_i^r(x) + (\nabla f_i^r(x), y) + \gamma/2\|y\|_2^2 + (2r)^{-1}\|y\|_2^2 \},$$

which reaches its minimum at $y = \frac{-r}{1 + r\gamma} \nabla f_i^r(x)$ and so equals to

$$f_i^r(x) + \frac{r}{1 + r\gamma} (\nabla f_i^r(x), \nabla f_i^r(x)) + \frac{r}{2(1 + r\gamma)} \|\nabla f_i^r(x)\|^2$$

$$= f_i^r(x) - \frac{r}{2(1 + r\gamma)} \|\nabla f_i^r(x)\|^2.$$

\[ \square \]

A.2 Convergence in argument

Lemma A.4 shows convergence in argument in the following sense: if the regularization parameter $r$ tends to zero, the argminimun $x^\ast_{r, \gamma} \in \mathcal{L}$ of $H^{r, \gamma}$ tends to the argminimun $x^\ast_{\gamma} \in S$ of $F^\gamma$. By Corollary A.2 we have $x^\ast_{r, \gamma} \in \mathcal{L}$ approximated by $x^k_{r, \gamma} \in (\mathbb{R}^d)^m$ for a sufficient number of iterations $k$.

**Lemma A.4.** Let $x^\ast_{r, \gamma} = \arg\min_{x \in \mathcal{L}} H^{r, \gamma}(x)$ for $H^{r, \gamma}$ defined in [13]. Let

$$\|\nabla f_i^r(x)\|_2 \leq K_\gamma \quad \forall x \in \{ y \in S \mid \|y - x^\ast_{\gamma}\|_2 \leq \zeta \}. \quad (18)$$

If $r$ is such that $\|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2 \leq \zeta$, then

$$\|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2 \leq \sqrt{\frac{\gamma m}{2\gamma}} K_\gamma. \quad (19)$$

**Proof.** Using (17) and strong convexity of $F^\gamma$ and $H^{r, \gamma}$ we have

$$F^\gamma(x^\ast_{\gamma}) \geq H^{r, \gamma}(x^\ast_{\gamma}) \geq H^{r, \gamma}(x^\ast_{r, \gamma}) + \frac{\gamma}{2(1 + r\gamma)} \|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2$$

$$\geq F^\gamma(x^\ast_{r, \gamma}) - \frac{r}{2(1 + r\gamma)} \|\nabla F^\gamma(x^\ast_{r, \gamma})\|_2^2 + \frac{\gamma}{2(1 + r\gamma)} \|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2$$

$$\geq F^\gamma(x^\ast_{r, \gamma}) - \frac{r}{2(1 + r\gamma)} mK_\gamma^2 + \frac{\gamma}{2(1 + r\gamma)} \|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2$$

$$\geq F^\gamma(x^\ast_{\gamma}) + \gamma/2\|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2 - \frac{r}{2(1 + r\gamma)} mK_\gamma^2 + \frac{\gamma}{2(1 + r\gamma)} \|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2$$

$$\geq F^\gamma(x^\ast_{\gamma}) + \frac{\gamma}{1 + r\gamma} \|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2 - \frac{r}{2(1 + r\gamma)} mK_\gamma^2.$$

Then $\frac{\gamma}{1 + r\gamma} \|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2 - \frac{r}{2(1 + r\gamma)} mK_\gamma^2 \leq 0$ and hence $\|x^\ast_{r, \gamma} - x^\ast_{\gamma}\|_2^2 \leq \frac{m}{\gamma} K_\gamma^2$. \[ \square \]

Combining Lemma A.4 with Corollary A.2 we get the following.
Remark 2. Let $\zeta > 0$ and $K_\zeta$ be such that
\[ \|\nabla f_i^r(x)\|_2 \leq K_\zeta \quad \forall x \in \{ y \in S \mid \|y - x_i^*\|_2 \leq \zeta \}. \]

If
\[ \sqrt{\frac{r \rho}{2 \gamma}} K_\zeta + \sqrt{C_1} \left(1 - \frac{\lambda_{\min}^+}{L_{\max}} \sqrt{\frac{r \gamma}{1 + r \gamma}} \right)^{k/2} \leq \zeta, \tag{20} \]
where $C_1 = \frac{(1+\tau^2)}{2} \gamma$, then both $x_{r,\gamma}^*$ and $x_{r,\gamma}^k$ belong to $\zeta$-neighborhood of $x_i^*$, i.e. $\|x_{r,\gamma}^* - x_i^*\|_2 \leq \zeta$ and $\|x_{r,\gamma}^k - x_i^*\|_2 \leq \zeta$.

A.3 Value approximation

Let $x_{r,\gamma}^* \in L$ be the only argminimum of $H^{r,\gamma}$ on the consensus space $L$, i.e.
\[ x_{r,\gamma}^* = \arg \min_{x \in S} H^{r,\gamma}(x). \tag{21} \]

In order to prove the value approximation (11) let us separate it into parts and estimate each of them:
\[
\begin{align*}
F^\gamma(x_{r,\gamma}^k) - F^\gamma(x_{r,\gamma}^*) & \leq F^\gamma(x_{r,\gamma}^k) - H^{r,\gamma}(x_{r,\gamma}^*) \tag{22a} \\
& \leq F^\gamma(x_{r,\gamma}^k) - H^{r,\gamma}(x_{r,\gamma}^*) + H^{r,\gamma}(x_{r,\gamma}^k) - H^{r,\gamma}(x_{r,\gamma}^*) \tag{22b} \\
& \quad + H^{r,\gamma}(x_{r,\gamma}^k) - F^\gamma(x_{r,\gamma}^*). \tag{22c}
\end{align*}
\]

The last addend is negative and can be eliminated:
\[ H^{r,\gamma}(x_{r,\gamma}^*) - F^\gamma(x_{r,\gamma}^*) \leq H^{r,\gamma}(x_{r,\gamma}^k) - F^\gamma(x_{r,\gamma}^*) \leq 0. \]

The rest are estimated in Lemmas [A.5] and [A.6] under additional assumptions.

Lemma A.5. Let $\|x_{r,\gamma}^* - x_i^*\|_2 \leq \zeta$ and let $\|x_{r,\gamma}^k - x_i^*\|_2 \leq \zeta$. If $\|\nabla f_i^r(x)\|_2 < K_\zeta$ for all $i$ and for all $x \in \{ y \in S \mid \|y - x_i^*\|_2 < \zeta \}$, then
\[ F^\gamma(x_{r,\gamma}^k) - H^{r,\gamma}(x_{r,\gamma}^k) \leq \frac{r}{2(1 + r \gamma)} m K_\zeta^2. \tag{23} \]

Proof. We cannot declare a uniform $K$ because $F^\gamma$ is not smooth. Nonetheless, assuming $x_{r,\gamma}^*$ and $x_{r,\gamma}^k$ belong to $\zeta$-neighborhood of $x_i^*$, we immediately obtain (23) from (17).

Lemma A.6. Let $\|\nabla f_i^r(x)\|_2 < K_\zeta$ for all $i$ and for all $x \in \{ y \in S \mid \|y - x_i^*\|_2 < \zeta \}$. It holds
\[
H^{r,\gamma}(x_{r,\gamma}^k) - H^{r,\gamma}(x_{r,\gamma}^*) \leq C_2 \left(1 - \frac{\lambda_{\min}^+}{L_{\max}} \sqrt{\frac{r \gamma}{1 + r \gamma}} \right)^{k/2}
\]
where
\[ C_2 = \frac{(1 + r \gamma) \sqrt{m K_\zeta}}{\sqrt{2 \gamma}} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + \frac{(1 + r \gamma)^2}{4r \gamma^2} \]

Proof. By $\frac{1}{2}$ smoothness of $H^{r,\gamma}$
\[
H^{r,\gamma}(x_{r,\gamma}^k) - H^{r,\gamma}(x_{r,\gamma}^*) \leq (\nabla H^{r,\gamma}(x_{r,\gamma}^*)^T)(z_{r,\gamma}^*) + \frac{1}{2}\|z_{r,\gamma}^* - x_{r,\gamma}^*\|^2_2
\]
where $z_{r,\gamma}^*$ is the limit of $z_{r,\gamma}^*$ and so it is the argminimum of $(H^{r,\gamma})^*$ on $L^*$. By (16) we have
\[
\frac{1}{2r} \|x_{r,\gamma}^k - x_{r,\gamma}^*\|^2_2 \leq \frac{1}{2r} C_1 \left(1 - \frac{\lambda_{\min}^+}{L_{\max}} \sqrt{\frac{r \gamma}{1 + r \gamma}} \right)^k \leq \frac{(1 + r \gamma)^2}{4r \gamma^2} \left(1 - \frac{\lambda_{\min}^+}{L_{\max}} \sqrt{\frac{r \gamma}{1 + r \gamma}} \right)^k
\]

14
Let us introduce an orthogonal projection matrix $P$ onto the subspace $L^\top$, i.e., it holds $Pv = \arg\min_{z \in L^\top} \{v - z\}$ for an arbitrary $v \in (\mathbb{R}^d)^n$. Then matrix $P$ is

$$P = \left( I_n - \frac{1}{n} 1_n 1_n^\top \right) \otimes I_d, \quad (24)$$

where $I_n$ denotes $n \times n$ identity matrix, $1_n = (1, \ldots, 1) \in \mathbb{R}^n$, $\otimes$ is a Kronecker product. Note that $P^\top P = P$.

Since $z_\infty^\infty \in L^\top$ and $x_\infty^\gamma \in L$, the first part simplifies to $(z_\infty^\infty P \nabla (H^\gamma)^\star (z_\infty^\gamma))$. We may use Lemma 2 in Kovalev et al. [2021b] to get the following estimation

$$\|P \nabla (H^\gamma)^\star (z_k^\gamma)\|^2_2 = \|\nabla (H^\gamma)^\star (z_k^\gamma)\|^2_2 \leq \frac{2}{\rho_{\min}^2} \left( (H^\gamma)^\star (z_k^\gamma) - (H^\gamma)^\star (z_{k+1}^\gamma) \right).$$

As $z_{k+1}^\gamma$ is a non-optimal point of Algorithm 1, this is not greater than

$$\frac{2}{\rho_{\min}^2} \left( (H^\gamma)^\star (z_k^\gamma) - (H^\gamma)^\star (z^\star) \right) \leq \frac{1+r\gamma}{\gamma^2 x_{\min}^2} \|z_k^\gamma - z^\star\|^2_2 \leq \frac{(1+r\gamma)^2}{\gamma^2} \frac{\lambda_{\max}^+}{\lambda_{\min}^-} \left( 1 - \frac{\lambda_{\min}^+}{\lambda_{\max}^-} \sqrt{1 + r\gamma} \right)$$

and the latter ones follow from the smoothness of $(H^\gamma)^\star$ and from the fact that the proof of Kovalev et al. [2021b] Theorem 1 actually covers the following chain of inequalities:

$$\|\nabla H^\star (z_k^\gamma) - x^\star\|^2_2 \leq \frac{1}{\mu^2} \|z_k^\gamma - z^\star\|^2_2 \leq C \left( 1 - \frac{\lambda_{\min}^+}{\lambda_{\max}^-} \sqrt{\frac{\mu}{L}} \right) \leq \frac{1}{2\mu^2} \left( 1 - \frac{\lambda_{\min}^+}{\lambda_{\max}^-} \sqrt{\frac{\mu}{L}} \right)^k.$$

By our assumption $\|z_\infty^\infty\|_2 = \|\nabla H^\gamma (x_\infty^\gamma)\|_2 < \sqrt{mK}\zeta$. Thus, we obtain

$$\leq \sqrt{mK}\zeta \sqrt{\frac{\mu}{2\gamma}} \sqrt{\frac{\lambda_{\min}^-}{\lambda_{\max}^+}} \left( 1 - \frac{\lambda_{\min}^+}{\lambda_{\max}^-} \sqrt{1 + r\gamma} \right)^{k/2} \leq \zeta,$$

$$\sqrt{\frac{r mK}{2\gamma}} + C_1 \left( 1 - \frac{\lambda_{\min}^-}{\lambda_{\max}^+} \sqrt{1 + r\gamma} \right)^{k/2} \leq \zeta,$$

$$C_2 \left( 1 - \frac{\lambda_{\min}^-}{\lambda_{\max}^+} \sqrt{1 + r\gamma} \right)^{k/2} \leq \zeta.$$

By Remark 2 and Lemmas A.5, A.6 we see that $F^\gamma (x^\gamma_{r,\gamma}) - F^\gamma (x^\gamma_{r,\gamma}) < \epsilon$ if

$$\forall i \\forall x \in \{y \in y - x^\gamma_{r,\gamma} \leq \zeta\} \|\nabla f_i^\gamma (x)\|_2 < K\zeta, \quad (25)$$

$$\sqrt{\frac{r mK}{2\gamma}} + C_1 \left( 1 - \frac{\lambda_{\min}^-}{\lambda_{\max}^+} \sqrt{1 + r\gamma} \right)^{k/2} \leq \zeta,$$

$$C_2 \left( 1 - \frac{\lambda_{\min}^-}{\lambda_{\max}^+} \sqrt{1 + r\gamma} \right)^{k/2} \leq \zeta.$$
To prove Theorem 3.3 we combine proved Theorem 3.1 with features of the entropy regularization of Wasserstein barycenter problem.

B.1 Entropy-regularized WB problem

Recall that for a fixed cost matrix $M$ we define the set of transport plans as

$$U(p, q) := \{X \in \mathbb{R}_{+}^{d \times d} | X1 = p, XT1 = q\}$$

and Wasserstein distance between two probability distributions $p$ and $q$ as

$$W(p, q) := \min_{X \in U(p, q)} \langle M, X \rangle.$$ 

The entropy-regularized (or smoothed) Wasserstein distance is defined as

$$W_{\gamma}(p, q) := \min_{X \in U(p, q)} \{\langle M, X \rangle - \gamma E(X)\},$$

(29)

where $\gamma > 0$ and

$$E(X) := -\sum_{i=1}^{d} \sum_{j=1}^{d} e(X_{ij}),$$

where $e(x) = \begin{cases} x \ln x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}.$

(30)

Then $W(p, q) = W_0(p, q)$.

Then the convex optimization problem (1) can be relaxed to the following $\gamma$ strongly convex optimization problem

$$\frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(\hat{p}_i) \rightarrow \min_{p \in S_1(d)} \frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(p),$$

(31)

where $W_{\gamma, q_i}(p) = W_{\gamma}(q_i, p)$. Solution of (31) is called the uniform Wasserstein barycenter of the family of $q_i \in S_1(d)$. Moreover, problem (31) admits a unique solution and approximates unregularized WB problem as follows.

**Remark 3.** Let $\gamma \leq \frac{\varepsilon}{4} \ln d$. If vectors $\hat{p}_i \in S_1(d)$ are such that

$$\frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(\hat{p}_i) - \min_{p \in S_1(d)} \frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(p) \leq \frac{\varepsilon}{2},$$

then

$$\frac{1}{m} \sum_{i=1}^{m} W_{q_i}(\hat{p}_i) - \min_{p \in S_1(d)} \frac{1}{m} \sum_{i=1}^{m} W_{q_i}(p) \leq \varepsilon.$$ 

Indeed, as entropy is bounded we have $W_{q_i}(p) \leq W_{\gamma, q_i}(p) \leq W_{q_i}(p) + 2\gamma \ln d$ for all $i$ and $p$. Then, for $p^* = \arg \min_{p \in S_1(d)} \frac{1}{m} \sum_{i=1}^{m} W_{q_i}(p)$ and $p^*_i = \arg \min_{p \in S_1(d)} \frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(p)$ it holds that

$$\frac{1}{m} \sum_{i=1}^{m} W_{q_i}(\hat{p}_i) - \frac{1}{m} \sum_{i=1}^{m} W_{q_i}(p^*)$$

$$\leq \frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(\hat{p}_i) - \frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(p^*) + 2\gamma \ln d$$

$$\leq \frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(\hat{p}_i) - \frac{1}{m} \sum_{i=1}^{m} W_{\gamma, q_i}(p^*) + \frac{\varepsilon}{2} \leq \varepsilon.$$ 


B.2 Legendre transforms

One particular advantage of entropy regularization of the Wasserstein distance is that it yields closed-form representations for the dual function $W^*_{\gamma,q}(\cdot)$ and for its gradient. Recall that the Fenchel-Legendre transform of (29) is defined as

$$W^*_{\gamma,q}(z) := \max_{p \in S_1(d)} \{\langle z, p \rangle - W_{\gamma,q}(p)\}.$$  \hfill (32)

**Theorem B.1** ([Cuturi and Peyré, 2015] [Theorem 2.4]). For $\gamma > 0$, the Fenchel-Legendre dual function $W^*_{\gamma,q}(z)$ is differentiable

$$W^*_{\gamma,q}(z) = \gamma \langle E(q) + \langle q, \ln K\alpha \rangle, x \rangle + \gamma \sum_{j=1}^m [q_j \ln \left( \sum_{i=1}^m \exp \left( \frac{1}{\gamma} \left( |z_i| - M_{ji} \right) \right) \right)$$

and its gradient $\nabla W^*_{\gamma,q}(z)$ is $1/\gamma$-Lipschitz in the 2-norm with

$$\nabla W^*_{\gamma,q}(z)_l = \sum_{j=1}^m q_j \frac{\exp \left( \frac{1}{\gamma} \left( |z_i| - M_{ji} \right) \right)}{\sum_{i=1}^m \exp \left( \frac{1}{\gamma} \left( |z_i| - M_{ij} \right) \right)}.$$ \hfill (34)

where $z \in \mathbb{R}^n$ and for brevity we denote $\alpha = \exp(z/\gamma)$ and $K = \exp(-M/\gamma)$.

Notice that to get back and obtain the approximated barycenter we can employ the following result (with $\lambda_i = 1$).

**Theorem B.2** ([Cuturi and Peyré, 2015] [Theorem 3.1]). The barycenter $p^*$ solving (31) satisfies

$$\forall i = 1, \ldots, m \quad p^*_i = \nabla W^*_{\gamma,q_i}(z^*_i),$$

where the set of $z^*_i$ constitutes any solution to any smoothed dual WB problem:

$$\min_{z_1, \ldots, z_m \in \mathbb{R}^d} \sum_{i=1}^m \lambda_i W^*_{\gamma,q_i}(z_i) \text{ s.t. } \sum_{i=1}^m \lambda_i z_i = 0.$$  \hfill (31)

Thus we can apply Theorem B.1 for the problem (31) with explicitly defined $\nabla W^*_{\gamma,q_i}$ and obtain $x^{n}_{\gamma,\gamma}$ that satisfies

$$\frac{1}{m} \sum_{i=1}^m W_{\gamma,q_i}(|x^n_{\gamma,\gamma}|_i) - \min_{p \in S_1(d)} \frac{1}{m} \sum_{i=1}^m W_{\gamma,q_i}(p) \leq \frac{r}{4(1 + r\gamma)} m K^2 + \frac{1}{2} C_2 \left( 1 - \frac{\lambda^+_{\min}}{T \lambda^+_{\max}} \sqrt{\frac{r\gamma}{1 + r\gamma}} \right)^{n/2} \leq \varepsilon/2.$$  \hfill (33)

By Remark 3 it proves

$$\left| \sum_{i=1}^m W_{\gamma_i}(|x^n_{\gamma_i}|_i) - \sum_{i=1}^m W_{\gamma_i}([p^*]_i) \right| \leq 2\gamma \ln d + \frac{r}{4(1 + r\gamma)} m K^2 + C \left( 1 - \frac{\lambda^+_{\min}}{T \lambda^+_{\max}} \sqrt{\frac{r\gamma}{1 + r\gamma}} \right)^{n/2} \leq \varepsilon,$$

for $C = \frac{1}{2} C_2 = \frac{(1 + r\gamma) m K \zeta}{2\sqrt{2\gamma}} \sqrt{\lambda^+_{\min} + \frac{(1 + r\gamma)^2}{8r\gamma^2}}$.

B.3 Parameter estimation

It remains to assign $\zeta > 0$ and $K = K_\zeta$, satisfying (35). Due to Assumption 2 such $\zeta$ and $K$ exist.

**Proposition B.3.** Let a set $\{q_i\}_{i=1}^m$ satisfies Assumption 2, let $p^*_i$ be the uniform Wasserstein barycenter of $\{q_i\}_{i=1}^m$, and let $\zeta \in (0, \min \{ \frac{1}{2}, \min_i \{||q_i||_1\} \}]$. For each $i = 1, \ldots, m$ the norm of the gradient $\|\nabla W_{\gamma,q_i}(\cdot)\|_2$ is uniformly bounded over $\{p \in S_1(d) \mid \|p - p^*_i\|_2 \leq \zeta\}$; and the bound $K_\rho$ is given in (35) for $\rho \leq \min \{ \frac{1}{2}, \min_i \{||q_i||_1\} \} - \zeta$.  \hfill (32)
We obtain Proposition B.3 as a combination of Lemma B.4 from [Bigot et al., 2019] and proved below Lemma B.5.

**Lemma B.4 ([Bigot et al., 2019], Lemma 3.5).** For any \( \rho \in (0, 1), q \in S_1(d), \) and \( p \in \{ x \in S_1(d) \mid \min_i x_i \geq \rho \} \) there is a bound: \( \| \nabla W_{\gamma,q}(p) \|_2^2 \leq K_\rho, \) where

\[
K_\rho = \sum_{j=1}^d \left( 2\gamma \ln d + \inf_i \sup_{l} |M_{jl} - M_{il}| - \gamma \ln \rho \right)^2. \tag{35}
\]

**Lemma B.5.** Let a set \( \{ q_i \}_{i=1}^m \) satisfies Assumption 2, let \( p_i^* \) be the uniform Wasserstein barycenter of \( \{ q_i \}_{i=1}^m. \) All components \( k \) of \( p_i^* \) have a uniform positive lower bound: \( [p_i^*]_k \geq \min\{ \frac{1}{\varepsilon}, \min_{i,l}[q_i] \}. \)

**Proof.** Let \( X_i^* \) denote the optimal transport plan between \( p_i^* \) and \( q_i. \) Assume the contrary: there is \( k \) such that \( [p_i^*]_k < \min\{ \frac{1}{\varepsilon}, \min_{i,l}[q_i] \}. \) Then there is another component \( n \) such that \( [p_i^*]_n > \min_{i,l}[q_i] \). Consider the vector \( p \) that consists of \( [p_i] = [p_i^*] \) except for the components \( [p]_n = [p_i^*]_n + \delta \) and \( [p]_l = [p_i^*]_l - \delta, \) where \( \delta > 0 \) is less than \( \min_{i,a \neq b} [X^*]_{a,b} \) of the optimal transport plan \( X_i^* \) between \( p_i^* \) and \( q_a. \). Because of the entropy, all these optimal transport plans contain only positive non-diagonal elements, so such a \( \delta \) exists.

Construct now non-optimal transport plans between \( p \) and each of \( q_i \) in order to get the contradiction with the assumption. Initially we have \( W_{\gamma,q_i}(p_i^*) = \langle C, X_i^* \rangle - \gamma X_i^* \ln X_i^* \). Consider the matrix \( X_i \) that differs from \( X_i^* \) only at the four elements:

\[
[X_i]_{kk} = [X_i^*]_{kk} + \frac{1}{2}\delta, \quad [X_i]_{kn} = [X_i^*]_{kn} + \frac{1}{2}\delta, \\
[X_i]_{nn} = [X_i^*]_{nn} + \frac{1}{2}\delta, \quad [X_i]_{nk} = [X_i^*]_{nk} + \frac{1}{2}\delta.
\]

Then \( X_i \) is a transport plan between \( p \) and \( q_i \) since its elements are positive and also \( X_i 1 = p \) and \( X_i^* = q_i. \) Using the monotonicity of entropy on the interval \((0, \frac{1}{\varepsilon})\) and the assumption that diagonal elements of the cost matrix \( C \) are zero, we get for each \( i: \)

\[
W_{\gamma,q_i}(p) \leq \langle C, X_i \rangle - \gamma X_i \ln X_i \\
= \langle C, X_i^+ \rangle - \gamma X_i^+ \ln X_i^+ + \frac{1}{2}\delta C_{kn} - \frac{1}{2}\delta C_{nk} \\
+ \langle [X_i]_{kk} \ln [X_i^*]_{kk} - [X_i^*]_{kk} \ln [X_i^+]_{kk} \rangle \\
+ \langle [X_i]_{kn} \ln [X_i^*]_{kn} - [X_i^*]_{kn} \ln [X_i^+]_{kn} \rangle \\
+ \langle [X_i]_{nn} \ln [X_i^*]_{nn} - [X_i^*]_{nn} \ln [X_i^+]_{nn} \rangle \\
< \langle C, X_i^+ \rangle - \gamma X_i^+ \ln X_i^+ + \frac{1}{2}\delta C_{kn} - \frac{1}{2}\delta C_{nk} \\
= \langle C, X_i^* \rangle - \gamma X_i^* \ln X_i^* = W_{\gamma,q_i}(p_i^*).
\]

The obtained inequalities \( W_{\gamma,q_i}(p) < W_{\gamma,q_i}(p_i^*) \) contradict to the fact that \( p_i^* \) is the barycenter; this proves the lemma. \( \Box \)