

# Decentralized Convex Optimization on Time-Varying Networks with Application to Wasserstein Barycenters

Olga Yufereva<sup>1</sup>, Michael Persiianov<sup>2</sup>, Pavel Dvurechensky<sup>3</sup>,  
Alexander Gasnikov<sup>2,4,5</sup>, Dmitry Kovalev<sup>6</sup>

<sup>1</sup>N. N. Krasovski Institute of Mathematics and Mechanics,  
Yekaterinburg, Russia.

<sup>2</sup>Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

<sup>3</sup>Weierstrass Institute for Applied Analysis and Stochastics, Berlin,  
Germany.

<sup>4</sup>Skoltech, Moscow, Russia.

<sup>5</sup>Institute of information transmission problems, Moscow, Russia.

<sup>6</sup>Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium.

Contributing authors: [olga.o.yufereva@gmail.com](mailto:olga.o.yufereva@gmail.com);  
[persiianov.mi@phystech.edu](mailto:persiianov.mi@phystech.edu); [pavel.dvurechensky@wias-berlin.de](mailto:pavel.dvurechensky@wias-berlin.de);  
[gasnikov@yandex.ru](mailto:gasnikov@yandex.ru); [dakovalev1@gmail.com](mailto:dakovalev1@gmail.com);

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 embed this problem into a class of non-smooth dual-friendly distributed optimization problems over time-varying networks, and develop a first-order method for this class. 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. In the experiments, we demonstrate the efficiency of the proposed algorithm when applied to the Wasserstein barycenter problem.

# 1 Introduction

Optimal transport (OT) [1, 2] is getting more and more attention from the machine learning and optimization community motivated by a long list of applications such as unsupervised learning, semi-supervised learning, clustering, text classification, image retrieval, and others, see [3] and references therein. Given a basis space (e.g., pixel grid) and a transportation cost function (e.g., squared Euclidean distance), the OT approach defines a distance between two objects (e.g., images), modeled as two probability measures on the basis space, as the minimal cost of transportation of the first measure to the second. Such distances, in particular, the Wasserstein distance, naturally capture the geometry of the data since they are invariant to shifts and rotations. In particular, the Fréché mean with respect to the Wasserstein distance, called Wasserstein barycenter (WB) [4], allows [5, 6] to reconstruct a template image from its random observations obtained by random shifts and rotations.

The benefits of the use of WB in real-world applications are sometimes outweighed by the large computational burden introduced by their definition. The computation of the Wasserstein distance is already a large-scale optimization problem, and to calculate the WB, one introduces a second optimization level as the WB minimizes the sum of Wasserstein distances to a set of probability measures. At this point, distributed optimization algorithms turned out to be efficient to scale-up the computations of the WB when the data is distributedly stored by a computational network [7–14]. On the other hand, the nature of the data-generating process may be distributed itself. In particular, it may be impossible to collect the data even in one datacenter, especially if the data processing has to respect the privacy of the individual data. Another example is a network of sensors that measure signals following some distributions and for the analysis purposes the whole network needs to find the WB of these distributions by peer-to-peer communications. In this case, decentralized algorithms are especially useful. Moreover, algorithms adapted to time-varying networks of sensors or computing devices are required. For example, some nodes of the network may be disconnected due to failures or, e.g., when the node is a satellite that observes certain area, the observation is available only for a certain period of time. At the same time, the development of decentralized distributed algorithms for general optimization problems is important on its own since the WB problem represents only one, yet important, example where such algorithms are efficient.

Summarizing, the WB problem is important for applications, yet requires to solve a large-scale optimization problem. For that problem, and other large-scale problems, decentralized distributed optimization algorithms on time-varying networks has to be developed. In this paper, we develop a general decentralized accelerated gradient method on time-varying networks and apply it to the WB problem.

## 1.1 Related work

In general, decentralized distributed optimization is an emerging and actively developed branch of optimization, see the recent survey [15]. Our main focus in this paper is on the setting of decentralized methods working on time-varying networks, for which in the smooth and strongly convex case efficient algorithms were recently proposed in

[16–18] (primal oracle) and [19] (dual oracle). Unfortunately, these methods are not directly applicable in our case since the WB problem is not smooth, not strongly convex, and has simple constraints. At the same time, the WB problem has tractable dual oracle, which motivated us to extend the ADOM algorithm of [19] to the non-smooth, non-strongly convex setting with simple constraints.

Starting with [8, 9] it was observed that decentralized methods with dual oracle are well suited for the WB problem. In the cycle of subsequent papers [9–13] different decentralized accelerated (randomized) algorithms were proposed for dual WB problem. In [20] the authors propose to reformulate the WB problem as a bilinear saddle-point problem (SPP). Decentralized algorithm for this problem was proposed in [14]. Unfortunately, all these algorithms are designed for static networks that do not change over time.

Moreover, their extensions to time-varying networks seem to be hardly possible. At the core of these algorithms lies the reformulation of the WB problem as a problem with linear constraints ensuring the consensus between the network nodes, and then solving the dual for that problem. If communication network changes over time, then the affine-consensus constraints also change over time, and so does the dual problem. This essentially requires to solve a family of dual problems, which is not possible by the accelerated gradient methods or the Mirror-Prox algorithm as in [9–14].

The recent work [21] considers the WB problem on time-varying networks and analyses a simple consensus method. The main difference with our paper is that they prove asymptotic convergence rather than convergence rates, and they consider possibly continuous measures on  $\mathbb{R}$  unlike our setting of discrete measures on  $\mathbb{R}^n$ . The paper [22] proposes Fenchel dual gradient methods for distributed convex optimization over time-varying networks. The main difference is that we propose an accelerated algorithm with better complexity, yet under a stronger assumption on the network (they assume the  $B$ -connectivity of the network).

## 1.2 Our contributions

Since the WB problem has an efficient dual oracle, a natural idea is to use ADOM [19] that is an optimal decentralized algorithm for smooth strongly convex unconstrained problems for time-varying networks with dual oracle. ADOM can be considered as projected accelerated algorithm with inexact consensus-based projection applied to specific dual reformulation of the distributed optimization problem. Since the WB problem

- (Smoothness) is not smooth;
- (Constraints) is not unconstrained; as the space of probability measures has simple constraints;
- (Strongly convex) is not strongly convex;

the direct application of ADOM to the WB problem is not possible. Moreover, not only WB problem, but also general non-smooth, non-strongly-convex constrained optimization problems lack efficient algorithms on time-varying networks.

The first main result of this paper is a generalization of ADOM for general  $\gamma$  strongly convex decentralized optimization problems with simple constraints on time-varying networks; solving it numerically requires  $\mathcal{O}\left(\frac{\lambda_{\max}}{\lambda_{\min}^+} \frac{1}{\sqrt{\gamma\varepsilon}} \ln \frac{1}{\gamma\varepsilon}\right)$  iterations. The second contribution is the application to the WB problem on time-varying networks; where the corresponding iteration number becomes  $\mathcal{O}\left(\frac{1}{\varepsilon} \ln \frac{1}{\varepsilon}\right)$ . The main ideas are the following. To obtain the strong convexity we use the regularization of the primal problem by the entropy [3]. To deal with the constraints and non-smoothness, we use special regularization of the dual problem. This regularization goes back to [23, 24] and by infimal convolution can be considered as the Moreau–Yosida smoothing of the primal problem [25, 26]. We emphasize that the proposed dual regularization (primal smoothing) was earlier investigated only for non time-varying networks [27]. For time-varying networks and problems with simple constraints the analysis is different.

### ***Paper organization***

Section 2 presents preliminaries and basic definitions for a general distributed optimization problem. Section 3 states our main results, i.e. the method, convergence rates and the parameter estimation. Section 4 introduces the Wasserstein barycenter problem and specifies main results for the particular case. Section 5 shows numerical experiments that illustrate and verify the theoretical results. All the proofs are in Appendices.

### ***Notation***

We utilize the following dimensions:

- $m$  for the number of individual devices (nodes),
- $d$  for the dimension of a data in each device.

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  $\mathbf{x} \in (\mathbb{R}^d)^m$  is denoted by  $[\mathbf{x}]_l$  which is the corresponding vector from  $\mathbb{R}^d$ .

Let  $\mathbf{1}$  denote a column vector with all entries equal to 1. The  $d$ -dimensional simplex is denoted  $S_1(d)$ , that means  $S_1(d) := \{p \in [0, 1]^d \mid p^\top \mathbf{1} = 1\}$ . 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  $\langle M, X \rangle := \sum_{i=1}^d \sum_{j=1}^d M_{ij} X_{ij}$ .

Abbreviation WB means Wasserstein barycenter and ADOM refers to Accelerated Decentralized Optimization Method proposed in [19].

## **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. For a convex closed set  $S$  and convex functions  $f_i$  *decentralized computation of the following*

optimization problem

$$\min_{x \in \mathcal{S}} \sum_{i=1}^m f_i(x) \quad (1)$$

requires numerical computation assuming that each function  $f_i$  is stored on the corresponding node  $i \in [m] := \{1, 2, \dots, m\}$ . Such approach brings us to an effective reformulation of the optimization problem.

### 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 (1) 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:

$$\min_{\mathbf{x} \in \mathcal{S}} F(\mathbf{x}) = \min_{\mathbf{x} \in \mathcal{S}} \sum_{i=1}^m f_i([\mathbf{x}]_i), \quad (2)$$

where  $\mathcal{S} = \{\mathbf{x} = ([\mathbf{x}]_1, \dots, [\mathbf{x}]_m) \in (S)^m \mid [\mathbf{x}]_1 = \dots = [\mathbf{x}]_m\}$ .

Here  $i$ -th component  $[\mathbf{x}]_i$  is a corresponding  $d$ -dimensional vector.

### 2.1.2 Time-varying communication network

We consider  $m$  distributed devices that seek to reach a consensus solution of an optimization problem. The devices are connected via an  $m$ -node network that changes over time. At each time step  $n$  we denote Laplacian of the corresponding network by  $\hat{W}_n$ . In general, it suffices to take any  $\hat{W}_n$  satisfying the following:

1.  $\hat{W}_n$  is symmetric and positive semi-definite,
2.  $[\hat{W}_n]_{i,j} \neq 0$  if and only if  $(i, j)$  are connected by the network,
3.  $\ker \hat{W}_n = \{(x_1, \dots, x_m) \mid x_1 = \dots = x_m\}$ .

Further we use *communication matrix* that is the block matrix  $\mathbf{W}_n = \hat{W}_n \otimes I_d$ . Hence, decentralized communication of each vector  $x_i$  stored on the  $i$ -th node at a time step  $n$  can be represented by multiplication of the  $md$ -dimensional vector  $(x_1, \dots, x_m)$  and matrix  $\mathbf{W}_n$ : indeed, if  $\mathbf{y} = \mathbf{W}_n \mathbf{x}$ , then it yields

$$[\mathbf{y}]_i = \sum_{j=1}^m [\hat{W}_n]_{i,j} [\mathbf{x}]_j = \sum_{j \in \mathcal{N}_i} [\hat{W}_n]_{i,j} [\mathbf{x}]_j,$$

where  $\mathcal{N}_i$  is the set of the neighboring nodes for the node  $i$  according to the communication network at  $n$ -th iteration. Thus, for each node  $i$ , vector  $[\mathbf{y}]_i$  is a linear combination of vectors  $[\mathbf{x}]_j$ , stored at the neighboring nodes  $j \in \mathcal{N}_i$ .

The considered algorithms require these communication matrices  $\mathbf{W}_n$  to have conditional numbers bounded for all  $n \in \{0, 1, 2, \dots\}$ . Namely, we utilize the following assumption.

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

$$\lambda_{\min}^+ \leq \lambda_{\min}^+(\mathbf{W}_n) \leq \lambda_{\max}(\mathbf{W}_n) \leq \lambda_{\max} \quad \forall n,$$

where  $\lambda_{\min}^+(\mathbf{W}_n)$  is the smallest positive eigenvalue of  $\mathbf{W}_n$  and  $\lambda_{\max}(\mathbf{W}_n)$  is the biggest one.

A condition number of the matrix  $\mathbf{W}_n$  is given as  $\frac{\lambda_{\max}(\mathbf{W}_n)}{\lambda_{\min}^+(\mathbf{W}_n)}$  and relates to the connectivity of a network; it appears in convergence rates of many decentralized algorithms.

### 3 Main results

One of our main features is that general convex functions of optimization problem can be defined on a convex set  $S \subseteq \mathbb{R}^d$  instead of the entire  $\mathbb{R}^d$ , that was crucial for duality of (10) and (11). First of all, we study the case of strongly convex functions (Theorem 1). Then we see that similarly we can approximate convex functions (Corollary 1). We obtain an important Theorem 2 by applying Theorem 1 to the Wasserstein barycenter problem; in this particular setup one can estimate parameters more precisely. In the next sections we introduce necessary definitions, state Theorem 2, and provide related numerical experiments. All the proofs are in the Appendices.

Consider  $\gamma$  strongly convex case, i.e. the following decentralized optimization problem

$$\mathbf{x}_\gamma^* = \arg \min_{\mathbf{x} \in \mathcal{S}} F^\gamma(\mathbf{x}) = \arg \min_{\mathbf{x} \in \mathcal{S}} \sum_{i=1}^m f_i^\gamma([\mathbf{x}]_i) = \arg \min_{x \in S} \sum_{i=1}^m f_i^\gamma(x), \quad (3)$$

$$\text{where } \mathcal{S} = \{\mathbf{x} = ([\mathbf{x}]_1, \dots, [\mathbf{x}]_m) \in (S)^m \subset (\mathbb{R}^d)^m \mid [\mathbf{x}]_1 = \dots = [\mathbf{x}]_m\},$$

functions  $f_i^\gamma$  are  $\gamma$  strongly convex, differentiable, and defined on a convex set  $S$ . Recall that  $i$ -th component  $[\mathbf{x}]_i$  is a corresponding  $d$ -dimensional vector.

**Theorem 1.** Let  $S \subset \mathbb{R}^d$  be a convex set, let functions  $f_i^\gamma: S \rightarrow \mathbb{R}$  of the problem (3) be  $\gamma$  strongly convex and differentiable, let  $\mathbf{W}_n$  be the  $n$ -th communication matrix satisfying Assumption 1 for some  $\lambda_{\min}^+, \lambda_{\max} > 0$ . For any  $r > 0$ , after  $n$  iterations of Algorithm 1 we obtain  $\mathbf{x}_{r,\gamma}^n = \nabla H^*(\mathbf{z}_g^n)$  that provides:

1. consensus condition approximation: for each  $i$  and  $j$

$$\left\| [\mathbf{x}_{r,\gamma}^n]_i - [\mathbf{x}_{r,\gamma}^n]_j \right\|_2^2 \leq C_1 \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}} \right)^n; \quad (4)$$

2. value approximation:

$$F^\gamma(\mathbf{x}_{r,\gamma}^n) - \min_{\mathbf{x} \in \mathcal{S}} F^\gamma(\mathbf{x}) \leq \frac{r}{2(1+r\gamma)} mK^2 + C_2 \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}} \right)^{n/2}; \quad (5)$$

where  $K$  is such that  $\|\nabla f_i^\gamma(x)\|_2 < K$  for each  $i$  and for all  $x$  from  $\varepsilon/\gamma$ -neighborhood of the solution  $\arg \min_{x \in S} \sum_{i=1}^m f_i^\gamma(x)$ . The parameters are  $C_1 = \frac{(1+r\gamma)^2}{2\gamma^2}$ ,  $C_2 = \frac{m(1+r\gamma)K}{\sqrt{2\gamma}} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^+}} + \frac{m(1+r\gamma)^2}{4r\gamma^2}$ .

*Proof.* See Appendix B. □

**Remark 1.** To reach  $\varepsilon$  approximation of (4) and (5) it suffices to take  $r \leq \frac{\varepsilon}{2mK^2}$ . Then the rate of the number of iterations is

$$n = \mathcal{O} \left( \frac{\lambda_{\max}}{\lambda_{\min}^+} \sqrt{\frac{1+r\gamma}{r\gamma}} \ln \frac{C_2}{\varepsilon} \right) = \mathcal{O} \left( \frac{\lambda_{\max}}{\lambda_{\min}^+} \frac{1}{\sqrt{\gamma\varepsilon}} \ln \frac{1}{\varepsilon} \right).$$

*Proof.* See Appendix B.5. □

---

### Algorithm 1 Modified ADOM

---

- 1: **input:**  $r > 0$ , for  $i = 1, \dots, m$ :  $f_i^\gamma: S \rightarrow \mathbb{R}$ ,  $(f_i^\gamma)^*(z) = \sup\{\langle z, x \rangle - f_i^\gamma(x) \mid x \in S\}$
  - 2: define  $\nabla h_i^*([\mathbf{z}]_i) = \nabla (f_i^\gamma)^*([\mathbf{z}]_i) + r[\mathbf{z}]_i$
  - 3: define  $\nabla H^*(\mathbf{z}) = (\nabla h_1^*([\mathbf{z}]_1), \dots, \nabla h_m^*([\mathbf{z}]_m))^\top$
  - 4: set  $\alpha = \frac{r}{2}$ ,  $\eta = \frac{2\lambda_{\min}^+ \sqrt{\gamma}}{7\lambda_{\max} \sqrt{r(1+r\gamma)}}$ ,  $\theta = \frac{\gamma}{\lambda_{\max}(1+r\gamma)}$ ,  $\sigma = \frac{1}{\lambda_{\max}}$ ,  $\tau = \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}$
  - 5: set  $\mathbf{z}^0 = \mathbf{0}$ ,  $\mathbf{z}_f^0 = \mathbf{z}^0$ ,  $\mathbf{m}^0 = \mathbf{0}$
  - 6: **for**  $n = 0, 1, 2, \dots$  **do**
  - 7:  $\mathbf{z}_g^n = \tau \mathbf{z}^n + (1 - \tau) \mathbf{z}_f^n$
  - 8:  $\Delta^n = \sigma \mathbf{W}_n (\mathbf{m}^n - \eta \nabla H^*(\mathbf{z}_g^n))$
  - 9:  $\mathbf{m}^{n+1} = \mathbf{m}^n - \eta \nabla H^*(\mathbf{z}_g^n) - \Delta^n$
  - 10:  $\mathbf{z}^{n+1} = \mathbf{z}^n + \eta \alpha (\mathbf{z}_g^n - \mathbf{z}^n) + \Delta^n$
  - 11:  $\mathbf{z}_f^{n+1} = \mathbf{z}_g^n - \theta \mathbf{W}_n \nabla H^*(\mathbf{z}_g^n)$
  - 12: **end for**
  - 13: **output:**  $\mathbf{x}_{r,\gamma}^n = \nabla H^*(\mathbf{z}_g^n)$
- 

**Corollary 1.** Let  $S$  be a convex set in  $\mathbb{R}^d$ , let  $f_i: S \rightarrow \mathbb{R}$  be differentiable convex functions for  $i = 1, \dots, m$ , and let  $\mathbf{W}_n$  be the  $n$ -th communication matrix satisfying Assumption 1 for some  $\lambda_{\min}^+, \lambda_{\max} > 0$ . Decentralized convex optimization problem  $\min_{x \in S} \sum_{i=1}^m f_i(x)$  over time-varying communication networks can be  $\varepsilon$  approximated numerically by Algorithm 1 applied for  $\gamma$  strongly convex regularizing functions<sup>1</sup>  $f_i^\gamma(x)$  that satisfy

$$0 \leq \min_{x \in S} \sum_{i=1}^m f_i(x) - \min_{x \in S} \sum_{i=1}^m f_i^\gamma(x) \leq \varepsilon/2 \quad (6)$$

---

<sup>1</sup>e.g., one can take  $f_i^\gamma(x) = f_i(x) + \frac{\gamma}{2} \|x\|_2^2$

if  $r < \frac{\varepsilon}{4mK^2}$  where  $K$  is such that  $\|\nabla f_i^\gamma(x)\|_2 < K$  for each  $i$  and for all  $x$  from  $\varepsilon/\gamma$ -neighborhood of the solution  $\arg \min_{x \in S} \sum_{i=1}^m f_i^\gamma(x)$ . Moreover, if  $\gamma = \sqrt{\varepsilon}$ , then the rate of the number of iterations is

$$n = \mathcal{O} \left( \frac{\lambda_{\max}}{\lambda_{\min}^+} \frac{1}{\varepsilon} \ln \frac{1}{\varepsilon} \right).$$

*Proof.* The condition  $r < \frac{\varepsilon}{4mK^2}$  follows that the right-hand side of the inequality (5) is less or equal  $\varepsilon/2$ , i.e.

$$\sum_{i=1}^m f_i^\gamma([\mathbf{x}_{r,\gamma}^n]_i) - \min_{x \in S} \sum_{i=1}^m f_i^\gamma(x) \leq \varepsilon/2.$$

Combining it with (6) we obtain  $\varepsilon$  approximation of the convex constrained optimization problem  $\min_{x \in S} \sum_{i=1}^m f_i(x)$ .  $\square$

## 4 Wasserstein Barycenter Problem

Wasserstein barycenter problem is a motivating and challenging convex constrained optimization problem that is convex, but not smooth and not strongly-convex. The problem belongs to the optimal transport theory which recently got various application in, e.g., texture mixing [28], statistical estimation of template models [29], graphics and machine learning (for regression, classification and generative modeling) [3]. Further we consider WB problem as a decentralized convex optimization problem and propose a computation method in Theorem 2.

### 4.1 Wasserstein distance

We provide here only necessary definitions and take into consideration only finite-supported distributions since we deal with numerical experiments. General theory, that begins with Wasserstein distance, can be found in [30].

Recall that we denote  $d$ -dimensional simplex by  $S_1(d)$  and it represents a set of possible probabilities distributions as  $S_1(d) = \{p \in [0, 1]^d \mid \sum_{i=1}^d [p]_i = 1\}$ . Consider two probability distributions  $p, q \in S_1(d)$  with support on a finite set of points  $\{\omega_i \in \mathbb{R}^n\}_{i=1}^d$  such that  $p(\omega_i) = p_i$  and  $q(\omega_i) = q_i$ . Then, a cost (loss) matrix  $M$  is such that its element  $[M]_{ij} \in \mathbb{R}_+$  represents the cost of moving a unit mass from  $\omega_i$  to  $\omega_j$ . So  $M$  is a non-negative symmetric matrix with zeros on the diagonal. It is often taken as the Euclidean distances matrix, i.e.  $[M]_{ij} = \|\omega_i - \omega_j\|_2^2$ . The set of *transport plans* is defined as

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

i.e. the set of probabilities measures on  $\mathbb{R}_+^{d \times d}$  with margins  $p$  and  $q$ . *Wasserstein distance* between two probability distributions defines as the following minimum among



component-wise multiplication of the cost matrix and transport plans:

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

## 4.2 Wasserstein barycenter

Wasserstein barycenter of a set of probability distributions  $q_1, \dots, q_m$  is a probability distribution itself that is defined as the solution to the following optimization problem

$$\min_{p \in S_1(d)} \sum_{i=1}^m \mathcal{W}_{q_i}(p), \quad (7)$$

where  $\mathcal{W}_{q_i}(p) := \mathcal{W}(q_i, p)$ . In distributed approach, each device (node) possesses its original distribution  $q_i$  and the corresponding function  $f_i(\cdot) = \mathcal{W}_{q_i}(\cdot)$ . The goal of the whole system is, by communicating with each other, to approximate the barycenter by an iterative algorithm. At each iteration each node computes a new guess for the barycenter distribution using current guesses from its neighbors. Typically, the first guess coincide with the original distribution  $q_i$  and the resulting distributions reach a consensus. It is known (see e.g. [5]) that the WB captures the mean structure of given data. On example of a dataset of hand-written digits ‘4’ of MNIST 784 [31] Figure 1 shows how local nodes’ guesses change and tend to global barycenter, which resembles a digit ‘4’ as well. That visually illustrates our theoretical result, Theorem 2. In this experiment communication networks are Erdős–Rényi random networks and change every 5 iterations.

## 4.3 Resulting algorithm for WB problem

To apply the proposed numerical scheme we use entropy regularization of Wasserstein distance  $\mathcal{W}$  and make the following assumption on the initial data.

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

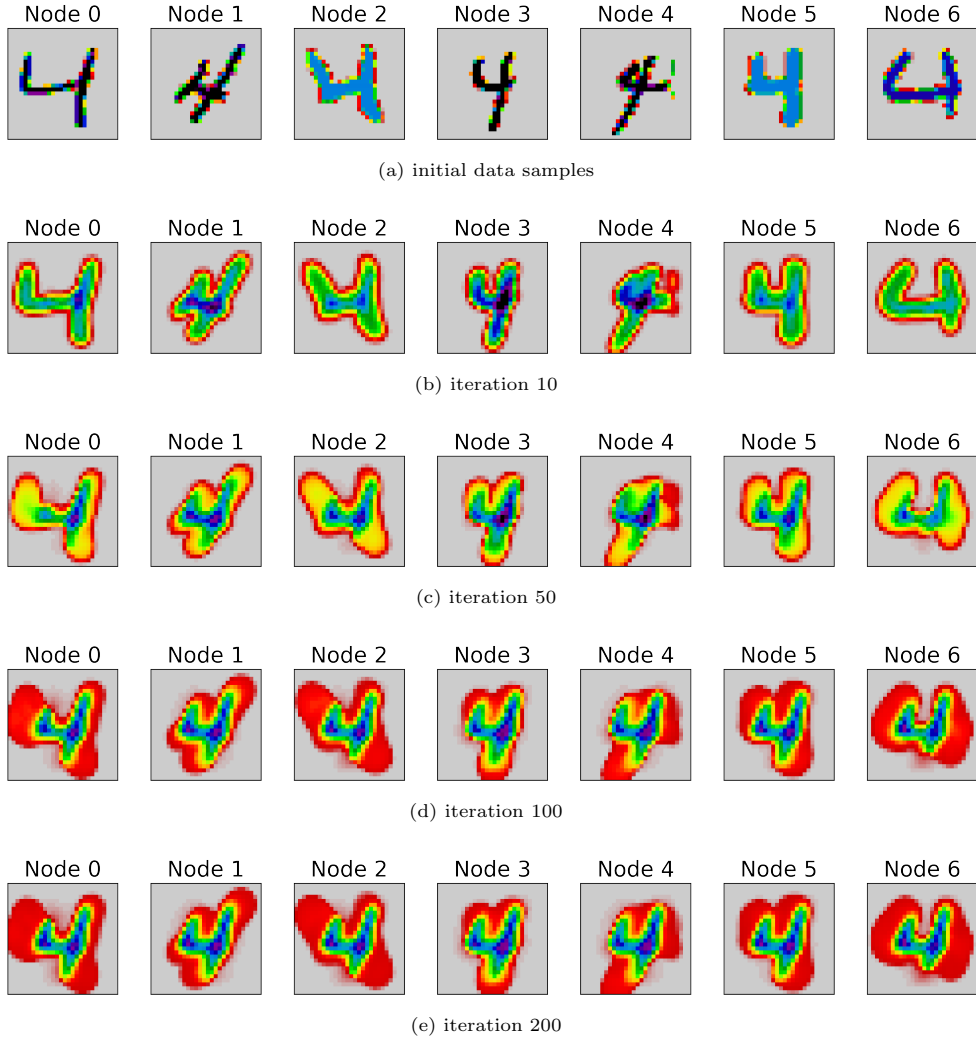
$$\min_{\substack{i=1, \dots, m \\ l=1, \dots, d}} [q_i]_l = \delta > 0.$$

It is not too restrictive as it only excludes zero probabilities of states, which can be done by a little distortion. The entropy regularized Wasserstein distance is defined as

$$\mathcal{W}_{\gamma, q}(p) := \mathcal{W}_{\gamma}(q, p) = \min_{X \in U(p, q)} \left\{ \langle M, X \rangle + \gamma \sum_{i=1}^d \sum_{j=1}^d X_{ij} \ln X_{ij} \right\}, \quad (8)$$

where  $x \ln x$  is assumed to equal zero if  $x = 0$ .

**Theorem 2.** *Let initial distributions  $q_i$  satisfy Assumption 2 and let  $p^*$  be their Wasserstein barycenter, i.e.  $p^*$  minimizes the problem (7). Let communication matrices  $\mathbf{W}_n$  satisfy Assumption 1 for some  $\lambda_{\min}^+, \lambda_{\max} > 0$ . If Algorithm 1 is applied for*



**Fig. 1:** Evolution of local data converging to Wasserstein barycenter of the handwritten 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  $\gamma = 0.03, r = 0.001$

entropy regularized Wasserstein distance functions  $f_i^\gamma = \mathcal{W}_{\gamma, q_i}$ , then functions  $\nabla h_i^*(z)$  are defined as

$$\nabla h_i^*(z) = \frac{r}{2}z + \sum_{j=1}^m [q]_j \frac{\exp(\frac{1}{\gamma}([z]_i - M_{ij}))}{\sum_{i=1}^m \exp(\frac{1}{\gamma}([z]_i - M_{ij}))}, \quad (9)$$

$$\nabla(H^{r, \gamma})^*(\mathbf{z}) = (\nabla h_1^*([z]_1), \dots, \nabla h_m^*([z]_m))^T,$$

and it suffices to take  $\gamma = \frac{1}{8}\varepsilon \ln d$ ,  $K^2 = \sum_{j=1}^d (2\gamma \ln d + \inf_i \sup_l |M_{jl} - M_{il}| - \gamma \ln \frac{\delta}{2})^2$ , and  $r = \frac{\varepsilon}{4mK^2}$  to  $\varepsilon$  approximate the solution  $p^*$  as follows

$$\begin{aligned} & \left| \sum_{i=1}^m \mathcal{W}_{q_i}([\mathbf{x}_{r,\gamma}^n]_i) - \sum_{i=1}^m \mathcal{W}_{q_i}(p^*) \right| \\ & \leq 2\gamma \ln d + \frac{r}{4(1+r\gamma)} mK^2 + C \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}} \right)^{n/2} \leq \varepsilon. \end{aligned}$$

Thus, a sufficient number of iterations of Algorithm 1 is

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

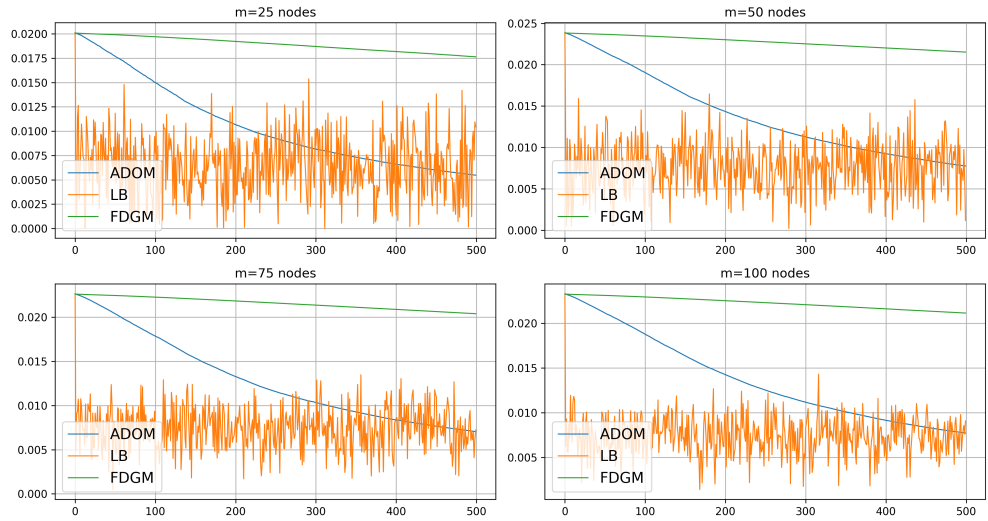
## 5 Numerical Experiments

We provide numerical experiments to demonstrate performance of the proposed method. We can fruitfully test our method on WB problem with an artificial set of univariate, discrete and truncated Gaussian distributions. For such a dataset, the resulting distribution (the zero-entropy Wasserstein barycenter) is described by an analytic formula. Namely, the barycenter is a Gaussian distribution which mean is the arithmetic average of the means of the given Gaussians and the standard deviation of the barycenter is the arithmetic average of the standard deviations of the given Gaussians.

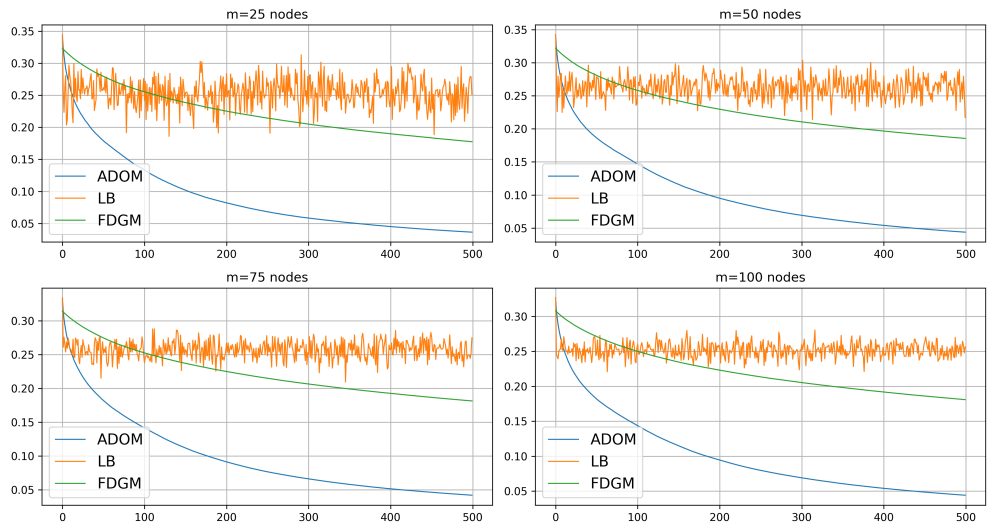
For all figures of this section we generated a dataset of truncated Gaussians. Each distribution size is 100, while a size of a dataset (i.e. the number of nodes) differs and is indicated at each figure. For entropy regularized Wasserstein distance we use normalized Euclidean cost matrix and entropy regularization parameter  $\gamma = 0.01$ . The regularization parameter  $r$  of the method is  $r = 0.001$ .

### 5.1 Comparison with other methods

To the best of our knowledge, we can compare our method (called here ADOM) with local barycenters method (LB) proposed in [21] and Fenchel dual gradient method (FDGM) proposed in [22]. They all are applicable for the WB problem on time-varying networks. Regardless of the analytical form of the LB algorithm, in order to implement it we need either to solve an optimization problem at each iteration or to use approximations, e.g. methods of [32]. Note that, for any particular setup, realization of FDGM is quite a problem, since the method is sensitive to the step size  $\alpha_n$ . ADOM negotiates limitations described above and reveals relatively stable convergence as it is presented at Figures 2–3, where we test the methods on cycle networks that change every iteration.



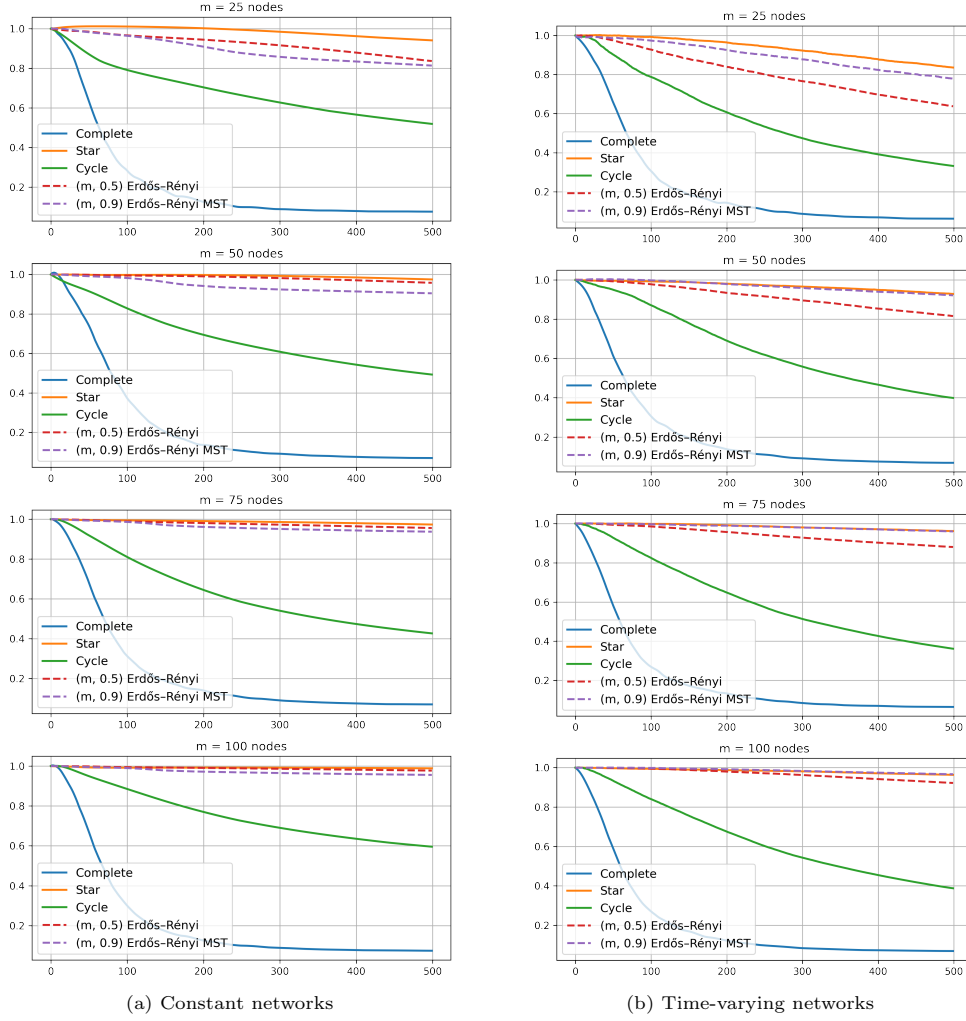
**Fig. 2:**  $\frac{1}{m} \left( \sum_{i=1}^m \mathcal{W}(q_i, [\mathbf{x}_{r,\gamma}^n]_i) - \sum_{i=1}^m \mathcal{W}(q_i, p^*) \right)$ -convergence comparison on cycle networks changing every iteration



**Fig. 3:** Consensus condition comparison of methods on cycle networks changing every iteration

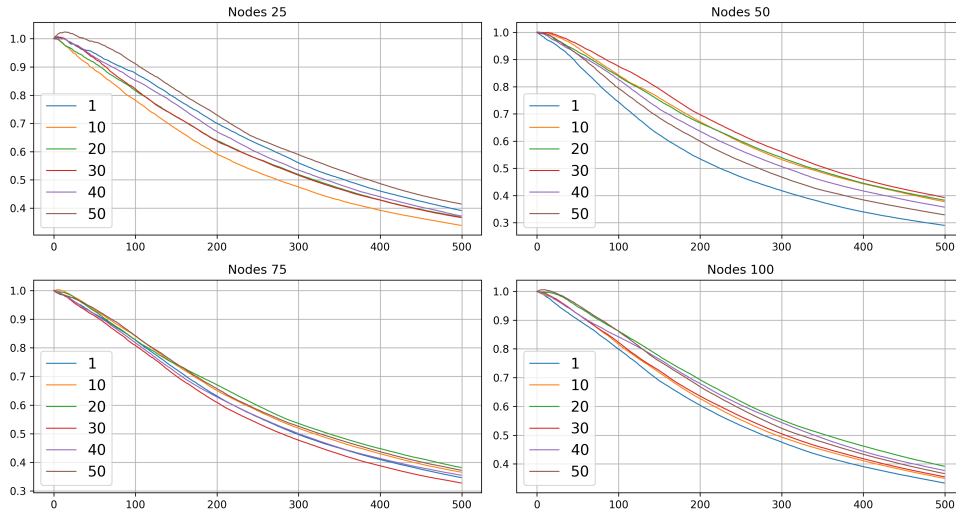
## 5.2 Performance rate

We can see the difference in error value, i.e. in  $\frac{1}{m} \left( \sum_{i=1}^m \mathcal{W}(q_i, [\mathbf{x}_{r,\gamma}^n]_i) - \sum_{i=1}^m \mathcal{W}(q_i, p^*) \right)$ , in two opposite cases: Figure 4a presents evolution of errors computed for constant networks of different topologies, while at Figure 4b networks change every iterations within indicated topology; the exception is complete network that cannot change. 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. 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.

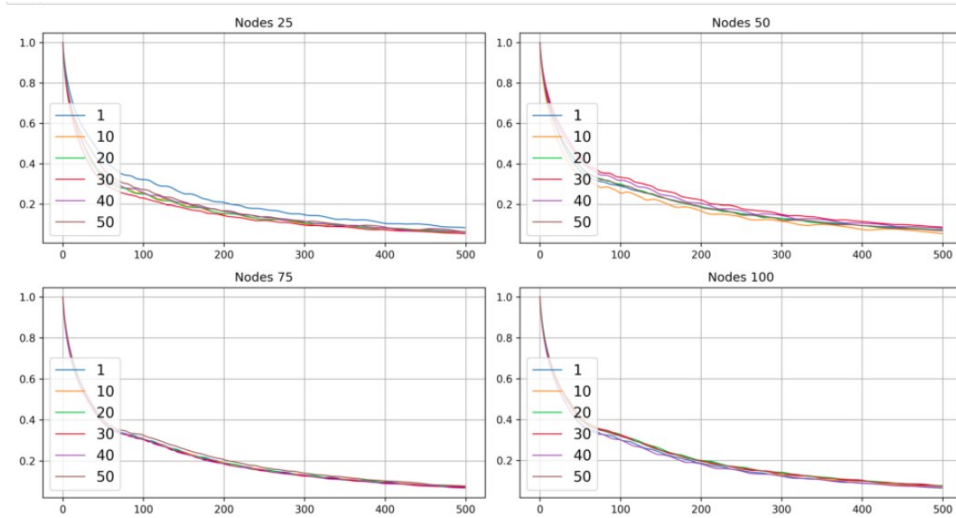


**Fig. 4:** Different network topologies: error over time

For the two ‘most efficient’ topologies that prove themselves at Figures 4a–4b we compute at Figure 5 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.



(a) Cycle networks



(b)  $(m, 0.9)$ -Erdős-Rényi networks

**Fig. 5:** Different number of iteration between network changing: errors over time

## Acknowledgments

The authors are grateful to Alexander Rogozin.

The work of A. Gasnikov in Section 4 of the paper was funded by Russian Science Foundation (project 18-71-10108).

The work of O. Yufereva in the rest 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 075-02-2023-913).

## References

- [1] Monge, G.: Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences de Paris (1781)
- [2] Kantorovich, L.: On the translocation of masses. (Doklady) Acad. Sci. URSS (N.S.) **37**, 199–201 (1942)
- [3] Peyré, G., Cuturi, M., *et al.*: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
- [4] Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis **43**(2), 904–924 (2011)
- [5] Cuturi, M., Doucet, A.: Fast computation of wasserstein barycenters. In: International Conference on Machine Learning, pp. 685–693 (2014). PMLR
- [6] Barrio, E., Gine, E., Matran, C.: Central limit theorems for the wasserstein distance between the empirical and the true distributions. The Annals of Probability **27**(2), 1009–1071 (1999)
- [7] Staib, M., Clatici, S., Solomon, J.M., Jegelka, S.: Parallel streaming wasserstein barycenters. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 2647–2658. Curran Associates, Inc., ??? (2017). <http://papers.nips.cc/paper/6858-parallel-streaming-wasserstein-barycenters.pdf>
- [8] Uribe, C.A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Nedić, A.: Distributed computation of wasserstein barycenters over networks. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 6544–6549 (2018). IEEE
- [9] Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., Nedich, A.: Decentralize and randomize: Faster algorithm for wasserstein barycenters. Advances in Neural Information Processing Systems **31** (2018)
- [10] Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Uribe, C.: On the complexity of approximating wasserstein barycenters. In: International Conference on Machine Learning, pp. 3530–3540 (2019). PMLR
- [11] Dvinskikh, D., Gorbunov, E., Gasnikov, A., Dvurechensky, P., Uribe, C.A.: On primal and dual approaches for distributed stochastic convex optimization over networks. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 7435–7440 (2019). IEEE



- [12] Krawtschenko, R., Uribe, C.A., Gasnikov, A., Dvurechensky, P.: Distributed optimization with quantization for computing wasserstein barycenters. arXiv preprint arXiv:2010.14325 (2020)
- [13] Dvinskikh, D.: Decentralized algorithms for wasserstein barycenters. PhD thesis, Humboldt Universitaet zu Berlin (Germany) (2021)
- [14] Rogozin, A., Beznosikov, A., Dvinskikh, D., Kovalev, D., Dvurechensky, P., Gasnikov, A.: Decentralized distributed optimization for saddle point problems. arXiv preprint arXiv:2102.07758 (2021)
- [15] Gorbunov, E., Rogozin, A., Beznosikov, A., Dvinskikh, D., Gasnikov, A.: In: Nikeghbali, A., Pardalos, P.M., Raigorodskii, A.M., Rassias, M.T. (eds.) *Recent Theoretical Advances in Decentralized Distributed Convex Optimization*, pp. 253–325. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-00832-0\\_8](https://doi.org/10.1007/978-3-031-00832-0_8)
- [16] Rogozin, A., Bochko, M., Dvurechensky, P., Gasnikov, A., Lukoshkin, V.: An accelerated method for decentralized distributed stochastic optimization over time-varying graphs. In: *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 3367–3373 (2021). <https://doi.org/10.1109/CDC45484.2021.9683110>
- [17] Li, H., Lin, Z.: Accelerated gradient tracking over time-varying graphs for decentralized optimization. arXiv preprint arXiv:2104.02596 (2021)
- [18] Kovalev, D., Gasanov, E., Gasnikov, A., Richtarik, P.: Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems* **34** (2021)
- [19] Kovalev, D., Shulgin, E., Richtárik, P., Rogozin, A.V., Gasnikov, A.: ADOM: accelerated decentralized optimization method for time-varying networks. In: *International Conference on Machine Learning*, pp. 5784–5793 (2021). PMLR
- [20] Dvinskikh, D., Tiapkin, D.: Improved complexity bounds in Wasserstein barycenter problem. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1738–1746 (2021). PMLR
- [21] Bishop, A.N., Doucet, A.: Network consensus in the wasserstein metric space of probability measures. *SIAM Journal on Control and Optimization* **59**(5), 3261–3277 (2021)
- [22] Wu, X., Lu, J.: Fenchel dual gradient methods for distributed convex optimization over time-varying networks. *IEEE Transactions on Automatic Control* **64**(11), 4629–4636 (2019) <https://doi.org/10.1109/TAC.2019.2901829>
- [23] Devolder, O., Glineur, F., Nesterov, Y.: Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM Journal on Optimization* **22**(2), 702–727 (2012)

- [24] Gasnikov, A.V., Gasnikova, E., Nesterov, Y.E., Chernov, A.: Efficient numerical methods for entropy-linear programming problems. *Computational Mathematics and Mathematical Physics* **56**(4), 514–524 (2016)
- [25] Rockafellar, R.T.: *Convex Analysis* vol. 11. Princeton university press, Princeton (1997)
- [26] Lemaréchal, C., Sagastizábal, C.: Practical aspects of the moreau–yosida regularization: Theoretical preliminaries. *SIAM journal on optimization* **7**(2), 367–385 (1997)
- [27] Uribe, C.A., Lee, S., Gasnikov, A., Nedić, A.: A dual approach for optimal algorithms in distributed optimization over networks. In: *2020 Information Theory and Applications Workshop (ITA)*, pp. 1–37 (2020). IEEE
- [28] Rabin, J., Peyré, G., Delon, J., Bernot, M.: Wasserstein barycenter and its application to texture mixing. In: *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446 (2011). Springer
- [29] Boissard, E., Le Gouic, T., Loubes, J.-M.: Distribution’s template estimate with wasserstein metrics. *Bernoulli* **21**(2), 740–759 (2015)
- [30] Villani, C.: *Optimal Transport: Old and New* vol. 338. Springer, Cham (2009)
- [31] LeCun, Y.: The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/> (1998)
- [32] Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N.T.H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D.J., Tavenard, R., Tong, A., Vayer, T.: Pot: Python optimal transport. *Journal of Machine Learning Research* **22**(78), 1–8 (2021)
- [33] Cuturi, M., Peyré, G.: A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences* (2015)
- [34] Bigot, J., Cazelles, E., Papadakis, N.: 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)

## A ADOM and its assumptions

The state of the art numerical computation method for time-varying networks, called ADOM, is developed in [19] and this subsection is to present its main objects. It has natural restrictions on the class of suitable problems and, e.g., Wasserstein barycenter problem lies beyond the requirements of this algorithm. So we modify ADOM to solve

more general optimization problems with restrictions. For the sake of consistency, we slightly change original notation and adduce below the results from [19].

In [19], optimization problem with the consensus condition is

$$\min_{\mathbf{x} \in \mathcal{R}} H(\mathbf{x}) = \min_{\mathbf{x} \in \mathcal{R}} \sum_{i=1}^m h_i([\mathbf{x}]_i), \quad (10)$$

$$\text{where } \mathcal{R} = \{\mathbf{x} = ([\mathbf{x}]_1, \dots, [\mathbf{x}]_m) \in (\mathbb{R}^d)^m \mid [\mathbf{x}]_1 = \dots = [\mathbf{x}]_m\},$$

where functions  $h_i: \mathbb{R}^d \rightarrow \mathbb{R}$  are assumed to be smooth and strongly convex. Problem (10) is equivalent to the following:

$$\min_{\mathbf{z} \in \mathcal{R}^\perp} H^*(\mathbf{z}), \quad (11)$$

$$\text{where } \mathcal{R}^\perp = \left\{ \mathbf{z} = ([\mathbf{z}]_1, \dots, [\mathbf{z}]_m) \in (\mathbb{R}^d)^m \mid \sum_{i=1}^m [\mathbf{z}]_i = 0 \right\},$$

where  $H^*$  is the Fenchel transform of the function  $H$  and  $\mathcal{R}^\perp$  is the orthogonal complement of  $\mathcal{R}$ , that exists since  $S = \mathbb{R}^d$  here.

**Theorem 3** ([19, Theorem 1]). *Let functions  $h_i: \mathbb{R}^d \rightarrow \mathbb{R}$  be  $L$  smooth and  $\mu$  strongly convex,  $\mathbf{x}^*$  be the solution of the optimization problem (10),  $\mathbf{W}_n$  be a communication matrix at the  $n$ -th iteration satisfying Assumption 1. Set parameters  $\alpha, \eta, \theta, \sigma, \tau$  of Algorithm 2 to  $\alpha = \frac{1}{2L}$ ,  $\eta = \frac{2\lambda_{\min}^+ \sqrt{\mu L}}{7\lambda_{\max}}$ ,  $\theta = \frac{\mu}{\lambda_{\max}}$ ,  $\sigma = \frac{1}{\lambda_{\max}}$ , and  $\tau = \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{\mu}{L}}$ . Then there exists  $C > 0$ , such that for Fenchel conjugate function  $H^*(\mathbf{z})$  from (11)*

$$\|\nabla H^*(\mathbf{z}_g^n) - \mathbf{x}^*\|_2^2 \leq C \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{\mu}{L}}\right)^n. \quad (12)$$

**Remark 2.** *Addressing details of the proof of Theorem 1 of [19] we see that there is a particular choice of the constant  $C$ , namely*

$$C = \max \left\{ \frac{2\tau}{\mu^2}, \frac{\tau(1-\tau)L}{\eta(1-\eta\alpha)\mu^2} \right\} = \frac{1}{\mu^2} \max \left\{ \frac{2\lambda_{\min}^+ \sqrt{\mu}}{7\lambda_{\max} \sqrt{L}}, \frac{1}{2} \right\} = \frac{1}{2\mu^2}. \quad (13)$$

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

---

**Algorithm 2** ADOM: Accelerated Decentralized Optimization Method
 

---

1: **input:**  $\nabla H^* : (\mathbb{R}^d)^m \rightarrow \mathbb{R}$ ,  $\mathbf{z}^0 \in \mathcal{R}^\perp$ ,  $\mathbf{m}^0 \in (\mathbb{R}^d)^\mathcal{V}$ ,  $\alpha, \eta, \theta, \sigma > 0$ ,  $\tau \in (0, 1)$   
 2: set  $\mathbf{z}_f^0 = \mathbf{z}^0$   
 3: **for**  $k = 0, 1, 2, \dots$  **do**  
 4:    $\mathbf{z}_g^n = \tau \mathbf{z}^n + (1 - \tau) \mathbf{z}_f^n$   
 5:    $\Delta^n = \sigma \mathbf{W}_n (\mathbf{m}^n - \eta \nabla H^*(\mathbf{z}_g^n))$   
 6:    $\mathbf{m}^{n+1} = \mathbf{m}^n - \eta \nabla H^*(\mathbf{z}_g^n) - \Delta^n$   
 7:    $\mathbf{z}^{n+1} = \mathbf{z}^n + \eta \alpha (\mathbf{z}_g^n - \mathbf{z}^n) + \Delta^n$   
 8:    $\mathbf{z}_f^{n+1} = \mathbf{z}_g^n - \theta \mathbf{W}_n \nabla H^*(\mathbf{z}_g^n)$   
 9: **end for**

---

## B Proof of Theorem 1

All the arguments below are applied under assumptions of Theorem 1, i.e. we assume that  $S \subset \mathbb{R}^d$  is a convex set,  $\mathbf{x} \in \mathcal{S}$  is equivalent to  $[\mathbf{x}]_i \in S$  for all  $i = 1, \dots, m$ , functions  $f_i^\gamma : S \rightarrow \mathbb{R}$  are  $\gamma$  strongly convex, and the output of Algorithm 1 is  $\mathbf{x}_{r,\gamma}^n = \nabla(H^{r,\gamma})^*(\mathbf{z}_g^n)$ . Denote also

$$\mathbf{x}_\gamma^* = (x_\gamma^*, \dots, x_\gamma^*) = \arg \min_{\mathbf{x} \in \mathcal{S}} F^\gamma(\mathbf{x}) = \arg \min_{x \in S} \sum_{i=1}^m f_i^\gamma(x).$$

### B.1 Derivation of $(H^{r,\gamma})^*$

In brief, in this subsection we show that functions  $h_i^{r,\gamma}$  from (14) are  $\frac{1}{r}$  smooth,  $\frac{\gamma}{1+r\gamma}$  strongly convex, and such that  $\nabla(H^{r,\gamma})^*$  from Line 3 of Algorithm 1 is the gradient of the conjugate function  $(H^{r,\gamma})^*$  of  $H^{r,\gamma} = \sum_{i=1}^m h_i^{r,\gamma}$  from (14). Then the consensus condition (4) becomes a corollary of Theorem 3 with  $L = \frac{1}{r}$  and  $\mu = \frac{\gamma}{1+r\gamma}$ .

From now on let functions  $h_i^{r,\gamma} : \mathbb{R}^d \rightarrow \mathbb{R}$  and  $H^{r,\gamma} : (\mathbb{R}^d)^m \rightarrow \mathbb{R}$  be

$$\begin{aligned}
 H^{r,\gamma}(\mathbf{x}) &= \sum_{i=1}^m h_i^{r,\gamma}([\mathbf{x}]_i), \text{ where} \\
 h_i^{r,\gamma}(x) &= \inf_{y \in S} \left\{ f_i^\gamma(y) + \frac{1}{2r} \|y - x\|_2^2 \right\}.
 \end{aligned} \tag{14}$$

Define their conjugate as  $(h_i^{r,\gamma})^*$  and  $(H^{r,\gamma})^*$ .

**Lemma 1.** *If functions  $h_i^{r,\gamma}$  and  $H^{r,\gamma}$  are defined by (14), then their Fenchel conjugate functions  $(h_i^{r,\gamma})^*$  and  $(H^{r,\gamma})^* : (\mathbb{R}^d)^m \rightarrow \mathbb{R}$  are*

$$\begin{aligned}
 (H^{r,\gamma})^*(\mathbf{z}) &= \sum_{i=1}^m (h_i^{r,\gamma})^*([\mathbf{z}]_i), \text{ where} \\
 (h_i^{r,\gamma})^*(z) &= (f_i^\gamma)^*(z) + \frac{r}{2} \|z\|_2^2.
 \end{aligned}$$

Moreover, its conjugate  $(H^{r,\gamma})^{**}$  coincides with  $H^{r,\gamma}$ .

*Proof.* The definition (14) is similar to Moreau–Yosida smoothing, but the tricky point is that the functions  $f_i^\gamma$  are defined on a convex set  $S$  instead of the  $\mathbb{R}^d$ . Let us introduce functions  $\tilde{f}_i^\gamma$  with domain  $\mathbb{R}^d$  as follows:

$$\tilde{f}_i^\gamma(x) = \begin{cases} f_i^\gamma(x) & \text{if } x \in S \\ +\infty, & \text{otherwise.} \end{cases} \quad (15)$$

Such  $\tilde{f}_i^\gamma$  are  $\gamma$  strongly convex as well. Moreover, substitution  $\tilde{f}_i^\gamma$  for  $f_i^\gamma$  affect neither primal  $h_i^{r,\gamma}$ :

$$h_i^{r,\gamma}(x) = \inf_{y \in S} \left\{ f_i^\gamma(y) + \frac{1}{2r} \|y - x\|_2^2 \right\} = \inf_{y \in \mathbb{R}^d} \left\{ \tilde{f}_i^\gamma(y) + \frac{1}{2r} \|y - x\|_2^2 \right\},$$

nor  $(f_i^\gamma)^*(z) + \frac{r}{2} \|z\|_2^2$ :

$$\begin{aligned} (f_i^\gamma)^*(z) + \frac{r}{2} \|z\|_2^2 &= \max_{x \in S} \{ \langle z, x \rangle - f_i^\gamma(x) \} + \frac{r}{2} \|z\|_2^2 \\ &= \max_{x \in \mathbb{R}^d} \{ \langle z, x \rangle - \tilde{f}_i^\gamma(x) \} + \frac{r}{2} \|z\|_2^2 = (\tilde{f}_i^\gamma)^*(z) + \frac{r}{2} \|z\|_2^2. \end{aligned}$$

For each  $i$  one can see that  $(h_i^{r,\gamma})^* = (f_i^\gamma)^*(z) + \frac{r}{2} \|z\|_2^2$  is the Fenchel conjugate of  $h_i^{r,\gamma}$  and vice versa. Indeed, for proper, convex and lower semicontinuous  $g_1, g_2: \mathbb{R}^d \rightarrow \mathbb{R}$  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  $H^{r,\gamma}$  will be

$$\begin{aligned} & \sup_{\mathbf{x} \in (\mathbb{R}^d)^m} \{ \langle \mathbf{z}, \mathbf{x} \rangle - H^{r,\gamma}(\mathbf{x}) \} \\ &= \sup_{\mathbf{x} \in (\mathbb{R}^d)^m} \left\{ \sum_{i=1}^m (\langle [\mathbf{z}]_i, [\mathbf{x}]_i \rangle - h_i^{r,\gamma}([\mathbf{x}]_i)) \right\} \quad (16) \\ &= \sum_{i=1}^m \sup_{[\mathbf{x}]_i \in \mathbb{R}^d} \{ \langle [\mathbf{z}]_i, [\mathbf{x}]_i \rangle - h_i^{r,\gamma}([\mathbf{x}]_i) \} \\ &= \sum_{i=1}^m (h_i^{r,\gamma})^*([\mathbf{z}]_i) = (H^{r,\gamma})^*(\mathbf{z}). \end{aligned}$$

In the same way one can see that  $H^{r,\gamma}$  and  $(H^{r,\gamma})^{**}$  coincide.  $\square$

**Remark 3.** For each  $i$  the function  $(h_i^{r,\gamma})^*$  from (14) is  $\left(\frac{1}{\gamma} + r\right)$  smooth and  $r$  strongly convex by definition, so we have  $h_i^{r,\gamma} = (h_i^{r,\gamma})^{**}$  being  $\frac{1}{r}$  smooth and  $\frac{\gamma}{1+r\gamma}$  strongly convex. In addition

$$\nabla(h_i^{r,\gamma})^*(z) = \nabla(f_i^\gamma)^*(z) + z$$

as stated in Line 3 of Algorithm 1. Then we can apply Algorithm 2 for  $L = r^{-1}$  smooth and  $\mu = \frac{\gamma}{1+r\gamma}$  strongly convex functions  $h_i^{r,\gamma}$  and get the values of  $\nabla(h_i^{r,\gamma})^*(z)$  as output.

Thus we construct a relaxation  $\min_{\mathbf{x} \in \mathcal{R}} H^{r,\gamma}(\mathbf{x})$  of the constrained convex optimization problem  $\min_{\mathbf{x} \in S} F^\gamma(\mathbf{x})$ .

**Corollary 2.** *Let a function  $H^{r,\gamma}$  be defined in (14) and let  $\mathbf{x}_{r,\gamma}^* = \arg \min_{\mathbf{x} \in \mathcal{R}} H^{r,\gamma}(\mathbf{x})$ .*

*Then applying Algorithm 2 for*

$$\nabla(h_i^{r,\gamma})^*(z) = (f_i^\gamma)^*(z) + rz$$

*we get by Theorem 3*

$$\|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_{r,\gamma}^n\|_2^2 \leq C \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^n, \quad (17)$$

*where  $\mathbf{x}_{r,\gamma}^n = \nabla(H^{r,\gamma})^*(\mathbf{z}_g^n)$  and*

$$C = \frac{(1+r\gamma)^2}{2\gamma^2}.$$

*Moreover, since  $\mathbf{x}_{r,\gamma}^* \in \mathcal{R}$ , i.e.  $[\mathbf{x}_{r,\gamma}^*]_i = [\mathbf{x}_{r,\gamma}^*]_j$  for all  $i$  and  $j$ , the consensus condition is approximated as follows*

$$\|[\mathbf{x}_{r,\gamma}^n]_i - [\mathbf{x}_{r,\gamma}^n]_j\|_2^2 \leq 2C \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^n.$$

## B.2 Value bounds on $H^{r,\gamma}$

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

**Lemma 2.** *Let functions  $h_i^{r,\gamma}$  be defined in (14). If  $x \in S$ , then for any  $r > 0$ , for each  $i = 1, \dots, m$  we have*

$$f_i^\gamma(x) - \frac{r}{2(1+r\gamma)} \|\nabla f_i^\gamma(x)\|_2^2 \leq h_i^{r,\gamma}(x) \leq f_i^\gamma(x). \quad (18)$$

*Proof.* The second inequality directly follows from the definition (14). To prove the first one we recall that  $f_i^\gamma$  is  $\gamma$  strongly convex and the following holds:

$$\begin{aligned} h_i^{r,\gamma}(x) &= \inf_{y \in S} \{f_i^\gamma(y) + (2r)^{-1} \|x - y\|_2^2\} \\ &= \inf_{y: (x-y) \in S} \{f_i^\gamma(x-y) + (2r)^{-1} \|y\|_2^2\} \\ &\geq \inf_{y: (x-y) \in S} \{f_i^\gamma(x) + \langle \nabla f_i^\gamma(x), -y \rangle + \gamma/2 \|y\|_2^2 + (2r)^{-1} \|y\|_2^2\} \\ &\geq \inf_{y \in \mathbb{R}^d} \{f_i^\gamma(x) + \langle \nabla f_i^\gamma(x), -y \rangle + \gamma/2 \|y\|_2^2 + (2r)^{-1} \|y\|_2^2\}, \end{aligned}$$

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

$$f_i^\gamma(x) + \frac{r}{1+r\gamma} \langle -\nabla f_i^\gamma(x), \nabla f_i^\gamma(x) \rangle + \frac{r}{2(1+r\gamma)} \|\nabla f_i^\gamma(x)\|_2^2$$

$$= f_i^\gamma(x) - \frac{r}{2(1+r\gamma)} \|\nabla f_i^\gamma(x)\|_2^2.$$

□

### B.3 Convergence in argument

Lemma 3 shows convergence in argument in the following sense: if the regularization parameter  $r$  tends to zero, the argminimum  $\mathbf{x}_{r,\gamma}^* \in \mathcal{R}$  of  $H^{r,\gamma}$  tends to the argminimum  $\mathbf{x}_\gamma^* \in \mathcal{S}$  of  $F^\gamma$ . By Corollary 2 we have  $\mathbf{x}_{r,\gamma}^* \in \mathcal{R}$  approximated by  $\mathbf{x}_{r,\gamma}^n \in (\mathbb{R}^d)^m$  for a sufficient number of iterations  $n$ .

**Lemma 3.** *Let  $\mathbf{x}_{r,\gamma}^* = \arg \min_{\mathbf{x} \in \mathcal{R}} H^{r,\gamma}(\mathbf{x})$  for  $H^{r,\gamma}$  defined in (14). Let*

$$\|\nabla F^\gamma(\mathbf{x})\|_2^2 \leq mK_\zeta^2 \quad \forall \mathbf{x} \in \{\mathbf{y} \in \mathcal{S} \mid \|\mathbf{y} - \mathbf{x}_\gamma^*\|_2 \leq \zeta\}. \quad (19)$$

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

$$\|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2 \leq \sqrt{\frac{rm}{2\gamma}} K_\zeta. \quad (20)$$

*Proof.* Using (18) and strong convexity of  $F^\gamma$  and  $H^{r,\gamma}$  we have

$$\begin{aligned} F^\gamma(\mathbf{x}_\gamma^*) &\geq H^{r,\gamma}(\mathbf{x}_\gamma^*) = \sum h_i^{r,\gamma}([\mathbf{x}_\gamma^*]_i) \\ &\geq \sum_{i=1}^m \left( h_i^{r,\gamma}(\mathbf{x}_{r,\gamma}^*) + \frac{\gamma}{2(1+r\gamma)} \|\mathbf{x}_{r,\gamma}^* - [\mathbf{x}_\gamma^*]_i\|_2^2 \right) \\ &= H^{r,\gamma}(\mathbf{x}_{r,\gamma}^*) + \frac{\gamma}{2(1+r\gamma)} \|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2^2 \\ &\geq F^\gamma(\mathbf{x}_{r,\gamma}^*) - \frac{r}{2(1+r\gamma)} \|\nabla F^\gamma(\mathbf{x}_{r,\gamma}^*)\|_2^2 + \frac{\gamma}{2(1+r\gamma)} \|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2^2 \\ &\geq F^\gamma(\mathbf{x}_{r,\gamma}^*) - \frac{r}{2(1+r\gamma)} mK_\zeta^2 + \frac{\gamma}{2(1+r\gamma)} \|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2^2 \\ &\geq F^\gamma(\mathbf{x}_\gamma^*) + \gamma/2 \|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2^2 - \frac{r}{2(1+r\gamma)} mK_\zeta^2 + \frac{\gamma}{2(1+r\gamma)} \|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2^2 \\ &\geq F^\gamma(\mathbf{x}_\gamma^*) + \frac{\gamma}{1+r\gamma} \|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2^2 - \frac{r}{2(1+r\gamma)} mK_\zeta^2. \end{aligned}$$

Then  $\frac{\gamma}{1+r\gamma} \|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2^2 - \frac{r}{2(1+r\gamma)} mK_\zeta^2 \leq 0$  and hence  $\|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2 \leq \frac{rm}{2\gamma} K_\zeta^2$ . □

Combining Lemma 3 with Corollary 2 we get the following.

**Remark 4.** *Let  $\zeta > 0$  and let  $K_\zeta$  be such that (19) holds. If*

$$\sqrt{\frac{rm}{2\gamma}} K_\zeta + \sqrt{C_1} \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}} \right)^{n/2} \leq \zeta,$$

where  $C_1 = \frac{(1+r\gamma)^2}{2\gamma^2}$ , then both  $\|\mathbf{x}_{r,\gamma}^* - \mathbf{x}_\gamma^*\|_2 \leq \zeta$  and  $\|\mathbf{x}_{r,\gamma}^n - \mathbf{x}_\gamma^*\|_2 \leq \zeta$  hold.

## B.4 Value approximation

Let  $\mathbf{x}_{r,\gamma}^* \in \mathcal{R}$  be the only argminimum of  $H^{r,\gamma}$  on the consensus space  $\mathcal{R}$ , i.e.

$$\mathbf{x}_{r,\gamma}^* = \arg \min_{\mathbf{x} \in \mathcal{S}} H^{r,\gamma}(\mathbf{x}). \quad (21)$$

In order to prove the value approximation (5) let us separate it into parts and estimate each of them:

$$F^\gamma(\mathbf{x}_{r,\gamma}^n) - F^\gamma(\mathbf{x}_\gamma^*) \quad (22a)$$

$$\leq F^\gamma(\mathbf{x}_{r,\gamma}^n) - H^{r,\gamma}(\mathbf{x}_{r,\gamma}^n) \quad (22b)$$

$$+ H^{r,\gamma}(\mathbf{x}_{r,\gamma}^n) - H^{r,\gamma}(\mathbf{x}_{r,\gamma}^*) \quad (22c)$$

$$+ H^{r,\gamma}(\mathbf{x}_{r,\gamma}^*) - F^\gamma(\mathbf{x}_\gamma^*). \quad (22d)$$

The last addend is negative and can be eliminated:

$$H^{r,\gamma}(\mathbf{x}_{r,\gamma}^*) - F^\gamma(\mathbf{x}_\gamma^*) \leq H^{r,\gamma}(\mathbf{x}_\gamma^*) - F^\gamma(\mathbf{x}_\gamma^*) \leq 0.$$

The rest are estimated in Lemmas 4 and 5 under additional assumptions.

**Lemma 4.** *Let  $\|\mathbf{x}_{r,\gamma}^n - \mathbf{x}_\gamma^*\|_2 \leq \zeta$ . If (19) holds, then*

$$F^\gamma(\mathbf{x}_{r,\gamma}^n) - H^{r,\gamma}(\mathbf{x}_{r,\gamma}^n) \leq \frac{r}{2(1+r\gamma)} m K_\zeta^2. \quad (23)$$

*Proof.* We cannot declare a uniform  $K$  instead of  $K_\zeta$  because  $F^\gamma$  is not smooth. Nonetheless, assuming  $\mathbf{x}_{r,\gamma}^n$  belong to  $\zeta$ -neighborhood of  $\mathbf{x}_\gamma^*$ , we immediately obtain from (18) and (19) that

$$F^\gamma(\mathbf{x}_{r,\gamma}^n) - H^{r,\gamma}(\mathbf{x}_{r,\gamma}^n) \leq \frac{r}{2(1+r\gamma)} \|\nabla F^\gamma(\mathbf{x}_{r,\gamma}^n)\|_2^2 \leq \frac{r}{2(1+r\gamma)} m K_\zeta^2.$$

□

**Lemma 5.** *Let (19) holds. Then*

$$H^{r,\gamma}(\mathbf{x}_{r,\gamma}^n) - H^{r,\gamma}(\mathbf{x}_{r,\gamma}^*) \leq C_2 \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2},$$

$$\text{where } C_2 = \frac{m(1+r\gamma)K_\zeta}{\sqrt{2}\gamma} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^+}} + \frac{m(1+r\gamma)^2}{4r\gamma^2}.$$

*Proof.* By  $\frac{m}{r}$  smoothness of  $H^{r,\gamma}$

$$\begin{aligned} H^{r,\gamma}(\mathbf{x}_{r,\gamma}^n) - H^{r,\gamma}(\mathbf{x}_{r,\gamma}^*) &\leq \langle \nabla H^{r,\gamma}(\mathbf{x}_{r,\gamma}^*), \mathbf{x}_{r,\gamma}^n - \mathbf{x}_{r,\gamma}^* \rangle + \frac{m}{2r} \|\mathbf{x}_{r,\gamma}^n - \mathbf{x}_{r,\gamma}^*\|_2^2 \\ &\leq \langle \nabla H^{r,\gamma}(\nabla(H^{r,\gamma})^*(\mathbf{z}_g^\infty)), \nabla(H^{r,\gamma})^*(\mathbf{z}_g^n) - \mathbf{x}_{r,\gamma}^* \rangle + \frac{m}{2r} \|\mathbf{x}_{r,\gamma}^n - \mathbf{x}_{r,\gamma}^*\|_2^2 \\ &\leq \langle \mathbf{z}_g^\infty, \nabla(H^{r,\gamma})^*(\mathbf{z}_g^n) - \mathbf{x}_{r,\gamma}^* \rangle + \frac{m}{2r} \|\mathbf{x}_{r,\gamma}^n - \mathbf{x}_{r,\gamma}^*\|_2^2, \end{aligned}$$



where  $\mathbf{z}_g^\infty$  is the limit of  $\mathbf{z}_g^n$  and so it is the argminimum of  $(H^{r,\gamma})^*$  on  $\mathcal{R}^\perp$ . By (17) we have

$$\frac{m}{2r} \|\mathbf{x}_{r,\gamma}^n - \mathbf{x}_{r,\gamma}^*\|_2^2 \leq \frac{m}{2r} C_1 \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^n = \frac{m(1+r\gamma)^2}{4r\gamma^2} \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^n$$

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

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

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

Since  $\mathbf{z}_g^\infty \in \mathcal{R}^\perp$  and  $\mathbf{x}_{r,\gamma}^* \in \mathcal{R}$ , the first part simplifies to  $\langle \mathbf{z}_g^\infty, \mathbf{P} \nabla (H^{r,\gamma})^*(\mathbf{z}_g^n) \rangle$ . We may use Lemma 2 in [19] to get the following estimation

$$\|\mathbf{P} \nabla (H^{r,\gamma})^*(\mathbf{z}_g^n)\|_2^2 = \|\nabla (H^{r,\gamma})^*(\mathbf{z}_g^n)\|_{\mathbf{P}}^2 \leq \frac{2}{\theta \lambda_{\min}^+} \left( (H^{r,\gamma})^*(\mathbf{z}_g^n) - (H^{r,\gamma})^*(\mathbf{z}_f^{n+1}) \right).$$

As  $\mathbf{z}_f^{n+1}$  is a non-optimal point of Algorithm 1, this is not greater than

$$\begin{aligned} & \frac{2}{\theta \lambda_{\min}^+} \left( (H^{r,\gamma})^*(\mathbf{z}_g^n) - (H^{r,\gamma})^*(\mathbf{z}^*) \right) \\ & \leq \frac{m(1+r\gamma)}{\gamma \theta \lambda_{\min}^+} \|\mathbf{z}_g^n - \mathbf{z}^*\|_2^2 = \frac{m(1+r\gamma)^2}{\gamma^2} \frac{\lambda_{\max}}{\lambda_{\min}^+} \|\mathbf{z}_g^n - \mathbf{z}^*\|_2^2 \\ & \leq \frac{m(1+r\gamma)^2}{2\gamma^2} \frac{\lambda_{\max}}{\lambda_{\min}^+} \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^n \end{aligned}$$

and the latter ones follow from the  $\frac{m(1+r\gamma)}{\gamma}$  smoothness of  $(H^{r,\gamma})^*$  and from the fact that the proof of [19, Theorem 1] actually covers the following chain of inequalities:

$$\|\nabla H^*(\mathbf{z}_g^n) - \mathbf{x}^*\|_2^2 \leq \frac{1}{\mu^2} \|\mathbf{z}_g^n - \mathbf{z}^*\|_2^2 \leq C \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{\mu}{L}}\right)^n = \frac{1}{2\mu^2} \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{\mu}{L}}\right)^n.$$

By our assumption  $\|\mathbf{z}_g^\infty\|_2 = \|\nabla H^{r,\gamma}(x_{r,\gamma}^*)\|_2 < \sqrt{m} K_\zeta$ . Thus, we obtain

$$\begin{aligned} & (H^{r,\gamma})^*(\mathbf{x}_{r,\gamma}^n) - (H^{r,\gamma})^*(\mathbf{x}_{r,\gamma}^*) \\ & \leq \sqrt{m} K_\zeta \frac{\sqrt{m(1+r\gamma)}}{\sqrt{2}\gamma} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^+}} \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2} + \frac{m(1+r\gamma)^2}{4r\gamma^2} \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^n \\ & \leq \left( \frac{m(1+r\gamma) K_\zeta}{\sqrt{2}\gamma} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^+}} + \frac{m(1+r\gamma)^2}{4r\gamma^2} \right) \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2} \\ & = C_2 \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2}. \end{aligned}$$

□

## B.5 Final compilation

This section completes the proof of Theorem 1 and shows Remark 1.

Recall that where  $C_1 = \frac{(1+r\gamma)^2}{2\gamma}$  and

$$C_2 = \frac{m}{2r}C_1 + \frac{m(1+r\gamma)K_\zeta}{\sqrt{2}\gamma} \frac{\lambda_{\max}}{\lambda_{\min}^+} = \frac{m(1+r\gamma)^2}{4r\gamma} + \frac{m(1+r\gamma)K_\zeta}{\sqrt{2}\gamma} \frac{\lambda_{\max}}{\lambda_{\min}^+}.$$

By Remark 4 and Lemmas 4, 5 we see that  $F^\gamma(\mathbf{x}_{r,\gamma}^n) - F^\gamma(\mathbf{x}_\gamma^*) < \varepsilon$  if

$$\forall \mathbf{x} \in \{\mathbf{y} \in \|\mathbf{y} - \mathbf{x}_\gamma^*\|_2 < \zeta\} \quad \|\nabla F^\gamma(\mathbf{x})\|_2^2 < mK_\zeta^2, \quad (25)$$

$$\sqrt{\frac{rm}{2\gamma}}K_\zeta + \sqrt{C_1} \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2} \leq \zeta, \quad (26)$$

$$\frac{r}{2(1+r\gamma)}mK_\zeta^2 \leq \varepsilon/2, \quad (27)$$

$$C_2 \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2} \leq \varepsilon/2. \quad (28)$$

Let  $\zeta = \sqrt{\varepsilon/\gamma}$  and let  $r \leq \frac{\varepsilon}{2mK_\zeta^2}$ . Then (27) holds. If (28) fulfills, then (26) follows from (27) and (28) as  $\sqrt{\frac{rm}{2\gamma}}K_\zeta \leq \sqrt{\frac{\varepsilon}{2\gamma}} \leq \zeta/\sqrt{2}$  and  $\sqrt{C_1} \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2} \leq \zeta/2$  since  $1 \leq \sqrt{C_1} \leq C_1 \leq C_2$  and  $\varepsilon \leq \sqrt{\varepsilon/\gamma} = \zeta$ . Thus, it suffices to assume

$$\begin{aligned} \forall i \quad \forall x \in \{y \in S \mid \|y - x_\gamma^*\|_2^2 \leq \varepsilon/\gamma\} \quad \|\nabla f_i^\gamma(x)\|_2 &\leq K, \\ r &\leq \frac{\varepsilon}{2mK^2}, \\ C_2 \left(1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}}\right)^{n/2} &\leq \varepsilon/2. \end{aligned}$$

So  $\varepsilon$  approximation requires a number of iteration

$$\mathcal{O} \left( \frac{\lambda_{\max}}{\lambda_{\min}^+} \sqrt{\frac{1+r\gamma}{r\gamma}} \ln \frac{C_2}{\varepsilon} \right) = \mathcal{O} \left( \frac{\lambda_{\max}}{\lambda_{\min}^+} \frac{1}{\sqrt{\gamma\varepsilon}} \ln \frac{1}{\varepsilon} \right).$$

## C Proof of Theorem 2

To prove Theorem 2 we combine proved Theorem 1 with features of the entropy regularization of the Wasserstein barycenter problem.

### C.1 Entropy regularized WB problem

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

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

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

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

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

$$\mathcal{W}_\gamma(p, q) := \min_{X \in U(p, q)} \{ \langle M, X \rangle - \gamma E(X) \}, \quad (29)$$

where  $\gamma > 0$  and

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

$$\text{where } e(x) = \begin{cases} x \ln x & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases} \quad (30)$$

So it seeks to minimize the transportation costs while maximizing the entropy. Moreover  $\mathcal{W}_\gamma(p, q) \rightarrow \mathcal{W}(p, q)$  as  $\gamma \rightarrow 0$ .

Then the convex optimization problem (7) can be relaxed to the following  $\gamma$  strongly convex optimization problem

$$\min_{p \in S_1(d)} \sum_{i=1}^m \mathcal{W}_{\gamma, q_i}(p), \quad (31)$$

where  $\mathcal{W}_{\gamma, q_i}(p) = \mathcal{W}_\gamma(q_i, p)$ . The argminimum of (31) is called the uniform Wasserstein barycenter [4, 5] of the family of  $q_1, \dots, q_m$ . Moreover, problem (31) admits a unique solution and approximates unregularized WB problem as follows.

**Remark 5.** Let  $\gamma \leq \frac{\varepsilon}{4} \ln d$ . If vectors  $\hat{p}_i \in S_1(d)$  are such that

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

then

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

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

$$\begin{aligned} & \sum_{i=1}^m \mathcal{W}_{q_i}(\hat{p}_i) - \sum_{i=1}^m \mathcal{W}_{q_i}(p^*) \\ & \leq \sum_{i=1}^m \mathcal{W}_{\gamma, q_i}(\hat{p}_i) - \sum_{i=1}^m \mathcal{W}_{\gamma, q_i}(p_\gamma^*) + 2\gamma \ln d \end{aligned}$$

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

## C.2 Legendre transforms

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

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

**Theorem 4** ([33][Theorem 2.4]). *For  $\gamma > 0$ , the Fenchel-Legendre dual function  $\mathcal{W}_{\gamma, q}^*(z)$  is differentiable*

$$\begin{aligned} \mathcal{W}_{\gamma, q}^*(z) &= \gamma (E(q) + \langle q, \ln \mathcal{K}\alpha \rangle) \\ &= -\gamma \langle q, \ln q \rangle + \gamma \sum_{j=1}^m [q]_j \ln \left( \sum_{i=1}^m \exp \left( \frac{1}{\gamma} ([z]_i - M_{ji}) \right) \right) \end{aligned} \quad (33)$$

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

$$\begin{aligned} \nabla \mathcal{W}_{\gamma, q}^*(z) &= \alpha \circ (\mathcal{K} \cdot q / (\mathcal{K}\alpha)) \in S_1(d), \\ [\nabla \mathcal{W}_{\gamma, q}^*(z)]_l &= \sum_{j=1}^m [q]_j \frac{\exp(\frac{1}{\gamma} ([z]_l - M_{lj}))}{\sum_{i=1}^m \exp(\frac{1}{\gamma} ([z]_i - M_{ij}))}. \end{aligned} \quad (34)$$

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

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

**Theorem 5** ([33][Theorem 3.1]). *The barycenter  $p^*$  solving (31) satisfies*

$$\forall i = 1, \dots, m \quad p^* = \nabla \mathcal{W}_{\gamma, q_i}^*(z_i^*),$$

where the set of  $z_i^*$  constitutes any solution of any smoothed dual WB problem:

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

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

$$\begin{aligned} & \sum_{i=1}^m \mathcal{W}_{\gamma, q_i}([\mathbf{x}_{r, \gamma}^n]_i) - \min_{p \in S_1(d)} \sum_{i=1}^m \mathcal{W}_{\gamma, q_i}(p) \\ & \leq \frac{r}{4(1+r\gamma)} mK^2 + \frac{1}{2} C_2 \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}} \right)^{n/2} \leq \varepsilon/2. \end{aligned}$$

By Remark 5 it proves

$$\begin{aligned} & \left| \sum_{i=1}^m \mathcal{W}_{q_i}([\mathbf{x}_{r,\gamma}^n]_i) - \sum_{i=1}^m \mathcal{W}_{q_i}([\mathbf{p}^*]_i) \right| \\ & \leq 2\gamma \ln d + \frac{r}{4(1+r\gamma)} mK^2 + C \left( 1 - \frac{\lambda_{\min}^+}{7\lambda_{\max}} \sqrt{\frac{r\gamma}{1+r\gamma}} \right)^n / 2 \leq \varepsilon, \end{aligned}$$

$$\text{for } C = \frac{1}{2}C_2 = \frac{(1+r\gamma)mK_\zeta}{2\sqrt{2}\gamma} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^+}} + \frac{(1+r\gamma)^2}{8r\gamma^2}.$$

### C.3 Parameter estimation

It remains to assign  $\zeta > 0$  and  $K = K_\zeta$  satisfying (25). Due to Assumption 2 such  $\zeta$  and  $K$  exist.

**Proposition 1.** *Let a set  $\{q_i\}_{i=1}^m$  satisfies Assumption 2, let  $p_\gamma^*$  be the uniform Wasserstein barycenter of  $\{q_i\}_{i=1}^m$ , and let  $\zeta \in (0, \min\{\frac{1}{e}, \min_{i,l}[q_i]_l\})$ . For each  $i = 1, \dots, m$  the norm of the gradient  $\|\nabla \mathcal{W}_{\gamma, q_i}(\cdot)\|_2^2$  is uniformly bounded over  $\{p \in S_1(d) \mid \|p - p_\gamma^*\|_2^2 \leq \zeta\}$ ; and the bound  $K_\rho$  is given in (35) for  $\rho \leq \min\{\frac{1}{e}, \min_{i,l}[q_i]_l\} - \zeta$ .*

We obtain Proposition 1 as a combination of Lemma 6 from [34] and proved below Lemma 7.

**Lemma 6** ([34, Lemma 3.5]). *For any  $\rho \in (0, 1)$ ,  $q \in S_1(d)$ , and  $p \in \{x \in S_1(d) \mid \min_l x_l \geq \rho\}$  there is a bound:  $\|\nabla \mathcal{W}_{\gamma, q}(p)\|_2^2 \leq K_\rho$ , where*

$$K_\rho = \sum_{j=1}^d \left( 2\gamma \ln d + \inf_i \sup_l |M_{jl} - M_{il}| - \gamma \ln \rho \right)^2. \quad (35)$$

**Lemma 7.** *Let a set  $\{q_i\}_{i=1}^m$  satisfies Assumption 2, let  $p_\gamma^*$  be the uniform Wasserstein barycenter of  $\{q_i\}_{i=1}^m$ . All components  $k$  of  $p_\gamma^*$  have a uniform positive lower bound:  $[p_\gamma^*]_k \geq \min\{\frac{1}{e}, \min_{i,l}[q_i]_l\}$ .*

*Proof.* Let  $X_i^*$  denote the optimal transport plan between  $p_\gamma^*$  and  $q_i$ . Assume the contrary: there is  $k$  such that  $[p_\gamma^*]_k < \min\{\frac{1}{e}, \min_{i,l}[q_i]_l\}$ . Then there is another component  $n$  such that  $[p_\gamma^*]_n > \min_i [q_i]_n > \min_{i,l}[q_i]_l$ . Consider the vector  $p$  that consists of  $[p]_i = [p_\gamma^*]_i$  except for the components  $[p]_n = [p_\gamma^*]_n + \delta$  and  $[p]_l = [p_\gamma^*]_l - \delta$ , where  $\delta > 0$  is less than  $\min_{i,a \neq b} [X_i^*]_{a,b}$  of the optimal transport plans  $X_i^*$  between  $p_\gamma^*$  and  $q_n$ . 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  $\mathcal{W}_{\gamma, q_i}(p_\gamma^*) = \langle C, X_i^* \rangle - \gamma X_i^* \ln X_i^*$ . Consider the matrix  $X_i$  that differs from  $X_i^*$  only at four elements:

$$\begin{aligned} [X_i]_{kk} &= [X_i^*]_{kk} + \frac{1}{2}\delta, & [X_i]_{kn} &= [X_i^*]_{kn} + \frac{1}{2}\delta, \\ [X_i]_{nn} &= [X_i^*]_{nn} + \frac{1}{2}\delta, & [X_i]_{nk} &= [X_i^*]_{nk} + \frac{1}{2}\delta. \end{aligned}$$

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

$$\begin{aligned}
\mathcal{W}_{\gamma, q_i}(p) &\leq \langle C, X_i \rangle - \gamma X_i \ln X_i \\
&= \langle C, X_i^* \rangle - \gamma X_i^* \ln X_i^* + \frac{1}{2} \delta C_{kn} - \frac{1}{2} \delta C_{nk} \\
&\quad + ([X_i]_{kk} \ln [X_i]_{kk} - [X_i^*]_{kk} \ln [X_i^*]_{kk}) \\
&\quad + ([X_i]_{kn} \ln [X_i]_{kn} - [X_i^*]_{kn} \ln [X_i^*]_{kn}) \\
&\quad + ([X_i]_{nk} \ln [X_i]_{nk} - [X_i^*]_{nk} \ln [X_i^*]_{nk}) \\
&\quad + ([X_i]_{nn} \ln [X_i]_{nn} - [X_i^*]_{nn} \ln [X_i^*]_{nn}) \\
&< \langle C, X_i^* \rangle - \gamma X_i^* \ln X_i^* + \frac{1}{2} \delta C_{kn} - \frac{1}{2} \delta C_{nk} \\
&= \langle C, X_i^* \rangle - \gamma X_i^* \ln X_i^* = \mathcal{W}_{\gamma, q_i}(p_\gamma^*).
\end{aligned}$$

The obtained inequalities  $\mathcal{W}_{\gamma, q_i}(p) < \mathcal{W}_{\gamma, q_i}(p_\gamma^*)$  contradict to the fact that  $p_\gamma^*$  is the barycenter; this proves the lemma.  $\square$