On the Complexity of Approximating Wasserstein Barycenters

Alexey Kroshnin [*] Darina Dvinskikh [†] Pavel Dvurechensky [‡]
Alexander Gasnikov [§] Nazarii Tupitsa [¶] César A. Uribe [‖]

May 13, 2019

Abstract

We study the complexity of approximating Wasserstein barycenter of m discrete measures, or histograms of size n by contrasting two alternative approaches, both using entropic regularization. The first approach is based on the Iterative Bregman Projections (IBP) algorithm for which our novel analysis gives a complexity bound proportional to $\frac{mn^2}{\varepsilon^2}$ to approximate the original non-regularized barycenter. Using an alternative accelerated-gradient-descent-based approach, we obtain a complexity proportional to $\frac{mn^{2.5}}{\varepsilon}$. As a byproduct, we show that the regularization parameter in both approaches has to be proportional to ε, which causes instability of both algorithms when the desired accuracy is high. To overcome this issue, we propose a novel proximal-IBP algorithm, which can be seen as a proximal gradient method, which uses IBP on each iteration to make a proximal step. We also consider the question of scalability of these algorithms using approaches from distributed optimization and show that the first algorithm can be implemented in a centralized distributed setting (master/slave), while the second one is amenable to a more general decentralized distributed setting with an arbitrary network topology.

Keywords: Optimal transport, Wasserstein barycenter, Sinkhorn's algorithm, Accelerated Gradient Descent, distributed optimization
AMS Classification: 90C25, 90C30, 90C06, 90C90.

[*]Institute for Information Transmission Problems RAS, Moscow; National Research University Higher School of Economics, Moscow; Université Claude Bernard Lyon 1, Villeurbanne, akroshnin@hse.ru

[†]Weierstrass Institute for Applied Analysis and Stochastics, Berlin; Institute for Information Transmission Problems RAS, Moscow, darina.dvinskikh@wias-berlin.de

[‡]Weierstrass Institute for Applied Analysis and Stochastics, Berlin; Institute for Information Transmission Problems RAS, Moscow, pavel.dvurechensky@wias-berlin.de

[§]Moscow Institute of Physics and Technology, Moscow; Institute for Information Transmission Problems RAS, Moscow; National Research University Higher School of Economics, Moscow, gasnikov@yandex.ru

[¶]Institute for Information Transmission Problems RAS, Moscow, tupitsa@phystech.edu

[‖]Massachusetts Institute of Technology, Cambridge, cauribe@mit.edu

Introduction

Optimal transport (OT) Monge [1781], Kantorovich [1942] is currently generating an increasing attraction in statistics, machine learning and optimization communities. Statistical procedures based on optimal transport are available Bigot et al. [2012], Del Barrio et al. [2015], Ebert et al. [2017], Le Gouic and Loubes [2017] as well as many applications in different areas of machine learning including unsupervised learning Arjovsky et al. [2017], Bigot et al. [2017], semi-supervised learning Solomon et al. [2014], clustering Ho et al. [2017], text classification Kusner et al. [2015]. Optimal transport distances lead to the concept of Wasserstein barycenter, which allows to define a mean of a set of complex objects, e.g. images, preserving their geometric structure Cuturi and Doucet [2014]. In this paper we focus on the computational aspects of optimal transport, namely on approximating Wasserstein barycenter of a set of histograms.

Starting with Altschuler et al. [2017], several groups of authors addressed the question of Wasserstein distance approximation complexity Chakrabarty and Khanna [2018], Dvurechensky et al. [2018b], Blanchet et al. [2018], Lin et al. [2019]. Implementable schemes are based on Sinkhorn's algorithm, which was first applied to OT in Cuturi [2013], and accelerated gradient descent proposed as an alternative in Dvurechensky et al. [2018b]. Much less is known about the complexity of approximating Wasserstein *barycenter*. The works Staib et al. [2017], Uribe et al. [2018], Dvurechensky et al. [2018a], are in some sense close, but do not provide an explicit answer. Following Genevay et al. [2016], Staib et al. [2017] use stochastic gradient descent and estimate the convergence rate of their algorithm. From their rate, one can obtain the iteration complexity $\frac{\kappa R^2}{\varepsilon^2}$ to achieve accuracy ε in approximation of the barycenter, where κ is some constant depending on the problem data, i.e. transportation cost matrices, R is some distance characterizing the solution of the dual problem. Dvurechensky et al. [2018a] consider regularized barycenter, but do not show, how to choose the regularization parameter to achieve ε-accuracy.

Following Dvurechensky et al. [2018b], we study two alternative approaches for approximating Wasserstein barycenter based on entropic regularization Cuturi [2013]. The first approach is based on Iterative Bregman Projection (IBP) algorithm Benamou et al. [2015], which can be considered as a general alternating projections algorithm and also as a generalization of the Sinkhorn's algorithm Sinkhorn [1974]. The second approach is based on constructing a dual problem and solving it by primal-dual accelerated gradient descent. For both approaches, we show, how the regularization parameter should be chosen in order to approximate the original, non-regularized barycenter.

We also address the question of scalability of computations in the Big Data regime, i.e. the size of histograms n and the number of histograms m are large. In this case the dataset of n histograms can be distributedly produced or stored in a network of agents/sensors/computers with the network structure given by an arbitrary connected graph. In a special case of centralized architecture, i.e. if there is a central "master" node surrounded by "slave" nodes, parallel algorithms such as Staib et al. [2017] can be applied. In a more general setup of arbitrary networks it makes sense to use decentralized distributed algorithms in the spirit of distributed optimization algorithms Nedić et al. [2017], Scaman et al. [2017].

Related Work

It is very hard to cover all the increasing stream of works on OT and we mention these books Villani [2008], Santambrogio [2015], Peyré and Cuturi [2018] as a starting point and the references therein. Approximation of Wasserstein barycenter was considered in Cuturi and Doucet [2014], Bonneel et al. [2015], Benamou et al. [2015], Staib et al. [2017], Puccetti et al. [2018], Claici et al. [2018], Uribe et al. [2018], Dvurechensky et al. [2018a] using different techniques as Sinkhorn-type algorithm, first-order methods, Newton-type methods. Considering the primal-dual approach based on accelerated gradient descent, our paper shares some similarities with Cuturi and Peyré [2016] with the main difference that we are focused on complexity and scalability of computations and explicitly analyze the algorithm applied to the dual problem.

There is a vast amount of literature on accelerated gradient descent with the canonical reference being Nesterov [1983]. Primal-dual extensions can be found in Lan et al. [2011], Tran-Dinh et al. [2018], Yurtsever et al. [2015], Chernov et al. [2016], Dvurechensky et al. [2016, 2017], Anikin et al. [2017], Nesterov et al. [2018], Lin et al. [2019]. We are focused on the extensions amenable to the decentralized distributed optimization, so that these algorithms can be scaled for large problems.

Distributed optimization algorithms were considered by many authors with the classical reference being Bertsekas and Tsitsiklis [1989]. Initial algorithms, such as Distributed Gradient Descent Nedic and Ozdaglar [2009], were relatively slow compared with their centralized counterparts. However, recent work has made significant advances towards a better understanding of the optimal rates of such algorithms and their explicit dependencies to the function and network parameters Lan et al. [2017], Scaman et al. [2017], Uribe et al. [2018]. These approaches has been extended to other scenarios such as time-varying graphs Rogozin et al. [2018], Maros and Jaldén [2018], Wu and Lu [2017]. The distributed setup is particularly interesting for machine learning applications on the big data regime, where the number of data points and the dimensionality is large, due to its flexibility to handle intrinsically distributed storage and limited communication, as well as privacy constraints He et al. [2018], Wai et al. [2018].

Our contributions

- We consider γ-regularized Wasserstein barycenter problem and obtain complexity bounds for finding an approximation to the regularized barycenter by two algorithms. The first one is Iterative Bregman Projections algorithm Benamou et al. [2015], for which we prove complexity proportional to $\frac{1}{\gamma\varepsilon}$ to achieve accuracy ε. The second one is based on accelerated gradient descent (AGD) and has complexity proportional to $\sqrt{\frac{n}{\gamma\varepsilon}}$. The benefit of the second algorithm is that it is better scalable and can be implemented in the decentralized distributed optimization setting over an arbitrary network.

- We show, how to choose the regularization parameter in order to find an ε-approximation for the non-regularized Wasserstein barycenter and find the resulting

complexity for IBP to be proportional to $\frac{mn^2}{\varepsilon^2}$ and for AGD to be proportional to $\frac{mn^{2.5}}{\varepsilon}$.

- As we can see from the complexity bounds for IBP and AGD, they depend on the regularization parameter γ quite badly regarding that this parameter has to be small leading to instability of the algorithms. To resolve the stability issue we propose a proximal-IBP method, which can be considered as a proximal method using IBP on each iteration to find the next iterate.

1 Problem Statement and Preliminaries

1.1 Notation

We define the probability simplex as $S_n(1) = \{q \in \mathbb{R}^n_+ \mid \sum_{i=1}^n q_i = 1\}$. Given two discrete measures p and q from $S_n(1)$ we introduce the set of their coupling measures as

$$\Pi(p, q) = \{\pi \in \mathbb{R}^{n\times n}_+ : \pi\mathbb{1} = p, \pi^\mathsf{T}\mathbb{1} = q\},$$

where $\mathbb{1}$ is a column of all ones. For coupling measure $\pi \in \mathbb{R}^{n\times n}_+$ we denote the negative entropy (up to an additive constant) as

$$H(\pi) = \sum_{i,j=1}^n \pi_{ij}\left(\ln \pi_{ij} - 1\right) = \langle \pi, \ln \pi - \mathbb{1}\mathbb{1}^\mathsf{T}\rangle.$$

Here and further by $\ln(A)$ ($\exp(A)$) we denote the element-wise logarithm (exponent) of matrix or vector A, and $\langle A, B\rangle \coloneqq \sum_{i,j=1}^n A_{ij}B_{ij}$ for any $A, B \in \mathbb{R}^{n\times n}$. For two matrices A and B we also define element-wise multiplication and element-wise division as $A \odot B$ and $\frac{A}{B}$ respectively. Kullback–Leibler divergence for measures $\pi, \pi' \in \mathbb{R}^{n\times n}_+$ is defined as the Bregman divergence associated with $H(\cdot)$:

$$KL(\pi|\pi') \coloneqq \sum_{i,j=1}^n \left(\pi_{ij}\ln\left(\frac{\pi_{ij}}{\pi'_{ij}}\right) - \pi_{ij} + \pi'_{ij}\right) = \langle \pi, \ln \pi - \ln \pi'\rangle + \langle \pi' - \pi, \mathbb{1}\mathbb{1}^\mathsf{T}\rangle.$$

We also define a symmetric cost matrix $C \in \mathbb{R}^{n\times n}_+$, which element c_{ij} corresponds to the cost of moving bin i to bin j. $\|C\|_\infty$ denotes the maximal element of this matrix.

We refer to $\lambda_{\max}(W)$ as the maximum eigenvalue of a symmetric matrix W, and $\lambda^+_{\min}(W)$ as the minimal non-zero eigenvalue, and define the condition number of matrix W as $\chi(W)\coloneqq \lambda_{\max}(W)/\lambda^+_{\min}(W)$. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ has L-Lipschitz-continuous gradient w.r.t. norm $\|\cdot\|$ if $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|$, $x, y \in \mathbb{R}^d$, where $\|\cdot\|_*$ is the dual norm defined by $\|g\|_* = \max_{\|x\|\le 1}\langle g, x\rangle$.

1.2 Wasserstein barycenters and entropic regularization

Given two probability measures $p, q \in S_n(1)$ and a cost matrix $C \in \mathbb{R}^{n\times n}$ we define optimal transportation distance between them as

$$\mathcal{W}(p, q) \coloneqq \min_{\pi\in\Pi(p,q)} \langle \pi, C\rangle. \tag{1}$$

For a given set of probability measures $\{p_1, \ldots, p_m\}$ and cost matrices $C_1, \ldots, C_m \in \mathbb{R}_+^{n \times n}$ we define their weighted barycenter with weights $w \in S_n(1)$ as a solution of the following convex optimization problem:

$$\min_{q \in S_n(1)} \sum_{l=1}^{m} w_l \mathcal{W}(p_l, q), \tag{2}$$

where $w_l \geq 0$, $l = 1, \ldots, m$ and

$$\sum_{l=1}^{m} w_l = 1.$$

We use c to denote $\max_{l=1,\ldots,m} \|C_l\|_\infty$. Using entropic regularization proposed in Cuturi [2013] we define regularized OT-distance for $\gamma \geq 0$:

$$\mathcal{W}_\gamma(p, q) \coloneqq \min_{\pi \in \Pi(p,q)} \left\{ \langle \pi, C \rangle + \gamma H(\pi) \right\}. \tag{3}$$

Respectively, one can consider regularized barycenter that is a solution of the following problem:

$$\min_{q \in S_n(1)} \sum_{l=1}^{m} w_l \mathcal{W}_\gamma(p_l, q). \tag{4}$$

2 Complexity of WB by Iterative Bregman Projections

In this section, we provide the theoretical analysis of the Iterative Bregman Projections algorithm Benamou et al. [2015] for the approximation of the regularized Wasserstein barycenter and obtain iteration complexity $O\left(c/(\gamma\varepsilon)\right)$ where $c = \max_{l=1,\ldots,m} \|C_l\|_\infty$. Then we estimate the bias introduced by regularization and estimate the value of γ to obtain an ε-approximation for the non-regularized barycenter. Combining this result with the iteration complexity of IBP, we obtain a complexity $\widetilde{O}\left(c^2 m n^2/\varepsilon^2\right)$ for approximating a non-regularized barycenter by the IBP algorithm. This algorithm can be implemented in a centralized distributed manner such that each node performs $\widetilde{O}\left(c^2 n^2/\varepsilon^2\right)$ arithmetic operations and the number of communication rounds is $\widetilde{O}\left(c^2/\varepsilon^2\right)$. We also introduce proximal-IBP algorithm and discuss its complexity and scalability.

2.1 Convergence of IBP for the regularized barycenter

In this subsection, we analyze Iterative Bregman Projection Algorithm [Benamou et al., 2015, Section 3.2] and analyze its complexity for solving problem (4). We reformulate this

problem as

$$\min_{\substack{q\in S_n(1),\\ \pi_l\in\Pi(p_l,q),l=1,\dots,m}} \sum_{l=1}^{m} w_l\big\{\langle \pi_l, C_l\rangle + \gamma H(\pi_l)\big\}$$

$$= \min_{\substack{\pi_l\mathbb{1}=p_l,\ \pi_l^{\mathsf{T}}\mathbb{1}=\pi_{l+1}^{\mathsf{T}}\mathbb{1},\\ \pi_l\in\mathbb{R}_+^{n\times n},\ l=1,\dots,m}} \sum_{l=1}^{m} w_l\big\{\langle \pi_l, C_l\rangle + \gamma H(\pi_l)\big\}, \tag{5}$$

and construct its dual (see Lemma 1). To solve the dual problem we reformulate the IBP algorithm as a blockwise minimization, as shown in Algorithm 2[1]. Notably, our reformulation of the IBP algorithm allows to solve simultaneously the primal and dual problem and has an adaptive stopping criterion (see line 7), which does not require to calculate any objective values.

Our first step is to recall the IBP algorithm from Benamou et al. [2015], formulate the dual problem for (5) and show that our Algorithm 2 solves this dual problem and is equivalent to the IBP algorithm.

Following the approach from Benamou et al. [2015] we present the problem (5) in a Kullback–Leibler projection form, i.e.,

$$\min_{\boldsymbol{\pi}\in\mathcal{C}_1\cap\mathcal{C}_2} \sum_{l=1}^{m} w_l KL\left(\pi_l | K_l\right), \tag{6}$$

where $K_l = \exp\left(-C_l/\gamma\right)$ and the affine convex sets $\mathcal{C}_1$ and $\mathcal{C}_2$ with

$$\begin{aligned}\mathcal{C}_1 &= \left\{\boldsymbol{\pi} = [\pi_1,\dots,\pi_m] : \forall l\ \pi_l\mathbb{1} = p_l\right\},\\ \mathcal{C}_2 &= \left\{\boldsymbol{\pi} = [\pi_1,\dots,\pi_m] : \pi_1^{\mathsf{T}}\mathbb{1} = \dots = \pi_m^{\mathsf{T}}\mathbb{1}\right\}.\end{aligned} \tag{7}$$

The IBP algorithm (Algorithm 1) consists in alternating projections to the sets $\mathcal{C}_1$ and $\mathcal{C}_2$ w.r.t. Kullback–Leibler divergence, and is a generalization of Sinkhorn's algorithm and a particular case of Dykstra's projection algorithm. This algorithm is equivalent to alternating minimization of the dual problem of (5) derived in Lemma 1, and leads to Algorithm 2.

[1] In the original paper Benamou et al. [2015] there were misprints in description of IBP. Correct author's variant can be found in `https://github.com/gpeyre/2014-SISC-BregmanOT`. In Algorithm 2 we use different notations. Firstly, our u corresponds to $\ln v$ from Benamou et al. [2015] and our v corresponds to $\ln u$ from Benamou et al. [2015]. Secondly, our transport plan matrix equals to transpose transport plan matrix from Benamou et al. [2015]. Thirdly, we build a little bit different dual problem, by introducing additional constraint $\sum_{l=1}^{m} w_l v_l = 0$, see Lemma 1. This allows us to simplify calculations in line 3 of Algorithm 2.

Algorithm 1 Iterative Bregman Projections, see Benamou et al. [2015]

Input: Cost matrices $C_1, \dots, C_m$, probability measures $p_1, \dots, p_m$, $\gamma > 0$, starting transport plans $\{\pi_l^0\}_{l=1}^m : \pi_l^0 \coloneqq \exp\left(-\frac{C_l}{\gamma}\right)$, $l = 1, \dots, m$

1: **for** t=0,1,... **do**
2: **if** $t \bmod 2 = 0$ **then**
3: $\boldsymbol{\pi}^{t+1} \coloneqq \underset{\boldsymbol{\pi} \in \mathcal{C}_1}{\operatorname{argmin}} \sum_{l=1}^m w_l KL\left(\pi_l | \pi_l^t\right)$
4: **else**
5: $\boldsymbol{\pi}^{t+1} \coloneqq \underset{\boldsymbol{\pi} \in \mathcal{C}_2}{\operatorname{argmin}} \sum_{l=1}^m w_l KL\left(\pi_l | \pi_l^t\right)$
6: **end if**
7: $t \coloneqq t + 1$
8: **end for**

Output: $\boldsymbol{\pi}^t$

Lemma 1. *The dual problem of* (5) *is (up to a multiplicative constant)*

$$\min_{\substack{\mathtt{u},\mathtt{v} \\ \sum_{l=1}^m w_l v_l = 0}} f(\mathtt{u}, \mathtt{v}) \coloneqq \sum_{l=1}^m w_l \big\{ \langle \mathbb{1}, B_l(u_l, v_l)\mathbb{1} \rangle - \langle u_l, p_l \rangle \big\}, \tag{8}$$

$\mathtt{u} = [u_1, \dots, u_m]$, $\mathtt{v} = [v_1, \dots, v_m]$, $u_l, v_l \in \mathbb{R}^n$, *and*

$$B_l(u_l, v_l) \coloneqq \operatorname{diag}\left(e^{u_l}\right) K_l \operatorname{diag}\left(e^{v_l}\right), K_l \coloneqq \exp\left(-\frac{C_l}{\gamma}\right).$$

Moreover, solution $\boldsymbol{\pi}_\gamma^*$ *to* (5) *is given by the formula*

$$[\boldsymbol{\pi}_\gamma^*]_l = B_l(u_l^*, v_l^*), \tag{9}$$

where $(\mathtt{u}^*, \mathtt{v}^*)$ *is the solution of the problem* (8).

Proof. The Lagrangian for (5) is equal to

$$\begin{aligned} L(\boldsymbol{\pi}; \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_{l=1}^m w_l \big\{ \langle \pi_l, C_l \rangle + \gamma H(\pi_l) \big\} + \sum_{l=1}^m \langle \lambda_l, \pi_l \mathbb{1} - p_l \rangle + \sum_{l=1}^m \langle \mu_l, \pi_{l+1}^\mathsf{T} \mathbb{1} - \pi_l^\mathsf{T} \mathbb{1} \rangle \\ &= \sum_{l=1}^m \left[w_l \big\{ \langle \pi_l, C_l \rangle + \gamma \langle \pi_l, \ln \pi_l - \mathbb{1}\mathbb{1}^\mathsf{T} \rangle \big\} + \langle \lambda_l, \pi_l \mathbb{1} - p_l \rangle + \langle \mu_{l-1} - \mu_l, \pi_l^\mathsf{T} \mathbb{1} \rangle \right], \end{aligned}$$

where $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_m]$, $\boldsymbol{\mu} = [\mu_1, \dots, \mu_m]$, $\lambda_l, \mu_l \in \mathbb{R}^n$ with convention $\mu_0 \equiv \mu_m \equiv 0$. Using the change of variables $u_l \coloneqq -\lambda_l / (w_l \gamma)$ and $v_l \coloneqq (\mu_l - \mu_{l-1}) / (w_l \gamma)$ we obtain

$$L(\boldsymbol{\pi}; \mathtt{u}, \mathtt{v}) = \gamma \sum_{l=1}^m w_l \left\{ \left\langle \pi_l, \frac{C_l}{\gamma} + \ln \pi_l - \mathbb{1}\mathbb{1}^\mathsf{T} - u_l \mathbb{1}^\mathsf{T} - \mathbb{1} v_l^\mathsf{T} \right\rangle + \langle u_l, p_l \rangle \right\}. \tag{10}$$

Notice that $\sum_{l=1}^m w_l v_l = 0$ and this condition allows uniquely reconstruct $\mu_1, \dots, \mu_m$. Then by min-max theorem

$$\begin{aligned}
&\min_{\substack{\boldsymbol{\pi}:\pi_l\in\mathbb{R}_+^{n\times n}\\ \pi_l\mathbb{1}=p_l,\ \pi_l^\mathsf{T}\mathbb{1}=\pi_{l+1}^\mathsf{T}\mathbb{1}}} \sum_{l=1}^m w_l\big\{\langle \pi_l, C_l\rangle + \gamma H(\pi_l)\big\} \\
&\qquad = \min_{\boldsymbol{\pi}:\pi_l\in\mathbb{R}_+^{n\times n}} \max_{\substack{\mathtt{u},\mathtt{v}\\ \sum_l w_l v_l = 0}} L(\boldsymbol{\pi};\mathtt{u},\mathtt{v}) = \max_{\substack{\mathtt{u},\mathtt{v}\\ \sum_l w_l v_l = 0}} \min_{\boldsymbol{\pi}:\pi_l\in\mathbb{R}_+^{n\times n}} L(\boldsymbol{\pi};\mathtt{u},\mathtt{v}) \\
&\qquad = \max_{\substack{\mathtt{u},\mathtt{v}\\ \sum_l w_l v_l = 0}} \gamma \sum_{l=1}^m w_l \left\{ \min_{\pi_l\in\mathbb{R}_+^{n\times n}} \left\langle \pi_l, \frac{C_l}{\gamma} + \ln \pi_l - \mathbb{1}\mathbb{1}^\mathsf{T} - u_l\mathbb{1}^\mathsf{T} - \mathbb{1}v_l^\mathsf{T} \right\rangle + \langle u_l, p_l\rangle \right\}.
\end{aligned}$$

By straightforward computations and the definition (1) of $B_l(u_l, v_l)$ we obtain

$$\begin{aligned}
&\min_{\pi_l\in\mathbb{R}_+^{n\times n}} \left\langle \pi_l, \frac{C_l}{\gamma} + \ln \pi_l - \mathbb{1}\mathbb{1}^\mathsf{T} - u_l\mathbb{1}^\mathsf{T} - \mathbb{1}v_l^\mathsf{T} \right\rangle \\
&\qquad\qquad = -\left\langle \mathbb{1}, \exp\left(u_l\mathbb{1}^\mathsf{T} - \frac{C_l}{\gamma} + \mathbb{1}v_l^\mathsf{T}\right)\mathbb{1} \right\rangle = -\left\langle \mathbb{1}, B_l(u_l, v_l)\mathbb{1}\right\rangle,
\end{aligned}$$

and the minimum is attained at point $\pi_l = B_l(u_l, v_l)$. Thus we have

$$\begin{aligned}
\min_{\substack{\boldsymbol{\pi}:\pi_l\in\mathbb{R}_+^{n\times n}\\ \pi_l\mathbb{1}=p_l,\ \pi_l^\mathsf{T}\mathbb{1}=\pi_{l+1}^\mathsf{T}\mathbb{1}}} \sum_{l=1}^m w_l\big\{\langle \pi_l, C_l\rangle + \gamma H(\pi_l)\big\} &= \max_{\substack{\mathtt{u},\mathtt{v}\\ \sum_l w_l v_l = 0}} \gamma \sum_{l=1}^m w_l \left\{-\langle \mathbb{1}, B_l(u_l, v_l)\mathbb{1}\rangle + \langle u_l, p_l\rangle + 1\right\} \\
&= -\gamma \min_{\substack{\mathtt{u},\mathtt{v}\\ \sum_l w_l v_l = 0}} \sum_{l=1}^m w_l \left\{\langle \mathbb{1}, B_l(u_l, v_l)\mathbb{1}\rangle - \langle u_l, p_l\rangle\right\}.
\end{aligned}$$

Consequently, the dual problem to (5) is equivalent to (8), and solution to (5) has the form $[\boldsymbol{\pi}_\gamma^*]_l = B_l(u_l^*, v_l^*)$. □

The following lemma shows that Algorithms 1 is equivalent to alternating minimization in dual problem (8) (what is a general fact for Dykstra's algorithm).

Lemma 2. *Sequence $\{\boldsymbol{\pi}^t\}_{t\geq 0}$ generated by Algorithm 1 has a form $\pi_l^t = B_l(u_l^t, v_l^t)$, where $B_l(\cdot,\cdot)$ is defined in* (1) *and*

$$\begin{aligned}
&\mathtt{u}^0 = \mathtt{v}^0 = 0, \\
&\mathtt{u}^{t+1} := \operatorname*{argmin}_{\mathtt{u}} f(\mathtt{u}, \mathtt{v}^t), \qquad && \mathtt{v}^{t+1} := \mathtt{v}^t \qquad && t \bmod 2 = 0, \qquad (11)\\
&\mathtt{v}^{t+1} := \operatorname*{argmin}_{\mathtt{v}:\sum_{l=1}^m w_l v_l = 0} f(\mathtt{u}^t, \mathtt{v}), \qquad && \mathtt{u}^{t+1} := \mathtt{u}^t \qquad && t \bmod 2 = 1. \qquad (12)
\end{aligned}$$

Proof. Let us prove it by induction. For $t = 0$ it is obviously true. Assume it holds for some $t \geq 0$. Then

$$KL\left(\pi_l|\pi_l^t\right) = H(\pi_l) - \langle \pi_l, \ln \pi_l^t\rangle + \langle \pi_l^t, \mathbb{1}\mathbb{1}^\mathsf{T}\rangle = H(\pi_l) + \left\langle \pi_l, \frac{C_l}{\gamma} - u_l^t\mathbb{1}^t - \mathbb{1}(v_l^t)^\mathsf{T}\right\rangle + \langle \pi_l^t, \mathbb{1}\mathbb{1}^\mathsf{T}\rangle.$$

Therefore,

$$\sum_{l=1}^{m} w_l KL\left(\pi_l | \pi_l^t\right) \to \min_{\boldsymbol{\pi} \in \mathcal{C}}$$

$$\iff \sum_{l=1}^{m} w_l \left(\langle C_l, \pi_l \rangle - \gamma \langle \pi_l \mathbb{1} - p_l, u_l^t \rangle - \gamma \langle \pi_l^\mathsf{T} \mathbb{1}, v_l^t \rangle\right) = L(\boldsymbol{\pi}; \mathtt{u}^t, \mathtt{v}^t) \to \min_{\boldsymbol{\pi} \in \mathcal{C}},$$

where L comes from (10).

Thus for even t

$$\boldsymbol{\pi}^{t+1} = \operatorname*{argmin}_{\boldsymbol{\pi} \in \mathcal{C}_1} L(\boldsymbol{\pi}; \mathtt{u}^t, \mathtt{v}^t),$$

Lagrangian for this problem has form $L(\boldsymbol{\pi}, \mathtt{u}, \mathtt{v}^t)$, hence the dual problem is

$$\gamma f(\mathtt{u}, \mathtt{v}^t) \to \min_{\mathtt{v}}.$$

and as $\mathtt{u}^{t+1} = \operatorname{argmin}_{\mathtt{u}} f(\mathtt{u}, \mathtt{v}^t)$,

$$\pi_l^{t+1} = B_l(u_l^{t+1}, v_l^t).$$

Similarly, for odd t

$$\boldsymbol{\pi}^{t+1} = \operatorname*{argmin}_{\boldsymbol{\pi} \in \mathcal{C}_2} L(\boldsymbol{\pi}; \mathtt{u}^t, \mathtt{v}^t),$$

with Lagrangian $L(\boldsymbol{\pi}, \mathtt{u}^t, \mathtt{v})$ and dual problem

$$\gamma f(\mathtt{u}^t, \mathtt{v}) \to \min_{\mathtt{v}: \sum_{l=1}^m w_l v_l = 0}.$$

Consequently,

$$\pi_l^{t+1} = B_l(u_l^{t+1}, v_l^t).$$

□

The next lemma gives us explicit recurrent expressions for $\mathtt{u}^t$ and $\mathtt{v}^t$ defined in the previous lemma. Equation (13) immediately follows from [Benamou et al., 2015, Proposition 1], and equation (14) is a reformulation of [Benamou et al., 2015, Proposition 2].

Lemma 3. *Equation* (11) *for even t is equivalent to*

$$u_l^{t+1} = u_l^t + \ln p_l - \ln\left(B_l\left(\mathtt{u}^t, \mathtt{v}^t\right)\mathbb{1}\right) = \ln p_l - \ln K_l e^{v_l^t}, \tag{13}$$

and equation (12) *for odd t is equivalent to*

$$v_l^{t+1} = v_l^t + \ln q^{t+1} - \ln q_l^t = \sum_{k=1}^{m} w_k \ln K_k^\mathsf{T} e^{u_k^t} - \ln K_l^\mathsf{T} e^{u_l^t}, \tag{14}$$

where $q_l^t \coloneqq B_l^\mathsf{T}(u_l^t, v_l^t)\mathbb{1}$, $q^{t+1} \coloneqq \exp\left(\sum_{l=1}^m w_l \ln q_l^t\right)$.

Algorithm 2 Dual Iterative Bregman Projection

Input: $C_1, \ldots, C_m$, $p_1, \ldots, p_m$, $\gamma > 0$, $\varepsilon' > 0$

1: $u_l^0 := 0$, $v_l^0 := 0$, $K_l := \exp\left(-\frac{C_l}{\gamma}\right)$, $l = 1, \ldots, m$
2: **repeat**
3: **if** $t \bmod 2 = 0$ **then**
4: $u_l^{t+1} := \ln p_l - \ln K_l e^{v_l^t}, \quad \mathtt{v}^{t+1} := \mathtt{v}^t$
5: **else**
6: $v_l^{t+1} := \sum_{k=1}^m w_k \ln K_k^\mathsf{T} e^{u_k^t} - \ln K_l^\mathsf{T} e^{u_l^t}, \quad \mathtt{u}^{t+1} := \mathtt{u}^t$
7: **end if**
8: $t := t + 1$
9: **until** $\sum_{l=1}^m w_l \|B_l^\mathsf{T}(u_l^t, v_l^t)\mathbb{1} - \bar{q}^t\|_1 \leq \varepsilon'$ and $\sum_{l=1}^m w_l \|B_l(u_l^t, v_l^t)\mathbb{1} - p_l\|_1 \leq \varepsilon'$, where $\bar{q}^t := \sum_{l=1}^m w_l B_l^\mathsf{T}(u_l^t, v_l^t)\mathbb{1}$

Output: $B_1(u_1^t, v_1^t), \ldots, B_m(u_m^t, v_m^t)$

Before we move to the analysis of the algorithm let us discuss the scalability of this algorithm by using centralized distributed computations framework. This framework includes a master and slave nodes. Each l-th slave node stores data p_l, C_l, K_l and variables u_l^t, v_l^t. On each iteration t, it calculates $K_l^\mathsf{T} e^{u_l^t}$ and sends it to the master node, which aggregates these products to the sum $\sum_{k=1}^m w_k \ln K_k^\mathsf{T} e^{u_k^t}$ and sends this sum back to the slave nodes. Based on this information, slave nodes update v_l^t and u_l^t. So, the main computational cost of multiplying a matrix by a vector, can be distributed on m slave nodes and the total working time will be smaller. It is not clear, how this algorithm can be implemented on a general network, for example when the data is produced by a distributed network of sensors without one master node. In contrast, as we illustrate in Section 3, the alternative accelerated-gradient-based approach can be implemented on an arbitrary network.

Theorem 1. *For given ε' Algorithm 2 stops in number of iterations N satisfying*

$$N \leq 4 + \frac{88 \max_l \|C_l\|_\infty}{\gamma \varepsilon'} = O\left(\frac{\max_l \|C_l\|_\infty}{\gamma \varepsilon'}\right).$$

It returns $B_1, \ldots, B_m$ s.t.

$$\sum_{l=1}^m w_l \|B_l^\mathsf{T}\mathbb{1} - \bar{q}\|_1 \leq \varepsilon', \quad \bar{q} = \sum_{l=1}^m w_l B_l \mathbb{1},$$

and

$$\sum_{l=1}^m w_l \left(\langle C_l, B_l\rangle + \gamma H(B_l)\right) - \sum_{l=1}^m w_l \left(\left\langle C_l, \pi^*_{\gamma,l}\right\rangle + \gamma H(\pi^*_{\gamma,l})\right) \leq \max_l \|C_l\|_\infty \varepsilon', \tag{15}$$

where $\boldsymbol{\pi}^*_\gamma = [\pi^*_{\gamma,1}, \ldots, \pi^*_{\gamma,m}]$ is the solution of problem (5).

Next lemmas are preliminaries to the proof of correctness and complexity bound for Algorithm 2.

The first lemma shows that dual potentials $e^{\mathtt{u}^t}$, $e^{\mathtt{v}^t}$ are bounded in Hilbert projective metric. However, we cannot prove that iterations of IBP are contracting, cf. Franklin and Lorenz [1989].

Lemma 4. *For any $t \geq 0$, $l = 1, \ldots, m$ it holds*

$$\max_j [v_l^t]_j - \min_j [v_l^t]_j \leq R_v, \qquad \max_j [v_l^*]_j - \min_j [v_l^*]_j \leq R_v,$$

where

$$R_v := \max_l \frac{\|C_l\|_\infty}{\gamma} + \sum_{k=1}^m w_k \frac{\|C_k\|_\infty}{\gamma}. \tag{16}$$

Proof. Bound can be derived in almost the same way as in Dvurechensky et al. [2018b]. For $t = 0$ it obviously holds. Let us denote by ν_l the minimal entry of K_l:

$$\nu_l := \min_{i,j} [K_l]_{ij} = e^{-\|C_l\|_\infty/\gamma}.$$

As $[K_l]_{ij} \leq 1$, we obtain for all $j = \overline{1,n}$

$$\ln \nu_l + \ln\langle \mathbb{1}, e^{u_l} \rangle \leq [\ln K_l^\mathsf{T} e^{u_l}]_j \leq \ln\langle \mathbb{1}, e^{u_l} \rangle,$$

therefore

$$\max_j [\ln K_l^\mathsf{T} e^{u_l}]_j - \min_j [\ln K_l^\mathsf{T} e^{u_l}]_j \leq -\ln \nu_l = \frac{\|C_l\|_\infty}{\gamma}.$$

Hence at any update of $\mathtt{v}^t$ it holds

$$\max_j [v_l^t]_j - \min_j [v_l^t]_j \leq \frac{\|C_l\|_\infty}{\gamma} + \sum_{k=1}^m w_k \frac{\|C_k\|_\infty}{\gamma} \leq R_v.$$

For solution to the problem $(\mathtt{u}^*, \mathtt{v}^*)$ condition (14) also holds, and consequently the solution also meets derived bounds. □

Let us define an excess function

$$\tilde{f}(\mathtt{u}, \mathtt{v}) := f(\mathtt{u}, \mathtt{v}) - f(\mathtt{u}^*, \mathtt{v}^*)$$

for further complexity analysis of Algorithm 2.

Lemma 5. *Let $\{\mathtt{u}^t, \mathtt{v}^t\}_{t \geq 0}$ be generated by Algorithm 2. Then for any even $t \geq 2$ we have*

$$\tilde{f}(\mathtt{u}^t, \mathtt{v}^t) \leq R_v \sum_{l=1}^m w_l \|q_l^t - \bar{q}^t\|_1. \tag{17}$$

Proof. Gradient inequality of any convex function g at point x^* reads as

$$g(x^*) \geq g(x) + \langle \nabla g(x), x^* - x \rangle, \quad \forall x \in \mathrm{dom}(g).$$

Applying the latter inequality to function f at point $(\mathtt{u}^*, \mathtt{v}^*)$ we obtain

$$\tilde{f}(\mathtt{u}^t, \mathtt{v}^t) = f(\mathtt{u}^t, \mathtt{v}^t) - f(\mathtt{u}^*, \mathtt{v}^*) \leq \sum_{l=1}^m w_l \langle u_l^t - u_l^*, B_l(u_l^t, v_l^t)\mathbb{1} - p_l \rangle + \sum_{l=1}^m w_l \langle v_l^t - v_l^*, q_l^t \rangle.$$

If $t \geq 2$ is even, then the first term in r.h.s. vanishes. Notice that $\langle q_l^t, \mathbb{1} \rangle = \langle \mathbb{1}, B_l(u_l^t, v_l^t)\mathbb{1} \rangle = \langle \mathbb{1}, p_l \rangle = 1$, thus

$$\begin{aligned}
\tilde{f}(\mathtt{u}^t, \mathtt{v}^t) &= \sum_{l=1}^m w_l \langle v_l^t - v_l^*, q_l^t \rangle = \sum_{l=1}^m w_l \langle v_l^t - v_l^*, q_l^t - \bar{q}^t \rangle \\
&= \sum_{l=1}^m w_l \langle (v_l^t - b_l^t \mathbb{1}) - (v_l^* - b_l^* \mathbb{1}), q_l^t - \bar{q}^t \rangle \\
&\leq \sum_{l=1}^m w_l \left(\|v_l^t - b_l^t \mathbb{1}\|_\infty + \|v_l^* - b_l^* \mathbb{1}\|_\infty \right) \|q_l^t - \bar{q}^t\|_1,
\end{aligned}$$

where

$$b_l^t \coloneqq \frac{\min_i [v_l^t]_i + \max_i [v_l^t]_i}{2}, \quad b_l^* \coloneqq \frac{\min_i [v_l^*]_i + \max_i [v_l^*]_i}{2}.$$

By Lemma 4 $\|v_l^t - b_l^t \mathbb{1}\|_\infty \leq R_v/2$ and $\|v_l^* - b_l^* \mathbb{1}\|_\infty \leq R_v/2$, therefore

$$\tilde{f}(\mathtt{u}^t, \mathtt{v}^t) \leq R_v \sum_{l=1}^m w_l \|q_l^t - \bar{q}^t\|_1.$$

□

Lemma 6. *For any odd $t \geq 1$ the following bound on the change of objective function $f(\cdot, \cdot)$ holds:*

$$f(\mathtt{u}^t, \mathtt{v}^t) - f(\mathtt{u}^{t+1}, \mathtt{v}^{t+1}) \geq \frac{1}{11} \sum_{l=1}^m w_l \|q_l^t - \bar{q}^t\|_1^2,$$

where $\bar{q}^t \coloneqq \sum_{l=1}^m w_l q_l^t$.

Proof. If t is odd, then $\mathtt{v}^{t+1}$ satisfies (14) and $\mathtt{u}^t = \mathtt{u}^{t+1}$. Therefore,

$$\begin{aligned} f(\mathtt{u}^t, \mathtt{v}^t) - f(\mathtt{u}^{t+1}, \mathtt{v}^{t+1}) &= \sum_{l=1}^m w_l \big\{ \langle q_l^t, \mathbb{1} \rangle - \langle q_l^{t+1}, \mathbb{1} \rangle \big\} = \big\langle \bar{q}^t - q^{t+1}, \mathbb{1} \big\rangle \\ &= \left\langle \bar{q}^t - \exp\left(\sum_{l=1}^m w_l q_l^t \right), \mathbb{1} \right\rangle \geq \frac{4}{11} \sum_{j=1}^n \frac{1}{[\bar{q}^t]_j} \sum_{l=1}^m w_l \big([q_l^t - \bar{q}^t]_j^- \big)^2 \\ &= \frac{4}{11} \sum_{l=1}^m w_l \sum_{j=1}^n \frac{\big([q_l^t - \bar{q}^t]_j^- \big)^2}{[\bar{q}^t]_j} \geq \frac{4}{11} \sum_{l=1}^m w_l \frac{\Big(\sum_j [q_l^t - \bar{q}^t]_j^- \Big)^2}{\sum_{j=1}^n [\bar{q}^t]_j} \\ &= \frac{1}{11} \sum_{l=1}^m w_l \| q_l^t - \bar{q}^t \|_1^2. \end{aligned}$$

Here we used equations $\langle q_l^t, \mathbb{1} \rangle = \langle \mathbb{1}, p_l \rangle = 1$, i.e. $q_l^t \in S_n(1)$ and thus $\bar{q}^t \in S_n(1)$, and the following fact: if $x \in \mathbb{R}_+^m$, $\bar{x} \coloneqq \sum_{l=1}^m w_l x_l$, then

$$\bar{x} - \prod_{l=1}^m x_l^{w_l} \geq \frac{4}{11} \sum_{l=1}^m w_l \frac{\big[(x_l - \bar{x})^- \big]^2}{\bar{x}}.$$

Indeed, let $\Delta_l \coloneqq x_l - \bar{x}$, then

$$\bar{x} - \prod_{l=1}^m x_l^{w_l} = \bar{x} - \exp\left\{ \sum_{l=1}^m w_l \ln(\bar{x} + \Delta_l) \right\} = \bar{x} \left(1 - \exp\left\{ \sum_{l=1}^m w_l \ln\left(1 + \frac{\Delta_l}{\bar{x}} \right) \right\} \right),$$

$$\sum_{l=1}^m w_l \ln\left(1 + \frac{\Delta_l}{\bar{x}} \right) \leq \sum_{l=1}^m w_l \left(\frac{\Delta_l}{\bar{x}} - \frac{(\Delta_l^-)^2}{2\bar{x}^2} \right) = - \sum_{l=1}^m w_l \frac{(\Delta_l^-)^2}{2\bar{x}^2}.$$

Notice that $\Delta_l^- \coloneqq \max\{-\Delta_l, 0\} = \max\{\bar{x} - x_l, 0\} \leq \bar{x}$, thus $\sum_{l=1}^m w_l \frac{(\Delta_l^-)^2}{\bar{x}^2} \leq 1$ and

$$\exp\left\{ -\frac{1}{2} \sum_{l=1}^m w_l \frac{(\Delta_l^-)^2}{\bar{x}^2} \right\} \leq 1 - \big(1 - e^{-1/2} \big) \sum_{l=1}^m w_l \frac{(\Delta_l^-)^2}{\bar{x}^2}.$$

Consequently,

$$\bar{x} - \prod_{l=1}^m x_l^{w_l} \geq \big(1 - e^{-1/2} \big) \sum_{l=1}^m w_l \frac{(\Delta_l^-)^2}{\bar{x}} \geq \frac{4}{11} \sum_{l=1}^m w_l \frac{(\Delta_l^-)^2}{\bar{x}}.$$

□

Proof of Theorem 1. First, notice that

$$H\big(B_l(u_l, v_l) \big) = \langle u_l, B_l(u_l, v_l) \mathbb{1} \rangle + \langle v_l, B_l^\mathsf{T}(u_l, v_l) \mathbb{1} \rangle - \langle \mathbb{1}, B_l(u_l, v_l) \mathbb{1} \rangle - \frac{1}{\gamma} \langle C_l, B_l(u_l, v_l) \rangle.$$

Since N is even, $B_l\mathbb{1} = p_l$ and therefore it holds

$$\begin{aligned}\sum_{l=1}^{m} w_l \left(\langle C_l, B_l\rangle + \gamma H(B_l)\right) &= \gamma \sum_{l=1}^{m} w_l \left(\langle u_l^N, B_l\mathbb{1}\rangle + \langle v_l^N, B_l^\mathsf{T}\mathbb{1}\rangle - \langle \mathbb{1}, B_l\mathbb{1}\rangle\right) \\ &= -\gamma \sum_{l=1}^{m} w_l \left(\langle \mathbb{1}, B_l\mathbb{1}\rangle - \langle u_l^N, p_l\rangle\right) + \gamma \sum_{l=1}^{m} w_l \langle v_l^N, q_l^N\rangle \\ &= -\gamma f(\mathtt{u}^N, \mathtt{v}^N) + \gamma \sum_{l=1}^{m} w_l \langle v_l^N, q_l^N - \bar{q}^N\rangle.\end{aligned}$$

Now Lemmas 1, 4, 5, and stopping criterion yield

$$\begin{aligned}&\sum_{l=1}^{m} w_l \left(\langle C_l, B_l\rangle + \gamma H(B_l)\right) - \sum_{l=1}^{m} w_l \left(\langle C_l, \pi_l^*\rangle + \gamma H(\pi_{\gamma,l}^*)\right) \\ &\qquad = \gamma\big(f(\mathtt{u}^*, \mathtt{v}^*) - f(\mathtt{u}^N, \mathtt{v}^N)\big) + \gamma \sum_{l=1}^{m} w_l \langle v_l^N, q_l^N - \bar{q}^N\rangle \\ &\qquad \le \gamma \sum_{l=1}^{m} w_l \frac{R_v}{2} \|q_l^N - \bar{q}^N\|_1 \le \frac{\gamma R_v}{2}\varepsilon' \le \max_l \|C_l\|_\infty \varepsilon'.\end{aligned}$$

Now let us prove the complexity bound. We will do it in two steps.

1. If $t \ge 2$ is even, then by Lemma 5

$$\tilde{f}(\mathtt{u}^t, \mathtt{v}^t) \le R_v \sum_{l=1}^{m} w_l \|q_l^t - \bar{q}^t\|_1$$

and since stopping criterion is not fulfilled,

$$\sum_{l=1}^{m} w_l \|q_l^t - \bar{q}^t\|_1 > \varepsilon'.$$

Inequality $\sum_{l=1}^{m} w_l \|q_l^t - \bar{q}^t\|_1^2 \ge \left(\sum_{l=1}^{m} w_l \|q_l^t - \bar{q}^t\|_1\right)^2$ together with Lemma 6 give us the following bound:

$$\begin{aligned}\tilde{f}(\mathtt{u}^t, \mathtt{v}^t) - \tilde{f}(\mathtt{u}^{t+1}, \mathtt{v}^{t+1}) &= f(\mathtt{u}^t, \mathtt{v}^t) - f(\mathtt{u}^{t+1}, \mathtt{v}^{t+1}) \\ &\ge \frac{1}{11} \max\left\{(\varepsilon')^2, \left(\frac{\tilde{f}(\mathtt{u}^t, \mathtt{v}^t)}{R_v}\right)^2\right\} = \max\left\{\frac{(\varepsilon')^2}{11}, \frac{1}{11R_v^2}\tilde{f}^2(\mathtt{u}^t, \mathtt{v}^t)\right\}.\end{aligned}$$

If t is odd then we have at least $\tilde{f}(\mathtt{u}^{t+1}, \mathtt{v}^{t+1}) \le \tilde{f}(\mathtt{u}^t, \mathtt{v}^t)$. To simplify derivation we define

$$\delta_t \coloneqq \tilde{f}(\mathtt{u}^t, \mathtt{v}^t). \tag{18}$$

Now we have two possibilities to estimate number of iteration. The first one is based on inequalities

$$\frac{1}{\delta_{t+1}} \geq \begin{cases} \frac{1}{\delta_t} + \frac{1}{11R_v^2}\frac{\delta_t}{\delta_{t+1}} \geq \frac{1}{\delta_t} + \frac{1}{11R_v^2}, & t \bmod 2 = 0, \\ \frac{1}{\delta_t}, & t \bmod 2 = 1. \end{cases}$$

Summation of these inequalities gives

$$\frac{1}{\delta_t} \geq \frac{1}{\delta_1} + \frac{t-2}{22R_v^2}$$

and hence

$$t \leq 2 + 22R_v^2\left(\frac{1}{\delta_t} - \frac{1}{\delta_1}\right). \tag{19}$$

The second estimate can be obtained from

$$\delta_{t+1} \leq \begin{cases} \delta_t - \frac{(\varepsilon')^2}{11}, & t \bmod 2 = 0, \\ \delta_t, & t \bmod 2 = 1. \end{cases}$$

Similarly, summation of these inequalities gives

$$\delta_t \geq \delta_t - \delta_{t+k} \geq \frac{k-1}{22}(\varepsilon')^2, \tag{20}$$

$$k \leq 1 + \frac{22\delta_t}{(\varepsilon')^2}. \tag{21}$$

2. To combine the two estimates (19) and (20), we consider a switching strategy parametrized by number $s \in (0, \delta_1)$. First t iterations we use (19), resulting in δ_t becomes below some s. Then, we use s as a starting point and estimate the remaining number of iteration by (20). The quantity s can be found from the minimization

$$N = t + k \leq 4 + \frac{2s}{(\varepsilon')^2} + 22R_v^2\left(\frac{1}{s} - \frac{1}{\delta_1}\right).$$

Minimizing the r.h.s. of the latter inequality in s leads to

$$N \leq \min_{0\leq s\leq\delta_1}\left\{4 + \frac{22s}{(\varepsilon')^2} + 22R_v^2\left(\frac{1}{s} - \frac{1}{\delta_1}\right)\right\} \leq 4 + \frac{44R_v}{\varepsilon'}. \tag{22}$$

The last inequality is obtained by the substitution $s = R_v\varepsilon'$ that is the solution to the minimization problem. Of course, the switching strategy is impossible if $\delta_1 < s$. But in this case (20) gives

$$N \leq 2 + \frac{22\delta_1}{(\varepsilon')^2} < 4 + \frac{44R_v}{\varepsilon'}. \tag{23}$$

In both cases (22) and (23) we have

$$N \leq 4 + \frac{44R_v}{\varepsilon'} \leq 4 + \frac{88\max_l\|C_l\|_\infty}{\gamma\varepsilon'} = O\left(\frac{\max_l\|C_l\|_\infty}{\gamma\varepsilon'}\right).$$

□

2.2 Approximating Non-regularized WB by IBP

To find an approximate non-regularized barycenter, i.e. solution to problem (4) with $\gamma = 0$, we apply Algorithm 2 with a suitable choice of γ and ε' and average marginals $q_1, \ldots, q_m$ with weights w_l, this leads to Algorithm 3.

Algorithm 3 Finding Wasserstein barycenter by IBP

Input: Accuracy ε; cost matrices $C_1, \ldots, C_m$; marginals $p_1, \ldots, p_m$

1: Set $\gamma := \frac{1}{4}\frac{\varepsilon}{\ln n}, \quad \varepsilon' := \frac{1}{4}\frac{\varepsilon}{\max_l \|C_l\|_\infty}$

2: Find $B_1 := B_1(u_1^t, v_1^t), \ldots, B_m := B_m(u_m^t, v_m^t)$ by Algorithm 2 with accuracy ε'

3: $q := \frac{1}{\sum_{l=1}^m w_l \langle \mathbb{1}, B_l \mathbb{1} \rangle} \sum_{l=1}^m w_l B_l^\mathsf{T} \mathbb{1}$

Output: q

Next theorem presents complexity bound for Algorithm 3.

Theorem 2. *For $\varepsilon > 0$, Algorithm 3 returns $q \in S_n(1)$ s.t.*

$$\sum_{l=1}^m w_l \mathcal{W}(p_l, q) - \sum_{l=1}^m w_l \mathcal{W}(p_l, q^*) \leq \varepsilon,$$

where q^* is a solution to non-regularized problem (4) with $\gamma = 0$. Moreover, it requires

$$O\left(\left(\frac{\max_l \|C_l\|_\infty}{\varepsilon}\right)^2 M_{m,n} \ln n + mn\right)$$

arithmetic operations, where $M_{m,n}$ is a time complexity of one iteration of Algorithm 2.

Remark 1. As each iteration of Algorithm 2 requires m matrix-vector multiplications, the general bound is $M_{m,n} = O(mn^2)$. However, for some specific form of matrices C_l it's possible to achieve better complexity, e.g. $M_{m,n} = O(mn \log n)$ via FFT[2] Peyré and Cuturi [2018], or $M_{m,n} = O\left(n \sum_l \operatorname{rank}(C_l)\right)$ for low-rank matrices.

Proof. Let $\boldsymbol{\pi}_\gamma^* = [\pi_{\gamma,1}^*, \ldots, \pi_{\gamma,m}^*]$ be a solution to (5) and $\boldsymbol{\pi}^* = [\pi_1^*, \ldots, \pi_m^*]$ be a solution to the non-regularized problem. Theorem 1 yields

$$\begin{aligned}
\sum_{l=1}^m w_l \langle C_l, B_l \rangle &\leq \sum_{l=1}^m w_l \left(\langle C_l, \pi_{\gamma,l}^* \rangle + \gamma H(\pi_{\gamma,l}^*) - \gamma H(B_l)\right) + \max_l \|C_l\|_\infty \varepsilon' \\
&\leq \sum_{l=1}^m w_l \left(\langle C_l, \pi_l^* \rangle + \gamma H(\pi_l^*) - \gamma H(B_l)\right) + \max_l \|C_l\|_\infty \varepsilon' \\
&\leq \sum_{l=1}^m w_l \mathcal{W}(p_l, q^*) + 2\gamma \ln n + \max_l \|C_l\|_\infty \varepsilon'.
\end{aligned}$$

[2]it is stable only for large enough γ, what is the case for proximal method, see Subsection 2.3

Here we used inequalities $-2\ln n \leq H(\pi)+1 \leq 0$ holding on $S_{n\times n}(1)$. Consider $\check{B}_l \in \Pi(p_l, q)$ s.t. $\|\check{B}_l - B_l\|_1 \leq \|B_l\mathbb{1} - p_l\|_1 + 2\sum_j [B_l^\mathsf{T}\mathbb{1} - q]_j^+$ for all $l = 1, \ldots, m$. Their existence immediately follows from the proof of Theorem 4 from Altschuler et al. [2017]. If stopping time t is even, $B_l\mathbb{1} = p_l$, therefore $q = \vec{q}^t$, and $\|B_l - \check{B}_l\|_1 \leq \|q_l^t - \vec{q}^t\|_1$. If t is odd, $B_l^\mathsf{T}\mathbb{1} = \vec{q}^t \leq q$ and $\|B_l - \check{B}_l\|_1 \leq \|B_l\mathbb{1} - p_l\|_1$. In both cases it follows from stopping criterion that

$$\sum_{l=1}^m w_l \|B_l - \check{B}_l\|_1 \leq \varepsilon'.$$

Since $\check{B}_l \in \Pi(p_l, q)$ for all $1 \leq l \leq m$, one has

$$\begin{aligned}
\sum_{l=1}^m w_l \mathcal{W}(p_l, q) &\leq \sum_{l=1}^m w_l \langle C_l, \check{B}_l \rangle \\
&\leq \sum_{l=1}^m w_l \langle C_l, B_l \rangle + \max_l \|C_l\|_\infty \sum_{l=1}^m w_l \|B_l - \check{B}_l\|_1 \\
&\leq \sum_{l=1}^m w_l \mathcal{W}(p_l, q^*) + 2\gamma \ln n + 2 \max_l \|C_l\|_\infty \varepsilon' \\
&\leq \sum_{l=1}^m w_l \mathcal{W}(p_l, q^*) + \varepsilon.
\end{aligned}$$

Complexity bound for the algorithm is a simple corollary of Theorem 1. □

Remark 2. Notice that according to the proof of Theorem 2, one can also reconstruct approximated optimal transportation plans $\check{B}_l$ between p_l and approximated barycenter q using Algorithm 2 from Altschuler et al. [2017].

2.3 Proximal IBP for Wasserstein barycenter problem

As we see from Theorems 1 and 2, to obtain an ε-approximation of the non-regularized barycenter, the regularization parameter γ should be chosen proportional to the desired accuracy ε, and the complexity of the IBP is inversely proportional to γ, which leads to large working time and instability issues. To overcome this obstacle we propose a novel proximal-IBP algorithm. Similarly to **?**, where this idea is used for Wasserstein distance, our method is inspired by proximal point algorithm with general Bregman divergence $V(x, y)$ Chen and Teboulle [1993]. The idea of this algorithm for minimization of a function $f(x)$ is to perform steps $x_{k+1} = \mathbf{prox}(x_k) = \arg\min_{x\in Q}\{f(x) + \gamma V(x, x_k)\}$. We use the KL-divergence as the Bregman divergence since in this case the proximal step leads to a similar problem to the entropic-regularized WB (4). Given the sets $\mathcal{C}_1, \mathcal{C}_2$ defined in (7), we define proximal

operator $\mathbf{prox} : \mathcal{C}_1 \cap \mathcal{C}_2 \to \mathcal{C}_1 \cap \mathcal{C}_2$ for function $\sum_{l=1}^m w_l \mathcal{W}_\gamma(p_l, q_l)$ as follows

$$\begin{aligned}\mathbf{prox}(\boldsymbol{\pi}^k) &= \underset{\boldsymbol{\pi} \in \mathcal{C}_1 \cap \mathcal{C}_2}{\operatorname{argmin}} \sum_{l=1}^m w_l \left[\langle C_l, \pi_l \rangle + \gamma KL(\pi_l | \pi_l^k)\right] \\ &= \underset{\boldsymbol{\pi} \in \mathcal{C}_1 \cap \mathcal{C}_2}{\operatorname{argmin}} \sum_{l=1}^m w_l KL\left(\pi_l | \pi_l^k \odot \exp\left(-\tfrac{C_l}{\gamma}\right)\right).\end{aligned}$$

The proximal gradient method has the following form

$$\boldsymbol{\pi}^{k+1} = \mathbf{prox}(\boldsymbol{\pi}^k). \tag{24}$$

Then we use Iterative Bregman Projection for finding the barycenter.

Algorithm 4 Finding Wasserstein barycenter by proximal IBP

Input: N — number of external iterations; ε' — precision for inner problem; $\gamma > 0$; $\mathtt{u}^{0,0} = 0$ $\mathtt{v}^{0,0} = 0$; starting transport plans $\{\pi_l^0\}_{l=1}^m : \pi_l^0 = \exp\left(-\frac{C_l}{\gamma}\right) \ \forall l = 1, \ldots, m$

1: **for** $k = 0, \ldots, N-1$ **do**
2: $K_l^k \coloneqq \pi_l^k \odot \exp(-\frac{C_l}{\gamma})$
3: Run Algorithm 2 with starting point $\mathtt{u}^{0,k}, \mathtt{v}^{0,k}$, kernels $\{K_l^k\}$ instead of $\{K_l\}$, and precision ε'.
4: Set $u_l^{0,k+1} = u_l^M$, $v_l^{0,k+1} = v_l^M$, where (u_l^M, v_l^M) is the last potentials generated by Algorithm 2.
5: Set $\pi_l^{k+1} = \check{B}_l \in \Pi(p_l, \bar{q}^M)$, obtained from $B_l(u_l^M, v_l^M)$ by Algorithm 2 from Altschuler et al. [2017].
6: **end for**

Output: $\bar{q}^M$, generated on the last inner iteration of Algorithm 2 on the last outer iteration.

We underline that in this setting, there is no need to choose γ to be small as it prescribed by Theorem 3. Algorithm 5 has two loops: external loop of proximal gradient step and inner loop of computing the next iterate $\boldsymbol{\pi}^t$ by IBP and as a byproduct an approximation q^t to the barycenter. The number of external iterations is proportional to γ/ε, see Chen and Teboulle [1993], and the complexity of inner loop is $O\left(\|C_l^t\|_\infty/(\gamma\varepsilon')\right)$. Slightly modifying algorithm Round and vectors p_l we can ensure that all $[\pi_l^t]_{ij} \gtrsim \varepsilon/n^2$, then it is enough to choose $\tilde{\varepsilon}$ proportional to ε^3/n^2, and inner loop complexity is $\widetilde{O}\left(n^2\|C_l\|_\infty^2/(\gamma\varepsilon^3)\right)$. However, experiments show that this estimate is too pessimistic, and in practice number of inner iterations is much smaller, see Section 4. Additional numerical experiments and more accurate theoretical analysis can be found in the follow-up paper Stonyakin et al. [2019].

In practice, one should try to find the optimal γ by using a restart procedure on the first external loop iteration. That is, we start with large enough γ and solve internal problem by IBP, then put $\gamma \coloneqq \gamma/2$ and solve internal problem once again. We stop this repeating procedure at the moment when the complexity of internal problem growth significantly. This

Algorithm 5 Finding Wasserstein barycenter by proximal IBP

Input: T — number of iterations, $\gamma > 0$, $\tilde{\varepsilon}$ — accuracy for inner problem, starting transport plans $\pi_l^0 \coloneqq \frac{1}{n} p_l^\mathsf{T} \mathbb{1} \quad \forall l = 1, \ldots, m$

1: **for** $t = 0, \ldots, T-1$ **do**

2: Run Algorithm 2 with cost matrices $C_l^t \coloneqq C_l - \gamma \ln \pi_l^t$, parameter γ and accuracy $\varepsilon' \propto \frac{\tilde{\varepsilon}}{\max_l \|C_l^t\|_\infty}$, and obtain matrices $B_1, \ldots, B_m$

3: $\pi_l^{t+1} \coloneqq \text{Round}\left(B_l, p_l, \bar{q}^{t+1}\right) \in \Pi(p_l, \bar{q}^{t+1})$, where Round is Algorithm 2 from Altschuler et al. [2017] and $\bar{q}^{t+1} \coloneqq \sum_{l=1}^m w_l B_l^\mathsf{T} \mathbb{1}$

4: **end for**

Output: $\bar{q}^T$

moment allows us to detect the optimal value of γ. On the next external iterations one may use this γ.

Algorithm 5 can be implemented in centralized distributed setting in the same way as Algorithm 2, see Subsection 2.1.

3 Complexity by Primal-Dual Accelerated Gradient Descent

In this subsection we consider Primal-Dual Accelerated Gradient Descent for approximating Wasserstein barycenter. First, we consider regularized barycenter, construct a dual problem to (4) and apply primal-dual accelerated gradient descent to solve it and approximate the regularized barycenter. Our dual problem is constructed via a matrix W, which can be quite general. We explain how the choice of this matrix is connected to distributed optimization and allows to implement the algorithm in the decentralized distributed setting. Then, we show, how the regularization parameter should be chosen in order to obtain an ε-approximation for the non-regularized Wasserstein barycenter, and estimate the complexity of the resulting algorithm. The proposed algorithms can be implemented in a decentralized distributed manner such that each node fulfils $\widetilde{O}(n^{2.5}/\varepsilon)$ arithmetic operations and the number of communication rounds is $\widetilde{O}(\sqrt{n}/\varepsilon)$.

3.1 Consensus view on Wasserstein barycenter problem

We rewrite the problem (4) in an equivalent way as

$$\min_{\substack{q_1, \ldots, q_m \in S_n(1) \\ q_1 = \cdots = q_m}} W_\gamma(\mathtt{p}, \mathtt{q}) \coloneqq \sum_{l=1}^m w_l \mathcal{W}_{\gamma(l)}(p_l, q_l), \tag{25}$$

where $\mathtt{p} = [p_1, \ldots, p_m]^\mathsf{T}$ and $\mathtt{q} = [q_1, \ldots, q_m]^\mathsf{T}$, we also use different regularizer $\gamma_l = \gamma(l)$ for l-th Wasserstein distance. Next we write a dual problem by dualizing the equality constraints

$q_1 = \cdots = q_m$. This can be done in many different ways and, following Lan et al. [2017], Scaman et al. [2017], Uribe et al. [2018], we do it by introducing a matrix $\bar{W} \in \mathbb{R}^{n\times n}$ which is a symmetric positive semi-definite matrix s.t. $\mathrm{Ker}(\bar{\mathrm{W}}) = \mathrm{span}(\mathbb{1})$. Then, defining $W = \bar{W} \otimes I_n$ and using the fact $q_1 = \cdots = q_m \iff \sqrt{W}\mathtt{q} = 0$, we equivalently rewrite problem (25) as

$$\max_{\substack{q_1,\ldots,q_m\in S_n(1),\\ \sqrt{W}\mathtt{q}=0}} -\sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q_l), \tag{26}$$

Dualizing the linear constraint $\sqrt{W}\mathtt{q} = 0$, we obtain the dual problem

$$\min_{\mathtt{u}\in\mathbb{R}^{mn}} \max_{\mathtt{q}\in\mathbb{R}^{nm}} \left\{\sum_{l=1}^{m}\langle u_l, [\sqrt{W}\mathtt{q}]_l\rangle - \sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q_l)\right\} = \min_{\mathtt{u}\in\mathbb{R}^{mn}} \sum_{l=1}^{m} w_l \mathcal{W}^*_{\gamma(l),p_l}([\sqrt{W}\mathtt{u}]_l/w_l) \tag{27}$$

where $\mathcal{W}^*_{\gamma(l),p_l}(\cdot)$ is the Fenchel–Legendre transform of $\mathcal{W}_{\gamma(l)}(p_l,\cdot)$, $[\sqrt{W}\mathtt{q}]_i$ and $[\sqrt{W}\mathtt{u}]_i$ represent the i-th n-dimensional block of vectors $\sqrt{W}\mathtt{q}$ and $\sqrt{W}\mathtt{u}$ respectively. Importantly, the objective in the dual problem (27) has L-Lipschitz-continuous gradient, where constant L is estimated below in Lemma 8. Since the dual problem is smooth, we apply distributed primal-dual accelerated gradient descent Algorithm 6 to solve the constructed pair of primal and dual problems.

Before we move to the theoretical analysis of the algorithm, let us discuss the scalability of Algorithm 6. Assume that we have an arbitrary network of agents given by connected undirected graph $G = (V, E)$ without self-loops with the set V of n vertices and the set of edges $E = \{(i, j) : i, j \in V\}$. Then matrix $\bar{W}$ can be chosen as the Laplacian matrix for this graph, which is such that a) $[\bar{W}]_{ij} = -1$ if $(i, j) \in E$, b) $[\bar{W}]_{ij} = \deg(i)$ if $i = j$, c) $[\bar{W}]_{ij} = 0$ otherwise. Here $\deg(i)$ is the degree of the node i, i.e., the number of neighbors of the node. We assume that an agent i can communicate with an agent j if and only if the edge $(i, j) \in E$. In particular, the Laplacian matrix for the star graph, which corresponds to the centralized distributed computations discussed in Section 2 is

$$\bar{W} : \{\forall i = 1, \ldots, m-1\ \bar{W}_{ii} = 1,\ \bar{W}_{im} = \bar{W}_{mi} = -1,\ \bar{W}_{mm} = m-1\}. \tag{28}$$

Algorithm 6 allows to perform calculations in an arbitrary connected undirected network of agents. This is in contrast to the IBP algorithm as discussed in Section 2.

For simplicity and comparison with the complexity of the IBP algorithm, we analyze the complexity of Algorithm 6 as if it is implemented on one machine, disregarding that it can be used for distributed setup.

Theorem 3. *Algorithm 6 after* $N = \frac{1}{\varepsilon}\sqrt{64\chi(\bar{W})mn\ln n\sum_{l=1}^{m} w_l^2\|C_l\|_\infty^2}$ *iterations generates an* ε*-solution of problem* (2)*, i.e. finds a vector* $\mathtt{q}^N = [q_1^\mathsf{T}, \ldots, q_m^\mathsf{T}]^\mathsf{T}$ *s.t.*

$$\sum_{l=1}^{m} w_l \mathcal{W}(p_l, q_l^N) - \sum_{l=1}^{m} w_l \mathcal{W}(p_l, q^*) \le \varepsilon, \quad \|\sqrt{W}\mathtt{q}^N\|_2 \le \varepsilon/2R, \tag{29}$$

Algorithm 6 Accelerated Distributed Computation of Wasserstein barycenter

Input: Each agent $l \in V$ is assigned its measure p_l and an upper bound L for the Lipschitz constant of the gradient of the dual objective.

1: Each agent finds $\tilde{p}_l \in S_n(1)$ s.t. $\|\tilde{p}_l - p_l\|_1 \leq \varepsilon/4$ and $\min_i [\tilde{p}_l]_i \geq \varepsilon/(8n)$ and redefine $p_l := \tilde{p}_l$. E.g., $\tilde{p}_l = \left(1 - \frac{\varepsilon}{8}\right)\left(p_l + \frac{\varepsilon}{n(8-\varepsilon)}\mathbb{1}\right)$.

2: All agents $l \in V$ set $\gamma(l) = \frac{\varepsilon}{4mw_l \ln n}$, $\eta_l^0 = \zeta_l^0 = \lambda_l^0 = q_l^0 = \mathbf{0} \in \mathbb{R}^n$, $A_0 = \alpha_0 = 0$ and N.

3: For each agent $l \in V$:

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

5: Find α_{k+1} as the largest root of the equation $A_{k+1} := A_k + \alpha_{k+1} = 2L\alpha_{k+1}^2$.

6: $\lambda_l^{k+1} = (\alpha_{k+1}\zeta_l^k + A_k\eta_l^k)/A_{k+1}$.

7: Calculate $\nabla \mathcal{W}^*_{\gamma(l),p_l}(\lambda_l^{k+1})$:
$[\nabla \mathcal{W}^*_{\gamma(l),p_l}(\lambda_l^{k+1})]_i = \sum_{j=1}^n [p_l]_j \frac{\exp(([\lambda_l^{k+1}]_i - [C_l]_{ij})/\gamma(l))}{\sum_{r=1}^n \exp(([\lambda_l^{k+1}]_r - [C_l]_{rj})/\gamma(l))}$, where $[\lambda]_i$ denotes $i-th$ component of a vector λ.

8: Share $\nabla \mathcal{W}^*_{\gamma(l),p_l}(\lambda_l^{k+1})$ with $\{j \mid (i,j) \in E\}$.

9: $\zeta_l^{k+1} = \zeta_l^k - \alpha_{k+1} \sum_{j=1}^m W_{lj} \nabla \mathcal{W}^*_{\gamma(l),p_j}(\lambda_j^{k+1})$. {Gradient step}

10: $\eta_l^{k+1} = (\alpha_{k+1}\zeta_l^{k+1} + A_k\eta_l^{k+1})/A_{k+1}$. {Extrapolation step}

11: $q_l^{k+1} = \frac{1}{A_{k+1}} \sum_{l=0}^{k+1} \alpha_i q_i(\lambda_l^{k+1}) = (\alpha_{k+1} q_i(\lambda_l^{k+1}) + A_k q_l^k)/A_{k+1}$,
where $q_l(\cdot) = \nabla \mathcal{W}^*_{\gamma(l),p_l}(\cdot)$ defined in step 7. {Primal update}

12: **end for**

Output: $\mathtt{q}^N = [q_1^\mathsf{T}, \ldots, q_m^\mathsf{T}]^\mathsf{T}$.

where q^* is an unregularized barycenter, i.e. is a solution to (2)*, and R is a bound on the solution to the dual problem. Moreover, the number of arithmetic operations is $O\left(N \cdot n(mn + \mathrm{nnz}(\bar{W}))\right)$.*

The proof is based on the complexity theorem of primal-dual accelerated gradient descent for a particular pair of primal-dual problems (25)–(27).

Theorem 4 (see [Dvurechensky et al., 2017, Theorem 2])**.** *Let accelerated primal-dual gradient descent be applied to the pair of problems* (25)–(27)*. Then the inequalities*

$$\sum_{l=1}^m w_l \mathcal{W}_{\gamma(l)}(p_l, q_l^N) - \sum_{l=1}^m \mathcal{W}_{\gamma(l)}(p_l, q^*) \leq \varepsilon/2, \quad \|\sqrt{W}\mathtt{q}^N\|_2 \leq \varepsilon/2R \tag{30}$$

hold no later than after $N = \sqrt{\frac{32LR^2}{\varepsilon}}$ iterations, where L is the Lipschitz constant of the gradient of the dual objective and R is such that $\|\mathtt{u}^*\|_2 \leq R$, $\mathtt{u}^*$ being an optimal dual solution.

Our next steps are to find the bounds for L in the next Lemma and R in Lemma 8.

Lemma 7. *Let in* (25) $\gamma(l) = \gamma/w_l$ *for some* $\gamma > 0$, *and* $\mathcal{W}^*_\gamma(\mathtt{u})$ *denote the dual objective in* (27). *Then its gradient is* $L = \lambda_{\max}(W)/\gamma$*-Lipschitz continuous w.r.t. 2-norm.*

Proof. Making the change of variable $[\boldsymbol{\lambda}]_l = [\sqrt{W}\mathtt{u}]_l/w_l$, by the chain rule, the i-th n-dimensional block of $\nabla\mathcal{W}^*_\gamma(\mathtt{u})$ is

$$\left[\nabla\mathcal{W}^*_\gamma(\mathtt{u})\right]_i = \left[\nabla\sum_{l=1}^m w_l\mathcal{W}^*_{\gamma(l),p_l}([\sqrt{W}\mathtt{u}]_l/w_l)\right]_i = \sum_{l=1}^m \sqrt{W}_{il}\nabla\mathcal{W}^*_{\gamma(l),p_l}(\lambda_l),\ i=1,\dots,m. \quad (31)$$

Thus,

$$\begin{aligned}
\|\nabla\mathcal{W}^*_\gamma(\mathtt{u}_1) - \nabla\mathcal{W}^*_\gamma(\mathtt{u}_2)\|_2^2 &\overset{(31)}{=} \left\|\sqrt{W}\begin{pmatrix}\nabla\mathcal{W}^*_{\gamma(1),p_1}([\boldsymbol{\lambda}_1]_1)\\ \dots \\ \nabla\mathcal{W}^*_{\gamma(m),p_m}([\boldsymbol{\lambda}_1]_m)\end{pmatrix} - \sqrt{W}\begin{pmatrix}\nabla\mathcal{W}^*_{\gamma(1),p_1}([\boldsymbol{\lambda}_2]_1)\\ \dots \\ \nabla\mathcal{W}^*_{\gamma(m),p_m}([\boldsymbol{\lambda}_2]_m)\end{pmatrix}\right\|_2^2\\
&\le (\lambda_{\max}(\sqrt{W}))^2\left\|\begin{matrix}\nabla\mathcal{W}^*_{\gamma(1),p_1}([\boldsymbol{\lambda}_1]_1) - \nabla\mathcal{W}^*_{\gamma(m),p_1}([\boldsymbol{\lambda}_2]_1)\\ \dots \\ \nabla\mathcal{W}^*_{\gamma(1),p_m}([\boldsymbol{\lambda}_1]_m) - \nabla\mathcal{W}^*_{\gamma(m),p_m}([\boldsymbol{\lambda}_2]_m)\end{matrix}\right\|_2^2\\
&= (\lambda_{\max}(\sqrt{W}))^2\sum_{i=1}^m\left\|\nabla\mathcal{W}^*_{\gamma(i),p_i}([\boldsymbol{\lambda}_1]_i) - \nabla\mathcal{W}^*_{\gamma(i),p_i}([\boldsymbol{\lambda}_2]_l)\right\|_2^2\\
&\le (\lambda_{\max}(\sqrt{W}))^2\sum_{i=1}^m\frac{1}{\gamma^2(i)}\left\|[\boldsymbol{\lambda}_1]_l - [\boldsymbol{\lambda}_2]_l\right\|_2^2\\
&= (\lambda_{\max}(\sqrt{W}))^2\sum_{i=1}^m\frac{1}{\gamma^2(i)}\left\|[\sqrt{W}(\mathtt{u}_1-\mathtt{u}_2)]_i/w_i\right\|_2^2,
\end{aligned}$$

where we used notation $[\boldsymbol{\lambda}]_i = [\sqrt{W}\mathtt{u}]_i/w_i$, the definition of matrix $\sqrt{W}$, $1/\gamma(i)$-Lipschitz continuity of $\nabla\mathcal{W}^*_{\gamma(i),p_i}(\cdot)$ for all $i = 1,\dots,m$ by [Cuturi and Peyré, 2016, Theorem 2.4]. Since $\gamma(i) = \gamma/w_i$, $i = 1,\dots,m$, we obtain

$$\|\nabla\mathcal{W}^*_\gamma(\mathtt{u}_1) - \nabla\mathcal{W}^*_\gamma(\mathtt{u}_2)\|_2^2 \le \frac{(\lambda_{\max}(\sqrt{W}))^2}{\gamma^2}\left\|\sqrt{W}(\mathtt{u}_1-\mathtt{u}_2)\right\|_2^2 \quad (32)$$

$$\le \frac{(\lambda_{\max}(\sqrt{W}))^4}{\gamma^2}\|\mathtt{u}_1-\mathtt{u}_2\|_2^2, \quad (33)$$

Since $(\lambda_{\max}(\sqrt{W}))^4 = (\lambda_{\max}(W))^2$, we get the statement of Lemma. □

The following Lemma is inspired by Lan et al. [2017].

Lemma 8. *Let* $\mathtt{q}^*_\gamma$ *be the optimal solution of problem* (4) *with minimal 2-norm, then there exists an optimal dual solution* $\boldsymbol{\lambda}^* = [\lambda^*_1,\dots,\lambda^*_m]$ *for problem* (27) *satisfying* $\|\boldsymbol{\lambda}^*\|_2 \le R$ *with*

$$R^2 = \frac{2n\sum_{l=1}^m w_l^2\|C_l\|_\infty^2}{\lambda^+_{\min}(W)}. \quad (34)$$

Here $\lambda^+_{min}(W)$ *is the minimal positive eigenvalue of the matrix* W.

Proof. Recall that $\mathcal{W}_\gamma(\mathtt{p},\cdot)$ denotes the objective value in the primal problem (25). Then

$$
\begin{aligned}
-\mathcal{W}_\gamma(\mathtt{p},\mathtt{q}^*_\gamma) &= \langle \boldsymbol{\lambda}^*, \sqrt{W}\mathtt{q}^*_\gamma\rangle - \mathcal{W}_\gamma(\mathtt{p},\mathtt{q}^*_\gamma) = \mathcal{W}^*_{\gamma,\mathtt{p}}(\sqrt{W}\boldsymbol{\lambda}^*) = \max_{\substack{q_l\in S_n(1),\\ l=1,\dots,m}} \{\langle \boldsymbol{\lambda}^*_l, \sqrt{W}\mathtt{q}_l\rangle - \mathcal{W}_\gamma(\mathtt{p},\mathtt{q})\} \\
&\geq \langle \boldsymbol{\lambda}^*, [\sqrt{W}\mathtt{q}]_l\rangle - \mathcal{W}_\gamma(\mathtt{p},\mathtt{q}) = \langle \boldsymbol{\lambda}^*, \sqrt{W}\mathtt{q} - \sqrt{W}\mathtt{q}^*_\gamma\rangle - \mathcal{W}_\gamma(\mathtt{p},\mathtt{q}) \\
&= \langle \sqrt{W}\boldsymbol{\lambda}^*, \mathtt{q}^*_\gamma - \mathtt{q}\rangle - \mathcal{W}_{\gamma,\mathtt{p}}(\mathtt{q}),
\end{aligned}
$$

where we used $W^\intercal = W$ and $\mathtt{q}^*_\gamma$ is the regularized barycenter.
From this inequality and using the convexity of $\mathcal{W}_\gamma(\mathtt{p},\mathtt{q})$ we have $\nabla\mathcal{W}_\gamma(\mathtt{p},\mathtt{q}^*) = -\sqrt{W}\boldsymbol{\lambda}^*$. Then we have

$$
\|\nabla\mathcal{W}_\gamma(\mathtt{p},\mathtt{q}^*)\|_2^2 = \|-\sqrt{W}\boldsymbol{\lambda}^*\|_2^2 = \langle\sqrt{W}\boldsymbol{\lambda}^*, \sqrt{W}\boldsymbol{\lambda}^*\rangle = \langle\boldsymbol{\lambda}^*, W\boldsymbol{\lambda}^*\rangle \geq \lambda^+_{\min}(W)\|\boldsymbol{\lambda}^*\|_2^2, \quad (35)
$$

where the last inequality holds due to $\boldsymbol{\lambda}^* \in \left(\mathrm{Ker}(\sqrt{\mathrm{W}})\right)^\perp$

Hence, we get

$$
\|\boldsymbol{\lambda}^*\|_2^2 \leq R^2 = \frac{\|\nabla\mathcal{W}_\gamma(\mathtt{p},\mathtt{q}^*_\gamma)\|_2^2}{\lambda^+_{\min}(W)} = \frac{\sum_{l=1}^m w_l^2\|\nabla\mathcal{W}_{\gamma(l)}(p_l,q^*_\gamma)\|_2^2}{\lambda^+_{\min}(W)} \quad (36)
$$

Let us now estimate $\|\nabla\mathcal{W}_{\gamma(l)}(p_l,q^*_\gamma)\|_2^2$. From (3), we can construct the dual problem to the regularized optimal transport problem

$$
\mathcal{W}_{\gamma(l)}(p,q) = \min_{\pi\in\Pi(p,q)}\{\langle\pi,C\rangle + \gamma(l)H(\pi)\} = \max_{\mu,\nu}\left\{-\langle\mu,p\rangle - \langle\nu,q\rangle - \frac{\gamma(l)}{e}e^{-\frac{\mu}{\gamma(l)}}e^{-\frac{C}{\gamma(l)}}e^{-\frac{\nu}{\gamma(l)}}\right\}
$$

By [Cuturi and Peyré, 2016, Proposition 2.3], $\nabla_q\mathcal{W}\gamma(l)(p,q) = -\nu^*$, where ν^* is the solution to the dual problem satisfying $\langle\nu^*,\mathbb{1}\rangle = 0$. Hence, $\min_{i=1,\dots,n}\nu^*_i \leq 0 \leq \max_{i=1,\dots,n}\nu^*_i$. As it follows from [Dvurechensky et al., 2018b, Lemma 1]

$$
\max_{i=1,\dots,n}\nu^*_i - \min_{i=1,\dots,n}\nu^*_i \leq \|C_l\|_\infty - \gamma(l)\ln\left(\min_i[p_l]_i\right) = \|C_l\|_\infty + \gamma(l)\ln\left(\frac{8n}{\varepsilon}\right),
$$

where we used the fact that the p_l was redefined in Algorithm 6 in such a way that $\min_i[p_l]_i \geq \frac{\varepsilon}{8n}$ and also that our variable ν^* and their dual variable u^* satisfy $u^* = -\frac{\nu^*}{\gamma(l)} - \frac{1}{2}$. Making the same arguments as in the proof [Lin et al., 2019, Lemma 3.2.], we obtain from the above two facts that

$$
\|\nu^*\|_2 \leq \|\nu^*\|_\infty\sqrt{n} \leq \|C_l\|_\infty\sqrt{n} + \frac{\gamma}{w_l}\ln\left(\frac{8n}{\varepsilon}\right),
$$

where we used that $\gamma(l) = \gamma/w_l$. Thus,

$$
\sum_{l=1}^m w_l^2\|\nabla\mathcal{W}_{\gamma(l)}(p_l,q^*_\gamma)\|_2^2 \leq 2n\sum_{l=1}^m w_l^2\|C_l\|_\infty^2 + 2\gamma^2\ln^2\left(\frac{8n}{\varepsilon}\right).
$$

Since γ is chosen proportional to ε which is small, we can neglect the second term in comparison with the first one.

□

Proof of Theorem 3. By Theorem 4, we have

$$\sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q_l^N) - \sum_{l=1}^{m} \mathcal{W}_{\gamma(l)}(p_l, q^*) \le \varepsilon/2, \quad \|\sqrt{W}\mathtt{q}^N\|_2 \le \varepsilon/2R \tag{37}$$

Since $KL(\pi|\theta) \ge 0$, we have

$$\sum_{l=1}^{m} w_l \mathcal{W}(p_l, q_l^N) - \sum_{l=1}^{m} w_l \mathcal{W}(p_l, q^*) \le \sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q_l^N) - \sum_{l=1}^{m} w_l \mathcal{W}(p_l, q^*). \tag{38}$$

By definition of the objective $\mathcal{W}_\gamma(\mathtt{p}, \cdot)$ in (25),

$$\begin{aligned}
\sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q^*) &= \min_{\substack{q_1=\cdots=q_m \\ q_1,\ldots q_m \in S_n(1)}} \sum_{l=1}^{m} w_l \min_{\pi \in \Pi(p_l, q_l)} \left\{ \sum_{i,j=1}^{n} C_{ij}\pi_{ij} + \gamma KL(\pi|\theta) \right\} \\
&\le \min_{\substack{q_1=\cdots=q_m \\ q_1,\ldots q_m \in S_1(n)}} \sum_{l=1}^{m} \left\{ \min_{\pi \in \Pi(p_l, q_l)} w_l \sum_{i,j=1}^{n} C_{ij}\pi_{ij} + w_l \max_{\pi \in \Pi(p_l, q_l)} \gamma KL(\pi|\theta) \right\} \\
&\le \min_{\substack{q_1=\cdots=q_m \\ q_1,\ldots q_m \in S_n(1)}} \sum_{l=1}^{m} \left\{ \min_{\pi \in \Pi(p_l, q_l)} \sum_{i,j=1}^{n} C_{ij}\pi_{ij} + 2\sum_{l=1}^{m} w_l \gamma(l) \ln n \right\} \\
&\le \min_{\substack{q_1=\cdots=q_m \\ q_1,\ldots q_m \in S_n(1)}} \sum_{l=1}^{m} \mathcal{W}(p_l, q_l) \ + 2\ln n \sum_{l=1}^{m} w_l \gamma(l) \\
&= \sum_{l=1}^{m} \mathcal{W}(p_l, q^*) + 2\ln n \sum_{l=1}^{m} w_l \gamma(l),
\end{aligned} \tag{39}$$

where we chose $\theta_{ij} = 1/n^2$ for all $i, j = 1, \ldots n$ so $KL(\pi|\theta) \in [0, 2\ln n]$.

Substituting (39) in (38) we get

$$\begin{aligned}
\sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q_l^N) - \sum_{l=1}^{m} w_l \mathcal{W}(p_l, q^*) \le &\sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q_l^N) \\
&- \sum_{l=1}^{m} w_l \mathcal{W}_{\gamma(l)}(p_l, q^*) + 2\ln n \sum_{l=1}^{m} w_l \gamma(l)
\end{aligned} \tag{40}$$

Using this inequality and (37) we get

$$\sum_{l=1}^{m} w_l \mathcal{W}(p_l, q_l^N) - \sum_{l=1}^{m} w_l \mathcal{W}(p_l, q^*) \le \varepsilon/2 + 2\ln n \sum_{l=1}^{m} w_l \gamma(l). \tag{41}$$

Since $\gamma(l) = \gamma/w_l$ with $\gamma = \varepsilon/(4m\ln n)$, we obtain that the inequality (37) holds. Combining the values of γ, L from Lemma 7, R from Lemma 8 with the estimate for N in Theorem

4 and the fact that $\chi(W) = \chi(\bar{W})$, we obtain an estimate for the number of iterations of the algorithm. Let us estimate the complexity of the algorithm. For each l we need to calculate the gradient $\mathcal{W}^*_{\gamma(l),p_l}(\cdot)$, which requires $O(n^2)$ arithmetic operations. To calculate $\sum_{j=1}^m W_{lj}\nabla\mathcal{W}^*_{\gamma(l),p_l)}(\lambda_j^{k+1})$ one needs $O(n \cdot \text{nnz}(\bar{W}_l))$ arithmetic operations, where $\text{nnz}(\bar{W}_l)$ is the number of non-zero elements in matrix $\bar{W}$ in the l-th row. More precisely, the dimension of $\nabla\mathcal{W}^*_{\gamma(l),p_l}(\cdot)$ is n and the matrix W_{lj} is diagonal for each $l, j = 1, \ldots, m$. Using definition of W we get that the complexity of calculating the gradient. Other operations require $O(n)$ operations. Hence, the complexity of one iteration is

$$O\left(mn^2 + \sum_{l=1}^m n \cdot \text{nnz}(\bar{W}_l)\right) = O\left(mn^2 + n \cdot \text{nnz}(\bar{W})\right) \tag{42}$$

and the total complexity follows from multiplying this value by N.

□

Let us make a couple of remarks. As for the choice of $\bar{W}$ one can show (by using graph sparsificators) that it can be chosen such that $\chi(W) = \chi(\bar{W}) = O(\text{Poly}(\ln(m)))$ and $\text{nnz}(\bar{W}) = O(m\text{Poly}(\ln(m)))$. For details on the graph sparsificators we refer to Vaidya [1990], Bern et al. [2006], Spielman and Teng [2014]. In the simple case of equal weights $w_l = \frac{1}{m}$, $l = 1, ..., m$ the complexity of approximating non-regularized barycenter by accelerated gradient descent can be estimated as $\widetilde{O}(mn^{2.5}/\varepsilon)$. In the distributed setting, each of m nodes makes $\widetilde{O}(n^{2.5}/\varepsilon)$ arithmetic operations, while the number of communications rounds is $\widetilde{O}(\sqrt{n}/\varepsilon)$. The case of general weights is interesting with respect to bootstrap procedure allowing to construct confidence sets for barycenter Ebert et al. [2017].

There is an interesting connection of our work with Cuturi and Peyré [2016]. Consider a particular case of equal weights $w_1 = \cdots = w_m = 1$ and matrix $\bar{W}$ corresponding to the star graph topology. Then the dual problem (25) has the form

$$\min_{\lambda_1,\ldots,\lambda_m\in\mathbb{R}^n}\left\{\sum_{l=1}^{m-1}\mathcal{W}^*_{\gamma,p_l}\left(\lambda_l - \frac{1}{m}\lambda_m\right) + \mathcal{W}^*_{\gamma,p_m}\left(\frac{m-1}{m}\lambda_m - \sum_{l=1}^{m-1}\lambda_l\right)\right\} \tag{43}$$

Changing the variable $\hat{\lambda}_l = \lambda_l - \frac{1}{m}\lambda_m$ we come to the following formulation

$$\min_{\hat{\lambda}_1,\ldots,\hat{\lambda}_{m-1}\in\mathbb{R}^n}\sum_{l=1}^{m-1}\mathcal{W}^*_{\gamma,p_l}(\hat{\lambda}_l) + \mathcal{W}^*_{\gamma,p_m}\left(-\sum_{l=1}^{m-1}\hat{\lambda}_l\right) \tag{44}$$

Hence, the approach presented in Cuturi and Peyré [2016], in Theorem 3.1 is the particular case for the approach described above, corresponding to the the star graph.

4 Numerical Analysis

In this section, we provide numerical analysis for the three algorithms for the computation of approximate Wasserstein barycenters. We compare their iteration performance for

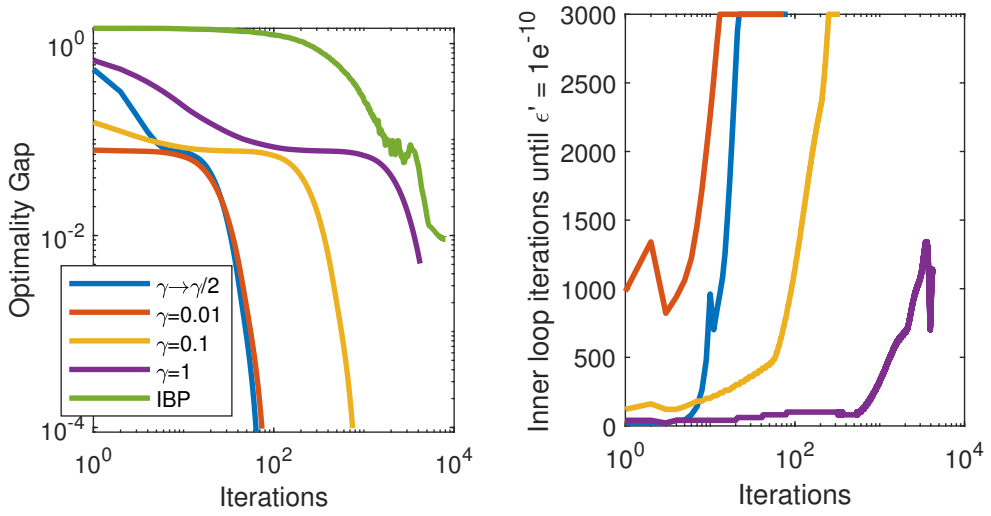

Figure 1: Distance to an optimal barycenter for the IBP method and the ProxIBP method.

the problem of computing the barycenter of a set of 15 discrete and truncated Gaussian distributions.

Figure 1 (Left) shows the distance to optimality versus the iteration count for the IBP method and the ProxIBP method. For the ProxIBP method, we show the performance for four different cases, namely: $\gamma = 1$, $\gamma = 0.1$, $\gamma = 0.01$, and varying with $\gamma_{k+1} = \gamma_k/2$, if $\gamma_k \geq 1e^{-3}$ ($\gamma_0 = 10$). Figure 1 (Right) shows the number of iterations required in the inner loop step of Algorithm 3 (Line 2) to reach the desired accuracy ε' for the same scenarios on γ. Results show that for smaller values of γ the inner problem requires larger number of iterations. Particularly for $\gamma = 1$ the inner problem is relatively computationally inexpensive, but the convergence of the overall method is slow. On the other side, with varying values of γ an accurate barycenter is found with low computational cost initially.

Figure 2 shows the performance of the primal-dual accelerated gradient descent method. Recall that this method is particularly suited for decentralized distributed approaches where the computation is performed over an arbitrary network. We show the distance to optimality and distance to consensus for the approximate barycenters generated by Algorithm 4.

Table 1 shows the numerical values of the optimality gap for a subset of the experiments shown above. The ProxIBP algorithm converges much faster than the other two, in exchange with higher computational loads per interation.

Conclusion

In this paper we show that IBP algorithm from Benamou et al. [2015] for Wasserstein barycenter problem can be implemented in a centralized distributed manner such that each node fulfils $\widetilde{O}\left(n^2/\varepsilon^2\right)$ arithmetic operations and the number of communication rounds is $\widetilde{O}\left(1/\varepsilon^2\right)$. We note that proper proximal envelope of this algorithm can sometimes give a

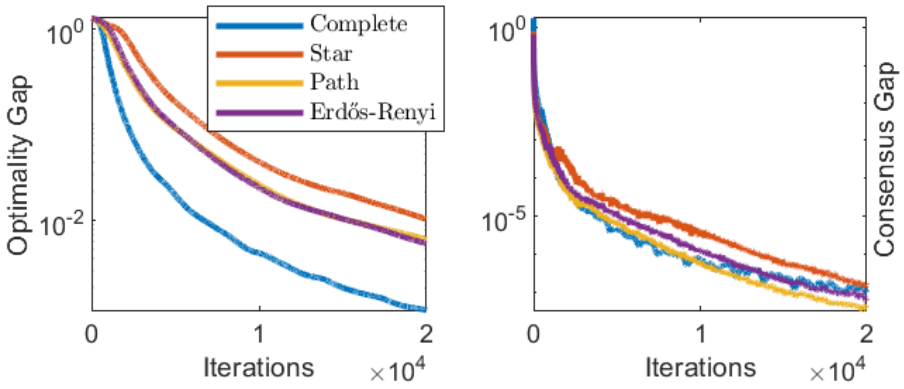

Figure 2: Optimality gap and consensus gap for the primal-dual accelerated gradient descent method for four classes of networks: complete, star, path and Erdős-Renyi random graph.

Table 1: Optimality Gap for the Approximate Barycenter

Iter.	ProxIBP $\gamma \to \gamma/2$	ProxIBP $\gamma = 0.01$	ProxIBP $\gamma = 1$	IBP	Algo. 4 Complete	Algo. 4 Erdős-Renyi
50	8.797e-4	1.779e-3	9.856e-2	0.2585	1.286	1.294
1000	4.17e-07	–	6.818e-2	0.2585	0.471	1.041
2000	–	–	4.201e-2	0.0741	0.111	0.463
3000	–	–	1.830e-2	0.0691	4.814e-2	0.226
4000	–	–	6.408e-3	0.0534	2.797e-2	0.135

significant acceleration. We also describe accelerated primal-dual gradient algorithm for the same problem. The proposed algorithm can be implemented in a more general decentralized distributed setting such that, to find an ε-approximation for the non-regularized barycenter, each node performs $\widetilde{O}(n^{2.5}/\varepsilon)$ arithmetic operations and the number of communication rounds is $\widetilde{O}(\sqrt{n}/\varepsilon)$.

Acknowledgments. This research was funded by Russian Science Foundation (project 18-71-10108).

References

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approxfimation algorithms for optimal transport via sinkhorn iteration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1961–1971. Curran Associates, Inc., 2017. arXiv:1705.09634.

A. S. Anikin, A. V. Gasnikov, P. E. Dvurechensky, A. I. Tyurin, and A. V. Chernov. Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. *Computational Mathematics and Mathematical Physics*, 57(8): 1262–1276, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv:1701.07875*, 2017.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Marshall Bern, John R Gilbert, Bruce Hendrickson, Nhat Nguyen, and Sivan Toledo. Support-graph preconditioners. *SIAM Journal on Matrix Analysis and Applications*, 27 (4):930–951, 2006.

Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

Jérémie Bigot, Thierry Klein, et al. Consistent estimation of a population barycenter in the wasserstein space. *ArXiv e-prints*, 2012.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.*, 53(1):1–26, 02 2017.

Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *arXiv:1810.07717*, 2018.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51 (1):22–45, Jan 2015. ISSN 1573-7683. doi: 10.1007/s10851-014-0506-3. URL `https://doi.org/10.1007/s10851-014-0506-3`.

Deeparnab Chakrabarty and Sanjeev Khanna. Better and simpler error analysis of the sinkhorn-knopp algorithm for matrix scaling. *arXiv:1801.02790*, 2018.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Alexey Chernov, Pavel Dvurechensky, and Alexander Gasnikov. Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In Yury Kochetov, Michael Khachay, Vladimir Beresnev, Evgeni Nurminski, and Panos Pardalos, editors, *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*, pages 391–403. Springer International Publishing, 2016.

Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic Wasserstein barycenters. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 999–1008. PMLR, 2018. URL `http://proceedings.mlr.press/v80/claici18a.html`.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR. URL `http://proceedings.mlr.press/v32/cuturi14.html`.

Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

Eustasio Del Barrio, Hélène Lescornel, and Jean-Michel Loubes. A statistical analysis of a deformation model with wasserstein barycenters : estimation procedure and goodness of fit test. *arXiv:1508.06465*, 2015.

Pavel Dvurechensky, Alexander Gasnikov, Evgenia Gasnikova, Sergey Matsievsky, Anton Rodomanov, and Inna Usik. Primal-dual method for searching equilibrium in hierarchical congestion population games. In *Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific*

School (DOOR 2016) Vladivostok, Russia, September 19 - 23, 2016, pages 584–595, 2016. arXiv:1606.08988.

Pavel Dvurechensky, Alexander Gasnikov, Sergey Omelchenko, and Alexander Tiurin. Adaptive similar triangles method: a stable alternative to sinkhorn's algorithm for regularized optimal transport. *arXiv:1706.07622*, 2017.

Pavel Dvurechensky, Darina Dvinskikh, Alexander Gasnikov, César A. Uribe, and Angelia Nedić. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, NeurIPS 2018, pages 10783–10793. Curran Associates, Inc., 2018a. arXiv:1806.03915.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, 2018b. arXiv:1802.04367.

Johannes Ebert, Vladimir Spokoiny, and Alexandra Suvorikova. Construction of non-asymptotic confidence sets in 2-Wasserstein space. *arXiv:1703.03658*, 2017.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717 – 735, 1989. ISSN 0024-3795. doi: http://dx.doi.org/10.1016/0024-3795(89)90490-4. URL `http://www.sciencedirect.com/science/article/pii/0024379589904904`. Special Issue Dedicated to Alan J. Hoffman.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

Lie He, An Bian, and Martin Jaggi. Cola: Decentralized linear learning. In *Advances in Neural Information Processing Systems*, pages 4541–4551, 2018.

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1501–1509. PMLR, 2017.

Leonid Kantorovich. On the translocation of masses. *Doklady Acad. Sci. USSR (N.S.)*, 37: 199–201, 1942.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. PMLR, 2015.

Guanghui Lan, Zhaosong Lu, and Renato D. C. Monteiro. Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126 (1):1–29, 2011.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *arXiv:1701.03961*, 2017.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

Tianyi Lin, Nhat Ho, and Michael I. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. `http://www-personal.umich.edu/~minhnhat/Arxiv_submission.pdf`, 2019.

Marie Maros and Joakim Jaldén. Panda: A dual linearly converging method for distributed optimization over time-varying undirected graphs. *arXiv preprint arXiv:1803.08328*, 2018.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, Wei Shi, and César A Uribe. Geometrically convergent distributed optimization with uncoordinated step-sizes. In *American Control Conference (ACC), 2017*, pages 3950–3955. IEEE, 2017.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *arXiv:1809.05895*, 2018.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *arXiv:1803.00567*, 2018.

Giovanni Puccetti, Ludger Rüschendorf, and Steven Vanduffel. On the computation of Wasserstein barycenters. `https://ssrn.com/abstract=3276147`, 2018.

Alexander Rogozin, César A Uribe, Alexander Gasnikov, Nikolay Malkovsky, and Angelia Nedić. Optimal distributed optimization on slowly time-varying graphs. *arXiv preprint arXiv:1805.06045*, 2018.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015. ISBN 9783319208282.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3027–3036, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/scaman17a.html`.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. II. *Proc. Amer. Math. Soc.*, 45:195–198, 1974.

Justin Solomon, Raif M. Rustamov, Leonidas Guibas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–306–I–314. PMLR, 2014.

Daniel A Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014.

Matthew Staib, Sebastian Claici, Justin M Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2647–2658. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6858-parallel-streaming-wasserstein-barycenters.pdf`.

Fedor Stonyakin, Darina Dvinskikh, Pavel Dvurechensky, Alexey Kroshnin, Olesya Kuznetsova, Artem Agafonov, Alexander Gasnikov, Alexander Tyurin, Cesar Uribe, Dmitry Pasechnyuk, and Sergei Artamonov. Gradient methods for problems with inexact model of the objective. *arXiv:1902.09001*, 2019.

Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28(1):96–134, 2018. doi: 10.1137/16M1093094. URL `https://doi.org/10.1137/16M1093094`. arXiv:1507.06243.

C. A. Uribe, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and A. Nedić. Distributed computation of wasserstein barycenters over networks. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6544–6549, 2018. arXiv:1803.02933.

César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. *arXiv preprint arXiv:1809.00710*, 2018.

P Vaidya. Solving linear equations with diagonally dominant matrices by constructing good preconditioners. Technical report, Technical report, Department of Computer Science, University of Illinois, 1990.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Hoi-To Wai, Nikolaos M Freris, Angelia Nedic, and Anna Scaglione. Sucag: Stochastic unbiased curvature-aided gradient method for distributed optimization. *arXiv preprint arXiv:1803.08198*, 2018.

Xuyang Wu and Jie Lu. Fenchel dual gradient methods for distributed convex optimization over time-varying networks. In *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*, pages 2894–2899. IEEE, 2017.

Alp Yurtsever, Quoc Tran-Dinh, and Volkan Cevher. A universal primal-dual convex optimization framework. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 3150–3158, Cambridge, MA, USA, 2015. MIT Press.