On the Complexity of Approximating Wasserstein Barycenters

Alexey Kroshnin [1 2 3] **Darina Dvinskikh** [4 1] **Pavel Dvurechensky** [4 1] **Alexander Gasnikov** [5 1 2] **Nazarii Tupitsa** [1 5]
César A. Uribe [6]

Abstract

We study the complexity of approximating the Wasserstein barycenter of m discrete measures, or histograms of size n, by contrasting two alternative approaches that use 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 mn^2/ε^2 to approximate the original non-regularized barycenter. On the other hand, using an approach based on accelerated gradient descent, we obtain a complexity proportional to mn^2/ε. 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.

Introduction

Optimal transport (OT) (Monge, 1781; Kantorovich, 1942) is becoming increasingly popular in the statistics, machine learning and optimization communities. Statistical methods based on optimal transport are readily available (Bigot et al., 2012; Del Barrio et al., 2015; Ebert et al., 2017; Le Gouic & Loubes, 2017), as well as many applications in 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), among others. 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 & Doucet, 2014). In this paper, we focus on the computational aspects of optimal transport, namely on the complexity approximating a Wasserstein barycenter of a set of histograms.

Starting with Altschuler et al. (2017), several groups of authors addressed the question of the Wasserstein distance approximation complexity (Chakrabarty & Khanna, 2018; Dvurechensky et al., 2018b; Blanchet et al., 2018; Lin et al., 2019). Implementable schemes based on Sinkhorn's algorithm were first applied to OT in Cuturi (2013), see also (Genevay et al., 2016). Also, accelerated gradient descent methods were 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; Dvurechensky et al., 2018a), are in some sense close, but do not provide an explicit answer.

Following Dvurechensky et al. (2018b), we study two alternative approaches for approximating Wasserstein barycenters based on entropic regularization (Cuturi, 2013). The first approach is based on the Iterative Bregman Projection (IBP) algorithm (Benamou et al., 2015), which can be considered as a general alternating projections algorithm. 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 in the Big Data regime, i.e., when the size of the 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 a network structure

[1]Institute for Information Transmission Problems RAS, Moscow, Russia [2]National Research University Higher School of Economics, Moscow, Russia [3]Université Claude Bernard Lyon 1, Villeurbanne, France [4]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany [5]Moscow Institute of Physics and Technology, Moscow, Russia [6]Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: Pavel Dvurechensky <pavel.dvurechensky@wias-berlin.de>, Alexey Kroshnin <akroshnin@hse.ru>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

given by an arbitrary connected graph. In a special case of a centralized architecture, i.e., if there is a central "master" node connected 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 (Scaman et al., 2017; Nedić et al., 2017).

Related Work. It is very hard to cover all the increasing stream of works on OT and we mention the books Villani (2008); Santambrogio (2015); Peyré & Cuturi (2018) as a starting point and the references therein. Approximation of Wasserstein barycenter was considered in Cuturi & 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). Considering the primal-dual approach based on accelerated gradient descent, our paper shares some similarities with (Cuturi & Peyré, 2016) with the main difference that we are focused on complexity and scalability of computations and explicitly analyzing 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; Gasnikov et al., 2016; Dvurechensky et al., 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 & Tsitsiklis (1989). Initial algorithms, such as Distributed Gradient Descent (Nedic & 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 have been extended to other scenarios such as time-varying graphs (Rogozin et al., 2018; Maros & Jaldén, 2018; Wu & 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. 1. We consider the 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 $1/(\gamma\varepsilon)$ to achieve accuracy ε. The second one is based on accelerated gradient descent (AGD) and has complexity proportional to $1/(\sqrt{\gamma\varepsilon})$. 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.

2. We show how to choose the regularization parameter in order to find an ε-approximation for the non-regularized Wasserstein barycenter. The resulting complexity for IBP is proportional to mn^2/ε^2 and for AGD is be proportional to $mn^{2.5}/\varepsilon$.

3. We solve the stability issues of the IBP and AGD approaches, present when the desired accuracy is high, or conversly when ε is small, by proposing a proximal-IBP method, which can be considered as a proximal method using IBP on each iteration to find the next iterate.

The full version of the paper with the proofs can be found as supplementary material and as (Kroshnin et al., 2019).

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 in $S_n(1)$, we introduce the set of coupling measures as

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

For coupling measure $\pi \in \mathbb{R}^{n\times n}_+$, we denote the negative entropy (up to an additive constant) as

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

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

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

We also define a cost matrix $C \in \mathbb{R}^{n\times n}_+$, which element c_{ij} corresponds to the cost of moving an element of 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 $\chi(W) = \lambda_{\max}(W)/\lambda^+_{\min}(W)$. 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 $C \in \mathbb{R}^{n\times n}$, following Cuturi (2013), we define entropy-regularized OT-distance for $\gamma \ge 0$:

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

For $\gamma = 0$ we use shortcut notation $\mathcal{W}(p, q)$ and refer to it as non-regularized distance. For a given set of probability measures $\{p_1, \dots, p_m\}$ and cost matrices $C_1, \dots, C_m \in \mathbb{R}^{n\times n}_+$ we define their weighted regularized barycenter with weights $w \in S_m(1)$ as a solution of the following problem:

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

where a solution of this problem for $\gamma = 0$ referred to as non-regularized barycenter.

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 $c \coloneqq \max_{l=1,\dots,m}\|C_l\|_\infty$. 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 $\widetilde{O}\left(c^2mn^2/\varepsilon^2\right)$ 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 $\widetilde{O}\left(c^2n^2/\varepsilon^2\right)$ 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 (2). We reformulate this problem as

$$\min_{\substack{q\in S_n(1),\\ \pi_l\in\Pi(p_l,q), l=1,\dots,m}} \sum_{l=1}^m w_l\left\{\langle \pi_l, C_l\rangle + \gamma H(\pi_l)\right\}$$
$$= \min_{\substack{\pi_l\mathbf{1}=p_l,\ \pi_l^\mathsf{T}\mathbf{1}=\pi_{l+1}^\mathsf{T}\mathbf{1},\\ \pi_l\in\mathbb{R}^{n\times n}_+,\ l=1,\dots,m}} \sum_{l=1}^m w_l\left\{\langle \pi_l, C_l\rangle + \gamma H(\pi_l)\right\}, \tag{3}$$

and construct its dual (see Lemma 2.1). To solve the dual problem we reformulate the IBP algorithm as a blockwise minimization, as shown in Algorithm 1 (this equivalence is a general fact for Dykstra's algorithm). 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). Following the approach of Benamou et al. (2015), we present problem (3) in a Kullback–Leibler projection form, i.e.,

$$\min_{\boldsymbol{\pi}\in\mathcal{C}_1\cap\mathcal{C}_2} \sum_{l=1}^m w_l KL\left(\pi_l|K_l\right), \tag{4}$$

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

$$\mathcal{C}_1 = \left\{\boldsymbol{\pi} = [\pi_1, \dots, \pi_m] : \forall l\ \pi_l\mathbf{1} = p_l\right\},$$
$$\mathcal{C}_2 = \left\{\boldsymbol{\pi} = [\pi_1, \dots, \pi_m] : \pi_1^\mathsf{T}\mathbf{1} = \dots = \pi_m^\mathsf{T}\mathbf{1}\right\}. \tag{5}$$

The IBP algorithm 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 (3) derived in Lemma 2.1, and leads to Algorithm 1.

Lemma 1. *The dual problem of (3) is (up to a multiplicative constant)*

$$\min_{\substack{\mathtt{u},\mathtt{v}\\ \sum_{l=1}^m w_l v_l = 0}} f(\mathtt{u}, \mathtt{v}) \coloneqq \sum_{l=1}^m w_l\left\{\langle \mathbf{1}, B_l(u_l, v_l)\mathbf{1}\rangle - \langle u_l, p_l\rangle\right\}, \tag{6}$$

$\mathtt{u} = [u_1, \dots, u_m]$, $\mathtt{v} = [v_1, \dots, v_m]$, $u_l, v_l \in \mathbb{R}^n$, *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 (3) is given by the formula $[\boldsymbol{\pi}^*_\gamma]_l = B_l(u_l^*, v_l^*)$, where $(\mathtt{u}^*, \mathtt{v}^*)$ is the solution of the problem (6).

Proof of Lemma is shown in the supplementary materials.

Algorithm 1 Dual Iterative Bregman Projection

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

$u_l^0 := 0, v_l^0 := 0, K_l := \exp\left(-\frac{C_l}{\gamma}\right), l = 1, \dots, m$

repeat

  if $t \bmod 2 = 0$ **then**

    $u_l^{t+1} := \ln p_l - \ln K_l e^{v_l^t}, \quad \mathtt{v}^{t+1} := \mathtt{v}^t$

  else

    $v_l^{t+1} := \sum_{k=1}^m w_k \ln K_k^\mathsf{T} e^{u_k^t} - \ln K_l^\mathsf{T} e^{u_l^t}, \quad \mathtt{u}^{t+1} := \mathtt{u}^t$

  end if

  $t := t + 1$

until $\sum_{l=1}^m w_l \|B_l^\mathsf{T}(u_l^t, v_l^t)\mathbf{1} - \bar{q}^t\|_1 \le \varepsilon'$ and $\sum_{l=1}^m w_l \|B_l(u_l^t, v_l^t)\mathbf{1} - p_l\|_1 \le \varepsilon'$, where $\bar{q}^t := \sum_{l=1}^m w_l B_l^\mathsf{T}(u_l^t, v_l^t)\mathbf{1}$

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

Before we move to the analysis of the algorithm let us discuss its scalability by using the centralized distributed computations framework. This framework includes a master and slave nodes. Each l-th slave node stores p_l, C_l, K_l and variables u_l^t, v_l^t. At each iteration t, a slave node calculates $K_l^\mathsf{T} e^{u_l^t}$ and sends it to the master node, which aggregates these products as $\sum_{k=1}^m w_k \ln K_k^\mathsf{T} e^{u_k^t}$ and sends this sum back to the slave nodes. Based on this information, slave nodes update v_l^t and u_l^t. Thus, 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 1 stops in number of iterations N satisfying*

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

Proof. Proof mainly follows ideas from (Dvurechensky et al., 2018b) for Sinkhorn algorithm. First, one can show that the following results hold:

1. for any $t \ge 0$, and $l = 1, \dots, m$

$$\max_j [v_l^t]_j - \min_j [v_l^t]_j \le R_v \le 2\frac{\max_l \|C_l\|_\infty}{\gamma}; \quad (7)$$

2. for any even $t \ge 2$ we have

$$\tilde{f}(\mathtt{u}^t, \mathtt{v}^t) := f(\mathtt{u}^t, \mathtt{v}^t) - f(\mathtt{u}^*, \mathtt{v}^*) \le R_v \sum_{l=1}^m w_l \|q_l^t - \bar{q}^t\|_1, \quad (8)$$

where $q_l^t := B_l^\mathsf{T}(u_l^t, v_l^t)\mathbf{1}$ and $\bar{q}^t := \sum_{l=1}^m w_l q_l^t$;

3. for any odd $t \ge 1$ the following bound on the change of objective function $f(\cdot, \cdot)$ holds:

$$f(\mathtt{u}^t, \mathtt{v}^t) - f(\mathtt{u}^{t+1}, \mathtt{v}^{t+1}) \ge \frac{1}{11}\left(\sum_{l=1}^m w_l \|q_l^t - \bar{q}^t\|_1\right)^2. \quad (9)$$

Now let us move to the proof of complexity bound. To simplify derivation we define $\delta_t := \tilde{f}(\mathtt{u}^t, \mathtt{v}^t)$. If $t \ge 2$ is even, then (8), (9), and the stopping criterion give us the following bound:

$$\delta_t - \delta_{t+1} \ge \max\left\{\frac{(\varepsilon')^2}{11}, \frac{(\delta_t)^2}{11 R_v^2}\right\}.$$

If t is odd then we have at least $\delta_{t+1} \le \delta_t$. These inequalities result in the following estimates:

$$t \le 2 + 22 R_v^2 \left(\frac{1}{\delta_t} - \frac{1}{\delta_1}\right), \quad (10)$$

$$k \le 1 + \frac{22(\delta_t - \delta_k)}{(\varepsilon')^2} \le 1 + \frac{22\delta_t}{(\varepsilon')^2}. \quad (11)$$

To combine the two estimates (10) and (11), we consider a switching strategy parametrized by number $s \in (0, \delta_1)$. First t iterations we use (10), resulting in δ_t becomes below some s. Then, we use s as a starting point and estimate the remaining number of iteration by (11). The quantity s can be found from the minimization

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

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

$$N \le 4 + \frac{44 R_v}{\varepsilon'} \le 4 + \frac{88 \max_l \|C_l\|_\infty}{\gamma \varepsilon'}.$$

□

2.2. Approximating Non-regularized WB by IBP

To find an approximate non-regularized barycenter, i.e. solution to problem (2) with $\gamma = 0$, we apply Algorithm 1 with a suitable choice of γ and ε' and average marginals $q_1, \dots, q_m$ with weights w_l, this leads to Algorithm 2.

Algorithm 2 Finding Wasserstein barycenter by IBP

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

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

Find $B_1 := B_1(u_1^t, v_1^t), \dots, B_m := B_m(u_m^t, v_m^t)$ by Algorithm 1 with accuracy ε'

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

Output: q

Theorem 2 presents complexity bound for Algorithm 2.

Theorem 2. *For $\varepsilon > 0$, Algorithm 2 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 (2) 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 1.

Remark 1. As each iteration of Algorithm 1 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[1] (Peyré & Cuturi, 2018), or $M_{m,n} = O\left(n \sum_l \operatorname{rank}(C_l)\right)$ for low-rank matrices.

Proof. Let $\boldsymbol{\pi}^* = [\pi_1^*, \dots, \pi_m^*]$ be a solution to the non-regularized problem. Equation (7) and duality yields

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

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. $\|\check{B}_l - B_l\|_1 \leq \|B_l \mathbf{1} - p_l\|_1 + 2\sum_j [B_l^\mathsf{T}\mathbf{1} - q]_j^+$ for all $l = 1, \dots, m$. Their existence immediately follows from the proof of Theorem 4 from (Altschuler et al., 2017). If stopping time t is even, $B_l \mathbf{1} = p_l$, therefore $q = \bar{q}^t$, and $\|B_l - \check{B}_l\|_1 \leq \|q_l^t - \bar{q}^t\|_1$. If t is odd, $B_l^\mathsf{T}\mathbf{1} = \bar{q}^t \leq q$ and $\|B_l - \check{B}_l\|_1 \leq \|B_l \mathbf{1} - p_l\|_1$. 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

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

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

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

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 (Xie et al., 2018), where this idea is used for Wasserstein distance, our method is inspired by proximal point algorithm with general Bregman divergence $V(x, y)$ (Chen & 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 (2). Given the sets $\mathcal{C}_1, \mathcal{C}_2$ defined in (5), 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

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

The proximal gradient method has the following form

$$\boldsymbol{\pi}^{k+1} = \mathbf{prox}(\boldsymbol{\pi}^k). \tag{12}$$

Then we use Iterative Bregman Projection for finding the barycenter.

We underline that in this setting, there is no need to choose γ to be small as it prescribed by Theorem 3. Algorithm 3 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 & 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. 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 $\gamma := \gamma/2$ and solve internal problem once

again. We stop this repeating procedure at the moment when the complexity of internal problem growth significantly. This moment allows us to detect the optimal value of γ. On the next external iterations one may use this γ.

Algorithm 3 can be implemented in centralized distributed setting in the same way as Algorithm 1.

Algorithm 3 Finding Wasserstein barycenter by proximal IBP

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

1: **for** $t = 0, \dots, T-1$ **do**
2: Run Algorithm 1 with cost matrices $C_l^t := C_l - \gamma \ln \pi_l^t$, parameter γ and accuracy $\varepsilon' \propto \frac{\tilde{\varepsilon}}{\max_l \|C_l^t\|_\infty}$, and obtain matrices $B_1, \dots, B_m$
3: $\pi_l^{t+1} := \text{Round}\left(B_l, p_l, \bar{q}^{t+1}\right) \in \Pi(p_l, \bar{q}^{t+1})$, where Round is Algorithm 2 from (Altschuler et al., 2017) and $\bar{q}^{t+1} := \sum_{l=1}^m w_l B_l^\top \mathbf{1}$
4: **end for**

Output: $\bar{q}^T$

3. Complexity by Primal-Dual Accelerated Gradient Descent

In this section, we consider the Primal-Dual Accelerated Gradient Descent method for approximating Wasserstein barycenters. First, we consider the regularized barycenter, construct a dual problem to (2) 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. These 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 the Wasserstein barycenter problem

We rewrite problem (2) in an equivalent way as

$$\min_{\substack{q_1,\dots,q_m \in S_n(1) \\ q_1 = \dots = q_m}} W_\gamma(\mathtt{p}, \mathtt{q}) := \sum_{l=1}^m w_l \mathcal{W}_{\gamma(l)}(p_l, q_l), \quad (13)$$

where $\mathtt{p} = [p_1, \dots, p_m]^T$ and $\mathtt{q} = [q_1, \dots, q_m]^T$, we also use different regularizer $\gamma_l = \gamma(l)$ for each l-th Wasserstein distance. Next, we write a dual problem by dualizing the equality constraints $q_1 = \dots = q_m$. 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. $\text{Ker}(\bar{\text{W}}) = \text{span}(\mathbf{1})$. Then, defining $W = \bar{W} \otimes I_n$ and using the fact $q_1 = \dots = q_m \iff \sqrt{W}\mathtt{q} = 0$, we equivalently rewrite problem (13) as

$$\max_{\substack{q_1,\dots,q_m \in S_n(1), \\ \sqrt{W}\mathtt{q} = 0}} -\sum_{l=1}^m w_l \mathcal{W}_{\gamma(l)}(p_l, q_l), \quad (14)$$

Therefore, we obtain the dual problem

$$\min_{\mathtt{u} \in \mathbb{R}^{mn}} \max_{\mathtt{q} \in \mathbb{R}^{nm}} \left\{ \sum_{l=1}^m \langle u_l, [\sqrt{W}\mathtt{q}]_l \rangle - \sum_{l=1}^m w_l \mathcal{W}_{\gamma(l)}(p_l, q_l) \right\}$$
$$= \min_{\mathtt{u} \in \mathbb{R}^{mn}} \sum_{l=1}^m w_l \mathcal{W}^*_{\gamma(l), p_l}([\sqrt{W}\mathtt{u}]_l / w_l), \quad (15)$$

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 (15) has L-Lipschitz-continuous gradient, where constant L is estimated below in Lemma 3. Since the dual problem is smooth, we apply primal-dual accelerated gradient descent Algorithm 4 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 4. 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

$$\bar{W} : \{\forall i = 1, \dots, m-1 \; \bar{W}_{ii} = 1, \; \bar{W}_{im} = \bar{W}_{mi} = -1,$$
$$\bar{W}_{mm} = m - 1\}. \quad (16)$$

Algorithm 4 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 4 as if it is implemented on one machine, disregarding that it can be used for distributed setup.

Algorithm 4 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., $\tilde{p}_l = \left(1 - \frac{\varepsilon}{8}\right)\left(p_l + \frac{\varepsilon}{n(8-\varepsilon)}\mathbf{1}\right)$ and sets 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.

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

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

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

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

6: 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.

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

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

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

10: $q_l^{k+1} = \frac{1}{A_{k+1}} \sum_{l=0}^{k+1} \alpha_i q_i(\lambda_l^{k+1}) = (\alpha_{k+1} q_i(\lambda_l^{k+1}) + A_k q_l^k)/A_{k+1}$, where $q_l(\cdot) = \nabla \mathcal{W}^*_{\gamma(l),p_l}(\cdot)$ defined in step 4. {Primal update}

11: **end for**

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

Theorem 3. *Algorithm 4 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) with $\gamma = 0$, i.e. finds a vector $\mathtt{q}^N = [q_1^\mathsf{T}, \ldots, q_m^\mathsf{T}]^\mathsf{T}$ s.t.*

$$\sum_{l=1}^m w_l \mathcal{W}(p_l, q_l^N) - \sum_{l=1}^m w_l \mathcal{W}(p_l, q^*) \leq \varepsilon, \text{ and}$$
$$\|\sqrt{W}\mathtt{q}^N\|_2 \leq \varepsilon/2R, \tag{17}$$

where q^* is an unregularized barycenter, i.e. is a solution to (2) with $\gamma = 0$, and R is a bound on the solution to the dual problem given in Lemma 3. The total number of arithmetic operations is $O\left(N \cdot n(mn + \mathrm{nnz}(\bar{W}))\right)$. Moreover, there exists a choice of matrix $\bar{W}$ such that, if, without loss of generality, $\|C_l\|_\infty \leq 1$, and the weights $w_l = 1/m$, $l = 1, ..., m$, the complexity of approximating non-regularized barycenter by Algorithm 4 is $\widetilde{O}(mn^{2.5}/\varepsilon)$.

The proof is based on the complexity theorem of primal-dual accelerated gradient descent for a particular pair of primal-dual problems (13)–(15).

Theorem 4 (see Theorem 2 from (Dvurechensky et al., 2017))**.** *Let accelerated primal-dual gradient descent be applied to the pair of problems (13)–(15). Then the inequalities*

$$\sum_{l=1}^m w_l \mathcal{W}_{\gamma(l)}(p_l, q_l^N) - \sum_{l=1}^m \mathcal{W}_{\gamma(l)}(p_l, q^*) \leq \varepsilon/2,$$
$$\|\sqrt{W}\mathtt{q}^N\|_2 \leq \varepsilon/2R \tag{18}$$

hold no later than after $N = \sqrt{32LR^2/\varepsilon}$ 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 3 inspired by (Lan et al., 2017).

Lemma 2. *Let in (13) $\gamma(l) = \gamma/w_l$ for some $\gamma > 0$, and $\mathcal{W}_\gamma^*(\mathtt{u})$ denote the dual objective in (15). Then its gradient is $L = \lambda_{\max}(W)/\gamma$-Lipschitz continuous w.r.t. 2-norm.*

Lemma 3. *Let $\mathtt{q}_\gamma^*$ be the optimal solution of problem (2) with minimal 2-norm, then there exists an optimal dual solution $\mathtt{u}^* = [u_1^*, \ldots, u_m^*]$ for problem (15) satisfying $\|\mathtt{u}^*\|_2 \leq R$ with*

$$R^2 = \frac{2n\sum_{l=1}^m w_l^2 \|C_l\|_\infty^2}{\lambda_{\min}^+(W)}. \tag{19}$$

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

Proof of Theorem 3. Using Theorem 4 and that $KL(\pi|\theta) \in [0, 2\ln n]$ we get the following inequality

$$\sum_{l=1}^m w_l \mathcal{W}(p_l, q_l^N) - \sum_{l=1}^m w_l \mathcal{W}(p_l, q^*) \leq \varepsilon/2$$
$$+ 2\ln n \sum_{l=1}^m w_l \gamma(l). \tag{20}$$

Since $\gamma(l) = \gamma/w_l$ with $\gamma = \varepsilon/(4m\ln n)$, we obtain that the inequality (17) holds. Combining the values of γ, L from Lemma 2, R from Lemma 3 with the estimate for N in Theorem 4 and the fact that $\chi(W) = \chi(\bar{W})$, we obtain the desired 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 $O(n \cdot \mathrm{nnz}(\bar{W}_l))$ arithmetic operations, where $\mathrm{nnz}(\bar{W}_l)$ 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

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

and the total complexity follows from multiplying this value by N. As for the choice of $\bar{W}$ one can show (by using graph sparsificators) that it can be chosen such that $\chi(W) = \chi(\bar{W}) = O(\mathrm{Poly}(\ln(m)))$ and $\mathrm{nnz}(\bar{W}) = O(m\mathrm{Poly}(\ln(m)))$. For details on the graph sparsificators we refer to (Vaidya, 1990; Bern et al., 2006; Spielman & Teng, 2014). Substituting the weights $w_l = 1/m$, $l = 1, ..., m$ to the bound for N, we obtain that the complexity of approximating non-regularized barycenter by Algorithm 4 is $\widetilde{O}(mn^{2.5}/\varepsilon)$. In the distributed setting, each of m nodes makes $\widetilde{O}(n^{2.5}/\varepsilon)$ arithmetic operations, while the number of communications rounds is $\widetilde{O}(\sqrt{n}/\varepsilon)$. □

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 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.

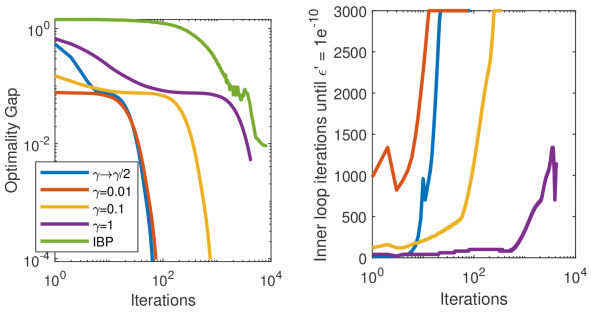

Figure 1. Distance to an optimal barycenter for the IBP method and the ProxIBP method.

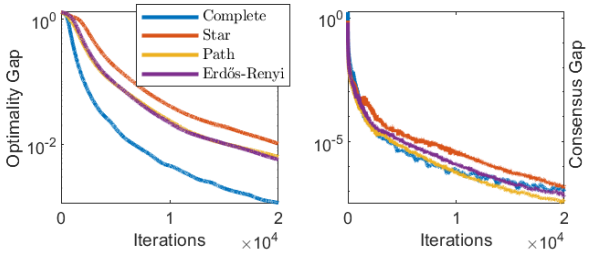

Figure 2. Optimality gap and consensus gap for the primal-dual accelerated gradient descent method for four classes of networks: complete, star, path and Erdős-Renyi random graph.

Conclusion

In this paper, we show that the IBP algorithm from Benamou et al. (2015) for the Wasserstein barycenter problem can be implemented in a centralized distributed manner such that each node requires $\widetilde{O}\left(n^2/\varepsilon^2\right)$ 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 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 each node fulfils $\widetilde{O}(n^{2.5}/\varepsilon)$ arithmetic operations and the number of communication rounds is $\widetilde{O}(\sqrt{n}/\varepsilon)$.

Table 1. Optimality Gap for the Approximate Barycenter

	ProxIBP			IBP	Algo. 4	
Iter.	$\gamma \to \gamma/2$	$\gamma = 0.01$	$\gamma = 1$		Complete	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

Acknowledgements

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

References

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

Anikin, A. S., Gasnikov, A. V., Dvurechensky, P. E., Tyurin, A. I., and Chernov, A. V. 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.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv:1701.07875*, 2017.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Bern, M., Gilbert, J. R., Hendrickson, B., Nguyen, N., and Toledo, S. Support-graph preconditioners. *SIAM Journal on Matrix Analysis and Applications*, 27(4):930–951, 2006.

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

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

Bigot, J., Gouet, R., Klein, T., and López, A. Geodesic PCA in the wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.*, 53(1):1–26, 02 2017.

Blanchet, J., Jambulapati, A., Kent, C., and Sidford, A. Towards optimal running times for optimal transport. *arXiv:1810.07717*, 2018.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. 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.

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

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

Chernov, A., Dvurechensky, P., and Gasnikov, A. Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In Kochetov, Y., Khachay, M., Beresnev, V., Nurminski, E., and Pardalos, P. (eds.), *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*, pp. 391–403. Springer International Publishing, 2016.

Claici, S., Chien, E., and Solomon, J. Stochastic Wasserstein barycenters. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 999–1008. PMLR, 2018. URL http://proceedings.mlr.press/v80/claici18a.html.

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

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 685–693, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/cuturi14.html.

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

Del Barrio, E., Lescornel, H., and Loubes, J.-M. A statistical analysis of a deformation model with wasserstein barycenters : estimation procedure and goodness of fit test. *arXiv:1508.06465*, 2015.

Dvurechensky, P., Gasnikov, A., Gasnikova, E., Matsievsky, S., Rodomanov, A., and Usik, I. 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*, pp. 584–595, 2016. arXiv:1606.08988.

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

Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedić, A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, NeurIPS 2018, pp. 10783–10793. Curran Associates, Inc., 2018a. arXiv:1802.04367.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1367–1376, 2018b. arXiv:1802.04367.

Ebert, J., Spokoiny, V., and Suvorikova, A. Construction of non-asymptotic confidence sets in 2-Wasserstein space. *arXiv:1703.03658*, 2017.

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

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3440–3448. Curran Associates, Inc., 2016.

He, L., Bian, A., and Jaggi, M. Cola: Decentralized linear learning. In *Advances in Neural Information Processing Systems*, pp. 4541–4551, 2018.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H. H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1501–1509. PMLR, 2017.

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

Kroshnin, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Tupitsa, N., and Uribe, C. On the complexity of approximating Wasserstein barycenter. *arXiv:1901.08686*, 2019.

Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 957–966. PMLR, 2015.

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

Lan, G., Lee, S., and Zhou, Y. Communication-efficient algorithms for decentralized and stochastic optimization. *arXiv:1701.03961*, 2017.

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

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

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

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

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

Nedić, A., Olshevsky, A., Shi, W., and Uribe, C. A. Geometrically convergent distributed optimization with uncoordinated step-sizes. In *American Control Conference (ACC), 2017*, pp. 3950–3955. IEEE, 2017.

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

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *arXiv:1809.05895*, 2018.

Peyré, G. and Cuturi, M. Computational optimal transport. *arXiv:1803.00567*, 2018.

Puccetti, G., Rüschendorf, L., and Vanduffel, S. On the computation of Wasserstein barycenters. https://ssrn.com/abstract=3276147, 2018.

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

Santambrogio, F. *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.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3027–3036, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Solomon, J., Rustamov, R. M., Guibas, L., and Butscher, A. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. I–306–I–314. PMLR, 2014.

Spielman, D. A. and Teng, S.-H. 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.

Staib, M., Claici, S., Solomon, J. M., and Jegelka, S. Parallel streaming wasserstein barycenters. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2647–2658. Curran Associates, Inc., 2017.

Stonyakin, F., Dvinskikh, D., Dvurechensky, P., Kroshnin, A., Kuznetsova, O., Agafonov, A., Gasnikov, A., Tyurin, A., Uribe, C., Pasechnyuk, D., and Artamonov, S. Gradient methods for problems with inexact model of the objective. *arXiv:1902.09001*, 2019a.

Stonyakin, F., Gasnikov, A., Tyurin, A., Pasechnyuk, D., Agafonov, A., Dvurechensky, P., Dvinskikh, D., Kroshnin, A., and Piskunova, V. Inexact model: A framework for optimization and variational inequalities. *arXiv:1902.00990*, 2019b.

Tran-Dinh, Q., Fercoq, O., and Cevher, V. 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.

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

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

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

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

Wai, H.-T., Freris, N. M., Nedic, A., and Scaglione, A. Sucag: Stochastic unbiased curvature-aided gradient method for distributed optimization. *arXiv preprint arXiv:1803.08198*, 2018.

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

Xie, Y., Wang, X., Wang, R., and Zha, H. A fast proximal point method for computing Wasserstein distance. *arXiv:1802.04307*, 2018.

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