Improved Complexity Bounds in Wasserstein Barycenter Problem

Darina Dvinskikh
Weierstrass Institute,
IITP RAS,
MIPT
darina.dvinskikh@wias-berlin.de

Daniil Tiapkin
HSE University
dtyapkin@hse.ru

Abstract

In this paper, we focus on computational aspects of the Wasserstein barycenter problem. We propose two algorithms to compute Wasserstein barycenters of m discrete measures of size n with accuracy ε. The first algorithm, based on mirror prox with a specific norm, meets the complexity of celebrated accelerated iterative Bregman projections (IBP), namely $\widetilde{O}(mn^2\sqrt{n}/\varepsilon)$, however, with no limitations in contrast to the (accelerated) IBP, which is numerically unstable under small regularization parameter. The second algorithm, based on area-convexity and dual extrapolation, improves the previously best-known convergence rates for the Wasserstein barycenter problem enjoying $\widetilde{O}(mn^2/\varepsilon)$ complexity.

1 Introduction

The theory of optimal transport (OT) provides a natural framework to compare objects that can be modeled as probability measures (images, videos, texts and etc.). Nowadays, the OT metric gains popularity in various fields such as statistics (Ebert et al., 2017; Bigot et al., 2012), machine learning (Arjovsky et al., 2017; Solomon et al., 2015), economics and finance (Rachev et al., 2011). However, the outstanding results of OT come with large computations. Indeed, to solve the OT problem between two discrete histograms of size n, one needs to make $\tilde{O}(n^3)$ arithmetic calculations (Tarjan, 1997; Peyré and Cuturi, 2018), e.g., by using simplex method or interior-point method. To overcome the computational issue, entropic regularization of the OT was proposed by Cuturi (2013). It enables an application of the Sinkhorn's algorithm, which is based on alternating minimization procedures and has $\widetilde{O}(n^2\|C\|^2_\infty/\varepsilon^2)$ convergence rate (Dvurechensky et al., 2018) to approximate a solution of OT with ε-precision. Here $C \in \mathbb{R}^{n\times n}_+$ is a ground cost matrix of transporting a unit of mass between probability measures, and the regularization parameter before negative entropy is of order ε. The Sinkhorn's algorithm can be accelerated to $\widetilde{O}\left(n^2\sqrt{n}\|C\|_\infty/\varepsilon\right)$ convergence rate (Guminov et al., 2019). In practice, the accelerated Sinkhorn's algorithm converges faster than the Sinkhorn's algorithm, and in theory, it has better dependence on ε but not on n. However, all entropy-regularized based approaches are numerically unstable when the regularizer parameter γ before negative entropy is small (this also means that precision ε is high as γ must be selected proportional to ε (Peyré and Cuturi, 2018; Kroshnin et al., 2019)). The recent work of Jambulapati et al. (2019) provides an optimal method for solving the OT problem, based on dual extrapolation (Nesterov, 2007) and area-convexity (Sherman, 2017), with convergence rate $\widetilde{O}(n^2\|C\|_\infty/\varepsilon)$. This method works without additional penalization and, moreover, it eliminates the term $\sqrt{n}$ in the bound for the accelerated Sinkhorn's algorithm. The rate $\widetilde{O}(n^2\|C\|_\infty/\varepsilon)$ was also obtained in a number of works of Blanchet et al. (2018); Allen-Zhu et al. (2017); Cohen et al. (2017).

The OT metric finds natural application to the Wasserstein barycenter (WB) problem. Regularizing each OT distance in the sum by negative entropy leads to presenting the WB problem as Kullback–Leibler projection that can be performed by the iterative Bregman projections (IBP) algorithm (Benamou et al., 2015). The IBP is an extension of the Sinkhorn's algorithm for m measures, and hence, its complexity is m times more than the Sinkhorn complexity, namely $\widetilde{O}\left(mn^2\|C\|^2_\infty/\varepsilon^2\right)$ (Kroshnin et al., 2019). An analog of the accelerated Sinkhorn's algorithm for the WB problem of m measures is the accelerated IBP algorithm with complexity $\widetilde{O}\left(mn^2\sqrt{n}\|C\|_\infty/\varepsilon\right)$ (Guminov et al., 2019), that is

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

also m times more than the accelerated Sinkhorn complexity. Another fast version of the IBP algorithm was recently proposed by Lin et al. (2020), named FastIBP with complexity $\widetilde{O}\left(mn^2\sqrt[3]{n}\|C\|_\infty^{4/3}/\varepsilon^{4/3}\right)$.

The main goal of this paper is providing an algorithm for the WB problem beating the complexity of the existing algorithms. To do so, we develop the idea of the paper of Jambulapati et al. (2019) that provides an optimal algorithm for the OT problem.

1.1 Contribution

Our first contribution is proposing an algorithm which does not suffer from a small value of the regularization parameter and, at the same time, has complexity not worse than the celebrated (accelerated) IBP. Our algorithm, running in $\widetilde{O}(mn^2\sqrt{n}/\varepsilon)$ wall-clock time, is based on mirror prox with specific prox-function.

The second contribution is providing an algorithm that has better complexity than the (accelerated) IBP. Motivated by the work of Jambulapati et al. (2019) proposing an optimal way of solving the OT problem with better complexity bounds than (accelerated) Sinkhorn, we develop an optimal algorithm for the WB problem of $\widetilde{O}(mn^2/\varepsilon)$ complexity. Our approach is based on rewriting the WB problem as a saddle-point problem and further application of the dual extrapolation scheme under the weaker convergence requirements of area-convexity.

We notice that the convergence rate obtained by our first algorithm is worse than the complexity of our second algorithm, however, in some sense, the first algorithm can be seen as a simplified version of the second algorithm and, hence, the first approach simplifies the understanding of the second approach.

In Table 1, we illustrate our contribution by comparing our algorithms with the most popular algorithms for the WB problem.

Paper Organisation. The structure of the paper is the following. In Section 2, we reformulate the WB problem as a saddle-point problem. Sections 3 and 4 present two our new algorithms to solve the WB problem.

Notation. Let $\Delta_n = \{p \in \mathbb{R}^n_+ : \sum_{i=1}^n p_i = 1\}$ be the probability simplex. We use bold symbol for column vector $\mathbf{x} = (x_1^\top, \cdots, x_m^\top)^\top \in \mathbb{R}^{mn}$, where $x_1, ..., x_m \in \mathbb{R}^n$. Then we refer to the i-th component of vector $\mathbf{x}$ as $x_i \in \mathbb{R}^n$ and to the j-th component of vector x_i as $[x_i]_j$. For two vectors x, y of the same size, denotations x/y and $x \odot y$ stand for the element-wise product and element-wise division respectively. When functions, such as *log* or *exp*, are used on vectors, they are always applied element-wise. For some norm $\|\cdot\|_\mathcal{X}$ on space $\mathcal{X}$, we define the dual norm $\|\cdot\|_{\mathcal{X}^*}$ on the dual space $\mathcal{X}^*$ in a usual way: $\|s\|_{\mathcal{X}^*} = \max_{x\in\mathcal{X}}\{\langle s, x\rangle : \|x\| \le 1\}$. For a prox-function $d(x)$, we define the corresponding Bregman divergence $B(x, y) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle$. We denote by I_n the identity matrix, and by $0_{n\times n}$ zeros matrix.

Table 1: Algorithms and their rates of convergence for the Wasserstein barycenter problem

Approach	Complexity
IBP (Kroshnin et al., 2019)	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty^2}{\varepsilon^2}\right)$
Accelerated IBP (Guminov et al., 2019)	$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
FastIBP (Lin et al., 2020)	$\widetilde{O}\left(\frac{mn^2\sqrt[3]{n}\|C\|_\infty^{4/3}}{\varepsilon\sqrt[3]{\varepsilon}}\right)$
Mirror prox with specific norm (This work)	$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
Dual extrapolation with area-convexity (This work)	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty}{\varepsilon}\right)$

2 Problem Statement

In this section, we recall the optimal transport (OT) problem, the Wasserstein barycenter (WB) problem, and reformulate them as saddle-point problems.

Given two histograms $p, q \in \Delta_n$ and ground cost $C \in \mathbb{R}_+^{n\times n}$, the OT problem is formulated as follows

$$W(p, q) = \min_{X\in\mathcal{U}(p,q)} \langle C, X\rangle, \tag{1}$$

where X is a transport plan from transport polytope $\mathcal{U} = \{X \in \mathbb{R}_+^{n\times n}, X\mathbf{1} = p, X^\top\mathbf{1} = q\}$. Let d be vectorized cost matrix of C, x be vectorized transport plan of X, $b = \begin{pmatrix} p \\ q \end{pmatrix}$, and $A = \{0, 1\}^{2n\times n^2}$ be an incidence matrix. As $\sum_{i,j=1}^n X_{ij} = 1$, we following by the paper of Jambulapati et al. (2019) rewrite (1) as

$$W(p, q) = \min_{x\in\Delta_{n^2}} \max_{y\in[-1,1]^{2n}} \{d^\top x + 2\|d\|_\infty (y^\top Ax - b^\top y)\}. \tag{2}$$

Given histograms $q_1, q_2, ..., q_m \in \Delta_n$, a WB of these measures is a solution of the following problem

$$p^* = \arg\min_{p\in\Delta_n} \frac{1}{m}\sum_{i=1}^{m} W(p,q_i). \quad (3)$$

Then, we rewrite the WB problem (3) using the reformulation (2) of OT as follows

$$\min_{p\in\Delta_n} \frac{1}{m}\sum_{i=1}^{m} \min_{x_i\in\Delta_{n^2}} \max_{y_i\in[-1,1]^{2n}} \{d^\top x_i + 2\|d\|_\infty \left(y_i^\top A x_i - b_i^\top y_i\right)\}, \quad (4)$$

where $b_i = (p^\top, q_i^\top)^\top$.

Next, we define spaces $\mathcal{X} \triangleq \prod^m \Delta_{n^2}\times\Delta_n$ and $\mathcal{Y}\triangleq[-1,1]^{2mn}$, where $\prod^m \Delta_{n^2}\times\Delta_n$ is a short form of $\underbrace{\Delta_{n^2}\times\ldots\times\Delta_{n^2}}_{m}\times\Delta_n$. Then we rewrite problem (4) for column vectors $\mathbf{x} = (x_1^\top,\ldots,x_m^\top,p^\top)^\top\in\mathcal{X}$ and $\mathbf{y} = (y_1^\top,\ldots,y_m^\top)^\top\in\mathcal{Y}$ as follows

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{m}\left\{\boldsymbol{d}^\top\mathbf{x} + 2\|d\|_\infty\left(\mathbf{y}^\top\boldsymbol{A}\mathbf{x} - \boldsymbol{c}^\top\mathbf{y}\right)\right\}, \quad (5)$$

where $\boldsymbol{d} = (d^\top,\ldots,d^\top,\mathbf{0}_n^\top)^\top$, $\boldsymbol{c} = (\mathbf{0}_n^\top, q_1^\top,\ldots,\mathbf{0}_n^\top,q_m^\top)^\top$ and $\boldsymbol{A} = \begin{pmatrix}\hat{A} & \mathcal{E}\end{pmatrix}\in\{-1,0,1\}^{2mn\times(mn^2+n)}$ with block-diagonal matrix $\hat{A}$ of m blocks

$$\hat{A} = \begin{pmatrix} A & 0_{2n\times n^2} & \cdots & 0_{2n\times n^2}\\ 0_{2n\times n^2} & A & \cdots & 0_{2n\times n^2}\\ \vdots & \vdots & \ddots & \vdots\\ 0_{2n\times n^2} & 0_{2n\times n^2} & \cdots & A\end{pmatrix}$$

and matrix

$$\mathcal{E}^\top = \left(\underbrace{\begin{pmatrix}-I_n & 0_{n\times n}\end{pmatrix}}_{-B_{\mathcal{E}}^\top}\underbrace{\begin{pmatrix}-I_n & 0_{n\times n}\end{pmatrix}}_{-B_{\mathcal{E}}^\top}\cdots\underbrace{\begin{pmatrix}-I_n & 0_{n\times n}\end{pmatrix}}_{-B_{\mathcal{E}}^\top}\right).$$

As objective $F(\mathbf{x},\mathbf{y})$ in (5) is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, problem (5) is a saddle-point problem. This means that we reformulated the WB problem (3) as saddle-point problem (5).

3 Mirror Prox for Wasserstein Barycenters

In this section, we present our first algorithm which does not improve the complexity of the state-of-the-art methods for the WB problem but has no limitations which other Sinkhorn-based-algorithms have. Moreover, this method contributes to a better understanding of our second approach. To present our results, we define the following setup which is used throughout this paper.

3.1 Setup

We endow space $\mathcal{Y}\triangleq[-1,1]^{2nm}$ with standard the Euclidean setup: the Euclidean ℓ_2-norm $\|\mathbf{y}\|_2$, prox-function $d_{\mathcal{Y}}(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2$, and the corresponding Bregman divergence $B_{\mathcal{Y}}(\mathbf{y},\breve{\mathbf{y}}) = \frac{1}{2}\|\mathbf{y}-\breve{\mathbf{y}}\|_2^2$.

For space $\mathcal{X}\triangleq\prod^m\Delta_{n^2}\times\Delta_n$, we choose the following specific norm $\|\mathbf{x}\|_{\mathcal{X}} = \sqrt{\sum_{i=1}^m\|x_i\|_1^2 + m\|p\|_1^2}$ for $\mathbf{x} = (x_1,\ldots,x_m,p)^T$, where $\|\cdot\|_1$ is the ℓ_1-norm (for $a\in\mathbb{R}^n, \|a\|_1 = \sum_{i=1}^n|a_i|$). We endow $\mathcal{X}$ with prox-function $d_{\mathcal{X}}(\mathbf{x}) = \sum_{i=1}^m\langle x_i,\ln x_i\rangle + m\langle p,\ln p\rangle$ and the corresponding Bregman divergence

$$B_{\mathcal{X}}(\mathbf{x},\breve{\mathbf{x}}) = \sum_{i=1}^m\langle x_i, \ln(x_i/\breve{x}_i)\rangle - \sum_{i=1}^m \mathbf{1}^\top(x_i-\breve{x}_i) + m\langle p,\ln(p/\breve{p})\rangle - m\mathbf{1}^\top(p-\breve{p}).$$

We also define $R_{\mathcal{X}}^2 = \sup_{\mathbf{x}\in\mathcal{X}} d_{\mathcal{X}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} d_{\mathcal{X}}(\mathbf{x})$ and $R_{\mathcal{Y}}^2 = \sup_{\mathbf{y}\in\mathcal{Y}} d_{\mathcal{Y}}(\mathbf{y}) - \min_{\mathbf{y}\in\mathcal{Y}} d_{\mathcal{Y}}(\mathbf{y})$.

Definition 3.1 *$f(x,y)$ is $(L_{\mathbf{xx}},L_{\mathbf{xy}},L_{\mathbf{yx}},L_{\mathbf{yy}})$-smooth if for any $\mathbf{x},\mathbf{x}'\in\mathcal{X}$ and $\mathbf{y},\mathbf{y}'\in\mathcal{Y}$,*

$$\|\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}}f(\mathbf{x}',\mathbf{y})\|_{\mathcal{X}^*} \le L_{\mathbf{xx}}\|\mathbf{x}-\mathbf{x}'\|_{\mathcal{X}},$$
$$\|\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y}')\|_{\mathcal{X}^*} \le L_{\mathbf{xy}}\|\mathbf{y}-\mathbf{y}'\|_{\mathcal{Y}},$$
$$\|\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y}')\|_{\mathcal{Y}^*} \le L_{\mathbf{yy}}\|\mathbf{y}-\mathbf{y}'\|_{\mathcal{Y}},$$
$$\|\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}}f(\mathbf{x}',\mathbf{y})\|_{\mathcal{Y}^*} \le L_{\mathbf{yx}}\|\mathbf{x}-\mathbf{x}'\|_{\mathcal{X}}.$$

3.2 Implementation and Complexity Bound

As problem (5) is a saddle-point problem, we will evaluate the quality of an algorithm that outputs a pair of solutions $(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})\in(\mathcal{X},\mathcal{Y})$ through the so-called duality gap

$$\max_{\mathbf{y}\in\mathcal{Y}} F(\widetilde{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x},\widetilde{\mathbf{y}}) \le \varepsilon. \quad (6)$$

Our first algorithm is based on mirror prox (MP) algorithm (Nemirovski, 2004) on space $\mathcal{Z}\triangleq\mathcal{X}\times\mathcal{Y}$ with prox-function $d_{\mathcal{Z}}(\mathbf{z}) = a_1 d_{\mathcal{X}}(\mathbf{x}) + a_2 d_{\mathcal{Y}}(\mathbf{y})$ and the corresponding Bregman divergence $B_{\mathcal{Z}}(\mathbf{z},\breve{\mathbf{z}}) = a_1 B_{\mathcal{X}}(\mathbf{x},\breve{\mathbf{x}}) + a_2 B_{\mathcal{Y}}(\mathbf{y},\breve{\mathbf{y}})$, where $a_1 = \frac{1}{R_{\mathcal{X}}^2}$, $a_2 = \frac{1}{R_{\mathcal{Y}}^2}$

$$\begin{pmatrix}\mathbf{u}^{k+1}\\ \mathbf{v}^{k+1}\end{pmatrix} = \arg\min_{\mathbf{z}\in\mathcal{Z}}\{\eta G(\mathbf{x}^k,\mathbf{y}^k)^\top\mathbf{z} + B_{\mathcal{Z}}(\mathbf{z},\mathbf{z}^k)\},$$
$$\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}\in\mathcal{Z}}\{\eta G(\mathbf{u}^{k+1},\mathbf{v}^{k+1})^\top\mathbf{z} + B_{\mathcal{Z}}(\mathbf{z},\mathbf{z}^k)\}.$$

Here η is a learning rate, $\mathbf{z}^1 = \arg\min_{\mathbf{z}\in\mathcal{Z}} d_{\mathcal{Z}}(\mathbf{z})$ and $G(\mathbf{x},\mathbf{y})$ is a gradient operator defined as follows

$$G(\mathbf{x},\mathbf{y}) = \begin{pmatrix}\nabla_{\mathbf{x}}F(\mathbf{x},\mathbf{y})\\ -\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y})\end{pmatrix} = \frac{1}{m}\begin{pmatrix}\boldsymbol{d} + 2\|d\|_\infty\boldsymbol{A}^\top\mathbf{y}\\ 2\|d\|_\infty(\boldsymbol{c}-\boldsymbol{A}\mathbf{x})\end{pmatrix}. \quad (7)$$

If $F(\mathbf{x},\mathbf{y})$ is $(L_{\mathbf{xx}}, L_{\mathbf{xy}}, L_{\mathbf{yx}}, L_{\mathbf{yy}})$-smooth, then to satisfy (6) with $\widetilde{\mathbf{x}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{u}^k$, $\widetilde{\mathbf{y}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{v}^k$ one needs to perform

$$N = \frac{4}{\varepsilon}\max\{L_{\mathbf{xx}}R_{\mathcal{X}}^2, L_{\mathbf{xy}}R_{\mathcal{X}}R_{\mathcal{Y}}, L_{\mathbf{yx}}R_{\mathcal{Y}}R_{\mathcal{X}}, L_{\mathbf{yy}}R_{\mathcal{Y}}^2\} \tag{8}$$

iterations of MP (Bubeck, 2014) with

$$\eta = 1/(2\max\{L_{\mathbf{xx}}R_{\mathcal{X}}^2, L_{\mathbf{xy}}R_{\mathcal{X}}R_{\mathcal{Y}}, L_{\mathbf{yx}}R_{\mathcal{Y}}R_{\mathcal{X}}, L_{\mathbf{yy}}R_{\mathcal{Y}}^2\}). \tag{9}$$

Algorithm 1 Mirror Prox for Wasserstein Barycenters

Input: measures $q_1, ..., q_m$, linearized cost matrix d, incidence matrix A, step η, starting points $p^1 = \frac{1}{n}\mathbf{1}_n$, $x_1^1 = ... = x_m^1 = \frac{1}{n^2}\mathbf{1}_{n^2}$, $y_1^1 = ... = y_m^1 = \mathbf{0}_{2n}$

$\alpha = 2\|d\|_\infty \eta n$, $\beta = 6\|d\|_\infty \eta \ln n/m$, $\gamma = 3\eta\ln n$.

for $k = 1, 2, \cdots, N-1$ **do**

  for $i = 1, 2, \cdots, m$ **do**

$$v_i^{k+1} = y_i^k + \alpha\left(Ax_i^k - \begin{pmatrix} p^k \\ q_i \end{pmatrix}\right),$$

  Project v_i^{k+1} onto $[-1,1]^{2n}$

$$u_i^{k+1} = \frac{x_i^k \odot \exp\left\{-\gamma\left(d + 2\|d\|_\infty A^\top y_i^k\right)\right\}}{\sum\limits_{l=1}^{n^2}[x_i^k]_l \exp\left\{-\gamma\left([d]_l + 2\|d\|_\infty [A^\top y_i^k]_l\right)\right\}}$$

  end for

$$s^{k+1} = \frac{p^k \odot \exp\left\{\beta\sum_{i=1}^m [y_i^k]_{1...n}\right\}}{\sum_{l=1}^n [p^k]_l \exp\left\{\beta\sum_{i=1}^m [y_i^k]_l\right\}}$$

  for $i = 1, 2, \cdots, m$ **do**

$$y_i^{k+1} = y_i^k + \alpha\left(Au_i^{k+1} - \begin{pmatrix} s^{k+1} \\ q_i \end{pmatrix}\right)$$

  Project y_i^{k+1} onto $[-1,1]^{2n}$

$$x_i^{k+1} = \frac{x_i^k \odot \exp\left\{-\gamma\left(d + 2\|d\|_\infty A^\top v_i^{k+1}\right)\right\}}{\sum\limits_{l=1}^{n^2}[x_i^k]_l \exp\left\{-\gamma\left([d]_l + 2\|d\|_\infty [A^\top v_i^{k+1}]_l\right)\right\}}$$

  end for

$$p^{k+1} = \frac{p^k \odot \exp\left\{\beta\sum_{i=1}^m [v_i^{k+1}]_{1...n}\right\}}{\sum_{l=1}^n [p^k]_l \exp\left\{\beta\sum_{i=1}^m [v_i^{k+1}]_l\right\}}$$

end for

Output: $\widetilde{\mathbf{u}} = \sum\limits_{k=1}^N \begin{pmatrix} u_1^k \\ \vdots \\ u_m^k \\ s^k \end{pmatrix}$, $\widetilde{\mathbf{v}} = \sum\limits_{k=1}^N \begin{pmatrix} v_1^k \\ \vdots \\ v_m^k \end{pmatrix}$

Lemma 3.2 *Objective $F(\mathbf{x},\mathbf{y})$ in (5) is $(L_{\mathbf{xx}}, L_{\mathbf{xy}}, L_{\mathbf{yx}}, L_{\mathbf{yy}})$-smooth with $L_{\mathbf{xx}} = L_{\mathbf{yy}} = 0$ and $L_{\mathbf{xy}} = L_{\mathbf{yx}} = 2\sqrt{2}\|d\|_\infty/m$.*

Proof. Let us consider bilinear function

$$f(\mathbf{x},\mathbf{y}) \triangleq \mathbf{y}^\top \boldsymbol{A}\mathbf{x}$$

that is equivalent to $F(\mathbf{x},\mathbf{y})$ from (5) up to multiplicative constant $2\|d\|_\infty/m$ and linear terms. As $f(\mathbf{x},\mathbf{y})$ is bilinear, $L_{\mathbf{xx}} = L_{\mathbf{yy}} = 0$ in Definition 3.1. Next we estimate $L_{\mathbf{xy}}$ and $L_{\mathbf{yx}}$. By the definition of $L_{\mathbf{xy}}$ and the spaces $\mathcal{X}, \mathcal{Y}$ defined in Setup 3.1 we have

$$\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}')\|_{\mathcal{X}^*} \le L_{\mathbf{xy}}\|\mathbf{y}-\mathbf{y}'\|_2.$$

Since $\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) = \boldsymbol{A}^\top\mathbf{y}$ we get

$$\|\boldsymbol{A}^\top(\mathbf{y}-\mathbf{y}')\|_{\mathcal{X}^*} \le L_{\mathbf{xy}}\|\mathbf{y}-\mathbf{y}'\|_2. \tag{10}$$

By the definition of dual norm we have

$$\|\boldsymbol{A}^\top(\mathbf{y}-\mathbf{y}')\|_{\mathcal{X}^*} = \max_{\|\mathbf{x}\|_{\mathcal{X}}\le 1}\langle \mathbf{x}, \boldsymbol{A}^\top(\mathbf{y}-\mathbf{y}')\rangle. \tag{11}$$

As $\langle \mathbf{x}, \boldsymbol{A}^\top(\mathbf{y}-\mathbf{y}')\rangle$ is a linear function, (10) can be rewritten using (11) as

$$L_{\mathbf{xy}} = \max_{\|\mathbf{y}-\mathbf{y}'\|_2\le 1}\max_{\|\mathbf{x}\|_{\mathcal{X}}\le 1}\langle \mathbf{x}, \boldsymbol{A}^\top(\mathbf{y}-\mathbf{y}')\rangle.$$

Making the change of variable $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{y}'$ and using the equality $\langle \mathbf{x}, \boldsymbol{A}^\top\tilde{\mathbf{y}}\rangle = \langle \boldsymbol{A}\mathbf{x}, \tilde{\mathbf{y}}\rangle$ we get

$$L_{\mathbf{xy}} = \max_{\|\tilde{\mathbf{y}}\|_2\le 1}\max_{\|\mathbf{x}\|_{\mathcal{X}}\le 1}\langle \boldsymbol{A}\mathbf{x}, \tilde{\mathbf{y}}\rangle. \tag{12}$$

By the same arguments we can get the same expression for $L_{\mathbf{yx}}$ up to rearrangement of maximums. Then since the ℓ_2-norm is the conjugate norm for the ℓ_2-norm , we rewrite (12) as follows

$$L_{\mathbf{xy}} = \max_{\|\mathbf{x}\|_{\mathcal{X}}\le 1}\|\boldsymbol{A}\mathbf{x}\|_2. \tag{13}$$

By the definition of matrix $\boldsymbol{A}$ we get

$$\|\boldsymbol{A}\mathbf{x}\|_2^2 = \sum_{i=1}^m \left\|Ax_i - \begin{pmatrix} p \\ 0 \end{pmatrix}\right\|_2^2 \le \sum_{i=1}^m \|Ax_i\|_2^2 + m\|p\|_2^2. \tag{14}$$

The last bound holds due to $\langle Ax_i, (p^\top, 0_n^\top)^\top\rangle \ge 0$ since the entries of A, x, p are non-zero. By the definition of vector $\mathbf{x}$ we have

$$\begin{aligned}\max_{\|\mathbf{x}\|_{\mathcal{X}}\le 1}\|\boldsymbol{A}\mathbf{x}\|_2^2 &= \max_{\|\mathbf{x}\|_{\mathcal{X}}^2\le 1}\|\boldsymbol{A}\mathbf{x}\|_2^2 \\ &= \max_{\sum_{i=1}^m \|x_i\|_1^2 + m\|p\|_1^2 \le 1}\|\boldsymbol{A}\mathbf{x}\|_2^2 \\ &\overset{(14)}{=} \max_{\alpha\in\Delta_{m+1}}\left(\sum_{i=1}^m \max_{\|x_i\|_1\le\sqrt{\alpha_i}}\|Ax_i\|_2^2 \right.\\ &\qquad \left. + \max_{\|p\|_1\le\sqrt{\frac{\alpha_{m+1}}{m}}} m\|p\|_2^2\right) \\ &= \max_{\alpha\in\Delta_{m+1}}\left(\sum_{i=1}^m \alpha_i \max_{\|x_i\|_1\le 1}\|Ax_i\|_2^2 \right.\\ &\qquad \left. + \max_{\|p\|_1\le 1}\alpha_{m+1}\|p\|_2^2\right).\end{aligned} \tag{15}$$

By the definition of incidence matrix A we get that $Ax_i = (h_1^\top, h_2^\top)^\top$,where h_1 and h_2 such that $\mathbf{1}^\top h_1 = \mathbf{1}^\top h_2 = \sum_{j=1}^{n^2} [x_i]_j = 1$ since $x_i \in \Delta_{n^2}$ $\forall i = 1, ..., m$. Thus,

$$\|Ax_i\|_2^2 = \|h_1\|_2^2 + \|h_2\|_2^2 \le \|h_1\|_1^2 + \|h_2\|_1^2 = 2. \quad (16)$$

For the second term in the r.h.s. of (15) we have

$$\max_{\|p\|_1 \le 1} \alpha_{m+1} \|p\|_2^2 \le \max_{\|p\|_1 \le 1} \alpha_{m+1} \|p\|_1^2 = \alpha_{m+1}. \quad (17)$$

Using (16) and (17) in (15) we get

$$\max_{\|\mathbf{x}\|_{\mathcal{X}} \le 1} \|\boldsymbol{A}\mathbf{x}\|_2^2 \le \max_{\alpha \in \Delta_{m+1}} \left(2 \sum_{i=1}^m \alpha_i + \alpha_{m+1} \right) \le \max_{\alpha \in \Delta_{m+1}} 2 \sum_{i=1}^{m+1} \alpha_i = 2.$$

Using this for (13) we have that $L_{\mathbf{xy}} = L_{\mathbf{yx}} = \sqrt{2}$. To get the constant of smoothness for function $F(\mathbf{x}, \mathbf{y})$ we multiply these constants by $2\|d\|_\infty / m$ and finish the proof. □

The next theorem gives the complexity bound for the MP algorithm for the WB problem with prox-function $d_{\mathcal{Z}}(\mathbf{z})$. For this particular problem formulated as a saddle-point problem (5), the MP has closed-form solutions presented in Algorithm 1.

Theorem 3.3 *Assume that $F(\mathbf{x}, \mathbf{y})$ in (5) is $(0, 2\sqrt{2}\|d\|_\infty/m, 2\sqrt{2}\|d\|_\infty/m, 0)$-smooth and $R_{\mathcal{X}} = \sqrt{3m \ln n}$, $R_{\mathcal{Y}} = \sqrt{mn}$. Then after $N = 8\|d\|_\infty \sqrt{6n \ln n}/\varepsilon$ iterations, Algorithm 1 with $\eta = \frac{1}{4\|d\|_\infty \sqrt{6n \ln n}}$ outputs a pair $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}) \in (\mathcal{X}, \mathcal{Y})$ such that*

$$\max_{\mathbf{y} \in \mathcal{Y}} F(\tilde{\mathbf{u}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} F(\mathbf{x}, \tilde{\mathbf{v}}) \le \varepsilon.$$

The total complexity of Algorithm 1 is

$$O\left(mn^2 \sqrt{n \ln n} \|d\|_\infty \varepsilon^{-1}\right).$$

Proof. By Lemma 3.2, $F(\mathbf{x}, \mathbf{y})$ is $(0, 2\sqrt{2}\|d\|_\infty/m, 2\sqrt{2}\|d\|_\infty/m, 0)$-smooth. Then the bound on duality gap follows from the direct substitution of the expressions for $R_{\mathcal{X}}$, $R_{\mathcal{Y}}$ and $L_{\mathbf{xx}}$, $L_{\mathbf{xy}}$, $L_{\mathbf{yx}}$, $L_{\mathbf{yy}}$ in (8) and (9).

The complexity of one iteration of Algorithm 1 is $O\left(mn^2\right)$ as the number of non-zero elements in matrix A is $2n^2$, and m is the number of vector-components in $\mathbf{y}$ and $\mathbf{x}$. Multiplying this by the number of iterations N, we get the last statement of the theorem. □

As d is the vectorized cost matrix of C, we may reformulate the complexity results of Theorem 3.3 with respect to C as $O\left(mn^2 \sqrt{n \ln n} \|C\|_\infty \varepsilon^{-1}\right)$.

4 Dual Extrapolation with Area-Convexity for Wasserstein Barycenters

In this section, we present our second algorithm which improves the complexity bounds for the WB problem.

4.1 General framework

We recall $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ is a space of pairs $(\mathbf{x}, \mathbf{y}), \mathbf{x} \in \mathcal{X}, \mathbf{y} \in Y$. Using this space, we can redefine our functions of pairs as a functions of a single argument from $\mathcal{Z}$, such as a gradient operator.

Now we use the main framework proposed by Sherman (2017) and developed by Jambulapati et al. (2019). The key idea is using a wider family of regularizers instead of strongly convex regularizers in dual extrapolation (Nesterov, 2007) for bilinear saddle-point problems. This family of such regularizers is called *area-convex regularizers* and can be defined as

Definition 4.1 *Regularizer r is called κ-area convex with respect to G if for any points $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathcal{Z}$*

$$\kappa \left(r(\boldsymbol{a}) + r(\boldsymbol{b}) + r(\boldsymbol{c}) - 3r\left(\frac{\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}}{3}\right)\right) \ge \langle G(\boldsymbol{a}) - G(\boldsymbol{b}), \boldsymbol{b} - \boldsymbol{c} \rangle.$$

Considering only differentiable regularizer, we are able to define a proximal operator using $r(\mathbf{z})$ as a prox-function and use dual extrapolation (Nesterov, 2007). In this condition, we have the following converge guarantees in terms of a number of iterations for any gradient operator $G(\mathbf{z})$ for bilinear saddle-point problems

Lemma 4.2 *(Jambulapati et al., 2019, Corollary 1) Let r be κ-area convex with respect to G. Let also for some $\boldsymbol{u}$, $\Theta \ge r(\boldsymbol{u}) - r(\bar{\mathbf{z}})$, where $\bar{\mathbf{z}} = \arg\min_{\mathbf{z} \in Z} r(\mathbf{z})$. Then the output $\tilde{\mathbf{w}}$ of Dual Extrapolation algorithm (2) with the proximal steps implemented with ε' additive error satisfies*

$$\langle G(\tilde{\mathbf{w}}), \tilde{\mathbf{w}} - \boldsymbol{u} \rangle \le 2\kappa\Theta/N + \varepsilon'.$$

If we choose $\Theta = \sup_{\mathbf{z} \in \mathcal{Z}} r(\mathbf{z}) - r(\bar{\mathbf{z}})$, we obtain the convergence guarantees in terms of duality gap (6).

4.2 Complexity bounds

For the WB problem, we define the regularizer as a generalization of the regularizer of Jambulapati et al.

(2019):

$$r(\mathbf{x},\mathbf{y}) = \frac{2\|d\|_\infty}{m}\Bigg(10\sum_{i=1}^{m}\langle x_i, \log x_i\rangle + 5m\langle p, \log p\rangle + \hat{x}^\top \hat{A}^\top(\mathbf{y})^2 - p^\top \mathcal{E}^\top(\mathbf{y})^2\Bigg), \quad (18)$$

where $\log x$ and $(x)^2$ are entry-wise, and $\hat{x} = (x_1^\top, \ldots, x_m^\top)^\top$. For this regularizer, area-convexity can be proven

Theorem 4.3 *r is 3-area-convex with respect to the gradient operator G.*

To compute the range of the regularizer, we can rewrite it in the following homogeneous manner

$$r(\mathbf{x},\mathbf{y}) = \frac{2\|d\|_\infty}{m}\Bigg(\sum_{i=1}^{m}\Big[10\langle x_i, \log x_i\rangle + \langle A x_i, (y_i)^2\rangle\Big] + \sum_{i=1}^{m}\Big[5\langle p, \log p\rangle + \langle B_{\mathcal{E}} p, (y_i)^2\rangle\Big]\Bigg).$$

Hence, using properties of spaces $\mathcal{X}$ and $\mathcal{Y}$, we obtain the following bound on the range of the regularizer

$$\Theta = \sup_{\mathbf{z}\in\mathcal{Z}} r(\mathbf{z}) - \inf_{\mathbf{z}\in\mathcal{Z}} r(\mathbf{z}) = 40\log n\|d\|_\infty + 6\|d\|_\infty.$$

The only question is how to compute a proximal step effectively. Formally, we are solving the following type of problem

$$H(\mathbf{x},\mathbf{y}) = \langle \boldsymbol{v}, \mathbf{x}\rangle + \langle \boldsymbol{u}, \mathbf{y}\rangle + r(\mathbf{x},\mathbf{y}). \quad (19)$$

It can be done using a simple alternating minimization scheme as in the case of Jambulapati et al. (2019).

Algorithm 2 Dual Extrapolation with area-convex r (General algorithm)

Input: area-convexity coefficient κ, regularizer r, gradient operator G, number of iterations N, starting point $\boldsymbol{s}^0 = 0$, $\bar{\mathbf{z}} = \arg\min_{\mathbf{z}\in\mathcal{Z}} r(\mathbf{z})$

for $k = 0, 1, 2, \cdots, N-1$ **do**
  $\mathbf{z}^k = \mathrm{prox}_{\bar{\mathbf{z}}}^r(\boldsymbol{s}^k)$
  $\mathbf{w}^k = \mathrm{prox}_{\bar{\mathbf{z}}}^r(\boldsymbol{s}^k + \frac{1}{\kappa}G(\mathbf{z}^k))$
  $\boldsymbol{s}^{k+1} = \boldsymbol{s}^k + \frac{1}{2\kappa}G(\mathbf{w}^k)$
end for

Output: $\widetilde{\mathbf{w}} = \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{w}^k$

Theorem 4.4 *Let at each iteration, Dual Extrapolation algorithm 2 calls Alternating minimization (AM) scheme 3 to make the proximal steps. Then for $N = \lceil\frac{4\kappa\Theta}{\varepsilon}\rceil$ iterations of Dual Extrapolation algorithm 2 running with regularizer (18) and $\kappa = 3$, AM scheme 3 accumulates additive error $\varepsilon/2$ running with*

$$M = 24\log\left(\left(\frac{88\|d\|_\infty}{\varepsilon^2} + \frac{4}{\varepsilon}\right)\Theta + \frac{36\|d\|_\infty}{\varepsilon}\right)$$

iterations in $O(mn^2\log\gamma)$ time, where $\gamma = \varepsilon^{-1}\|d\|_\infty \log n$.

The complete algorithm is presented in Algorithm 3 and is referred as $\mathtt{AM}(M, \mathbf{v}, \mathbf{u})$.

Algorithm 3 Alternating minimization for (19)

Input: number of iterations M, $\mathbf{v} = (v_1^\top, \ldots, v_m^\top, v_{m+1}^\top)^\top$, $\mathbf{u} = (u_1^\top, \ldots, u_m^\top)^\top$, starting points $p^0 = \frac{1}{n}\mathbf{1}_n$, $x_1^0 = \ldots = x_m^0 = \frac{1}{n^2}\mathbf{1}_{n^2}$, $y_1^0 = \ldots = y_m^0 = \mathbf{0}_{2n}$

for $t = 0, 1, 2, \ldots, M-1$ **do**
  for $i = 1, 2, \ldots, m$ **do**
    $\gamma_i = \frac{m}{20\|d\|_\infty} v_i + \frac{1}{10} A^\top (y_i^t)^2$
    $x_i^{k+1} = \frac{\exp(-\gamma_i)}{\sum_{j=1}^{n^2}[\exp(-\gamma_i)]_j}$
  end for
  $\gamma_{m+1} = \frac{1}{10\|d\|_\infty} v_{m+1} + \frac{1}{5m}\sum_{j=1}^{m}[y_j^k]_{1,\ldots,n}$;
  $p^{k+1} = \frac{\exp(-\gamma_{m+1})}{\sum_{j=1}^{n}[\exp(-\gamma_{m+1})]_j}$
  for $i = 1, 2, \ldots, m$ **do**
    $[y_i^{k+1}]_{1,\ldots,n} = -\frac{m}{4\|d\|_\infty}\frac{[u_i]_{1,\ldots,n}}{[Ax_i^{k+1}]_{1,\ldots,n} + p^{k+1}}$
    $[y_i^{k+1}]_{n+1,\ldots,2n} = -\frac{m}{4\|d\|_\infty}\frac{[u_i]_{n+1,\ldots,2n}}{[Ax_i^{k+1}]_{n+1,\ldots,2n}}$
    Project y_i^{k+1} onto $[-1,1]^{2n}$
  end for
end for

Output: $\mathbf{x}^k = \begin{pmatrix} x_1^k \\ \vdots \\ x_m^k \\ p^k \end{pmatrix}$, $\mathbf{y}^k = \begin{pmatrix} y_1^k \\ \vdots \\ y_m^k \end{pmatrix}$

The proof of the correctness of this procedure can be found in the supplementary material to this paper. It consists of three main parts: the required details from the proof of Jambulapati et al. (2019) to obtain a linear convergence, bound on the time for each substep, and the bound on the initial error for our setup of proximal steps.

Overall, for the particular WB problem (5), we obtain the required complexity bound by combination of Lemma 4.2, Theorem 4.3 and Theorem 4.4. The final algorithm for this problem is Algorithm 4.

Theorem 4.5 *Dual Extrapolation algorithm 4 after*

$$N = 12\Theta/\varepsilon = (480\log n\|d\|_\infty + 72\|d\|_\infty)/\varepsilon$$

Algorithm 4 Dual Extrapolation for Wasserstein Barycenters

Input: measures $q_1, \dots, q_m$, linearized cost matrix d, incidence matrix A, area-convexity coefficient κ, starting points $\boldsymbol{s}_{\mathbf{x}}^0 = 0_{mn^2+n}, \boldsymbol{s}_{\mathbf{y}}^0 = 0_{2mn}$

$\nabla_{\mathbf{x}} r(\bar{\mathbf{z}}) = \frac{10\|d\|_\infty}{m}((-4\log n + 2)\mathbf{1}_{mn^2},\ m(-\log n + 1)\mathbf{1}_n)$

$\nabla_{\mathbf{y}} r(\bar{\mathbf{z}}) = 0_{2mn}$

$\Theta = 40\|d\|_\infty \log n + 6\|d\|_\infty$

$M = 24\log\left(\left(\frac{88\|d\|_\infty}{\varepsilon^2} + \frac{4}{\varepsilon}\right)\Theta + \frac{36\|d\|_\infty}{\varepsilon}\right)$

for $k = 0, 1, 2, \dots, N-1$ **do**

$\quad \mathbf{v} = \boldsymbol{s}_{\mathbf{x}}^k - \nabla_{\mathbf{x}} r(\bar{\mathbf{z}}),\ \mathbf{u} = \boldsymbol{s}_{\mathbf{x}}^k - \nabla_{\mathbf{y}} r(\bar{\mathbf{z}})$

$\quad \mathbf{z}_{\mathbf{x}}^k, \mathbf{z}_{\mathbf{y}}^k = \mathtt{AM}(M, \mathbf{v}, \mathbf{u}))$

$\quad \mathbf{v} = \mathbf{v} + \dfrac{1}{\kappa m}(\boldsymbol{b} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{z}_{\mathbf{y}}^k)$

$\quad \mathbf{u} = \mathbf{u} + \dfrac{2\|d\|_\infty}{\kappa m}(\boldsymbol{c} - \boldsymbol{A}\mathbf{z}_{\mathbf{x}}^k)$

$\quad \mathbf{w}_{\mathbf{x}}^k, \mathbf{w}_{\mathbf{y}}^k = \mathtt{AM}(M, \mathbf{v}, \mathbf{u})$

$\quad \boldsymbol{s}_{\mathbf{x}}^{k+1} = \boldsymbol{s}_{\mathbf{x}}^k + \dfrac{1}{2\kappa m}(\boldsymbol{b} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{w}_{\mathbf{y}}^k)$

$\quad \boldsymbol{s}_{\mathbf{y}}^{k+1} = \boldsymbol{s}_{\mathbf{y}}^k + \dfrac{\|d\|_\infty}{\kappa m}(\boldsymbol{c} - \boldsymbol{A}\mathbf{w}_{\mathbf{x}}^k)$

end for

Output: $\widetilde{\mathbf{w}}_{\mathbf{x}} = \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{w}_{\mathbf{x}}^k,\ \widetilde{\mathbf{w}}_{\mathbf{y}} = \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{w}_{\mathbf{y}}^k$

iterations outputs a pair $(\widetilde{\mathbf{w}}_{\mathbf{x}}, \widetilde{\mathbf{w}}_{\mathbf{y}}) \in (\mathcal{X}, \mathcal{Y})$ *such that the duality gap* (6) *becomes less then* ε. *It can be done in wall-clock time*

$$\widetilde{O}(mn^2\|d\|_\infty \varepsilon^{-1}).$$

Proof. The required number of iterations to obtain $\varepsilon/2$ precision follows from the choice of 3-area-convex regularizer r (follows from Lemma 4.3) and Lemma 4.2. For each step we need to do two proximal steps, that can be done in $O(mn^2 \log\gamma)$ time by Theorem 4.4. As a result, we have an algorithm with $O(mn^2\|d\|_\infty\varepsilon^{-1}\log n \log\gamma) = \widetilde{O}(mn^2\|d\|_\infty\varepsilon^{-1})$ time complexity. □

In terms of the initial cost matrix C, we obtain $\tilde{O}(mn^2\|C\|_\infty\varepsilon^{-1})$ complexity.

5 Numerical experiments

There are two goals of this section: compare the convergence of two our proposed algorithms, and prove the instability of entropy-regularized based approaches in contrast to our algorithms when a high precision for the WB problem is desired. The experiments are performed on CPU using the MNIST dataset, the notMNIST dataset and Gaussian distributions.

MNIST and notMNIST. In the paper, we mentioned that when a high-precision ε of calculating Wasserstein barycenters is desired, the iterative Bregman projections (IBP) algorithm with regularizing parameter γ is numerically unstable (as γ must be selected proportional to ε (Peyré and Cuturi, 2018; Kroshnin et al., 2019)) in contrast to Algorithm 1 (Mirror Prox for WB) and Algorithm 4 (Dual Extrapolation for WB). Now we support this statement by computing Wasserstein barycenters of hand-written digits '5' from the MNIST dataset and letters 'A' in a variety of fonts from the notMNIST dataset. Figure 1 illustrates the results obtained by the proposed algorithms in comparison with the IBP algorithm from the POT Python library with small values of regularizing parameter ($\gamma = 10^{-3}; 10^{-5}$).

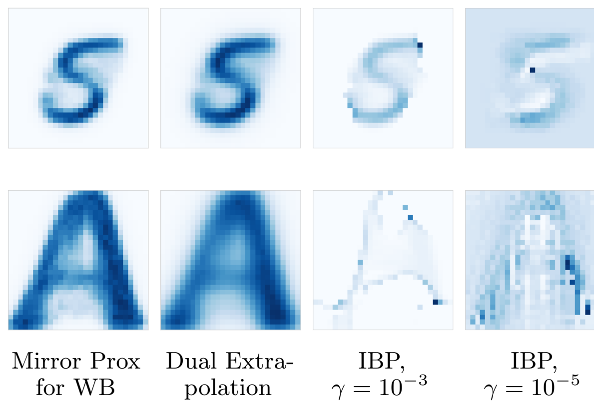


Figure 1: Wasserstein barycenters of hand-written digits '5' (first row), and Wasserstein barycenters of letters 'A' (second row) computed by Algorithm 1 (Mirror Prox for WB), Algorithm 4 (Dual Extrapolation for WB) and the IBP algorithm with small values of the regularizing parameter.

Gaussian measures. To compare the convergence of the proposed algorithms, we randomly generated 10 Gaussian measures with equally spaced support of 100 points in $[-10, 10]$, mean from $[-5, 5]$ and variance from $[0.8, 1.8]$. We studied the convergence of calculated barycenters to the theoretical true barycenter (Delon and Desolneux, 2020). Figure 2 presents the convergence with respect to the function optimality gap $\frac{1}{m}\sum_{i=1}^m \mathcal{W}(p, q_i) - \frac{1}{m}\sum_{i=1}^m \mathcal{W}(p^*, q_i)$. Here p^* is the true barycenter. Despite the fact that Algorithm 4 has better complexity bound, Algorithm 1 has better convergence in practice. The slope ration -1 for the convergence of Algorithm 1 in log-scale perfectly fits the theoretical dependence of working time (iteration number N) on the desired accuracy ε ($N \sim \varepsilon^{-1}$ from Theorem 3.3). For Algorithm 4, this slope ratio -1 is achieved only after a number of iterations but this is due to the need of solving practically computationally costly subproblems.

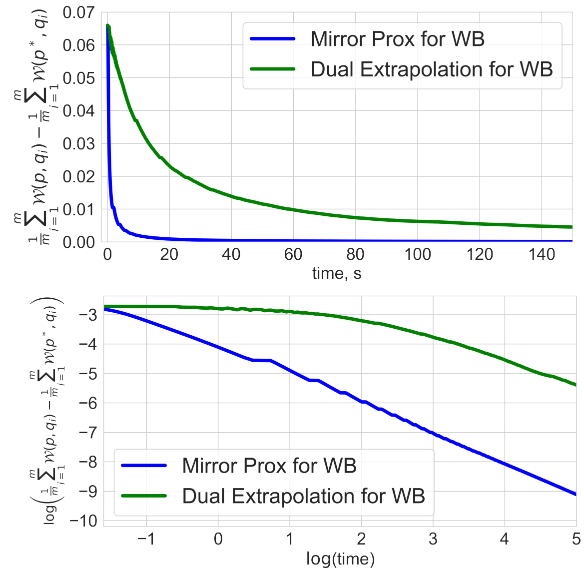
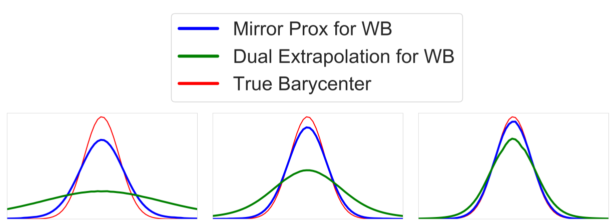


Figure 2: Convergence of Algorithm 1 (Mirror Prox for WB) and Algorithm 4 (Dual Extrapolation for WB) to the true barycenter of Gaussian measures w.r.t the function optimality gap $\frac{1}{m}\sum_{i=1}^m \mathcal{W}(p, q_i) - \frac{1}{m}\sum_{i=1}^m \mathcal{W}(p^*, q_i)$. Here p^* is the true barycenter.

Figure 3 illustrates the convergence of the barycenters obtained by Algorithms 1 and 4 to the true barycenter.



After 20 seconds After 100 seconds After 500 seconds

Figure 3: Convergence of the barycenters obtained by Algorithm 1 (Mirror Prox for WB) and Algorithm 4 (Dual Extrapolation for WB) to the true barycenter of Gaussian measures.

Next, we compare the convergence of the barycenters obtained by Algorithms 1 and 4 with the barycenter obtained by the IBP algorithm. Figure 4 demonstrates better approximations of the true Gaussian barycenter by Algorithms 1 and 4 compared to the γ-regularized IBP barycenter. The regularization parameter for the IBP algorithm (from the POT python library) is taken as smallest as possible under which the IBP still works since the smaller γ, the closer regularized IBP barycenter is to the true barycenter.

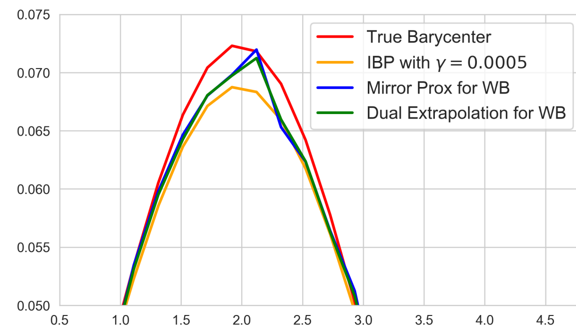


Figure 4: Convergence of the barycenters obtained by Algorithm 1 (Mirror Prox for WB), Algorithm 4 (Dual Extrapolation for WB), and the IBP to the true barycenter of Gaussian measures.

6 Conclusion

In this work, we provided two algorithms which have theoretical and practical interests. The main theoretical value is obtaining $\sqrt{n}$ faster algorithm for approximating Wasserstein barycenters of discrete measures with support n. The main practical value is the opportunity to calculate Wasserstein barycenters with a high desired precision that is not possible by using entropy-regularized based approaches.

Acknowledgements

The research of Section 3 is supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) No. 075-00337-20-03, project No. 0714-2020-0005. The work of Section 4 was prepared within the framework of the HSE University Basic Research Program. The research of Section 5 is supported by the Russian Science Foundation (project 18-71-10108). The work of D. Tiapkin was fulfilled in Sirius, Sochi https://ssopt.org (August 2020), the work was initiated by A.Gasnikov.

References

Allen-Zhu, Z., Li, Y., Oliveira, R., and Wigderson, A. (2017). Much faster algorithms for matrix scaling. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 890–901. https://arxiv.org/abs/1704.02315.

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

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

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

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

Bubeck, S. (2014). Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 15.

Cohen, M. B., Madry, A., Tsipras, D., and Vladu, A. (2017). Matrix scaling and balancing via box constrained newton's method and interior point methods. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 902–913. https://arxiv.org/abs/1704.02310.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

Delon, J. and Desolneux, A. (2020). A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970.

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

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

Guminov, S., Dvurechensky, P., and Gasnikov, A. (2019). Accelerated alternating minimization. *arXiv preprint arXiv:1906.03622*.

Jambulapati, A., Sidford, A., and Tian, K. (2019). A direct $\tilde{O}(1/\varepsilon)$ iteration parallel algorithm for optimal transport. In *Advances in Neural Information Processing Systems*, pages 11359–11370.

Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., and Uribe, C. (2019). On the complexity of approximating Wasserstein barycenters. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3530–3540. arXiv:1901.08686.

Lin, T., Ho, N., Chen, X., Cuturi, M., and Jordan, M. I. (2020). Fixed-support wasserstein barycenters: Computational hardness and fast algorithm.

Nemirovski, A. (2004). Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Nesterov, Y. (2007). Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344.

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

Rachev, S. T., Stoyanov, S. V., and Fabozzi, F. J. (2011). *A probability metrics approach to financial risk measures*. John Wiley & Sons.

Sherman, J. (2017). Area-convexity, l_∞ regularization, and undirected multicommodity flow. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 452–460.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66.

Tarjan, R. E. (1997). Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177.