Stochastic approximation versus sample average approximation for Wasserstein barycenters

Darina Dvinskikh

Published online: 09 Sep 2021.

Submit your article to this journal ⭷

Article views: 1084

View related articles ⭷

View Crossmark data ⭷

Citing articles: 2 View citing articles ⭷

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS | Check for updates

Stochastic approximation versus sample average approximation for Wasserstein barycenters

Darina Dvinskikh [a,b,c]

[a]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany; [b]Moscow Institute of Physics and Technology, Dolgoprudny, Russia; [c]Institute for Information Transmission Problems, Moscow, Russia

ABSTRACT

In the machine learning and optimization community, there are two main approaches for the convex risk minimization problem, namely the Stochastic Approximation (SA) and the Sample Average Approximation (SAA). In terms of the oracle complexity (required number of stochastic gradient evaluations), both approaches are considered equivalent on average (up to a logarithmic factor). The total complexity depends on a specific problem, however, starting from the work [A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM. J. Opt. 19 (2009), pp. 1574–1609] it was generally accepted that the SA is better than the SAA. We show that for the Wasserstein barycenter problem, this superiority can be inverted. We provide a detailed comparison by stating the complexity bounds for the SA and SAA implementations calculating barycenters defined with respect to optimal transport distances and entropy-regularized optimal transport distances. As a byproduct, we also construct confidence intervals for the barycenter defined with respect to entropy-regularized optimal transport distances in the ℓ_2-norm. The preliminary results are derived for a general convex optimization problem given by the expectation to have other applications besides the Wasserstein barycenter problem.

1. Introduction

In this paper, we consider the problem of finding a barycenter of discrete random probability measures. We refer to optimal transport (OT) metrics which provides a successful framework to compare objects that can be modelled as probability measures (images, videos, texts, etc.). Transport-based distances have gained popularity in various fields such as statistics [13,27], unsupervised learning [4], signal and image analysis [76], computer vision [66], text classification [51], economics and finance [62] and medical imaging [38,77]. Moreover, a lot of statistical results are known about optimal transport distances [47,72,78].

The success of optimal transport led to an increasing interest in Wasserstein barycenters (WBs). Wasserstein barycenters are used in Bayesian computations [74], texture mixing

---

CONTACT Darina Dvinskikh ✉ darina.dvinskikh@wias-berlin.de

[61], clustering (*k*-means for probability measures) [21], shape interpolation and color transferring [71], statistical estimation of template models [15] and neuroimaging [38]. For discrete random probability measures from the probability simplex Δ_n (*n* is the size of support) with distribution $\mathbb{P}$, a Wasserstein barycenter is introduced through a notion of Fréchet mean [30]

$$\min_{p \in \Delta_n} \mathbb{E}_{q \sim \mathbb{P}} W(p, q). \tag{1}$$

If a solution of (1) exists and is unique, then it is referred to as the population barycenter of distribution $\mathbb{P}$. Here $W(p, q)$ is optimal transport between measures p and q

$$W(p, q) = \min_{\pi \in U(p,q)} \langle C, \pi \rangle, \tag{2}$$

where $C \in \mathbb{R}_+^{n \times n}$ is a symmetric transportation cost matrix and $U(p, q) \triangleq \{\pi \in \mathbb{R}_+^{n \times n} : \pi \mathbf{1} = p, \pi^T \mathbf{1} = q\}$ is transport polytope.[1]

In [18], the entropic regularization of optimal transport problem (2) was proposed to improve its statistical properties [9,47] and to reduce the computational complexity from $\tilde{O}(n^3)$ (*n* is the size of the support of the measures) to $n^2 \min\{\tilde{O}(\frac{1}{\varepsilon}), \tilde{O}(\sqrt{n})\}$ arithmetic operations[2]

$$W_\gamma(p, q) \triangleq \min_{\pi \in U(p,q)} \{\langle C, \pi \rangle - \gamma E(\pi)\}. \tag{3}$$

Here $\gamma > 0$ and $E(\pi) \triangleq -\langle \pi, \log \pi \rangle$ is the entropy. Since $E(\pi)$ is 1-strongly concave on Δ_{n^2} in the ℓ_1-norm, the objective in (3) is γ-strongly convex with respect to π in the ℓ_1-norm on Δ_{n^2}, and hence, problem (3) has a unique optimal solution. Moreover, $W_\gamma(p, q)$ is γ-strongly convex with respect to p in the ℓ_2-norm on Δ_n [8, Theorem 3.4]. Another particular advantage of the entropy-regularized optimal transport (3) is a closed-form representation for its dual function [1,19] defined by the Fenchel–Legendre transform of $W_\gamma(p, q)$ as a function of p

$$W_{\gamma,q}^*(u) = \max_{p \in \Delta_n} \{\langle u, p \rangle - W_\gamma(p, q)\} = \gamma \left(E(q) + \langle q, \log(K\beta) \rangle \right),$$

where $\beta = \exp(u/\gamma)$, $K = \exp(-C/\gamma)$ and functions log or exp are applied element-wise. Hence, the gradient of dual function $W_{\gamma,q}^*(u)$ is also represented in a closed-form [19]

$$\nabla W_{\gamma,q}^*(u) = \beta \odot \left(K \cdot q/(K\beta) \right) \in \Delta_n,$$

where symbols $\odot$ and / stand for the element-wise product and element-wise division respectively.

Background on the SA and the SAA. Let us consider a general stochastic convex minimization problem

$$\min_{x \in X \subseteq \mathbb{R}^n} F(x) \triangleq \mathbb{E} f(x, \xi), \tag{4}$$

where function f is convex in x ($x \in X$, X is a convex set) and $\mathbb{E} f(x, \xi)$ is the expectation of f with respect to $\xi \in \Xi$. Such kind of problems arise in many applications of data science [67,69] (e.g. risk minimization) and mathematical statistics [73] (e.g. maximum likelihood estimation). There are two competing approaches based on Monte Carlo sampling

techniques to solve (4): the Stochastic Approximation (SA) [64] and the Sample Average Approximation (SAA). The SAA approach replaces the objective in problem (4) with its sample average approximation

$$\min_{x \in X} \hat{F}(x) \triangleq \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i),$$ (5)

where $\xi_1, \xi_2, \ldots, \xi_m$ are the realizations of a random variable ξ. The number of realizations m is adjusted by a desired precision. The total working time of both approaches to solve problem (4) with the average precision ε in the non-optimality gap in term of the objective function (i.e. to find x^N such that $\mathbb{E}F(x^N) - \min_{x \in X} F(x) \leq \varepsilon$), depends on a specific problem. However, it was generally accepted [54] that the SA approach is better than the SAA approach. Stochastic gradient (mirror) descent, an implementation of the SA approach [44], gives the following estimation for the number of iterations (that is equivalent to the sample size of $\xi_1, \xi_2, \xi_3, \ldots, \xi_m$)

$$m = O\left(\frac{M^2 R^2}{\varepsilon^2}\right).$$ (6)

Here we considered the minimal assumptions (non-smoothness) for the objective $f(x, \xi)$

$$\|\nabla f(x, \xi)\|_2^2 \leq M^2, \quad \forall x \in X, \xi \in \Xi.$$ (7)

Whereas the application of the SAA approach requires the following sample size [70]:

$$m = \widetilde{O}\left(\frac{nM^2 R^2}{\varepsilon^2}\right),$$

that is n times more (n is the problem dimension) than the sample size in the SA approach. This estimate was obtained under the assumptions that problem (5) is solved exactly. This is one of the main drawback of the SAA approach. However, if the objective $f(x, \xi)$ is λ-strongly convex in x, the sample sizes are equal up to logarithmic terms

$$m = O\left(\frac{M^2}{\lambda \varepsilon}\right).$$

Moreover, in this case, for the SAA approach, it suffices to solve problem (5) with accuracy [68]

$$\varepsilon' = O\left(\frac{\varepsilon^2 \lambda}{M^2}\right).$$ (8)

Therefore, to eliminate the linear dependence on n in the SAA approach for a non-strongly convex objective, regularization $\lambda = \frac{\varepsilon}{2R^2}$ should be used [68].

Let us suppose that $f(x, \xi)$ in (4) is convex but not strongly convex in x (possibly, λ-strongly convex but with very small $\lambda \ll \frac{\varepsilon}{R^2}$). Here $R = \|x^1 - x^*\|_2$ is the Euclidean distance between starting point x^1 and the solution x^* of (4) which corresponds to the minimum of this norm (if the solution is not the only one). Then, the problem (4) can be replaced by

$$\min_{x \in X} \mathbb{E}f(x, \xi) + \frac{\varepsilon}{2R^2} \|x - x^1\|_2^2.$$ (9)

The empirical counterpart of (9) is

$$\min_{x \in X} \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i) + \frac{\varepsilon}{2R^2} \|x - x^1\|_2^2, \tag{10}$$

where the sample size m is defined in (6). Thus, in the case of non-strongly objective, a regularization equates the sample size of both approaches.

1.1. Contribution and related work

The SA and SAA approaches. This paper is inspired by the work [54], where it is stated that the SA approach outperforms the SAA approach for a certain class of convex stochastic problems. Our aim is to show that for the Wasserstein barycenter problem, this superiority can be inverted. We provide a detailed comparison by stating the complexity bounds for implementations of the SA and SAA approaches for the Wasserstein barycenter problem. As a byproduct, we also construct a confidence interval for the barycenter defined w.r.t. entropy-regularized OT.

Sample size. We also estimate the sample size of measures to calculate an approximation for Fréchet mean of a probability distribution with a given precision.

Consistency and rates of convergence. The consistency of empirical barycenter as an estimator of true Wasserstein barycenter (defined by the notion of Fréchet mean) as the number of measures tends to infinity was studied in many papers, e.g. [12,52,57,63], under some conditions for the process generated the measures. Moreover, the authors of [15] provide the rate of this convergence but under restrictive assumption on the process (it must be from admissible family of deformations, i.e. it is a gradient of a convex function). Without any assumptions on generating process, the rate of convergence was obtained in [11], however, only for measures with one-dimensional support. For some specific types of metrics and measures, the rates of convergence were also provided in works [17,37,49]. Our results were obtained under the condition of discreteness of the measures. We can always achieve this condition through additional preprocessing (discretization of measures).

Penalization of barycenter problem. For a general convex (but not strongly convex) optimization problem, the empirical minimization may fail in offline approach despite the guaranteed success of an online approach if no regularization was introduced [68]. The limitations of the SAA approach for non-strongly convex case are also discussed in [39,70]. Our contribution includes introducing a new regularization for population Wasserstein barycenter problem that improves the complexity bounds for the standard penalty (squared norm penalty) [68]. This regularization relies on the Bregman divergence and peox-function introduced in [6].

1.2. Preliminaries

Notation. Let $\Delta_n = \{a \in \mathbb{R}^n_+ \mid \sum_{l=1}^{n} a_l = 1\}$ be the probability simplex. Then we refer to the jth component of vector x_i as $[x_i]_j$. The notation $[n]$ means $1, 2, \ldots, n$. 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 $\| \cdot \|$ on space $\mathcal{X}$, we define the dual

norm $\|\cdot\|_*$ on the dual space $\mathcal{X}^*$ in a usual way $\|s\|_* = \max_{x \in \mathcal{X}}\{\langle x, s \rangle : \|x\| \leq 1\}$. We denote by I_n the identity matrix, and by $0_{n \times n}$ we denote zeros matrix. For a positive semi-definite matrix A, we denote its smallest positive eigenvalue by $\lambda_{\min}^+(A)$. We use denotation $O(\cdot)$ when we want to indicate the complexity hiding constants, to hide also logarithms, we use denotation $\widetilde{O}(\cdot)$.

Definition 1.1: A function $f(x, \xi) : X \times \Xi \to \mathbb{R}$ is M-Lipschitz continuous in x w.r.t. a norm $\|\cdot\|$ if it satisfies

$$|f(x, \xi) - f(y, \xi)| \leq M\|x - y\|, \quad \forall x, y \in X, \ \forall \xi \in \Xi.$$

Definition 1.2: A function $f : X \times \Xi \to \mathbb{R}$ is γ-strongly convex in x w.r.t. a norm $\|\cdot\|$ if it is continuously differentiable and it satisfies

$$f(x, \xi) - f(y, \xi) - \langle \nabla f(y, \xi), x - y \rangle \geq \frac{\gamma}{2}\|x - y\|^2, \quad \forall x, y \in X, \ \forall \xi \in \Xi.$$

Definition 1.3: The Fenchel–Legendre conjugate for a function $f : (X, \Xi) \to \mathbb{R}$ w.r.t. x is

$$f^*(u, \xi) \triangleq \sup_{x \in X}\{\langle x, u \rangle - f(x, \xi)\}, \quad \forall \xi \in \Xi.$$

1.3. Paper organization

The structure of the paper is as follows. In Section 2, we give a background on the SA and SAA approaches and derive preliminary results. Section 4 presents the comparison of the SA and SAA approaches for the problem of Wasserstein barycenter defined w.r.t. regularized optimal transport distances. Finally, Section 5 gives the comparison of the SA and SAA approaches for the problem of Wasserstein barycenter defined w.r.t. (unregularized) optimal transport distances.

2. Strongly convex optimization problem

We start with preliminary results stated for a general stochastic strongly convex optimization problem of form

$$\min_{x \in X \subseteq \mathbb{R}^n} F(x) \triangleq \mathbb{E}f(x, \xi), \tag{11}$$

where $f(x, \xi)$ is γ-strongly convex with respect to x. Let us define $x^* = \arg\min_{x \in X} F(x)$.

2.1. The SA approach: stochastic gradient descent

The classical SA algorithm for problem (11) is presented by stochastic gradient descent (SGD) method. We consider the SGD with inexact oracle given by $g_\delta(x, \xi)$ such that

$$\forall x \in X, \xi \in \Xi, \qquad \|\nabla f(x, \xi) - g_\delta(x, \xi)\|_2 \leq \delta. \tag{12}$$

Then the iterative formula of SGD can be written as ($k = 1, 2, \ldots, N$.)

$$x^{k+1} = \Pi_X\left(x^k - \eta_k g_\delta(x^k, \xi^k)\right). \tag{13}$$

Here $x^1 \in X$ is starting point, Π_X is the projection onto X, η_k is a stepsize. For a γ-strongly convex $f(x, \xi)$ in x, the stepsize η_k can be taken as $\frac{1}{\gamma k}$ to obtain optimal rate $O\left(\frac{1}{\gamma N}\right)$.

A good indicator of the success of an algorithm is the *regret*

$$Reg_N \triangleq \sum_{k=1}^{N} \left(f(x^k, \xi^k) - f(x^*, \xi^k) \right).$$

It measures the value of the difference between a made decision and the optimal decision on all the rounds. The work [46] gives a bound on the excess risk of the output of an online algorithm in terms of the average regret.

Theorem 2.1 ([46, Theorem 2]): *Let* $f : X \times \Xi \to [0, B]$ *be* γ*-strongly convex and M-Lipschitz w.r.t. x. Let* $\tilde{x}^N \triangleq \frac{1}{N} \sum_{k=1}^{N} x^k$ *be the average of online vectors* $x^1, x^2, \ldots, x^N$. *Then with probability at least* $1 - 4\beta \log N$

$$F(\tilde{x}^N) - F(x^*) \leq \frac{Reg_N}{N} + 4\sqrt{\frac{M^2 \log(1/\beta)}{\gamma}} \frac{\sqrt{Reg_N}}{N} + \max\left\{\frac{16M^2}{\gamma}, 6B\right\} \frac{\log(1/\beta)}{N}.$$

For the update rule (13) with $\eta_k = \frac{1}{\gamma k}$, this theorem can be specified as follows.

Theorem 2.2: *Let* $f : X \times \Xi \to [0, B]$ *be* γ*-strongly convex and M-Lipschitz w.r.t. x. Let* $\tilde{x}^N \triangleq \frac{1}{N} \sum_{k=1}^{N} x^k$ *be the average of outputs generated by iterative formula (13) with* $\eta_k = \frac{1}{\gamma k}$. *Then, with probability at least* $1 - \alpha$ *the following holds:*

$$F(\tilde{x}^N) - F(x^*) \leq \frac{3\delta D}{2} + \frac{3(M^2 + \delta^2)}{N\gamma}(1 + \log N)$$

$$+ \max\left\{\frac{18M^2}{\gamma}, 6B + \frac{2M^2}{\gamma}\right\} \frac{\log(4 \log N/\alpha)}{N}.$$

where $D = \max_{x', x'' \in X} \|x' - x''\|_2$ *and* δ *is defined by (12).*

Proof: The proof mainly relies on Theorem 2.1 and estimating the regret for iterative formula (13) with $\eta_k = \frac{1}{\gamma k}$.

From γ-strongly convexity of $f(x, \xi)$ in x, it follows for $x^k, x^* \in X$

$$f(x^*, \xi^k) \geq f(x^k, \xi^k) + \langle \nabla f(x^k, \xi^k), x^* - x^k \rangle + \frac{\gamma}{2}\|x^* - x^k\|_2.$$

Adding and subtracting the term $\langle g_\delta(x^k, \xi^k), x^* - x^k \rangle$ we get using Cauchy–Schwarz inequality and (12)

$$f(x^*, \xi^k) \geq f(x^k, \xi^k) + \langle g_\delta(x^k, \xi^k), x^* - x^k \rangle + \frac{\gamma}{2}\|x^* - x^k\|_2$$

$$+ \langle \nabla f(x^k, \xi^k) - g_\delta(x^k, \xi^k), x^* - x^k \rangle$$

$$\geq f(x^k, \xi^k) + \langle g_\delta(x^k, \xi^k), x^* - x^k \rangle + \frac{\gamma}{2}\|x^* - x^k\|_2 + \delta\|x^* - x^k\|_2. \quad (14)$$

From the update rule (13) for x^{k+1} we have

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2 &= \|\Pi_X(x^k - \eta_k g_\delta(x^k, \xi^k)) - x^*\|_2 \\
&\leq \|x^k - \eta_k g_\delta(x^k, \xi^k) - x^*\|_2 \\
&\leq \|x^k - x^*\|_2^2 + \eta_k^2 \|g_\delta(x^k, \xi^k)\|_2^2 - 2\eta_k \langle g_\delta(x^k, \xi^k), x^k - x^* \rangle.
\end{aligned}
$$

From this it follows

$$
\langle g_\delta(x^k, \xi^k), x^k - x^* \rangle \leq \frac{1}{2\eta_k}(\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2) + \frac{\eta_k}{2}\|g_\delta(x^k, \xi^k)\|_2^2.
$$

Together with (14) we get

$$
\begin{aligned}
f(x^k, \xi^k) - f(x^*, \xi^k) \leq\ & \frac{1}{2\eta_k}(\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2) \\
& - \left(\frac{\gamma}{2} + \delta\right)\|x^* - x^k\|_2 + \frac{\eta_k^2}{2}\|g_\delta(x^k, \xi^k)\|_2^2.
\end{aligned}
$$

Summing this from 1 to N, we get using $\eta_k = \frac{1}{\gamma k}$

$$
\begin{aligned}
\sum_{k=1}^N f(x^k, \xi^k) - f(x^*, \xi^k) \leq\ & \frac{1}{2}\sum_{k=1}^N \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} + \gamma + \delta\right)\|x^* - x^k\|_2 \\
& + \frac{1}{2}\sum_{k=1}^N \eta_k\|g_\delta(x^k, \xi^k)\|_2^2 \\
\leq\ & \frac{\delta}{2}\sum_{k=1}^N \|x^* - x^k\|_2 + \frac{1}{2}\sum_{k=1}^N \eta_k\|g_\delta(x^k, \xi^k)\|_2^2. \quad (15)
\end{aligned}
$$

From Lipschitz continuity of $f(x, \xi)$ w.r.t. to x it follows that $\|\nabla f(x, \xi)\|_2 \leq M$ for all $x \in X, \xi \in \Xi$. Thus, using that for all $a, b, (a + b)^2 \leq 2a^2 + 2b^2$ it follows

$$
\|g_\delta(x, \xi)\|_2^2 \leq 2\|\nabla f(x, \xi)\|_2^2 + 2\delta^2 = 2M^2 + 2\delta^2
$$

From this and (15) we bound the regret as follows

$$
\begin{aligned}
Reg_N \triangleq \sum_{k=1}^N f(x^k, \xi^k) - f(x^*, \xi^k) \leq\ & \frac{\delta}{2}\sum_{k=1}^N \|p^* - p^k\|_2 + (M^2 + \delta^2)\sum_{k=1}^N \frac{1}{\gamma k} \\
\leq\ & \frac{1}{2}\delta DN + \frac{M^2 + \delta^2}{\gamma}(1 + \log N). \quad (16)
\end{aligned}
$$

Here the last bound takes place due to the sum of harmonic series. Then for (16) we can use Theorem 2.1. Firstly, we simplify it rearranging the terms using that $\sqrt{ab} \leq \frac{a+b}{2}$

$$
F(\tilde{x}^N) - F(x^*) \leq \frac{Reg_N}{N} + 4\sqrt{\frac{M^2 \log(1/\beta)}{N\gamma}}\sqrt{\frac{Reg_N}{N}} + \max\left\{\frac{16M^2}{\gamma}, 6B\right\}\frac{\log(1/\beta)}{N}
$$

$$\leq \frac{3Reg_N}{N} + \frac{2M^2 \log(1/\beta)}{N\gamma} + \max\left\{\frac{16M^2}{\gamma}, 6B\right\}\frac{\log(1/\beta)}{N}$$

$$= \frac{3Reg_N}{N} + \max\left\{\frac{18M^2}{\gamma}, 6B + \frac{2M^2}{\gamma}\right\}\frac{\log(1/\beta)}{N}.$$

Then we substitute (16) in this inequality and make the change $\alpha = 4\beta \log N$ and get with probability at least $1 - \alpha$

$$F(\tilde{x}^N) - F(x^*) \leq \frac{3\delta D}{2} + \frac{3(M^2 + \delta^2)}{N\gamma}(1 + \log N)$$

$$+ \max\left\{\frac{18M^2}{\gamma}, 6B + \frac{2M^2}{\gamma}\right\}\frac{\log(4\log N/\alpha)}{N}.$$

■

2.2. Preliminaries on the SAA approach

The SAA approach replaces the objective in (11) with its sample average

$$\min_{x \in X} \hat{F}(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f(x, \xi_i), \tag{17}$$

where each $f(x, \xi_i)$ is γ-strongly convex in x. Let us define the empirical minimizer of (17) $\hat{x}^* = \arg\min_{x \in X} \hat{F}(x)$, and $\hat{x}_{\varepsilon'}$ such that

$$\hat{F}(\hat{x}_{\varepsilon'}) - \hat{F}(\hat{x}^*) \leq \varepsilon'. \tag{18}$$

The next theorem gives a bound on the excess risk for problem (17) in the SAA approach.

Theorem 2.3: *Let $f : X \times \Xi \to [0, B]$ be γ-strongly convex and M-Lipschitz w.r.t. x in the ℓ_2-norm. Let $\hat{x}_{\varepsilon'}$ satisfies (18) with precision ε'. Then, with probability at least $1 - \alpha$ we have*

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{2M^2}{\gamma}\varepsilon'} + \frac{4M^2}{\alpha\gamma m}.$$

Let $\varepsilon' = O\left(\frac{\gamma\varepsilon^2}{M^2}\right)$ and $m = O\left(\frac{M^2}{\alpha\gamma\varepsilon}\right)$. Then, with probability at least $1 - \alpha$ the following holds:

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \varepsilon \quad \text{and} \quad \|\hat{x}_{\varepsilon'} - x^*\|_2 \leq \sqrt{2\varepsilon/\gamma}.$$

The proof of this theorem mainly relies on the following theorem.

Theorem 2.4: *[68, Theorem 6] Let $f(x, \xi)$ be γ-strongly convex and M-Lipschitz w.r.t. x in the ℓ_2-norm. Then, with probability at least $1 - \alpha$ the following holds:*

$$F(\hat{x}^*) - F(x^*) \leq \frac{4M^2}{\alpha\gamma m},$$

where m is the sample size.

Proof: For any $x \in X$, the following holds:

$$F(x) - F(x^*) = F(x) - F(\hat{x}^*) + F(\hat{x}^*) - F(x^*). \tag{19}$$

From Theorem 2.4 with probability at least $1 - \alpha$ the following holds:

$$F(\hat{x}^*) - F(x^*) \leq \frac{4M^2}{\alpha \gamma m}.$$

Then from this and (19) we have with probability at least $1 - \alpha$

$$F(x) - F(x^*) \leq F(x) - F(\hat{x}^*) + \frac{4M^2}{\alpha \gamma m}. \tag{20}$$

From Lipschitz continuity of $f(x, \xi)$ it follows, that for ant $x \in X, \xi \in \Xi$ the following holds:

$$|f(x, \xi) - f(\hat{x}^*, \xi)| \leq M\|x - \hat{x}^*\|_2.$$

Taking the expectation of this inequality w.r.t. ξ we get

$$\mathbb{E}|f(x, \xi) - f(\hat{x}^*, \xi)| \leq M\|x - \hat{x}^*\|_2.$$

Then we use Jensen's inequality $(g(\mathbb{E}(Y)) \leq \mathbb{E}g(Y))$ for the expectation, convex function g and a random variable Y. Since the module is a convex function we get

$$|\mathbb{E}f(x, \xi) - \mathbb{E}f(\hat{x}^*, \xi)| = |F(x) - F(\hat{x}^*)| \leq \mathbb{E}|f(x, \xi) - f(\hat{x}^*, \xi)| \leq M\|x - \hat{x}^*\|_2.$$

Thus we have

$$|F(x) - F(\hat{x}^*)| \leq M\|x - \hat{x}^*\|_2. \tag{21}$$

From strong convexity of $f(x, \xi)$ in x, it follows that the average of $f(x, \xi_i)$'s, that is $\hat{F}(x)$, is also γ-strongly convex in x. Thus we get for any $x \in X, \xi \in \Xi$

$$\|x - \hat{x}^*\|_2 \leq \sqrt{\frac{2}{\gamma}(\hat{F}(x) - \hat{F}(\hat{x}^*))}. \tag{22}$$

By using (21) and (22) and taking $x = \hat{x}_{\varepsilon'}$ in (20), we get the first statement of the theorem

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{2M^2}{\gamma}(\hat{F}(\hat{x}_{\varepsilon'}) - \hat{F}(\hat{x}^*))} + \frac{4M^2}{\alpha \gamma m} \leq \sqrt{\frac{2M^2}{\gamma}\varepsilon'} + \frac{4M^2}{\alpha \gamma m}. \tag{23}$$

Then from the strong convexity we have

$$\|\hat{x}_{\varepsilon'} - x^*\|_2 \leq \sqrt{\frac{2}{\gamma}\left(\sqrt{\frac{2M^2}{\gamma}\varepsilon'} + \frac{4M^2}{\alpha \gamma m}\right)}. \tag{24}$$

Equating (23) to ε, we get the expressions for the sample size m and auxiliary precision ε'. Substituting both of these expressions in (24) we finish the proof. ∎

3. Non-strongly convex optimization problem

Now we consider a non-strongly convex optimization problem of type

$$\min_{x \in X \subseteq \mathbb{R}^n} F(x) \triangleq \mathbb{E}f(x, \xi), \tag{25}$$

where $f(x, \xi)$ is Lipschitz continuous in x. Let us define $x^* = \arg\min_{x \in X} F(x)$.

3.1. The SA approach: stochastic mirror descent

We consider stochastic mirror descent (MD) with inexact oracle [34,44,54][3]. For a prox-function $d(x)$ and the corresponding Bregman divergence $B_d(x, x^1)$, the proximal mirror descent step is

$$x^{k+1} = \arg\min_{x \in X} \left(\eta \left\langle g_\delta(x^k, \xi^k), x \right\rangle + B_d(x, x^k) \right). \tag{26}$$

We consider the simplex setup: prox-function $d(x) = \langle x, \log x \rangle$. Here and below, functions such as $< roman > log < /roman >$ or exp are always applied element-wise. The corresponding Bregman divergence is given by the Kullback–Leibler divergence

$$\mathrm{KL}(x, x^1) = \langle x, \log(x/x^1) \rangle - \mathbf{1}^\top (x - x^1).$$

Then a starting point is taken as $x^1 = \arg\min_{x \in \Delta_n} d(x) = (1/n, \dots, 1/n)$.

Theorem 3.1: *Let* $R^2 \triangleq \mathrm{KL}(x^*, x^1) \le \log n$ *and* $D = \max_{x', x'' \in \Delta_n} \|x' - x''\|_1 = 2$. *Let* $f : X \times \Xi \to \mathbb{R}^n$ *be* M_∞*-Lipschitz w.r.t. x in the ℓ_1-norm. Let* $\check{x}^N \triangleq \frac{1}{N} \sum_{k=1}^N x^k$ *be the average of outputs generated by iterative formula (26) with* $\eta = \frac{\sqrt{2}R}{M_\infty \sqrt{N}}$. *Then, with probability at least $1 - \alpha$ we have*

$$F(\check{x}^N) - F(x^*) \le \frac{M_\infty(3R + 2D\sqrt{\log(\alpha^{-1})})}{\sqrt{2N}} + \delta D = O\left(\frac{M_\infty \sqrt{\log(n/\alpha)}}{\sqrt{N}} + 2\delta \right).$$

Proof: For MD with prox-function function $d(x) = \langle x \log x \rangle$ the following holds for any $x \in \Delta_n$ [44, Equation 5.13]

$$\eta \langle g_\delta(x^k, \xi^k), x^k - x \rangle \le \mathrm{KL}(x, x^k) - \mathrm{KL}(x, x^{k+1}) + \frac{\eta^2}{2} \|g_\delta(x^k, \xi^k)\|_\infty^2$$

$$\le \mathrm{KL}(x, x^k) - \mathrm{KL}(x, x^{k+1}) + \eta^2 M_\infty^2.$$

Then by adding and subtracting the terms $\langle F(x), x - x^k \rangle$ and $\langle \nabla f(x, \xi^k), x - x^k \rangle$ in this inequality, we get using Cauchy–Schwarz inequality the following:

$$\eta \langle \nabla F(x^k), x^k - x \rangle$$
$$\le \eta \langle \nabla f(x^k, \xi^k) - g_\delta(x^k, \xi^k), x^k - x \rangle$$
$$\quad + \eta \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x \rangle + \mathrm{KL}(x, x^k) - \mathrm{KL}(x, x^{k+1}) + \eta^2 M_\infty^2$$
$$\le \eta \delta \max_{k=1,\dots,N} \|x^k - x\|_1 + \eta \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x \rangle$$

$$+ \text{KL}(x, x^k) - \text{KL}(x, x^{k+1}) + \eta^2 M_\infty^2. \tag{27}$$

Then using convexity of $F(x^k)$ we have

$$F(x^k) - F(x) \leq \eta \langle \nabla F(x^k), x^k - x \rangle$$

Then we use this for (27) and sum for $k = 1, \dots, N$ at $x = x^*$

$$\eta \sum_{k=1}^{N} F(x^k) - F(x^*) \leq \eta \delta N \max_{k=1,\dots,N} \|x^k - x^*\|_1 + \eta \sum_{k=1}^{N} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle$$

$$+ \text{KL}(x^*, x^1) - \text{KL}(x^*, x^{N+1}) + \eta^2 M_\infty^2 N$$

$$\leq \eta \delta ND + \eta \sum_{k=1}^{N} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle + R^2 + \eta^2 M_\infty^2 N \tag{28}$$

where we used $\text{KL}(x^*, x^1) \leq R^2$ and $\max_{k=1,\dots,N} \|p^k - p^*\|_1 \leq D$. Then using convexity of $F(x^k)$ and the definition of output $\breve{x}^N$ in (28) we have

$$F(\breve{x}^N) - F(x^*) \leq \delta D + \frac{1}{N} \sum_{k=1}^{N} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle + \frac{R^2}{\eta N} + \eta M_\infty^2. \tag{29}$$

Next we use the Azuma–Hoeffding's [45] inequality and get for all $\beta \geq 0$

$$\mathbb{P} \left(\sum_{k=1}^{N+1} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle \leq \beta \right) \geq 1 - \exp \left(-\frac{2\beta^2}{N(2M_\infty D)^2} \right) = 1 - \alpha. \tag{30}$$

Here we used that $\langle \nabla F(p^k) - \nabla f(x^k, \xi^k), x^* - x^k \rangle$ is a martingale-difference and

$$\left| \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^* - x^k \rangle \right| \leq \|\nabla F(x^k) - \nabla W(p^k, q^k)\|_\infty \|x^* - x^k\|_1$$

$$\leq 2M_\infty \max_{k=1,\dots,N} \|x^k - x^*\|_1 \leq 2M_\infty D.$$

Thus, using (30) for (29) we have that with probability at least $1 - \alpha$

$$F(\breve{x}^N) - F(x^*) \leq \delta D + \frac{\beta}{N} + \frac{R^2}{\eta N} + \eta M_\infty^2. \tag{31}$$

Then, expressing β through α and substituting $\eta = \frac{R}{M_\infty} \sqrt{\frac{2}{N}}$ to (31) (such η minimize the r.h.s. of (31)), we get

$$F(\breve{x}^N) - F(x^*) \leq \delta D + \frac{M_\infty D \sqrt{2 \log(1/\alpha)}}{\sqrt{N}} + \frac{M_\infty R}{\sqrt{2N}} + \frac{M_\infty R \sqrt{2}}{\sqrt{N}}$$

$$\leq \delta D + \frac{M_\infty (3R + 2D\sqrt{\log(1/\alpha)})}{\sqrt{2N}}.$$

Using $R = \sqrt{\log n}$ and $D = 2$ in this inequality, we obtain

$$F(\check{x}^N) - F(x^*) \leq \frac{M_\infty(3\sqrt{\log n} + 4\sqrt{\log(1/\alpha)})}{\sqrt{2N}} + 2\delta. \tag{32}$$

We raise this to the second power, use that for all $a, b \geq 0$, $2\sqrt{ab} \leq a + b$ and then extract the square root. We obtain the following:

$$\sqrt{\left(3\sqrt{\log n} + 4\sqrt{\log(1/\alpha)}\right)^2} = \sqrt{9\log n + 16\log(1/\alpha) + 24\sqrt{\log n}\sqrt{\log(1/\alpha)}}$$

$$\leq \sqrt{18\log n + 32\log(1/\alpha)}.$$

Using this for (32), we get the statement of the theorem

$$F(\check{x}^N) - F(x^*) \leq \frac{M_\infty\sqrt{18\log n + 32\log(1/\alpha)}}{\sqrt{2N}} + 2\delta = O\left(\frac{M_\infty\sqrt{\log(n/\alpha)}}{\sqrt{N}} + 2\delta\right).$$

∎

3.2. Penalization in the SAA approach

In this section, we study the SAA approach for non-strongly convex problem (25). We regularize this problem by 1-strongly convex w.r.t. x penalty function $r(x, x^1)$ in the ℓ_2-norm

$$\min_{x \in X \subseteq \mathbb{R}^n} F_\lambda(x) \triangleq \mathbb{E}f(x, \xi) + \lambda r(x, x^1) \tag{33}$$

and we prove that the sample sizes in the SA and SAA approaches will be equal up to logarithmic terms. The empirical counterpart of problem (33) is

$$\min_{x \in X} \hat{F}_\lambda(x) \triangleq \frac{1}{m} \sum_{i=1}^m f(x, \xi_i) + \lambda r(x, x^1). \tag{34}$$

Let us define $\hat{x}_\lambda = \arg\min_{x \in X} \hat{F}_\lambda(x)$. The next lemma proves the statement from [68] on boundness of the population sub-optimality in terms of the square root of empirical sub-optimality.

Lemma 3.2: *Let $f(x, \xi)$ be convex and M-Lipschitz continuous w.r.t ℓ_2-norm. Then for any $x \in X$ with probability at least $1 - \delta$ the following holds:*

$$F_\lambda(x) - F_\lambda(x_\lambda^*) \leq \sqrt{\frac{2M_\lambda^2}{\lambda}\left(\hat{F}_\lambda(x) - \hat{F}_\lambda(\hat{x}_\lambda)\right)} + \frac{4M_\lambda^2}{\alpha\lambda m},$$

where $x_\lambda^* = \arg\min_{x \in X} F_\lambda(x)$, $M_\lambda \triangleq M + \lambda\mathcal{R}^2$ and $\mathcal{R}^2 = r(x^*, x^1)$.

Proof: Let us define $f_\lambda(x, \xi) \triangleq f(x, \xi) + \lambda r(x, x^1)$. As $f(x, \xi)$ is M-Lipschitz continuous, $f_\lambda(x, \xi)$ is also Lipschitz continuous with $M_\lambda \triangleq M + \lambda\mathcal{R}^2$. From the Jensen's inequality for

the expectation, and the module as a convex function, we get that $F_\lambda(x)$ is also M_λ-Lipschitz continuous

$$|F_\lambda(x) - F_\lambda(\hat{x}_\lambda)| \le M_\lambda \|x - \hat{x}_\lambda\|_2, \quad \forall x \in X. \tag{35}$$

From λ-strong convexity of $f(x, \xi)$, we obtain that $\hat{F}_\lambda(x)$ is also λ-strongly convex

$$\|x - \hat{x}_\lambda\|_2^2 \le \frac{2}{\lambda}\left(\hat{F}_\lambda(x) - \hat{F}_\lambda(\hat{x}_\lambda)\right), \quad \forall x \in X.$$

From this and (35) it follows

$$F_\lambda(x) - F_\lambda(\hat{x}_\lambda) \le \sqrt{\frac{2M_\lambda^2}{\lambda}\left(\hat{F}_\lambda(x) - \hat{F}_\lambda(\hat{x}_\lambda)\right)}. \tag{36}$$

For any $x \in X$ and $x_\lambda^* = \arg\min_{x \in X} F_\lambda(x)$ we consider

$$F_\lambda(x) - F_\lambda(x_\lambda^*) = F_\lambda(x) - F_\lambda(\hat{x}_\lambda) + F_\lambda(\hat{x}_\lambda) - F_\lambda(x_\lambda^*). \tag{37}$$

From [68, Theorem 6], we have with probability at least $1 - \alpha$

$$F_\lambda(\hat{x}_\lambda) - F_\lambda(x_\lambda^*) \le \frac{4M_\lambda^2}{\alpha\lambda m}.$$

Using this and (36) for (37), we obtain with probability at least $1 - \alpha$

$$F_\lambda(x) - F_\lambda(x_\lambda^*) \le \sqrt{\frac{2M_\lambda^2}{\lambda}\left(\hat{F}_\lambda(x) - \hat{F}_\lambda(\hat{x}_\lambda)\right)} + \frac{4M_\lambda^2}{\alpha\lambda m}. \qquad \blacksquare$$

The next theorem proves eliminating the linear dependence on n in the sample size of the regularized SAA approach for a non-strongly convex objective (see estimate (6)), and estimates the auxiliary precision for the regularized SAA problem (8).

Theorem 3.3: *Let $f(x, \xi)$ be convex and M-Lipschitz continuous w.r.t x and let $\hat{x}_{\varepsilon'}$ be such that*

$$\frac{1}{m}\sum_{i=1}^m f(\hat{x}_{\varepsilon'}, \xi_i) + \lambda r(\hat{x}_{\varepsilon'}, x^1) - \arg\min_{x \in X}\left\{\frac{1}{m}\sum_{i=1}^m f(x, \xi_i) + \lambda r(x, x^1)\right\} \le \varepsilon'.$$

To satisfy

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \le \varepsilon$$

with probability at least $1 - \alpha$, we need to take $\lambda = \varepsilon/(2\mathcal{R}^2)$,

$$m = \frac{32M^2\mathcal{R}^2}{\alpha\varepsilon^2},$$

where $\mathcal{R}^2 = r(x^*, x^1)$. The precision ε' is defined as

$$\varepsilon' = \frac{\varepsilon^3}{64M^2\mathcal{R}^2}.$$

Proof: From Lemma 3.2 we get for $x = \hat{x}_{\varepsilon'}$

$$F_\lambda(\hat{x}_{\varepsilon'}) - F_\lambda(x_\lambda^*) \leq \sqrt{\frac{2M_\lambda^2}{\lambda}\left(\hat{F}_\lambda(\hat{x}_{\varepsilon'}) - \hat{F}_\lambda(\hat{x}_\lambda)\right)} + \frac{4M_\lambda^2}{\alpha\lambda m}$$

$$= \sqrt{\frac{2M_\lambda^2}{\lambda}\varepsilon' + \frac{4M_\lambda^2}{\alpha\lambda m}}, \tag{38}$$

where we used the definition of $\hat{x}_{\varepsilon'}$ from the statement of the this theorem. Then we subtract $F(x^*)$ in both sides of (38) and get

$$F_\lambda(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{2M_\lambda^2\varepsilon'}{\lambda} + \frac{4M_\lambda^2}{\alpha\lambda m}} + F_\lambda(x_\lambda^*) - F(x^*). \tag{39}$$

Then we use

$$F_\lambda(x_\lambda^*) \triangleq \min_{x \in X}\left\{F(x) + \lambda r(x, x^1)\right\}$$

$$\leq F(x^*) + \lambda r(x^*, x^1)$$

The inequality holds for any $x \in X$,

$$= F(x^*) + \lambda\mathcal{R}^2$$

where $\mathcal{R} = r(x^*, x^1)$. Then from this and (39) and the definition of $F_\lambda(\hat{x}_{\varepsilon'})$ in (33) we get

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{2M_\lambda^2}{\lambda}\varepsilon' + \frac{4M_\lambda^2}{\alpha\lambda m}} - \lambda r(\hat{x}_{\varepsilon'}, x^1) + \lambda\mathcal{R}^2$$

$$\leq \sqrt{\frac{2M_\lambda^2\varepsilon'}{\lambda} + \frac{4M_\lambda^2}{\alpha\lambda m}} + \lambda\mathcal{R}^2. \tag{40}$$

Assuming $M \gg \lambda\mathcal{R}^2$ and choosing $\lambda = \varepsilon/(2\mathcal{R}^2)$ in (40), we get the following:

$$F(\hat{x}_{\varepsilon'}) - F(x^*) = \sqrt{\frac{4M^2\mathcal{R}^2\varepsilon'}{\varepsilon} + \frac{8M^2\mathcal{R}^2}{\alpha m\varepsilon}} + \varepsilon/2. \tag{41}$$

Equating the first term and the second term in the r.h.s. of (41) to $\varepsilon/4$ we obtain the rest statements of the theorem including $F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \varepsilon$. ∎

4. Fréchet mean with respect to entropy-regularized optimal transport

In this section, we consider the problem of finding population barycenter of discrete measures. We define the population barycenter of distribution $\mathbb{P}$ with respect to entropy-regularized transport distances

$$\min_{p \in \Delta_n} W_\gamma(p) \triangleq \mathbb{E}_q W_\gamma(p, q), \qquad q \sim \mathbb{P}. \tag{42}$$

4.1. Properties of entropy-regularized optimal transport

Entropic regularization of transport distances [18] improves their statistical properties [9,47] and reduces their computational complexity. Entropic regularization has shown good results in generative models [35], multi-label learning [31], dictionary learning [65], image processing [19,60], neural imaging [38].

Let us firstly remind optimal transport problem between histograms $p, q \in \Delta_n$ with cost matrix $C \in \mathbb{R}_+^{n \times n}$

$$W(p,q) \triangleq \min_{\pi \in U(p,q)} \langle C, \pi \rangle, \tag{43}$$

where

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

Remark 1 (Connection with the ρ-Wasserstein distance): When for $\rho \geq 1$, $C_{ij} = \mathrm{d}(x_i, x_j)^\rho$ in (43), where $\mathrm{d}(x_i, x_j)$ is a distance on support points x_i, x_j of space X, then $W(p,q)^{1/\rho}$ is known as the ρ-Wasserstein distance.

Nevertheless, all the results of this paper are based only on the assumptions that the matrix $C \in \mathbb{R}_+^{n \times n}$ is symmetric and non-negative.

Following [18], we introduce entropy-regularized optimal transport problem

$$W_\gamma(p,q) \triangleq \min_{\pi \in U(p,q)} \{\langle C, \pi \rangle - \gamma E(\pi)\}, \tag{44}$$

where $\gamma > 0$ and $E(\pi) \triangleq -\langle \pi, \log \pi \rangle$ is the entropy. Since $E(\pi)$ is 1-strongly concave on Δ_n in the ℓ_1-norm, the objective in (44) is γ-strongly convex with respect to π in the ℓ_1-norm on Δ_n, and hence problem (44) has a unique optimal solution. Moreover, $W_\gamma(p,q)$ is γ-strongly convex with respect to p in the ℓ_2-norm on Δ_n [8, Theorem 3.4].

One particular advantage of the entropy-regularized optimal transport is a closed-form representation for its dual function [1,19] defined by the Fenchel–Legendre transform of $W_\gamma(p,q)$ as a function of p

$$W_{\gamma,q}^*(u) = \max_{p \in \Delta_n} \left\{ \langle u, p \rangle - W_\gamma(p,q) \right\} = \gamma \left(E(q) + \langle q, \log(K\beta) \rangle \right)$$

$$= \gamma \left(-\langle q, \log q \rangle + \sum_{j=1}^n [q]_j \log \left(\sum_{i=1}^n \exp \left(([u]_i - C_{ji})/\gamma \right) \right) \right) \tag{45}$$

where $\beta = \exp(u/\gamma)$, $K = \exp(-C/\gamma)$ and $[q]_j$ is jth component of vector q. Functions such as $<roman> log </roman>$ or exp are always applied element-wise for vectors. Hence, the gradient of dual function $W_{\gamma,q}^*(u)$ is also represented in a closed form [19]

$$\nabla W_{\gamma,q}^*(u) = \beta \odot \left(K \cdot q/(K\beta) \right) \in \Delta_n,$$

where symbols $\odot$ and $/$ stand for the element-wise product and element-wise division respectively. This can be also written as

$$\forall l = 1, \dots, n \qquad [\nabla W_{\gamma,q}^*(u)]_l = \sum_{j=1}^n [q]_j \frac{\exp \left(([u]_l - C_{lj})/\gamma \right)}{\displaystyle\sum_{i=1}^n \exp \left(([u]_i - C_{ji})/\gamma \right)}. \tag{46}$$

The dual representation of $W_\gamma(p, q)$ is

$$W_\gamma(p, q) = \min_{\pi \in U(p,q)} \sum_{i,j=1}^{n} \left(C_{ij} \pi_{i,j} + \gamma \pi_{i,j} \log \pi_{i,j} \right)$$

$$= \max_{u,v \in \mathbb{R}^n} \left\{ \langle u, p \rangle + \langle v, q \rangle - \gamma \sum_{i,j=1}^{n} \exp\left(([u]_i + [v]_j - C_{ij})/\gamma - 1 \right) \right\}$$

$$= \max_{u \in \mathbb{R}^n} \left\{ \langle u, p \rangle - \gamma \sum_{j=1}^{n} [q]_j \log \left(\frac{1}{[q]_j} \sum_{i=1}^{n} \exp\left(([u]_i - C_{ij})/\gamma \right) \right) \right\}. \quad (47)$$

Any solution $\begin{pmatrix} u^* \\ v^* \end{pmatrix}$ of (47) is a subgradient of $W_\gamma(p, q)$ [58, Proposition 4.6]

$$\nabla W_\gamma(p, q) = \begin{pmatrix} u^* \\ v^* \end{pmatrix}. \quad (48)$$

We consider u^* and v^* such that $\langle u^*, \mathbf{1} \rangle = 0$ and $\langle v^*, \mathbf{1} \rangle = 0$ (u^* and v^* are determined up to an additive constant).

The next theorem [8] describes the Lipschitz continuity of $W_\gamma(p, q)$ in p on probability simplex Δ_n restricted to

$$\Delta_n^\rho = \left\{ p \in \Delta_n : \min_{i \in [n]} p_i \geq \rho \right\},$$

where $0 < \rho < 1$ is an arbitrary small constant.

Theorem 4.1 ([8, Theorem 3.4, Lemma 3.5]): • *For any $q \in \Delta_n$, $p \in \Delta_n$, $W_\gamma(p, q)$ is γ-strongly convex w.r.t. p in the ℓ_2-norm*
• *For any $q \in \Delta_n$, $p \in \Delta_n^\rho$ and $0 < \rho < 1$, $\|\nabla_p W_\gamma(p, q)\|_2 \leq M$, where*

$$M = \sqrt{\sum_{j=1}^{n} \left(2\gamma \log n + \inf_{i \in [n]} \sup_{l \in [n]} |C_{jl} - C_{il}| - \gamma \log \rho \right)^2}.$$

We roughly take $M = O(\sqrt{n}\|C\|_\infty)$ since for all $i, j \in [n]$, $C_{ij} > 0$, we get

$$M \overset{[8]}{=} O\left(\sqrt{\sum_{j=1}^{n} \left(\inf_{i \in [n]} \sup_{l \in [n]} |C_{jl} - C_{il}| \right)^2} \right)$$

$$= O\left(\sqrt{\sum_{j=1}^{n} \sup_{l \in [n]} C_{jl}^2} \right) = O\left(\sqrt{n} \sup_{j,l \in [n]} C_{jl} \right) = O\left(\sqrt{n} \sup_{j \in [n]} \sum_{l \in [n]} C_{jl} \right) = O\left(\sqrt{n}\|C\|_\infty \right).$$

Thus, we suppose that $W_\gamma(p, q)$ and $W(p, q)$ are Lipschitz continuous with almost the same Lipschitz constant M in the ℓ_2-norm on Δ_n^ρ. Moreover, by the same arguments, for

the Lipschitz continuity in the ℓ_1-norm: $\|\nabla_p W_\gamma(p, q)\|_\infty \leq M_\infty$, we can roughly estimate $M_\infty = O(\|C\|_\infty)$ by taking maximum instead of the square root of the sum.

In what follows, we use Lipschitz continuity of $W_\gamma(p, q)$ and $W(p, q)$ for measures from Δ_n keeping in mind that adding some noise and normalizing the measures makes them belong to Δ_n^ρ. We also notice that if the measures are from the interior of Δ_n then their barycenter will be also from the interior of Δ_n.

4.2. The SA approach: stochastic gradient descent

For problem (42), as a particular case of problem (4), stochastic gradient descent method can be used. From Equation (48), it follows that an approximation for the gradient of $W_\gamma(p, q)$ with respect to p can be calculated by Sinkhorn algorithm [3,26,58] through the computing dual variable u with δ-precision

$$\|\nabla_p W_\gamma(p, q) - \nabla_p^\delta W_\gamma(p, q)\|_2 \leq \delta, \quad \forall q \in \Delta_n. \tag{49}$$

Here denotation $\nabla_p^\delta W_\gamma(p, q)$ means an inexact stochastic subgradient of $W_\gamma(p, q)$ with respect to p. Algorithm 3 combines stochastic gradient descent given by iterative formula (13) for $\eta_k = \frac{1}{\gamma k}$ with the Sinkhorn algorithm (Algorithm 1) and Algorithm 2 making the projection onto the simplex Δ_n.

Algorithm 1 Sinkhorn's algorithm [57] for calculating $\nabla_p^\delta W_\gamma(p^k, q^k)$.

1: **procedure** SINKHORN(p, q, C, γ)
2:     $a^1 \leftarrow (1/n, ..., 1/n), b^1 \leftarrow (1/n, ..., 1/n)$
3:     $K \leftarrow \exp(-C/\gamma)$
4:     **while** not converged **do**
5:         $a \leftarrow p/(Kb)$
6:         $b \leftarrow q/(K^\top a)$
7:     **end while**
8:     **return** $\gamma \log(a)$ ▷ Sinkhorn scaling $a = e^{u/\gamma}$
9: **end procedure**

Algorithm 2 Euclidean Projection $\Pi_{\Delta_n}(p) = \arg\min_{v \in \Delta_n} \|p - v\|_2$ onto Simplex Δ_n [24].

1: **procedure** PROJECTION$(w \in \mathbb{R}^n)$
2:     Sort components of w in decreasing manner: $r_1 \geq r_2 \geq \cdots \geq r_n$.
3:     Find $\rho = \max\left\{j \in [n] : r_j - \frac{1}{j}\left(\sum_{i=1}^j r_i - 1\right)\right\}$
4:     Define $\theta = \frac{1}{\rho}(\sum_{i=1}^\rho r_i - 1)$
5:     For all $i \in [n]$, define $p_i = \max\{w_i - \theta, 0\}$.
6:     **return** $p \in \Delta_n$
7: **end procedure**

Algorithm 3 Projected Online Stochastic Gradient Descent for WB (PSGDWB).

Require: starting point $p^1 \in \Delta_n$, realization q^1, δ, γ.

  1: **for** $k = 1, 2, 3, \ldots$ **do**

  2:      $\eta_k = \frac{1}{\gamma k}$

  3:      $\nabla_p^\delta W_\gamma(p^k, q^k) \leftarrow$ Sinkhorn(p^k, q^k, C, γ) or the accelerated Sinkhorn [40]

  4:      $p^{(k+1)/2} \leftarrow p^k - \eta_k \nabla_p^\delta W_\gamma(p^k, q^k)$

  5:      $p^{k+1} \leftarrow$ Projection$(p^{(k+1)/2})$

  6:      Sample q^{k+1}

  7: **end for**

Ensure: $p^1, p^2, p^3 \ldots$

For Algorithm 3 and problem (42), Theorem 2.2 can be specified as follows

Theorem 4.2: *Let $\tilde{p}^N \triangleq \frac{1}{N} \sum_{k=1}^{N} p^k$ be the average of N online outputs of Algorithm 3 run with δ. Then, with probability at least $1 - \alpha$ the following holds:*

$$W_\gamma(\tilde{p}^N) - W_\gamma(p_\gamma^*) = O\left(\frac{M^2 \log(N/\alpha)}{\gamma N} + \delta\right),$$

where $p_\gamma^* \triangleq \arg\min_{p \in \Delta_n} W_\gamma(p)$.

Let Algorithm 3 run with $\delta = O(\varepsilon)$ and $N = \widetilde{O}\left(\frac{M^2}{\gamma \varepsilon}\right) = \widetilde{O}\left(\frac{n\|C\|_\infty^2}{\gamma \varepsilon}\right)$. Then, with probability at least $1 - \alpha$

$$W_\gamma(\tilde{p}^N) - W_\gamma(p_\gamma^*) \leq \varepsilon \quad \text{and} \quad \|\tilde{p}^N - p_\gamma^*\|_2 \leq \sqrt{2\varepsilon/\gamma}.$$

The total complexity of Algorithm 3 is

$$\widetilde{O}\left(\frac{n^3 \|C\|_\infty^2}{\gamma \varepsilon} \min\left\{\exp\left(\frac{\|C\|_\infty}{\gamma}\right)\left(\frac{\|C\|_\infty}{\gamma} + \log\left(\frac{\|C\|_\infty}{\kappa \varepsilon^2}\right)\right), \sqrt{\frac{n\|C\|_\infty^2}{\kappa \gamma \varepsilon^2}}\right\}\right),$$

where $\kappa \triangleq \lambda_{\min}^+(\nabla^2 W_{\gamma,q}^*(u^*))$.

Proof: We estimate the co-domain (image) of $W_\gamma(p, q)$

$$\max_{p,q \in \Delta_n} W_\gamma(p, q) = \max_{p,q \in \Delta_n} \min_{\substack{\pi \in \mathbb{R}_+^{n \times n}, \\ \pi \mathbf{1} = p, \\ \pi^T \mathbf{1} = q}} \sum_{i,j=1}^{n} (C_{ij} \pi_{ij} + \gamma \pi_{ij} \log \pi_{ij})$$

$$\leq \max_{\substack{\pi \in \mathbb{R}_+^{n \times n}, \\ \sum_{i,j=1}^{n} \pi_{ij} = 1}} \sum_{i,j=1}^{n} (C_{ij} \pi_{ij} + \gamma \pi_{ij} \log \pi_{ij}) \leq \|C\|_\infty.$$

Therefore, $W_\gamma(p, q) : \Delta_n \times \Delta_n \to [-2\gamma \log n, \|C\|_\infty]$. Then we apply Theorem 2.2 with $B = \|C\|_\infty$ and $D = \max_{p', p'' \in \Delta_n} \|p' - p''\|_2 = \sqrt{2}$, and we sharply get

$$W_\gamma(\tilde{p}^N) - W_\gamma(p_\gamma^*) = O\left(\frac{M^2 \log(N/\alpha)}{\gamma N} + \delta\right),$$

Equating each terms in the r.h.s. of this equality to $\varepsilon/2$ and using $M = O(\sqrt{n}\|C\|_\infty)$, we get the expressions for N and δ. The statement $\|\tilde{p}^N - p_\gamma^*\|_2 \leq \sqrt{2\varepsilon/\gamma}$ follows directly from strong convexity of $W_\gamma(p,q)$ and $W_\gamma(p)$.

The proof of algorithm complexity follows from the complexity of the Sinkhorn's algorithm. To state the complexity of the Sinkhorn's algorithm, we first define $\tilde{\delta}$ as the accuracy in function value of the inexact solution u of maximization problem in (47). Using this we formulate the number of iteration of the Sinkhorn's [16,29,50,75]

$$\tilde{O}\left(\exp\left(\frac{\|C\|_\infty}{\gamma}\right)\left(\frac{\|C\|_\infty}{\gamma} + \log\left(\frac{\|C\|_\infty}{\tilde{\delta}}\right)\right)\right). \tag{50}$$

The number of iteration for the accelerated Sinkhorn's can be improved [40]

$$\tilde{O}\left(\sqrt{\frac{n\|C\|_\infty^2}{\gamma \varepsilon'}}\right). \tag{51}$$

Here ε' is the accuracy in the function value, which is the expression $\langle u, p\rangle + \langle v, q\rangle - \gamma \sum_{i,j=1}^n \exp((-C_{ji} + u_i + v_j)/\gamma - 1)$ under the maximum in (47). From strong convexity of this objective on the space orthogonal to eigenvector $\mathbf{1}_n$ corresponds to the eigenvalue 0 for this function, it follows that

$$\varepsilon' \geq \frac{\gamma}{2}\|u - u^*\|_2^2 = \frac{\kappa}{2}\delta, \tag{52}$$

where $\kappa \triangleq \lambda_{\min}^+(\nabla^2 W_{\gamma,q}^*(u^*))$. From [8, Proposition A.2.], for the eigenvalue of $\nabla^2 W_{\gamma,q}^*(u^*)$ it holds that $0 = \lambda_n(\nabla^2 W_{\gamma,q}^*(u^*)) < \lambda_k(\nabla^2 W_{\gamma,q}^*(u^*))$ for all $k = 1, \ldots, n - 1$. Inequality (52) holds due to $\nabla_p^\delta W_\gamma(p,q) := u$ in Algorithm 3 and $\nabla_p W_\gamma(p,q) \triangleq u^*$ in (48). Multiplying both of estimates (50) and (51) by the complexity of each iteration of the (accelerated) Sinkhorn's algorithm $O(n^2)$ and the number of iterations $N = \tilde{O}\left(\frac{M^2}{\gamma \varepsilon}\right)$ (measures) of Algorithm 3, and taking the minimum, we get the last statement of the theorem. ∎

Next, we study the practical convergence of projected stochastic gradient descent (Algorithm 3). Using the fact that the true Wasserstein barycenter of one-dimensional Gaussian measures has closed-form expression for the mean and the variance [22], we study the convergence to the true barycenter of the generated truncated Gaussian measures. Figure 1 illustrates the convergence in the 2-Wasserstein distance within 40 s.

4.3. The SAA approach

The empirical counterpart of problem (42) is the (empirical) Wasserstein barycenter problem

$$\min_{p \in \Delta_n} \frac{1}{m} \sum_{i=1}^m W_\gamma(p, q_i), \tag{53}$$

where $q_1, q_2, \ldots, q_m$ are some realizations of random variable with distribution $\mathbb{P}$.
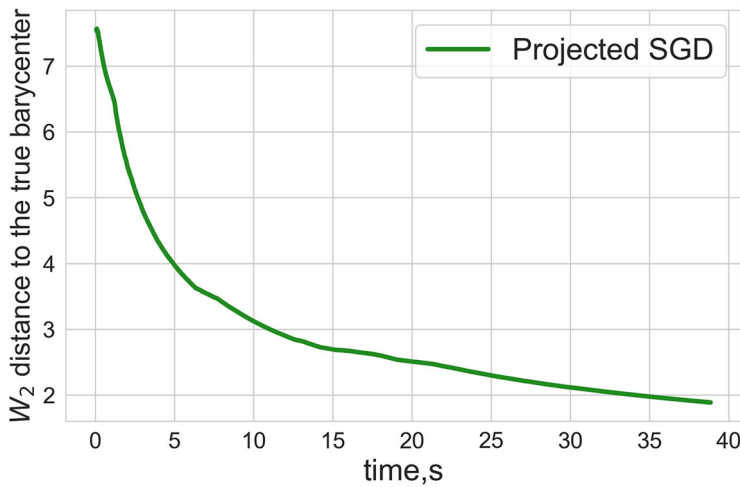
Figure 1. Convergence of projected stochastic gradient descent to the true barycenter of 2×10^4 Gaussian measures in the 2-Wasserstein distance.

Let us define $\hat{p}_\gamma^m \triangleq \arg\min_{p \in \Delta_n} \frac{1}{m} \sum_{i=1}^{m} W_\gamma(p, q_i)$ and its ε'-approximation $\hat{p}_{\varepsilon'}$ such that

$$\frac{1}{m} \sum_{i=1}^{m} W_\gamma(\hat{p}_{\varepsilon'}, q_i) - \frac{1}{m} \sum_{i=1}^{m} W_\gamma(\hat{p}_\gamma^m, q_i) \leq \varepsilon'. \tag{54}$$

For instance, $\hat{p}_{\varepsilon'}$ can be calculated by the IBP algorithm [7] or the accelerated IBP algorithm [40]. The next theorem specifies Theorem 2.3 for the Wasserstein barycenter problem (53).

Theorem 4.3: *Let $\hat{p}_{\varepsilon'}$ satisfies (54). Then, with probability at least $1 - \alpha$*

$$W_\gamma(\hat{p}_{\varepsilon'}) - W_\gamma(p_\gamma^*) \leq \sqrt{\frac{2M^2}{\gamma} \varepsilon'} + \frac{4M^2}{\alpha \gamma m},$$

where $p_\gamma^* \triangleq \arg\min_{p \in \Delta_n} W_\gamma(p)$. Let $\varepsilon' = O\left(\frac{\varepsilon^2 \gamma}{n \|C\|_\infty^2}\right)$ and $m = O\left(\frac{M^2}{\alpha \gamma \varepsilon}\right) = O\left(\frac{n \|C\|_\infty^2}{\alpha \gamma \varepsilon}\right)$. Then, with probability at least $1 - \alpha$

$$W_\gamma(\hat{p}_{\varepsilon'}) - W_\gamma(p_\gamma^*) \leq \varepsilon \quad \text{and} \quad \|\hat{p}_{\varepsilon'} - p_\gamma^*\|_2 \leq \sqrt{2\varepsilon/\gamma}.$$

The total complexity of the accelerated IBP computing $\hat{p}_{\varepsilon'}$ is

$$\widetilde{O}\left(\frac{n^4 \|C\|_\infty^4}{\alpha \gamma^2 \varepsilon^2}\right).$$

Proof: From Theorem 2.3 we get the first statement of the theorem

$$W_\gamma(\hat{p}_{\varepsilon'}) - W_\gamma(p_\gamma^*) \leq \sqrt{\frac{2M^2}{\gamma} \varepsilon'} + \frac{4M^2}{\alpha \gamma m}.$$
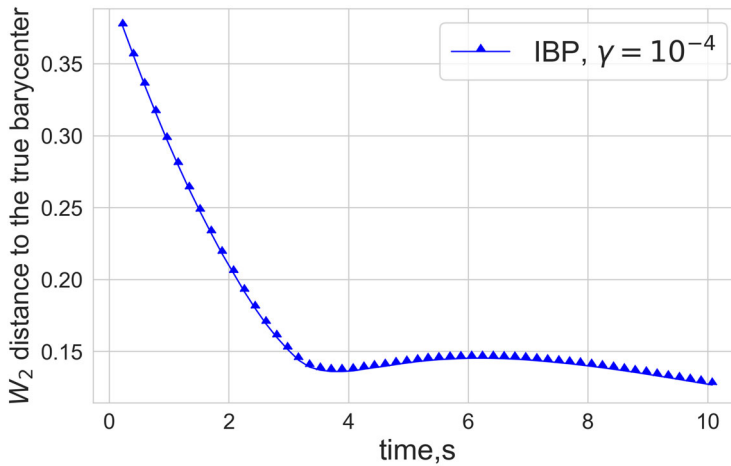
Figure 2. Convergence of the Iterative Bregman Projections to the true barycenter of 2×10^4 Gaussian measures in the 2-Wasserstein distance.

From [40] we have that complexity of the accelerated IBP is

$$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\sqrt{\gamma \varepsilon'}} \right).$$

Substituting the expression for m and the expression for ε' from Theorem 2.3

$$\varepsilon' = O\left(\frac{\varepsilon^2 \gamma}{M^2} \right), \qquad m = O\left(\frac{M^2}{\alpha \gamma \varepsilon} \right)$$

to this equation we get the final statement of the theorem and finish the proof. ∎

Next, we study the practical convergence of the Iterative Bregman Projections on truncated Gaussian measures. Figure 2 illustrates the convergence of the barycenter calculated by the IBP algorithm to the true barycenter of Gaussian measures in the 2-Wasserstein distance within 10 s. For the convergence to the true barycenter w.r.t. the 2-Wasserstein distance in the SAA approach, we refer to [15], however, considering the convergence in the ℓ_2-norm (Theorem 4.3) allows to obtain better convergence rate in comparison with the bounds for the 2-Wasserstein distance.

4.4. Comparison of the SA and the SAA for the WB problem

Now we compare the complexity bounds for the SA and the SAA implementations solving problem (42). We skip the high probability details since we can fixed α (say $\alpha = 0.05$) in the all bounds. Moreover, based on [68], we assume that in fact all bounds of this paper have logarithmic dependence on α which is hidden in $\widetilde{O}(\cdot)$ [28,48].

Table 1 presents the total complexity of the numerical algorithms implementing the SA and the SAA approaches. When γ is not too large, the complexity in the first row of the

Table 1. Total complexity of the SA and the SAA implementations for $\min_{p \in \Delta_n} \mathbb{E}_q W_\gamma(p, q)$.

Algorithm	Complexity
Projected SGD (SA)	$\widetilde{O}\left(\frac{n^3 \|C\|_\infty^2}{\gamma \varepsilon} \min\{\exp\left(\frac{\|C\|_\infty}{\gamma}\right)\left(\frac{\|C\|_\infty}{\gamma} + \log\left(\frac{\|C\|_\infty}{\kappa \varepsilon^2}\right)\right), \sqrt{\frac{n\|C\|_\infty^2}{\kappa \gamma \varepsilon^2}}\} \right)$
Accelerated IBP (SAA)	$\widetilde{O}\left(\frac{n^4 \|C\|_\infty^4}{\gamma^2 \varepsilon^2} \right)$

table is achieved by the second term under the minimum, namely

$$\widetilde{O}\left(\frac{n^3 \sqrt{n} \|C\|_\infty^3}{\gamma \sqrt{\gamma \kappa} \varepsilon^2} \right),$$

where $\kappa \triangleq \lambda_{\min}^+(\nabla^2 W_{\gamma,q}^*(u^*))$. This is typically bigger than the SAA complexity when $\kappa \ll \gamma/n$. Hereby, the SAA approach may outperform the SA approach provided that the regularization parameter γ is not too large.

From the practical point of view, the SAA implementation converges much faster than the SA implementation. Executing the SAA algorithm in a distributed manner only enhances this superiority since for the case when the objective is not Lipschitz smooth, the distributed implementation of the SA approach is not possible. This is the case of the Wasserstein barycenter problem, indeed, the objective is Lipschitz continuous but not Lipschitz smooth.

5. Fréchet mean with respect to optimal transport

Now we are interested in finding a Fréchet mean with respect to optimal transport

$$\min_{p \in \Delta_n} W(p) \triangleq \mathbb{E}_q W(p, q). \tag{55}$$

5.1. The SA approach with regularization: stochastic gradient descent

The next theorem explains how the solution of strongly convex problem (42) approximates a solution of convex problem (55) under the proper choice of the regularization parameter γ.

Theorem 5.1: Let $\tilde{p}^N \triangleq \frac{1}{N} \sum_{k=1}^N p^k$ be the average of N online outputs of Algorithm 3 run with $\delta = O(\varepsilon)$ and $N = \widetilde{O}(\frac{n\|C\|_\infty^2}{\gamma \varepsilon})$. Let $\gamma = \varepsilon/(2\mathcal{R}^2)$ with $\mathcal{R}^2 = 2 \log n$. Then, with probability at least $1 - \alpha$ the following holds:

$$W(\tilde{p}^N) - W(p^*) \leq \varepsilon,$$

where p^* is a solution of (55).

The total complexity of Algorithm 3 with the accelerated Sinkhorn is

$$\widetilde{O}\left(\frac{n^3 \sqrt{n} \|C\|_\infty^3}{\gamma \sqrt{\gamma \kappa} \varepsilon^2} \right) = \widetilde{O}\left(\frac{n^3 \sqrt{n} \|C\|_\infty^3}{\varepsilon^3 \sqrt{\varepsilon \kappa}} \right).$$

where $\kappa \triangleq \lambda_{\min}^+(\nabla^2 W_{\gamma,q}^*(u^*))$.

Proof: The proof of this theorem follows from Theorem 4.2 and the following [33,50,58]:

$$W(p) - W(p^*) \le W_\gamma(p) - W_\gamma(p^*) + 2\gamma \log n \le W_\gamma(p) - W_\gamma(p_\gamma^*) + 2\gamma \log n,$$

where $p \in \Delta_n$, $p^* = \arg\min_{p \in \Delta_n} W(p)$. The choice $\gamma = \frac{\varepsilon}{4 \log n}$ ensures the following:

$$W(p) - W(p^*) \le W_\gamma(p) - W_\gamma(p_\gamma^*) + \varepsilon/2, \quad \forall p \in \Delta_n.$$

This means that solving problem (42) with $\varepsilon/2$ precision, we get a solution of problem (55) with ε precision.

When γ is not too large, Algorithm 3 uses the accelerated Sinkhorns algorithm (instead of Sinkhorns algorithm). Thus, using $\gamma = \frac{\varepsilon}{4 \log n}$ and meaning that ε is small, we get the complexity according to the statement of the theorem. ∎

5.2. The SA approach: stochastic mirror descent

Now we propose an approach to solve problem (55) without an additional regularization. The approach is based on mirror descent given by the iterative formula (26). We use simplex setup which provides a closed form solution for (26). Algorithm 4 presents the application of mirror descent to problem (55), where the gradient of $W(p^k, q^k)$ can be calculated using dual representation of optimal transport [58] by any LP solver exactly:

$$W(p, q) = \max_{\substack{(u,v) \in \mathbb{R}^n \times \mathbb{R}^n, \\ u_i + v_j \le C_{ij}, \forall i,j \in [n]}} \left\{ \langle u, p \rangle + \langle v, q \rangle \right\}. \tag{56}$$

Then

$$\nabla_p W(p, q) = u^*,$$

where u^* is a solution of (56) such that $\langle u^*, \mathbf{1} \rangle = 0$.

The next theorem estimates the complexity of Algorithm 4

Algorithm 4 Stochastic mirror descent for the Wasserstein barycenter problem.

Require: starting point $p^1 = (1/n, ..., 1/n)^T$, number of measures N, $q^1, ..., q^N$, accuracy of gradient calculation δ

1: $\eta = \frac{\sqrt{2 \log n}}{\|C\|_\infty \sqrt{N}}$
2: **for** $k = 1, \ldots, N$ **do**
3:     Calculate $\nabla_{p^k} W(p^k, q^k)$ solving dual LP by any LP solver
4:

$$p^{k+1} = \frac{p^k \odot \exp\left(-\eta \nabla_{p^k} W(p^k, q^k)\right)}{\sum_{j=1}^{n} [p^k]_j \exp\left(-\eta \left[\nabla_{p^k} W(p^k, q^k)\right]_j\right)}$$

5: **end for**
Ensure: $\check{p}^N = \frac{1}{N} \sum_{k=1}^{N} p^k$

Theorem 5.2: *Let $\check{p}^N$ be the output of Algorithm 4 processing N measures. Then, with probability at least $1 - \alpha$ we have*

$$W(\check{p}^N) - W(p^*) = O\left(\frac{\|C\|_\infty \sqrt{\log(n/\alpha)}}{\sqrt{N}}\right),$$

Let Algorithm 4 run with $N = \widetilde{O}\left(\frac{M_\infty^2 R^2}{\varepsilon^2}\right) = \widetilde{O}\left(\frac{\|C\|_\infty^2}{\varepsilon^2}\right)$, $R^2 \triangleq \mathrm{KL}(p^1, p^*) \le \log n$. Then, with probability at least $1 - \alpha$

$$W(\check{p}^N) - W(p^*) \le \varepsilon.$$

The total complexity of Algorithm 4 is

$$\widetilde{O}\left(\frac{n^3 \|C\|_\infty^2}{\varepsilon^2}\right).$$

Proof: From Theorem 3.1 and using $M_\infty = O(\|C\|_\infty)$, we have

$$W(\check{p}^N) - W(p^*) = O\left(\frac{\|C\|_\infty \sqrt{\log(n/\alpha)}}{\sqrt{N}} + 2\delta\right). \tag{57}$$

Notice, that $\nabla_{p^k} W(p^k, q^k)$ can be calculated exactly by any LP solver. Thus we take $\delta = 0$ in (57) and get the first statement of the theorem.

The second statement of the theorem directly follows from this and the condition $W(\check{p}^N) - W(p^*) \le \varepsilon$.

To get the complexity bounds we notice that the complexity for calculating $\nabla_p W(p^k, q^k)$ is $\widetilde{O}(n^3)$ [2,20,23,32], multiplying this by $N = O(\|C\|_\infty^2 R^2/\varepsilon^2)$ with $R^2 \triangleq \mathrm{KL}(p^*, p^1) \le \log n$, we get the last statement of the theorem.

$$\widetilde{O}(n^3 N) = \widetilde{O}\left(n^3 \left(\frac{\|C\|_\infty R}{\varepsilon}\right)^2\right) = \widetilde{O}\left(n^3 \left(\frac{\|C\|_\infty}{\varepsilon}\right)^2\right).$$

∎

Next we compare the SA approaches with and without regularization of optimal transport in problem (55). Entropic regularization of optimal transport leads to strong convexity of regularized optimal transport in the ℓ_2-norm, hence, the Euclidean setup should be used. Regularization parameter $\gamma = \frac{\varepsilon}{4 \log n}$ ensures ε-approximation for the unregularized solution. In this case, we use stochastic gradient descent with Euclidean projection onto simplex since it converges faster for strongly convex objective. For non-regularized problem we can significantly use the simplex prox structure, indeed, we can apply stochastic mirror descent with simplex setup (the Kullback–Leibler divergence as the Bregman divergence) with Lipschitz constant $M_\infty = O(\|C\|_\infty)$ that is $\sqrt{n}$ better than Lipschitz constant in the Euclidean norm $M = O(\sqrt{n}\|C\|_\infty)$.

We studied the convergence of stochastic mirror descent (Algorithm 4) and stochastic gradient descent (Algorithm 3) in the 2-Wasserstein distance within 10^4 iterations (processing of 10^4 probability measures). Figure 3 confirms better convergence of stochastic mirror descent than projected stochastic gradient descent as stated in their theoretical complexity (Theorems 5.1 and 5.2).
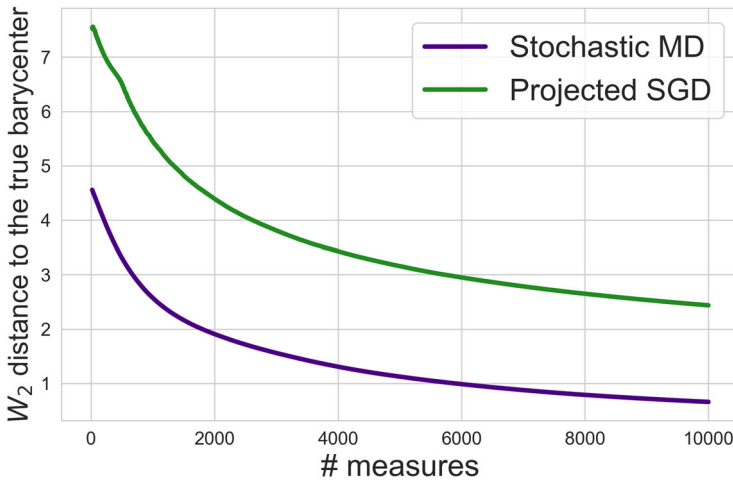
Figure 3. Convergence of projected stochastic gradient descent, and stochastic mirror descent to the true barycenter of 2×10^4 Gaussian measures in the 2-Wasserstein distance.

5.3. The SAA approach

Similarly for the SA approach, we provide the proper choice of the regularization parameter γ in the SAA approach so that the solution of strongly convex problem (42) approximates a solution of convex problem (55).

Theorem 5.3: *Let $\hat{p}_{\varepsilon'}$ satisfy*

$$\frac{1}{m} \sum_{i=1}^{m} W_{\gamma}(\hat{p}_{\varepsilon'}, q^i) - \frac{1}{m} \sum_{k=1}^{m} W_{\gamma}(\hat{p}_{\gamma}^*, q^i) \leq \varepsilon',$$

where $\quad \hat{p}_{\gamma}^* = \arg\min_{p \in \Delta_n} \frac{1}{m} \sum_{i=1}^{m} W_{\gamma}(p, q^i), \quad \varepsilon' = O(\frac{\varepsilon^2 \gamma}{n\|C\|_{\infty}^2}), \quad m = O(\frac{n\|C\|_{\infty}^2}{\alpha \gamma \varepsilon}), \quad$ *and* $\gamma = \varepsilon/(2\mathcal{R}^2)$ *with* $\mathcal{R}^2 = 2\log n$. *Then, with probability at least* $1 - \alpha$ *the following holds:*

$$W(\hat{p}_{\varepsilon'}) - W(p^*) \leq \varepsilon.$$

The total complexity of the accelerated IBP computing $\hat{p}_{\varepsilon'}$ is

$$\widetilde{O}\left(\frac{n^4 \|C\|_{\infty}^4}{\alpha \varepsilon^4}\right).$$

Proof: The proof follows from Theorem 4.3 and the proof of Theorem 5.1 with $\gamma = \varepsilon/(4\log n)$. ∎

5.4. Penalization of the WB problem

For the population Wasserstein barycenter problem, we construct 1-strongly convex penalty function in the ℓ_1-norm based on Bregman divergence. We consider the following

prox-function [6]:

$$d(p) = \frac{1}{2(a-1)} \|p\|_a^2, \quad a = 1 + \frac{1}{2 \log n}, \qquad p \in \Delta_n$$

that is 1-strongly convex in the ℓ_1-norm. Then Bregman divergence $B_d(p, p^1)$ associated with $d(p)$ is

$$B_d(p, p^1) = d(p) - d(p^1) - \langle \nabla d(p^1), p - p^1 \rangle.$$

$B_d(p, p^1)$ is 1-strongly convex w.r.t. p in the ℓ_1-norm and $\tilde{O}(1)$-Lipschitz continuous in the ℓ_1-norm on Δ_n. One of the advantages of this penalization compared to the negative entropy penalization proposed in [5,10] is that we get the upper bound on the Lipschitz constant, the properties of strong convexity in the ℓ_1-norm on Δ_n remain the same. Moreover, this penalization contributes to the better wall-clock time complexity than quadratic penalization [10] since the constants of Lipschitz continuity for $W(p, q)$ with respect to the ℓ_1-norm is $\sqrt{n}$ better than with respect to the ℓ_2-norm but $R^2 = \|p^* - p^1\|_2^2 \leq \|p^* - p^1\|_1^2 \leq \sqrt{2}$ and $R_d^2 = B_d(p^*, p^1) = O(\log n)$ are equal up to a logarithmic factor.

The regularized SAA problem is as follows:

$$\min_{p \in \Delta_n} \left\{ \frac{1}{m} \sum_{k=1}^m W(p, q^k) + \lambda B_d(p, p^1) \right\}. \tag{58}$$

The next theorem is a particular case of Theorem (3.3) for the population WB problem (55) with $r(p, p^1) = B_d(p, p^1)$.

Theorem 5.4: *Let $\hat{p}_{\varepsilon'}$ be such that*

$$\frac{1}{m} \sum_{k=1}^m W(\hat{p}_{\varepsilon'}, q^k) + \lambda B_d(\hat{p}_{\varepsilon'}, p^1) - \min_{p \in \Delta_n} \left\{ \frac{1}{m} \sum_{k=1}^m W(p, q^k) + \lambda B_d(p, p^1) \right\} \leq \varepsilon'. \tag{59}$$

To satisfy

$$W(\hat{p}_{\varepsilon'}) - W(p^*) \leq \varepsilon.$$

with probability at least $1 - \alpha$, we need to take $\lambda = \varepsilon/(2R_d^2)$ and

$$m = \tilde{O}\left(\frac{\|C\|_\infty^2}{\alpha \varepsilon^2} \right),$$

where $R_d^2 = B_d(p^*, p^1) = O(\log n)$. The precision ε' is defined as

$$\varepsilon' = \tilde{O}\left(\frac{\varepsilon^3}{\|C\|_\infty^2} \right).$$

The total complexity of Mirror Prox computing $\hat{p}_{\varepsilon'}$ is

$$\tilde{O}\left(\frac{n^2 \sqrt{n} \|C\|_\infty^5}{\varepsilon^5} \right).$$

Proof: The proof is based on saddle-point reformulation of the WB problem. Further, we provide the explanation how to do this. First we rewrite the OT as [42]

$$W(p, q) = \min_{x \in \Delta_{n^2}} \max_{y \in [-1,1]^{2n}} \{d^\top x + 2\|d\|_\infty (y^\top A x - b^\top y)\}, \tag{60}$$

where $b = (p^\top, q^\top)^\top$, d is vectorized cost matrix of C, x be vectorized transport plan of X, and $A = \{0, 1\}^{2n \times n^2}$ is an incidence matrix. Then we reformulate the WB problem as a saddle-point problem [25]

$$\min_{\substack{p \in \Delta^n, \\ \mathbf{x} \in \mathcal{X} \triangleq \underbrace{\Delta_{n^2} \times \ldots \times \Delta_{n^2}}_{m}}} \max_{\mathbf{y} \in [-1,1]^{2mn}} \frac{1}{m} \left\{ \boldsymbol{d}^\top \mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A} \mathbf{x} - \mathbf{b}^\top \mathbf{y} \right) \right\}, \tag{61}$$

where $\mathbf{x} = (x_1^\top, \ldots, x_m^\top)^\top$, $\mathbf{y} = (y_1^\top, \ldots, y_m^\top)^\top$, $\mathbf{b} = (p^\top, q_1^\top, \ldots, p^\top, q_m^\top)^\top$, $\boldsymbol{d} = (d^\top, \ldots, d^\top)^\top$, and $\boldsymbol{A} = \text{diag}\{A, \ldots, A\} \in \{0, 1\}^{2mn \times mn^2}$ is block-diagonal matrix. Similarly to (61) we reformulate (58) as a saddle-point problem

$$\min_{\substack{p \in \Delta^n, \mathbf{y} \in [-1,1]^{2mn} \\ \mathbf{x} \in \mathcal{X}}} \max f_\lambda(\mathbf{x}, p, \mathbf{y}) \triangleq \frac{1}{m} \left\{ \boldsymbol{d}^\top \mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A} \mathbf{x} - \mathbf{b}^\top \mathbf{y} \right) \right\} + \lambda B_d(p, p^1)$$

The gradient operator for $f(\mathbf{x}, p, \mathbf{y})$ is defined by

$$G(\mathbf{x}, p, \mathbf{y}) = \begin{pmatrix} \nabla_\mathbf{x} f \\ \nabla_p f \\ -\nabla_\mathbf{y} f \end{pmatrix} = \frac{1}{m} \begin{pmatrix} \boldsymbol{d} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{y} \\ -2\|d\|_\infty \{[y_i]_{1\cdots n}\}_{i=1}^m + \lambda(\nabla d(p) - \nabla d(p^1)) \\ 2\|d\|_\infty (\boldsymbol{A} \mathbf{x} - \mathbf{b}) \end{pmatrix}, \tag{62}$$

where $[d(p)]_i = \frac{1}{a-1} \|p\|_a^{2-a} [p]_i^{a-1}$.

To get the complexity of MP, we use the same reasons as in [25] with (62). The total complexity is

$$\widetilde{O} \left(\frac{mn^2 \sqrt{n} \|C\|_\infty}{\varepsilon'} \right)$$

Then we use Theorem 3.3 and get the expressions for m, ε' with $\lambda = \varepsilon/(2R_d^2)$, where $R_d^2 = B_d(p^*, p^1)$. The number of measures is

$$m = \frac{32M_\infty^2 R_d^2}{\alpha \varepsilon^2} = \widetilde{O} \left(\frac{\|C\|_\infty^2}{\alpha \varepsilon^2} \right).$$

The precision ε' is defined as

$$\varepsilon' = \frac{\varepsilon^3}{64M_\infty^2 R_d^2} = O \left(\frac{\varepsilon^3}{\|C\|_\infty^2} \right).$$

∎

Table 2. Total complexity of the SA and the SAA implementations for $\min_{p \in \Delta_n} \mathbb{E}_q W(p, q)$.

Algorithm	Theorem	Complexity
Projected SGD (SA) with $\gamma = \frac{\varepsilon}{4 \log n}$	5.1	$\widetilde{O}\left(\frac{n^3 \sqrt{n} \|C\|_\infty^3}{\varepsilon^3 \sqrt{\varepsilon \kappa}}\right)$
Stochastic MD (SA)	5.2	$\widetilde{O}\left(\frac{n^3 \|C\|_\infty^2}{\varepsilon^2}\right)$
Accelerated IBP (SAA) with $\gamma = \frac{\varepsilon}{4 \log n}$	5.3	$\widetilde{O}\left(\frac{n^4 \|C\|_\infty^4}{\varepsilon^4}\right)$
Mirror Prox with $B_d(p^*, p^1)$ penalization (SAA)	5.4	$\widetilde{O}\left(\frac{n^2 \sqrt{n} \|C\|_\infty^5}{\varepsilon^5}\right)$

5.5. Comparison of the SA and the SAA for the WB problem

Now we compare the complexity bounds for the SA and SAA implementations solving problem (55). We also skip the high probability details as in fact all bounds of this paper have logarithmic dependence on α [28,48]. Table 2 presents the total complexity for the numerical algorithms.

For the SA algorithms, which are Stochastic MD and Projected SGD, we can conclude the following: non-regularized approach (Stochastic MD) uses simplex prox structure and gets better complexity bounds, indeed Lipschitz constant in the ℓ_1-norm is $M_\infty = O(\|C\|_\infty)$, whereas Lipschitz constant in the Euclidean norm is $M = O(\sqrt{n}\|C\|_\infty)$. The practical comparison of Stochastic MD (Algorithm 4) and Projected SGD (Algorithm 3) can be found in Figure 3.

For the SAA approaches (Accelerated IBP and Mirror Prox with specific penalization), we enclose the following: entropy-regularized approach (Accelerated IBP) has better dependence on ε than penalized approach (Mirror Prox with specific penalization), however, worse dependence on n. Using Dual Extrapolation method for the WP problem from paper [25] instead of Mirror Prox allows to omit $\sqrt{n}$ in the penalized approach.

Using Dual Extrapolation method for the WP problem [25] instead of Mirror Prox, we can omit $\sqrt{n}$ in the last row of Table 2. One of the main advantages of the SAA approach is the possibility to perform it in a decentralized manner in contrast to the SA approach, which cannot be executed in a decentralized manner or even in distributed or parallel fashion for non-smooth objective [36]. This is the case of the Wasserstein barycenter problem, indeed, the objective is Lipschitz continuous but not Lipschitz smooth.

Notes

1. When for $\rho \geq 1$, $C_{ij} = d(x_i, x_j)^\rho$ in (2), where $d(x_i, x_j)$ is a distance on support points x_i, x_j, then $W(p, q)^{1/\rho}$ is known as the ρ-Wasserstein distance. Nevertheless, all the results of this thesis are based only on the assumptions that the matrix $C \in \mathbb{R}_+^{n \times n}$ is symmetric and non-negative. Thus optimal transport problem defined in (2) is a more general than the Wasserstein distances.
2. The estimate $n^2 \min\{\tilde{O}(\frac{1}{\varepsilon}), \tilde{O}(\sqrt{n})\}$ is the best known theoretical estimate for solving OT problem [14,42,53,59]. The best known practical estimates are $\sqrt{n}$ times worse (see [40] and references therein).

3. By using dual averaging scheme [55] we can rewrite Algorithm 4 in online regime [41,56] without including N in the stepsize policy. Note, that mirror descent and dual averaging scheme are very close to each other [43].

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Notes on contributor

Darina Dvinskikh received BSc and MSc degrees from the Faculty of Control and Applied Mathematics of Moscow Institute of Physics and Technology in 2016 and 2018 respectively. She obtained a PhD from Humboldt-Universität zu Berlin in 2021.

ORCID

Darina Dvinskikh http://orcid.org/0000-0003-1757-1021

References

[1] M. Agueh and G. Carlier, *Barycenters in the Wasserstein space*, SIAM Journal on Mathematical Analysis 43 (2011), pp. 904–924. https://doi.org/10.1137/100805741.

[2] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, *Network flows: Theory, Algorithms, and Applications 526*, 1993.

[3] J. Altschuler, J. Weed, and P. Rigollet, *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*, in *Advances in Neural Information Processing Systems 30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Curran Associates, Inc., 2017, pp. 1961–1971. Available at http://papers.nips.cc/paper/6792-near-linear-time-approximation-algorithms-for-optimal-transport-via-sinkhorn-iteration.pdf, arXiv:1705.09634.

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

[5] M. Ballu, Q. Berthet, and F. Bach, *Stochastic optimization for regularized Wasserstein estimators*, preprint (2020), arXiv:2002.08695.

[6] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, Society for Industrial and Applied Mathematics, 2001. Available at http://epubs.siam.org/doi/abs/10.1137/1.9780898718829.

[7] J.D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyr, *Iterative Bregman projections for regularized transportation problems*, SIAM. J. Sci. Comput. 37 (2015), pp. A1111–A1138.

[8] J. Bigot, E. Cazelles, and N. Papadakis, *Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration*, Inf. Infer. J. IMA 8 (2019), pp. 719–755.

[9] J. Bigot, E. Cazelles, and N. Papadakis, *Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications*, (2019).

[10] J. Bigot, E. Cazelles, and N. Papadakis, *Penalization of barycenters in the Wasserstein space*, SIAM J. Math. Anal. 51 (2019), pp. 2261–2285.

[11] J. Bigot, R. Gouet, T. Klein, and A. Lopez, et al. *Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line*, Electron. J. Stat. 12 (2018), pp. 2253–2289.

[12] J. Bigot and T. Klein, *Characterization of barycenters in the Wasserstein space by averaging optimal transport maps*, preprint (2012), arXiv:1212.2562.

[13] J. Bigot and T. Klein, et al. *Consistent estimation of a population barycenter in the Wasserstein space*, preprint (2012), ArXiv e-prints.

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

[15] E. Boissard, T. Le Gouic, and J.M. Loubes, et al. *Distributions template estimate with Wasserstein metrics*, Bernoulli 21 (2015), pp. 740–759.

[16] G. Carlier, *On the linear convergence of the multi-marginal Sinkhorn algorithm*, (2021).

[17] S. Chewi, T. Maunu, P. Rigollet, and A.J. Stromme, *Gradient descent algorithms for Bures–Wasserstein barycenters*, preprint (2020), arXiv:2001.01700.

[18] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, in *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, eds., Curran Associates, Inc., 2013, pp. 2292–2300. Available at http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf.

[19] M. Cuturi and G. Peyré, *A smoothed dual approach for variational Wasserstein problems*, SIAM J. Imaging Sci. 9 (2016), pp. 320–343.

[20] D. Dadush and S. Huiberts, *A friendly smoothed analysis of the simplex method*, in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 390–403.

[21] E. Del Barrio, J.A. Cuesta-Albertos, C. Matrán, and A. Mayo-Íscar, *Robust clustering tools based on optimal transportation*, Stat. Comput. 29 (2019), pp. 139–160.

[22] J. Delon and A. Desolneux, *A Wasserstein-type distance in the space of Gaussian mixture models*, SIAM J. Imaging. Sci. 13 (2020), pp. 936–970.

[23] Y. Dong, Y. Gao, R. Peng, I. Razenshteyn, and S. Sawlani, *A study of performance of optimal transport*, preprint (2020), arXiv:2005.01182.

[24] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections onto the l 1-ball for learning in high dimensions*, in *Proceedings of the 25th international conference on machine learning*, 2008, pp. 272–279.

[25] D. Dvinskikh and D. Tiapkin, *Improved complexity bounds in wasserstein barycenter problem,* in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021, pp. 1738–1746.

[26] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, *Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorns algorithm*, in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, eds., Vol. 80, 2018, pp. 1367–1376. arXiv:1802.04367.

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

[28] V. Feldman and J. Vondrák, *High probability generalization bounds for uniformly stable algorithms with nearly optimal rate*, preprint (2019). arXiv:1902.10710.

[29] J. Franklin and J. Lorenz, *On the scaling of multidimensional matrices*, in *Linear Algebra and its Applications 114*, 1989, pp. 717–735. Available at http://www.sciencedirect.com/science/article/pii/0024379589904904, Special Issue Dedicated to Alan J. Hoffman.

[30] M. Fréchet, *Les éléments aléatoires de nature quelconque dans un espace distancié*, in *Annales de l'institut Henri Poincaré 10*, 1948, pp. 215–310.

[31] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T.A. Poggio, *Learning with a Wasserstein loss*, in *Advances in Neural Information Processing Systems*, 2015, pp. 2053–2061.

[32] H.N. Gabow and R.E. Tarjan, *Faster scaling algorithms for general graph matching problems*, J. ACM38 (1991), pp. 815–853.

[33] A. Gasnikov, P. Dvurechensky, D. Kamzolov, Y. Nesterov, V. Spokoiny, P. Stetsyuk, A. Suvorikova, and A. Chernov, *Universal method with inexact oracle and its applications for searching equillibriums in multistage transport problems*, preprint (2015), arXiv:1506.00292.

[34] A.V. Gasnikov, A.A. Lagunovskaya, I.N. Usmanova, and F.A. Fedorenko, *Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex*, Autom. Remote Contr. 77 (2016), pp. 2018–2034. Available at http://dx.doi.org/10.1134/S0005117916110114, arXiv:1412.3890.

[35] A. Genevay, G. Peyré, and M. Cuturi, *Learning generative models with Sinkhorn divergences*, preprint (2017), arXiv:1706.00292.

[36] E. Gorbunov, D. Dvinskikh, and A. Gasnikov, *Optimal decentralized distributed algorithms for stochastic convex optimization*, preprint (2019), arXiv:1911.07363.

[37] T.L. Gouic, Q. Paris, P. Rigollet, and A.J. Stromme, *Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space*, preprint (2019), arXiv:1908.00828.

[38] A. Gramfort, G. Peyré, and M. Cuturi, *Fast optimal transport averaging of neuroimaging data*, in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 261–272.

[39] V. Guigues, A. Juditsky, and A. Nemirovski, *Non-asymptotic confidence bounds for the optimal value of a stochastic program*, Opt. Methods Softw. 32 (2017), pp. 1033–1058. Available at https://doi.org/10.1080/10556788.2017.1350177.

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

[41] E. Hazan, et al. *Introduction to online convex optimization*, Found. Trends Opt. 2 (2019), pp. 157–325.

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

[43] A. Juditsky, J. Kwon, and É. Moulines, *Unifying mirror descent and dual averaging*, preprint (2019). arXiv:1910.13742.

[44] A. Juditsky and A. Nemirovski, *First order methods for non-smooth convex large-scale optimization, i: General purpose methods*, in *Optimization for Machine Learning*, S.W. Suvrit Sra Sebastian Nowozin, ed., Cambridge, MA: MIT Press, 2012, pp. 121–184.

[45] A. Juditsky, P. Rigollet, and A.B. Tsybakov, et al. *Learning by mirror averaging*, Ann. Stat. 36 (2008), pp. 2183–2206.

[46] S.M. Kakade and A. Tewari, *On the generalization ability of online strongly convex programming algorithms*, in *Advances in Neural Information Processing Systems*, 2009, pp. 801–808.

[47] M. Klatt, C. Tameling, and A. Munk, *Empirical regularized optimal transport: Statistical theory and applications*, SIAM J. Math. Data Sci. 2 (2020), pp. 419–443.

[48] Y. Klochkov and N. Zhivotovskiy, *Stability and deviation optimal risk bounds with convergence rate O(1/n)*, (2021), arXiv preprint. arXiv:2103.12024.

[49] A. Kroshnin, V. Spokoiny, and A. Suvorikova, *Statistical inference for Bures–Wasserstein barycenters*, preprint (2019), arXiv:1901.00226.

[50] A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe, *On the Complexity of Approximating Wasserstein Barycenters*, in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, eds., Vol. 97, 2019, pp. 3530–3540. arXiv:1901.08686.

[51] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, *From word embeddings to document distances*, in *International conference on machine learning*, PMLR, 2015, pp. 957–966.

[52] T. Le Gouic and J.M. Loubes, *Existence and consistency of Wasserstein barycenters*, Probab. Theory Relat. Fields 168 (2017), pp. 901–917.

[53] Y.T. Lee and A. Sidford, *Path finding methods for linear programming: Solving linear programs in $\tilde{O}(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow*, in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014 Oct, pp. 424–433.

[54] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Opt. 19 (2009), pp. 1574–1609. Available at https://doi.org/10.1137/070704277.

[55] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, in *Mathematical Programming 120*, (2009), pp. 221–259. Available at https://doi.org/10.1007/s10107-007-0149-x, First appeared in 2005 as CORE discussion paper 2005/67.

[56] F. Orabona, *A modern introduction to online learning*, preprint (2019), arXiv:1912.13213.

[57] V.M. Panaretos and Y. Zemel, *Statistical aspects of Wasserstein distances*, Annu. Rev. Stat. Appl. 6 (2019), pp. 405–431.

[58] G. Peyré and M. Cuturi, et al. *Computational optimal transport*, Foundations and Trends in Machine Learning 11 (2019), pp. 355–607.

[59] K. Quanrud, *Approximating optimal transport with linear programs*, preprint (2018), arXiv:1810.05957.

[60] J. Rabin and N. Papadakis, *Convex color image segmentation with optimal transport distances*, in *International Conference on Scale Space and Variational Methods in Computer Vision*, Springer, 2015, pp. 256–269.

[61] J. Rabin, G. Peyré, J. Delon, and M. Bernot, *Wasserstein barycenter and its application to texture mixing*, in *International Conference on Scale Space and Variational Methods in Computer Vision*, Springer, 2011, pp. 435–446.

[62] S.T. Rachev, S.V. Stoyanov, and F.J. Fabozzi, *A Probability Metrics Approach to Financial Risk Measures*, John Wiley & Sons, 2011.

[63] G. Rios, J. Backhoff-Veraguas, J. Fontbona, and F. Tobar, *Bayesian learning with Wasserstein barycenters*, ArXiv e-prints (2018).

[64] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat. 22 (1951), pp. 400–407.

[65] A. Rolet, M. Cuturi, and G. Peyré, *Fast dictionary learning with a smoothed Wasserstein loss*, in *Artificial Intelligence and Statistics*, 2016, pp. 630–638.

[66] Y. Rubner, C. Tomasi, and L.J. Guibas, *A metric for distributions with applications to image databases*, in *6th International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, IEEE, 1998, pp. 59–66.

[67] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.

[68] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Stochastic convex optimization*, in *COLT*, 2009.

[69] A. Shapiro, D. Dentcheva, and A. Ruszczyski, *Lectures on Stochastic Programming*, Society for Industrial and Applied Mathematics, 2009, Available at http://epubs.siam.org/doi/abs/10.1137/1.9780898718751.

[70] A. Shapiro and A. Nemirovski, *On complexity of stochastic programming problems*, in *Continuous optimization*, Springer, 2005, pp. 111–146.

[71] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, *Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains*, ACM Trans. Graph. 34 (2015), pp. 66.

[72] M. Sommerfeld and A. Munk, *Inference for empirical Wasserstein distances on finite spaces*, J. R. Stat. Soc. Ser. B 80 (2018), pp. 219–238.

[73] V. Spokoiny, et al. *Parametric estimation. finite sample theory*, Ann. Stat. 40 (2012), pp. 2877–2909.

[74] S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson, *WASP: Scalable Bayes via barycenters of subset posteriors*, in *Artificial Intelligence and Statistics*, PMLR, (2015), pp. 912–920.

[75] F.S. Stonyakin, D. Dvinskikh, P. Dvurechensky, A. Kroshnin, O. Kuznetsova, A. Agafonov, A. Gasnikov, A. Tyurin, C.A. Uribe, D. Pasechnyuk, and S. Artamonov, *Gradient Methods for Problems with Inexact Model of the Objective*, in *Mathematical Optimization Theory and Operations Research*, M. Khachay, Y. Kochetov, and P. Pardalos, eds., Cham. Springer International Publishing, 2019, pp. 97–114. arXiv:1902.09001.

[76] M. Thorpe, S. Park, S. Kolouri, G.K. Rohde, and D. Slepcev, *A transportation L^p distance for signal analysis*, J. Math. Imaging. Vis. 59 (2017), pp. 187–210.

[77] W. Wang, J.A. Ozolek, D. Slepcev, A.B. Lee, C. Chen, and G.K. Rohde, *An optimal transportation approach for nuclear structure-based pathology*, IEEE. Trans. Med. Imaging 30 (2010), pp. 621–631.

[78] J. Weed and F. Bach, et al. *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance*, Bernoulli 25 (2019), pp. 2620–2648.