Decentralized Algorithms for Wasserstein Barycenters

DISSERTATION

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
im Fach Mathematik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Darina Dvinskikh

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr. Vladimir Spokoiny
2. Prof. Dr. Bernhard Schmitzer
3. Prof. Dr. Jonathan Niles-Weed

Tag der mündlichen Prüfung: 18. August 2021

Abstract

In this thesis, we consider the Wasserstein barycenter problem of discrete probability measures as well as the population Wasserstein barycenter problem given by a Fréchet mean from computational and statistical sides.

The statistical focus is estimating the sample size of measures needed to calculate an approximation of a Fréchet mean (barycenter) of probability distributions with a given precision. For empirical risk minimization approaches, the question of the regularization is also studied along with proposing a new regularization which contributes to the better complexity bounds in comparison with the quadratic regularization.

The computational focus is developing decentralized algorithms for calculating Wasserstein barycenters: dual algorithms and saddle point algorithms. The motivation for dual approaches is closed-forms for the dual formulation of entropy-regularized Wasserstein distances and their derivatives, whereas the primal formulation has a closed-form expression only in some cases, e.g., for Gaussian measures. Moreover, the dual oracle returning the gradient of the dual representation for entropy-regularized Wasserstein distance can be computed for a cheaper price in comparison with the primal oracle returning the gradient of the (entropy-regularized) Wasserstein distance. The number of dual oracle calls in this case will be also less, i.e., the square root of the number of primal oracle calls. Furthermore, in contrast to the primal objective, the dual objective has Lipschitz continuous gradient due to the strong convexity of regularized Wasserstein distances. Moreover, we study saddle-point formulation of the non-regularized Wasserstein barycenter problem which leads to the bilinear saddle-point problem. This approach also allows us to get optimal complexity bounds and it can be easily presented in a decentralized setup.

Keywords: optimal transport, Wasserstein barycenter, stochastic optimization, decentralized optimization, distributed optimization, primal-dual methods, first-order oracle.

Zusammenfassung

In dieser Arbeit beschäftigen wir uns mit dem Wasserstein Baryzentrumproblem diskreter Wahrscheinlichkeitsmaße sowie mit dem population Wasserstein Baryzentrumproblem gegeben von a Fréchet Mittelwerts von der rechnerischen und statistischen Seiten.

Der statistische Fokus liegt auf der Schätzung der Stichprobengröße von Maßen zur Berechnung einer Annäherung des Fréchet Mittelwerts (Baryzentrum) der Wahrscheinlichkeitsmaße mit einer bestimmten Genauigkeit. Für empirische Risikominimierung (ERM) wird auch die Frage der Regularisierung untersucht zusammen mit dem Vorschlag einer neuen Regularisierung, die zu den besseren Komplexitätsgrenzen im Vergleich zur quadratischen Regularisierung beiträgt.

Der Rechenfokus liegt auf der Entwicklung von dezentralen Algorithmen zur Berechnung von Wasserstein Baryzentrum: duale Algorithmen und Sattelpunktalgorithmen. Die Motivation für duale Optimierungsmethoden ist geschlossene Formen für die duale Formulierung von entropie-regulierten Wasserstein Distanz und ihren Derivaten, während, die primale Formulierung nur in einigen Fällen einen Ausdruck in geschlossener Form hat, z.B. für Gauß-Maße. Außerdem kann das duale Orakel, das den Gradienten der dualen Darstellung für die entropie-regulierte Wasserstein Distanz zurückgibt, zu einem günstigeren Preis berechnet werden als das primale Orakel, das den Gradienten der (entropie-regulierten) Wasserstein Distanz zurückgibt. Die Anzahl der dualen Orakelrufe ist in diesem Fall ebenfalls weniger, nämlich die Quadratwurzel der Anzahl der primalen Orakelrufe. Im Gegensatz zum primalen Zielfunktion, hat das duale Zielfunktion Lipschitz-stetig Gradient aufgrund der starken Konvexität regulierter Wasserstein Distanz. Außerdem untersuchen wir die Sattelpunktformulierung des (nicht regulierten) Wasserstein Baryzentrum, die zum Bilinearsattelpunktproblem führt. Dieser Ansatz ermöglicht es uns auch, optimale Komplexitätsgrenzen zu erhalten, und kann einfach in einer dezentralen Weise präsentiert werden.

Stichwörter: optimaler Transport, Wasserstein Baryzentrum, stochastische Optimierung, dezentrale Optimierung, primal-duale Optimierungsmethoden erster Ordnung, Orakel erster Ordnung.

To my family

Acknowledgements

First of all, I would thank my advisor, Vladimir Spokoiny, for his support, his advice and warm research meetings during my PhD at Weierstrass Institute in Berlin. Especially, I am grateful for his trust and mathematical freedom which he gave me in choosing a research direction provided me with valuable advice. It was an honor for me to work in his research group and and attend the group seminars where we got to know the research field of each group member and prominent visiting scientists from all over the world.

I thank the employees and researchers in the Weierstrass Institute who are always ready to help with organizational working issues. Especially, I thank Pavel Dvurechensky for his helpful advice and rewarding meeting that we have together. Many thanks to Franz Besold for his help with teaching assistants in the statistical seminars at Humboldt-Universität zu Berlin.

I am also very grateful to Alexander Gasnikov for his fruitful ideas and explanations which had a great influence on the content of this thesis.

I also want to thank warmly all the researches, I worked with, for the opportunity to carry out researches jointly. In particular, thanks to Angelia Nedić, César A. Uribe, Daniil Tiapkin, Eduard Gorbunov, Alexander Rogozin.

The research of Chapter 2 was supported by the Russian Science Foundation (project 18-71-10108), https://rscf.ru/project/18-71-10108/. The research of Chapter 4 was 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 research of Chapter 5 was funded by RFBR 19-31-51001.

Contents

List of Figures

List of Tables

Notations

- $\Delta_n = \{a \in \mathbb{R}^n_+ \mid \sum_{l=1}^n a_l = 1\}$ is the probability simplex.
- $I_{n\times n}$ is the identity matrix of size $n \times n$.
- $0_{n\times n}$ is zeros matrix of size $n \times n$.
- $\mathbf{1}_n$ is the vector of ones of size n.
- $[n]$ is the sequence of integer number from 1 to n.
- Capital symbols, e.g., A, B, are used for matrices.
- Bold capital symbols, e.g., $\mathbf{A}, \mathbf{B}$, are used for block-matrices.
- Bold small symbol, e.g., $\mathbf{x} = (x_1^\top, \cdots, x_m^\top)^\top \in \mathbb{R}^{mn}$ is the column vector of vectors $x_1, ..., x_m \in \mathbb{R}^n$.
- We refer to the i-th component of vector $\mathbf{x}$ as $x_i \in \mathbb{R}^n$.
- $[x]_j$ is j-th component of vector x.
- $\langle \cdot\, , \cdot \rangle$ is the usual Euclidean dot-product between vectors. For two matrices of the same size A and B, $\langle A, B\rangle = \mathrm{tr(AB)}$ is the Frobenius dot-product.
- $\|s\|_* = \max_{x\in X}\{\langle x, s\rangle : \|x\| \leq 1\}$ is the dual norm for some norm $\|x\|$, $x \in X$. In particular, for the ℓ_p-norm, its dual norm is ℓ_q-norm, where $\frac{1}{p} + \frac{1}{q} = 1$.
- For two vectors x, y (or matrices A, B) of the same size, x/y (A/B) and $x \odot y$ $(A \odot B)$ stand for the element-wise product and element-wise division respectively. When used on vectors, functions such as log or exp are always applied element-wise.
- For prox-function $d(x)$, the corresponding Bregman divergence is $B(x, y) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle$.
- $\lambda_{\max}(W)$ is the maximum eigenvalue of a symmetric matrix W
- $\lambda^+_{\min}(W)$ is the minimal non-zero eigenvalue of a symmetric matrix W
- $\chi(W) = \frac{\lambda_{\max}(W)}{\lambda^+_{\min}(W)}$ is the condition number of matrix W

- $O(\cdot)$ is the notation for an upper bound on the growth rate hiding constants.
- $\widetilde{O}(\cdot)$ is the notation for an upper bound on the growth rate hiding logarithms.

Mathematical Preliminaries

Definition 0.0.1 (M-Lipschitz). *A function $f : X \times \Xi \to \mathbb{R}$ is M-Lipschitz continious with respect to $x \in X$ in norm $\|\cdot\|$ if it satisfies*

$$|f(x,\xi) - f(y,\xi)| \leq M\|x - y\|, \qquad \forall x, y \in X,\ \forall \xi \in \Xi. \tag{1}$$

From Eq. (1) it follows that

$$\|\nabla_x f(x,\xi)\|_* \leq M, \qquad \forall x \in X,\ \forall \xi \in \Xi,$$

where $\nabla_x f(x,\xi)$ is a subgradient of $f(x,\xi)$ with respect to x (Shapiro et al., 2009).

Definition 0.0.2 (L-smoothness). *A function $f : X \times \Xi \to \mathbb{R}$ is L-Lipschitz smooth, or has L-Lipschitz continuous gradient, with respect to norm $\|\cdot\|_X$ if $f(x,\xi)$ is continuously differentiable with respect to x and its gradient satisfies Lipschitz condition*

$$\|\nabla_x f(x,\xi) - \nabla_y f(y,\xi)\|_* \leq L\|x - y\|, \quad \forall x, y \in X,\ \forall \xi \in \Xi. \tag{2}$$

From Eq. (2) it follows that

$$f(y,\xi) \leq f(x,\xi) + \langle \nabla_x f(x,\xi), y - x\rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in X,\ \forall \xi \in \Xi. \tag{3}$$

Definition 0.0.3 (γ-strong convexity). *A function $f : X \times \Xi \to \mathbb{R}$ is γ-strongly convex with respect to x in norm $\|\cdot\|_X$ if it is continuously differential and it satisfies*

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

Definition 0.0.4 (Dual Function). *The Fenchel–Legendre conjugate for a function $f : (X, \Xi) \to \mathbb{R}$ is*

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

Theorem 0.0.5. *(Kakade et al., 2009, Theorem 6 (Strong/Smooth Duality)) Assume that f is a closed and convex function on $X = \mathbb{R}^n$. Then f is γ-strongly convex w.r.t. a norm $\|\cdot\|_X$ if and only if f^* is $\frac{1}{\gamma}$–Lipschitz smooth w.r.t. the dual norm $\|\cdot\|_{X^*}$.*

Theorem 0.0.6. *(Nesterov, 2005, Theorem 1) Assume that function $f(x)$ is continuous and γ-strongly convex w.r.t. a norm $\|\cdot\|$. Then $\varphi(u) = \max_{x \in X}\{\langle Ax, u\rangle - f(x)\}$ is $\frac{\lambda_{\max}(A^\top A)}{\gamma}$–Lipschitz smooth w.r.t. the dual norm $\|\cdot\|_*$.*

Chapter 1

Introduction

1.1 Background on Optimal Transport

Optimal transport problem is closely related to the notion of *linear programming*. Linear programming (LP) is the science of theoretical and numerical analysis and solving extremal (e.g., maximization or maximization) problems defined by systems of linear equations and inequalities. A lot of mathematicians made contributions to the development of linear programming, including T. Koopmans, G.B. Danzig (a founder of the simplex method, 1949) and I.I. Dikin (a founder of the interior points method, 1967), but the priority belongs to the Soviet mathematician and economist L. V. Kantorovich (Kantorovich, 1960), who was the first who discovered that a wide class of the most important production problems can be described mathematically and solved numerically (1939).

Particular and important cases of linear programming problems are network flow problem, multicommodity flow problem, and *optimal transport* (OT) problem. The history of optimal transport begins with the French mathematician G. Monge (Monge, 1781), who proposed a complicated theory of describing an optimal mass transportation in a geometric way. Inspired by the problem of resource allocation, L.V. Kantorovich introduced relaxations which allowed him to formulate the transport problem as linear programming problem, and as a consequence, to apply linear programming methods to solve it. The main relaxation was based on the refusing of deterministic nature of transportation (a mass from the source point could only be transferred to one target point) and introducing a probabilistic transport. To do so, a coupling matrix was introduced instead of Monge maps. Admissible couplings (also known as transportation polytope) of all coupling matrices with marginals discrete source μ and discrete target ν can be written as follows

$$U(\mu, \nu) \triangleq \{\pi \in \mathbb{R}_+^{n_2 \times n_1} : \pi \mathbf{1}_{n_1} = \mu, \pi^T \mathbf{1}_{n_2} = \nu\}.$$

Here π is a coupling (transport plan) (π_{ij} describes the amount of mass moving from source bin i towards target bin j) Thus, the problem of optimal transport between μ and ν under a symmetric transportation cost matrix $C \in \mathbb{R}_+^{n \times n}$, called

also as the Monge–Kantorovich problem, is formulated as follows

$$\min_{\pi \in U(\mu,\nu)} \langle C, \pi \rangle. \tag{1.1}$$

Moreover, Kantorovich formulated an infinite-dimension analog of optimal transport problem (1.1) between probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ under transportation cost function $c(x, y)$

$$\min_{\pi \in \mathcal{U}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y),$$

where

$$\mathcal{U}(\mu, \nu) \triangleq \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : T_{\mathcal{X}\#} = \mu, T_{\mathcal{Y}\#} = \nu\}.$$

Here $T_{\mathcal{X}\#}$ and $T_{\mathcal{Y}\#}$ are the push-forwards. Furthermore, the replacement of Monge's maps by couplings and infinite-dimension formulation of optimal transport allowed Kantorovich and G. S. Rubinstein to introduce Kantorovich–Rubinstein distance in the space of probability measures. Nowadays, it is often referred to as Wasserstein distance. Namely, ρ-Wasserstein distance ($\rho \geq 1$) between probability measures $\mu, \nu \in \mathcal{P}(X)$ is defined as follows

$$\mathcal{W}_\rho(\mu, \nu) \triangleq \left(\min_{\pi \in \mathcal{U}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \mathtt{d}(x, y)^\rho d\pi(x, y) \right)^{1/\rho}, \tag{1.2}$$

where it was assumed that $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = \mathtt{d}(x, y)^\rho$ is a distance on $\mathcal{X}$.

For multivariate Gaussian measures, the 2-Wasserstein distance has a closed-form solution and is expressed through Bures metric (Bures, 1969) which is used to compare quantum states in quantum physics.

Nowadays, optimal transport metric provides a successful framework to compare objects that can be modeled as probability measures (images, videos, texts and etc.). Transport based distances, especially 1-Wasserstein distance (EMD), have gained popularity in various fields such as statistics (Ebert et al., 2017; Bigot et al., 2012), unsupervised learning (Arjovsky et al., 2017), signal and image analysis (Thorpe et al., 2017), computer vision (Rubner et al., 1998), text classification (Kusner et al., 2015), economics and finance (Rachev et al., 2011) and medical imaging (Wang et al., 2010; Gramfort et al., 2015). A lot of statistical results are known about optimal transport (Wasserstein) distances (Sommerfeld and Munk, 2018; Weed et al., 2019; Klatt et al., 2020).

1.2 Background on Wasserstein Barycenters

The success of optimal transport led to an increasing interest in *Wasserstein barycenters*. In (Agueh and Carlier, 2011), the notion of a Wasserstein barycenter was introduced in the Wasserstein space (space $\mathcal{P}_2(\mathcal{X})$ of probability measures with finite second moment supported on a convex domain $\mathcal{X}$) similarly to the barycenter

of points in the Euclidean space by replacing the squared Euclidean distance with the squared 2-Wasserstein distance. Namely, a Wasserstein barycenter of a set of probability measures $\nu_1, \nu_2, ..., \nu_m$ is defined as follows

$$\min_{\mu \in \mathcal{P}_2(\mathcal{X})} \sum_{i=1}^{m} \lambda_i \mathcal{W}_2^2(\mu, \nu_i), \tag{1.3}$$

where the λ_i's are positive weights summing to 1.

Wasserstein barycenters are used in Bayesian computations (Srivastava et al., 2015), texture mixing (Rabin et al., 2011), clustering (k-means for probability measures) (Del Barrio et al., 2019), shape interpolation and color transferring (Solomon et al., 2015), statistical estimation of template models (Boissard et al., 2015) and neuroimaging (Gramfort et al., 2015).

1.3 Background on Population Wasserstein barycenter

For random probability measures with distribution $\mathbb{P}$ supported on $P_2(\mathcal{X})$, population Wasserstein barycenter is introduced through a notion of a Fréchet mean (Fréchet, 1948)

$$\min_{p \in P_2(\mathcal{X})} \mathbb{E}_{q \sim \mathbb{P}} W(p, q) = \min_{p \in P_2(\mathcal{X})} \int_{P_2(\mathcal{X})} W(p, q) d\mathbb{P}(q). \tag{1.4}$$

For identically distributed measures, problem (1.3) can be interpreted as an empirical counterpart of problem (1.4). If a solution of (1.4) exists and is unique, then it is referred to as the population barycenter of distribution $\mathbb{P}$.

1.4 Overview of the Thesis

In this thesis, we consider the Wasserstein barycenter problem of discrete probability measures as well as the population Wasserstein barycenter problem given by a Fréchet mean. The main focus of this thesis is computational aspect of the Wasserstein barycenter problem: deriving first-order methods to compute Wasserstein barycenters. Dual first-order methods rely on the fact that regularized optimal transport by negative entropy with $\gamma > 0$, that is

$$W_\gamma(p, q) = \min_{\pi \in U(p,q)} \left\{ \langle C, \pi \rangle + \gamma \langle \pi, \log \pi \rangle \right\},$$

has a dual closed-form representation defined by the Fenchel–Legendre transform w.r.t. $p \in \Delta_n$ (Agueh and Carlier, 2011; Cuturi and Peyré, 2016):

$$\begin{aligned} W^*_{\gamma,q}(u) &= \max_{p \in \Delta_n} \left\{ \langle u, p \rangle - W_\gamma(p, q) \right\} \\ &= \gamma \left(-\langle q, \log q \rangle + \sum_{j=1}^{n} [q]_j \log \left(\sum_{i=1}^{n} \exp\left(([u]_i - C_{ji})/\gamma \right) \right) \right), \end{aligned} \tag{1.5}$$

where $[q]_j$ and $[u]_i$ are the j-th and i-th components of q and u respectively, and C_{ji} is the entry of matrix C. The gradient of dual function $W^*_{\gamma,q}(u)$ is Lipschitz continuous and has also a closed-form solution

$$[\nabla W^*_{\gamma,q}(u)]_l = \sum_{j=1}^{n} [q]_j \frac{\exp\left(([u]_l - C_{lj})/\gamma\right)}{\sum_{\ell=1}^{n} \exp\left(([u]_\ell - C_{\ell j})/\gamma\right)}, \tag{1.6}$$

for all $l = 1, ..., n$.

A saddle point approach for Wasserstein barycenter problem relies on the fact that non-regularized optimal transport (1.1) has a bilinear saddle-point representation (Jambulapati et al., 2019):

$$W(p,q) = \min_{x \in \Delta_{n^2}} \max_{y \in [-1,1]^{2n}} \left\{ \langle d, x \rangle + 2\|d\|_\infty \left(y^\top A x - \left\langle \begin{pmatrix} p \\ q \end{pmatrix}, y \right\rangle \right) \right\}.$$

Here d is the vectorized cost matrix C, $x \in \Delta_{n^2}$ is the vectorized transport plan π, and

$$A \triangleq \begin{pmatrix} I_{n\times n} & \otimes & \mathbf{1}_n^\top \\ \mathbf{1}_n^\top & \otimes & I_{n\times n} \end{pmatrix} = \{0,1\}^{2n \times n^2}$$

is the incidence matrix.

Decentralized formulations of the Wasserstein barycenter problem both for the saddle-point and dual representations are based on introducing artificial constraint $p_1 = p_2 = ... = p_m \in \mathbb{R}^n$ which is further replaced with affine constraint $\mathbf{W}\mathbf{p} = 0$ (in the saddle-point approach) and $\sqrt{\mathbf{W}}\mathbf{p} = 0$ (in the dual approach), where $\mathbf{p} = (p_1^\top, ..., p_m^\top)^\top$ is column vector and $\mathbf{W}$ is referred as the communication matrix for a decentralized system. From the definition of matrix $\mathbf{W}$ it follows that

$$\sqrt{\mathbf{W}}\mathbf{p} = 0 \iff \mathbf{W}\mathbf{p} = 0 \iff p_1 = p_2 = ... = p_m.$$

The affine constraint $\mathbf{W}\mathbf{p} = 0$ (or $\sqrt{\mathbf{W}}\mathbf{p} = 0$) is brought to the objective via the Fenchel–Legendre transform. Thus, for the primal Wasserstein barycenter problem defined w.r.t. entropy-regularized optimal transport

$$\min_{p \in \Delta_n} \frac{1}{m} \sum_{i=1}^{m} W_\gamma(p, q_i) = \min_{\substack{p_1 = ... = p_m, \\ p_1, ..., p_m \in \Delta_n}} \frac{1}{m} \sum_{i=1}^{m} W_\gamma(p_i, q_i) = \min_{\substack{\sqrt{\mathbf{W}}\mathbf{p} = 0, \\ p_1, ..., p_m \in \Delta_n}} \frac{1}{m} \sum_{i=1}^{m} W_\gamma(p_i, q_i),$$

we can construct the corresponding dual Wasserstein barycenter problem:

$$\min_{\mathbf{y} \in \mathbb{R}^{nm}} W^*_{\gamma,\mathbf{q}}(\sqrt{\mathbf{W}}\mathbf{y}) \triangleq \frac{1}{m} \sum_{i=1}^{m} W^*_{\gamma,q_i}(m[\sqrt{\mathbf{W}}\mathbf{y}]_i), \tag{1.7}$$

where $\mathbf{q} = (q_1^\top, \cdots, q_m^\top)^\top$, and $\mathbf{y} = (y_1^\top, \cdots, y_m^\top)^\top \in \mathbb{R}^{nm}$ is the Lagrangian dual multiplier. As the primal function is strongly convex, then the dual function is L-Lipschitz smooth, or has Lipschitz continuous gradient. The constant L for $W^*_{\gamma,\mathbf{q}}(\sqrt{\mathbf{W}}\mathbf{y})$ is defined via communication matrix $\mathbf{W}$ and regularization parameter γ. Hence, accelerated gradient descent-based method can be used, which is optimal

in terms of the number of iterations and oracle calls. For simplicity, the decentralized procedure solving dual problem (1.7) can be demonstrated on the gradient descent as follows

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L}\nabla W^*_{\gamma,\mathbf{q}}(\sqrt{\mathbf{W}}\mathbf{y}^k) = \mathbf{y}^k - \frac{1}{L}\sqrt{\mathbf{W}}\mathbf{p}(\sqrt{\mathbf{W}}\mathbf{y}^k).$$

Without change of variable, it is unclear how to execute this procedure in a distributed fashion. Let $\mathbf{u} := \sqrt{\mathbf{W}}\mathbf{y}$, then the gradient step multiplied by $\sqrt{\mathbf{W}}$ can be rewritten as

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \frac{1}{L}\mathbf{W}\mathbf{p}(\mathbf{u}^k),$$

where $[\mathbf{p}(\boldsymbol{u})]_i = p_i(u_i) = \nabla W^*_{\gamma,q_i}(u_i)$ from (1.6), $i = 1, ..., m$. This procedure can be performed in a decentralized manner on a distributed network. The vector $\mathbf{W}\mathbf{p}(\boldsymbol{u})$ naturally defines communications with neighboring nodes due to the structure of communication matrix $\mathbf{W}$ as the elements of communication matrix are zero for non-neighboring nodes. Moreover, in the dual approach which is based on gradient method, the randomization of $\nabla W^*_{\gamma,q_i}(u_i)$ can be used to reduce the complexity of calculating the true gradient, that is $O(n^2)$ arithmetic operations, by calculating its stochastic approximation of $O(n)$ arithmetic operations. The randomization for the true gradient (1.6) is achieved by taking the j-th term in the sum with probability $[q]_j$

$$[\nabla W^*_{\gamma,q}(u,\xi)]_l = \frac{\exp\left(([u]_l - C_{l\xi})/\gamma\right)}{\sum_{\ell=1}^n \exp\left(([u]_\ell - C_{\ell\xi})/\gamma\right)}, \qquad \forall l = 1, ..., n.$$

where we replaced index j by ξ to underline its randomness. This is the motivation for considering the first-order methods with stochastic oracle.

For greater generality, we derive the methods for a general convex minimization problem where the objective is given by the sum of functions, and for a general stochastic convex minimization problem where the objective is given by its expectation. These two problems are generalizations of problems (1.3) and (1.4). The reason for this generality is obtaining the results of other interests than Wasserstein barycenter problem.

Thus, we consider a general stochastic convex optimization problem whose objective is given by its expectation (problem (1.4) is a particular case of this problem)

$$\min_{x \in X \subseteq \mathbb{R}^n} F(x) \triangleq \mathbb{E} f(x, \xi), \tag{1.8}$$

where $\mathbb{E}f(x,\xi)$ is the expectation with respect to random variable ξ from set Ξ, $f(x,\xi)$ is convex in x on convex set X. Such kind of problems arise in many machine learning applications (Shalev-Shwartz and Ben-David, 2014) (e.g., empirical risk minimization) and statistical applications (Spokoiny et al., 2012) (e.g., maximum likelihood estimation). We will say that an output x^N of an algorithm is an ε-solution of problem (1.8) if the following holds with probability at least $1 - \beta$

$$F(x^N) - \min_{x \in X} F(x) \leq \varepsilon.$$

The complexity of an algorithm is measured by the number of iterations and the number of oracle calls. We consider the (stochastic) first-order oracle, i.e., the oracle which for a given realization $\xi \in \Xi$, returns the gradient (subgradient) of $f(x, \xi)$ calculated with respect to $x \in X$. For the dual first-order methods, we use the dual (stochastic) first-order oracle returning the gradient of the dual to $f(x, \xi)$ function given by the Fenchel–Legendre transform of $f(x, \xi)$.

We also consider a general convex optimization problem whose objective is given by the sum of convex functions (problem (1.3) is a particular case of this problem)

$$\min_{x \in X \subseteq \mathbb{R}^n} f(x) \triangleq \frac{1}{m} \sum_{i=1}^{m} f_i(x). \tag{1.9}$$

Problems of type (1.9) can be effectively solved in a distributed manner on a computational network. In the last decade, distributed optimization became especially popular with the release of the book (Bertsekas and Tsitsiklis, 1997) and due to the emergence of big data and rapid growth of problem sizes. The idea of distributed calculations is simple: every node (computational unit of some connected undirected graph (network)), assigned by its private function f_i, calculates the gradient of the private function and simultaneously communicates with its neighbors by exchanging messages at each communication round.

For primal approaches, the lower and upper bounds on communications rounds and (stochastic) primal oracle calls of ∇f_i per node i are known, as well as the methods matching these lower bounds. We refer to works (Scaman et al., 2017; Li et al., 2018; Uribe et al., 2017) describing these bounds for Lipschitz smooth deterministic objective. For non-smooth (deterministic and stochastic) objective, we appeal to (Lan et al., 2017; Scaman et al., 2018). In the stochastic Lipschitz smooth case, the optimal bound on the number of communication rounds was obtained in (Dvinskikh and Gasnikov, 2021), the optimal bound on the number of stochastic oracle calls was gained in (Rogozin et al., 2021a). Tables 1.1 and 1.2 summarize the results for deterministic and stochastic primal oracles respectively. In these tables, factor $\widetilde{O}(\sqrt{\chi})$ is responsible for the consensus time, i.e., the number of communication rounds required to reach the consensus in the considered network; σ^2 and σ_ψ^2 are the sub-Gaussian variance for $\nabla f_i(x_i, \xi_i)$ and $\nabla \psi_i(\lambda_i, \xi_i)$ respectively, where $\nabla \psi_i(\lambda_i, \xi_i)$ is the dual function to $\nabla f_i(x_i, \xi_i)$ with respect to x_i.

For deterministic dual oracle, the bounds are also known: Scaman et al. (2017) provided the results for strongly convex and smooth primal objective, the bounds for non-smooth but strongly convex primal objective were obtained in (Uribe et al., 2018, 2020). Stochastic dual oracle was not actively studied and optimal bounds on the number of stochastic dual oracle calls were not obtained. We leverage this gap and derive primal-dual decentralized algorithms which are optimal in terms of the number of dual (stochastic) oracle calls and the number of communication rounds. Table 1.3 summarizes the results for deterministic dual oracle. Table 1.4 demonstrates one of the contributions of this thesis: optimal bounds for stochastic dual oracle. The case of non-smooth but strongly convex primal objective in Table 1.4 corresponds to the Wasserstein barycenter problem defined with respect to

Table 1.1: Optimal bounds on the number of communication rounds and deterministic oracle calls of $\nabla f_i(x_i)$ per node

Property of f_i	μ-strongly convex, L-smooth	L-smooth	μ-strongly convex, M-Lipschitz	M-Lipschitz
Number of communication rounds	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$	$\widetilde{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\chi\right)$	$O\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$	$O\left(\sqrt{\frac{M^2R^2}{\varepsilon^2}}\chi\right)$
Number of oracle calls of $\nabla f_i(x_i)$ per node i	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\right)$	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$O\left(\frac{M^2}{\mu\varepsilon}\right)$	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Table 1.2: Optimal bounds on the number of communication rounds and stochastic oracle calls of $\nabla f_i(x_i, \xi_i)$ per node

Property of f_i	μ-strongly convex, L-smooth	L-smooth	μ-strongly convex, $\mathbb{E}\|\nabla f_i(x_i,\xi_i)\|_2^2 \leq M^2$	$\mathbb{E}\|\nabla f_i(x_i,\xi_i)\|_2^2 \leq M^2$
Number of communication rounds	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$	$\widetilde{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\chi\right)$	$O\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$	$O\left(\sqrt{\frac{M^2R^2}{\varepsilon^2}}\chi\right)$
Number of oracle calls of $\nabla f_i(x_i,\xi_i)$ per node i	$\widetilde{O}\left(\max\left\{\frac{\sigma^2}{m\mu\varepsilon}, \sqrt{\frac{L}{\mu}}\right\}\right)$	$O\left(\max\left\{\frac{\sigma^2R^2}{m\varepsilon^2}, \sqrt{\frac{LR^2}{\varepsilon}}\right\}\right)$	$O\left(\frac{M^2}{\mu\varepsilon}\right)$	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

entropy-regularized optimal transport. This is one of the motivation to consider the dual oracle since the dual representation (the Fenchel–Legendre transformation) of the entropy-regularized optimal transport and its derivatives can be presented in closed-forms.

Table 1.3: The optimal bounds for dual deterministic oracle

Property of f_i	μ-strongly convex, L-smooth	μ-strongly convex, $\|\nabla f_i(x^*)\|_2 \leq M$
The number of communication rounds	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}\chi(W)}\right)$	$O\left(\sqrt{\frac{M^2}{\mu\varepsilon}\chi(W)}\right)$
The number of oracle calls of $\nabla\psi_i(\lambda_i)$ per node i	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}\chi(W)}\right)$	$O\left(\sqrt{\frac{M^2}{\mu\varepsilon}\chi(W)}\right)$

Table 1.4: The optimal bounds for dual stochastic (unbiased) oracle

Property of f_i	μ-strongly convex, L-smooth	μ-strongly convex, $\|\nabla f_i(x^*)\|_2 \leq M$
The number of communication rounds	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi(W)\right)$	$O\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi(W)\right)$
The number of oracle calls of $\nabla\psi_i(\lambda_i, \xi_i)$ per node i	$\widetilde{O}\left(\max\left\{\frac{M^2\sigma_\psi^2}{\varepsilon^2}\chi(W), \sqrt{\frac{L}{\mu}}\chi(W)\right\}\right)$	$O\left(\max\left\{\frac{M^2\sigma_\psi^2}{\varepsilon^2}\chi(W), \sqrt{\frac{M^2}{\mu\varepsilon}}\chi(W)\right\}\right)$

1.4.1 Thesis Structure

The dissertation consists of 5 Chapters:

In Chapter 2, we study the two main approaches in machine learning and optimization community for convex risk minimization problem, namely, the Stochastic Approximation (SA) and the Sample Average Approximation (SAA) also known as the Monte Carlo approach. 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 the specific problem, however, starting from work (Nemirovski et al., 2009) it was generally accepted that the SA is better than the SAA. We show that for the Wasserstein barycenter problem, this superiority can be swapped. We provide the detailed comparison with stating the complexity bounds for the SA and the SAA implementations calculating Fréchet mean defined with respect to optimal transport distances and Fréchet mean defined with respect to entropy-regularized optimal transport distances. As a byproduct, we also construct confidence intervals for population barycenter defined with respect to entropy-regularized optimal transport distances in the ℓ_2-norm. Here we propose a new regularization for the the SAA approach which contributs to a better convergence rate in comparison with the quadratic regularization. The preliminary results were derived for a general convex optimization problem given by the expectation so that they can be applied to a wider range of problems other than the Wasserstein barycenter problem.

In Chapter 3, we introduce a decentralized dual algorithm to minimize the sum of strongly convex functions on a network of agents (nodes). This algorithm is based on accelerated gradient descent and it allows to obtain optimal bounds on the number of communication rounds and oracle calls of dual objective per node. The results can be naturally applied for the Wasserstein barycenter problem as the dual formulation of entropy-regularized Wasserstein distances and their derivatives have closed-form representations.

In Chapter 4, we provide saddle point approach to compute unregularized Wasserstein barycenters with no limitations in contrast to the regularized-based methods, which are numerically unstable under a small value of the regularization

parameter. The approach is based on the saddle-point problem reformulation and the application of mirror prox algorithm with a specific norm. We also show how the algorithm can be executed in a decentralized manner. The complexity of the proposed methods meets the best known results in the decentralized and non-decentralized setting.

Chapter 5 has interests other than Wasserstein barycenters. The purpose of this Chapter is obtaining the optimal bounds on the number of communication rounds and oracle calls for the gradient of the dual objective per node in the problem of minimizing the sum of strongly convex functions with Lipschitz continuous gradients. Thus, this Chapter complements Chapter 3 for the case of additionally Lipschitz smooth (stochastic) objectives.

1.5 Main Contributions

- *Statistical issue:* statistical study of the Wasserstein barycenter problem
  (a) Estimating the sample size of measures needed to calculate an approximation for a Fréchet mean (barycenter) of a probability distribution with a given precision
  (b) Proposing a new regularization for risk minimization approach (also known as the SAA approach) which contributes to better convergence rate in comparison with quadratic regularization
- *Computational issue:* proposing decentralized (stochastic) algorithms with optimal convergence rates
  (a) Obtaining optimal bounds on the number of communication rounds and dual oracle calls for the gradient of the dual (stochastic) objective per node in decentralized optimization for minimizing the sum of strongly convex functions, possibly with Lipschitz continuous gradients
  (b) Developing decentralized algorithms with the best known bounds for the problem of calculating Wasserstein barycenters of a set of discrete measures

1.6 Bibliographic Notes

The contribution of this thesis is based on the following papers.

- Chapter 2 is based on the work (Dvinskikh, 2020) accepted to the journal 'Optimization Methods and Software'
- Chapter 3 is partially based on the results of joint paper with Eduard Gorbunov, Alexander Gasnikov, Pavel Dvurechensky and César A. Uribe (Dvinskikh et al., 2019) published in the proceedings of the 58th Conference on

Decision and Control (CDC, 2019 IEEE), on a part of the results of joint paper with Pavel Dvurechensky, Alexander Gasnikov, Angelia Nedić and César A. Uribe (Dvurechensky et al., 2018a) published in the proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), and on a part of the results of joint paper with Alexey Kroshnin, Nazarii Tupitsa, Pavel Dvurechensky, Alexander Gasnikov and César A. Uribe (Kroshnin et al., 2019) published in the proceedings of the 36th International Conference on Machine Learning

- Chapter 4 partially uses the results from joint paper with Daniil Tiapkin (Dvinskikh and Tiapkin, 2021) published in the proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS, 2021). Besides, this Chapter contains a part of the results from arXiv preprint (Rogozin et al., 2021b) with Alexander Rogozin, Alexander Beznosikov, Dmitry Kovalev, Pavel Dvurechensky and Alexander Gasnikov

- The results of Chapter 5 are from joint paper with Alexander Gasnikov (Dvinskikh and Gasnikov, 2021) published in the Journal of Inverse and Ill-posed Problems, 2021

Chapter 2

Two Approaches: Stochastic Approximation (SA) and Sample Average Approximation (SAA).

This Chapter is inspired by the work (Nemirovski et al., 2009) stated that the SA approach outperforms the SAA approach for certain class of convex stochastic problems. We show that for the Wasserstein barycenter problem, this superiority can be inverted. We provide detailed comparison with stating the complexity bounds for the SA and the SAA implementations calculating Fréchet mean defined with respect to optimal transport distances and entropy-regularized optimal transport distances. The preliminary results are derived for a general convex optimization problem given by the expectation for interest other than the Wasserstein barycenter problem.

Background on the SA and the SAA and Convergence Rates. We consider the stochastic convex minimization problem

$$\min_{x \in X \subseteq \mathbb{R}^n} F(x) \triangleq \mathbb{E} f(x, \xi), \tag{2.1}$$

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 (Shalev-Shwartz and Ben-David, 2014; Shapiro et al., 2009) (e.g., risk minimization) and mathematical statistics (Spokoiny et al., 2012) (e.g., maximum likelihood estimation). There are two competing approaches based on Monte Carlo sampling techniques to solve (2.1): the Stochastic Approximation (SA) (Robbins and Monro, 1951) and the Sample Average Approximation (SAA). The SAA approach replaces the objective in problem (2.1) with its sample average approximation (SAA) problem

$$\min_{x \in X} \hat{F}(x) \triangleq \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i), \tag{2.2}$$

where $\xi_1, \xi_2, ..., \xi_m$ are the realizations of a random variable ξ. The number of realizations m is adjusted by the desired precision. The total working time of both approaches to solve problem (2.1) with the average precision ε in the non-optimality gap in term of the objective function (i.e., to find x^N such that $\mathbb{E}F(x^N) - \min\limits_{x \in X} F(x) \leq \varepsilon$), depends on the specific problem. However, it was generally accepted (Nemirovski et al., 2009) that the SA approach is better than the SAA approach. Stochastic gradient (mirror) descent, an implementation of the SA approach (Juditsky and Nemirovski, 2012), gives the following estimation for the number of iterations (that is equivalent to the sample size of $\xi_1, \xi_2, \xi_3, ..., \xi_m$)

$$m = O\left(\frac{M^2 R^2}{\varepsilon^2}\right). \tag{2.3}$$

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

$$\|\nabla f(x, \xi)\|_2^2 \leq M^2, \quad \forall x \in X, \xi \in \Xi. \tag{2.4}$$

Whereas, the application of the SAA approach requires the following sample size (Shapiro and Nemirovski, 2005)

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

that is n times more (n is the problem's dimension) than the sample size in the SA approach. This estimate was obtained under the assumptions that problem (2.2) 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 (2.2) with accuracy (Shalev-Shwartz et al., 2009)

$$\varepsilon' = O\left(\frac{\varepsilon^2 \lambda}{M^2}\right). \tag{2.5}$$

Therefore, to eliminate the linear dependence on n in the SAA approach for a non-strongly convex objective, regularization $\lambda = \frac{\varepsilon}{R^2}$ should be used (Shalev-Shwartz et al., 2009).

Let us suppose that $f(x, \xi)$ in (2.1) is convex but non-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 (2.1) which corresponds to the minimum of this norm (if the solution is not the only one). Then, the problem (2.1) can be replaced by

$$\min_{x \in X} \mathbb{E} f(x, \xi) + \frac{\varepsilon}{2R^2} \|x - x^1\|_2^2. \tag{2.6}$$

The empirical counterpart of (2.6) is

$$\min_{x\in X} \frac{1}{m}\sum_{i=1}^{m} f(x,\xi_i) + \frac{\varepsilon}{2R^2}\|x-x^1\|_2^2, \tag{2.7}$$

where the sample size m is defined in (2.3). Thus, in the case of non-strongly convex objective, a regularization equates the sample size of both approaches.

2.1 Strongly Convex Optimization Problem

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

$$\min_{x\in X\subseteq\mathbb{R}^n} F(x) \triangleq \mathbb{E}f(x,\xi), \tag{2.8}$$

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

2.1.1 The SA Approach: Stochastic Gradient Descent

The classical SA algorithm for problem (2.8) is presented by stochastic gradient descent (SGD) method. We consider the SGD with inexect oracle given by $g_\delta(x,\xi)$ such that

$$\forall x\in X, \xi\in\Xi, \qquad \|\nabla f(x,\xi) - g_\delta(x,\xi)\|_2 \le \delta. \tag{2.9}$$

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

$$x^{k+1} = \Pi_X\left(x^k - \eta_k g_\delta(x^k,\xi^k)\right). \tag{2.10}$$

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, stepsize η_k can be taken as $\frac{1}{\gamma k}$ to obtain optimal rate $O(\frac{1}{\gamma N})$.

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 (Kakade and Tewari, 2009) gives a bound on the excess risk of the output of an online algorithm in terms of the average regret.

Theorem 2.1.1. *(Kakade and Tewari, 2009, 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 $x^1, x^2, ..., x^N$. Then with probability at least $1-4\beta\log N$*

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

For the update rule (2.10) with $\eta_k = \frac{1}{\gamma k}$, this theorem can be specify as follows.

Theorem 2.1.2. *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 (2.10) with $\eta_k = \frac{1}{\gamma k}$. Then, with probability at least $1-\beta$ the following holds*

$$F(\tilde{x}^N) - F(x^*) \leq \frac{3\delta D}{2} + \frac{3(M^2+\delta^2)}{N\gamma}(1+\log N) + \max\left\{\frac{18M^2}{\gamma}, 6B + \frac{2M^2}{\gamma}\right\}\frac{\log(4\log N/\beta)}{N}.$$

where $D = \max_{x',x''\in X}\|x' - x''\|_2$ and δ is defined by (2.9).

Proof. The proof mainly relies on Theorem 2.1.1 and estimating the regret for iterative formula (2.10) with $\eta_k = \frac{1}{\gamma k}$.

From γ-strongly convexity in x of $f(x,\xi)$, 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 (2.9)

$$\begin{aligned} f(x^*,\xi^k) &\geq f(x^k,\xi^k) + \langle g_\delta(x^k,\xi^k), x^* - x^k\rangle + \frac{\gamma}{2}\|x^* - x^k\|_2 \\ &\quad + \langle \nabla f(x^k,\xi^k) - g_\delta(x^k,\xi^k), x^* - x^k\rangle \\ &\geq f(x^k,\xi^k) + \langle g_\delta(x^k,\xi^k), x^* - x^k\rangle + \frac{\gamma}{2}\|x^* - x^k\|_2 + \delta\|x^* - x^k\|_2. \end{aligned} \quad (2.11)$$

From the update rule (2.10) 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 (2.11) we get

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

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

$$\sum_{k=1}^{N} f(x^k, \xi^k) - f(x^*, \xi^k) \leq \frac{1}{2} \sum_{k=1}^{N} \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} + \gamma + \delta \right) \|x^* - x^k\|_2 + \frac{1}{2} \sum_{k=1}^{N} \eta_k \|g_\delta(x^k, \xi^k)\|_2^2 \leq \frac{\delta}{2} \sum_{k=1}^{N} \|x^* - x^k\|_2 + \frac{1}{2} \sum_{k=1}^{N} \eta_k \|g_\delta(x^k, \xi^k)\|_2^2. \tag{2.12}$$

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 (2.12) we bound the regret as follows

$$Reg_N \triangleq \sum_{k=1}^{N} f(x^k, \xi^k) - f(x^*, \xi^k) \leq \frac{\delta}{2} \sum_{k=1}^{N} \|p^* - p^k\|_2 + (M^2 + \delta^2) \sum_{k=1}^{N} \frac{1}{\gamma k} \leq \frac{1}{2} \delta D N + \frac{M^2 + \delta^2}{\gamma} (1 + \log N). \tag{2.13}$$

Here the last bound takes place due to the sum of harmonic series. Then for (2.13) we can use Theorem 2.1.1. 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 (2.13) in this inequality and making change $\beta = 4\beta \log N$ and get with probability at least $1 - \beta$

$$F(\tilde{x}^N) - F(x^*) \leq \frac{3\delta D}{2} + \frac{3(M^2 + \delta^2)}{N\gamma} (1 + \log N) + \max\left\{ \frac{18M^2}{\gamma}, 6B + \frac{2M^2}{\gamma} \right\} \frac{\log(4 \log N/\beta)}{N}.$$

□

2.1.2 Preliminaries on the SAA Approach

The SAA approach replaces the objective in (2.8) with its sample average

$$\min_{x\in X} \hat{F}(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f(x,\xi_i), \tag{2.14}$$

where each $f(x,\xi_i)$ is γ-strongly convex in x. Let us define the empirical minimizer of (2.14) $\hat{x}^* = \arg\min_{x\in X} \hat{F}(x)$, and $\hat{x}_{\varepsilon'}$ such that

$$\hat{F}(\hat{x}_{\varepsilon'}) - \hat{F}(\hat{x}^*) \le \varepsilon'. \tag{2.15}$$

The next theorem gives a bound on the excess risk for problem (2.14) in the SAA approach.

Theorem 2.1.3. *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 (2.15) with precision ε'. Then, with probability at least $1-\beta$ we have*

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

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

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

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

Theorem 2.1.4. *(Shalev-Shwartz et al., 2009, 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 $1-\beta$ the following holds*

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

where m is the sample size.

Proof of Theorem 2.1.3. For any $x \in X$, the following holds

$$F(x) - F(x^*) = F(x) - F(\hat{x}^*) + F(\hat{x}^*) - F(x^*). \tag{2.16}$$

From Theorem 2.1.4 with probability at least $1-\beta$ the following holds

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

Then from this and (2.16) we have with probability at least $1-\beta$

$$F(x) - F(x^*) \leq F(x) - F(\hat{x}^*) + \frac{4M^2}{\beta\gamma m}. \tag{2.17}$$

From Lipschitz continuity of $f(x,\xi)$ it follows, that for any $x \in X, \xi \in \Xi$ 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

$$|F(x) - F(\hat{x}^*)| \leq M\|x - \hat{x}^*\|_2. \tag{2.18}$$

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 $x \in X, \xi \in \Xi$

$$\|x - \hat{x}^*\|_2 \leq \sqrt{\frac{2}{\gamma}(\hat{F}(x) - \hat{F}(\hat{x}^*))}. \tag{2.19}$$

By using (2.18) and (2.19) and taking $x = \hat{x}_{\varepsilon'}$ in (2.17), we get the first statement of the theorem

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{2M^2}{\gamma}(\hat{F}(\hat{x}_{\varepsilon'}) - \hat{F}(\hat{x}^*))} + \frac{4M^2}{\beta\gamma m} \leq \sqrt{\frac{2M^2}{\gamma}\varepsilon'} + \frac{4M^2}{\beta\gamma m}. \tag{2.20}$$

Then from the strong convexity we have

$$\|\hat{x}_{\varepsilon'} - x^*\|_2 \leq \sqrt{\frac{2}{\gamma}\left(\sqrt{\frac{2M^2}{\gamma}\varepsilon'} + \frac{4M^2}{\beta\gamma m}\right)}. \tag{2.21}$$

Equating (2.20) to ε, we get the expressions for the sample size m and auxiliary precision ε'. Substituting both of these expressions in (2.21) we finish the proof. □

2.2 Non-Strongly Convex Optimization Problem

Now we consider non-strongly convex optimization problem

$$\min_{x \in X \subseteq \mathbb{R}^n} F(x) \triangleq \mathbb{E}f(x,\xi), \tag{2.22}$$

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

2.2.1 The SA Approach: Stochastic Mirror Descent

We consider stochastic mirror descent (MD) with inexact oracle (Nemirovski et al., 2009; Juditsky and Nemirovski, 2012; Gasnikov et al., 2016).[1] For a prox-function $d(x)$ and the corresponding Bregman divergence $B_d(x, x^1)$, the proximal mirror descent step is

$$x^{k+1} = \arg\min_{x \in X} \left(\eta \left\langle g_\delta(x^k, \xi^k), x \right\rangle + B_d(x, x^k) \right). \tag{2.23}$$

We consider the simplex setup: prox-function $d(x) = \langle x, \log x \rangle$. Here and below, functions such as log 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 the starting point is taken as $x^1 = \arg\min_{x \in \Delta_n} d(x) = (1/n, ..., 1/n)$.

Theorem 2.2.1. *Let* $R^2 \triangleq \mathrm{KL}(x^*, x^1) \leq \log n$ *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* $\breve{x}^N \triangleq \frac{1}{N} \sum_{k=1}^N x^k$ *be the average of outputs generated by iterative formula* (2.23) *with* $\eta = \frac{\sqrt{2}R}{M_\infty \sqrt{N}}$. *Then, with probability at least* $1 - \beta$ *we have*

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

Proof. For MD with prox-function function $d(x) = \langle x \log x \rangle$ the following holds for any $x \in \Delta_n$ (Juditsky and Nemirovski, 2012, Eq. 5.13)

$$\begin{aligned} \eta \langle g_\delta(x^k, \xi^k), x^k - x \rangle &\leq \mathrm{KL}(x, x^k) - \mathrm{KL}(x, x^{k+1}) + \frac{\eta^2}{2} \|g_\delta(x^k, \xi^k)\|_\infty^2 \\ &\leq \mathrm{KL}(x, x^k) - \mathrm{KL}(x, x^{k+1}) + \eta^2 M_\infty^2. \end{aligned}$$

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

$$\begin{aligned} \eta \langle \nabla F(x^k), x^k - x \rangle &\leq \eta \langle \nabla f(x^k, \xi^k) - g_\delta(x^k, \xi^k), x^k - x \rangle \\ &\quad + \eta \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x \rangle \\ &\quad + \mathrm{KL}(x, x^k) - \mathrm{KL}(x, x^{k+1}) + \eta^2 M_\infty^2 \\ &\leq \eta \delta \max_{k=1,...,N} \|x^k - x\|_1 + \eta \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x \rangle \\ &\quad + \mathrm{KL}(x, x^k) - \mathrm{KL}(x, x^{k+1}) + \eta^2 M_\infty^2. \end{aligned} \tag{2.24}$$

[1] By using dual averaging scheme (Nesterov, 2009) we can rewrite Alg. 4 in online regime (Hazan et al., 2016; Orabona, 2019) without including N in the stepsize policy. Note, that mirror descent and dual averaging scheme are very close to each other (Juditsky et al., 2019).

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 (2.24) and sum for $k = 1, ..., N$ at $x = x^*$

$$\begin{aligned}
&\eta \sum_{k=1}^{N} F(x^k) - F(x^*) \\
&\quad \leq \eta \delta N \max_{k=1,...,N} \|x^k - x^*\|_1 + \eta \sum_{k=1}^{N} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle \\
&\quad + \mathrm{KL}(x^*, x^1) - \mathrm{KL}(x^*, x^{N+1}) + \eta^2 M_\infty^2 N \\
&\quad \leq \eta \delta N D + \eta \sum_{k=1}^{N} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle + R^2 + \eta^2 M_\infty^2 N. \qquad (2.25)
\end{aligned}$$

Where we used $\mathrm{KL}(x^*, x^1) \leq R^2$ and $\max\limits_{k=1,...,N} \|p^k - p^*\|_1 \leq D$. Then using convexity of $F(x^k)$ and the definition of output $\breve{x}^N$ in (2.25) we have

$$F(\breve{x}^N) - F(x^*) \leq \delta D + \frac{1}{N} \sum_{k=1}^{N} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle + \frac{R^2}{\eta N} + \eta M_\infty^2. \qquad (2.26)$$

Next we use the Azuma–Hoeffding's (Juditsky et al., 2008) inequality and get for all $\beta \geq 0$

$$\mathbb{P}\left(\sum_{k=1}^{N+1} \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^k - x^* \rangle \leq \beta \right) \geq 1 - \exp\left(-\frac{2\beta^2}{N(2M_\infty D)^2} \right) = 1 - \beta. \qquad (2.27)$$

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

$$\begin{aligned}
\left| \langle \nabla F(x^k) - \nabla f(x^k, \xi^k), x^* - x^k \rangle \right| &\leq \|\nabla F(x^k) - \nabla W(p^k, q^k)\|_\infty \|x^* - x^k\|_1 \\
&\leq 2M_\infty \max_{k=1,...,N} \|x^k - x^*\|_1 \leq 2M_\infty D.
\end{aligned}$$

Thus, using (2.27) for (2.26) we have that with probability at least $1 - \beta$

$$F(\breve{x}^N) - F(x^*) \leq \delta D + \frac{\beta}{N} + \frac{R^2}{\eta N} + \eta M_\infty^2. \qquad (2.28)$$

Then, expressing β through β and substituting $\eta = \frac{R}{M_\infty}\sqrt{\frac{2}{N}}$ to (2.28) (such η minimize the r.h.s. of (2.28)), we get

$$\begin{aligned}
F(\breve{x}^N) - F(x^*) &\leq \delta D + \frac{M_\infty D \sqrt{2\log(1/\beta)}}{\sqrt{N}} + \frac{M_\infty R}{\sqrt{2N}} + \frac{M_\infty R \sqrt{2}}{\sqrt{N}} \\
&\leq \delta D + \frac{M_\infty (3R + 2D\sqrt{\log(1/\beta)})}{\sqrt{2N}}.
\end{aligned}$$

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

$$F(\breve{x}^N) - F(x^*) \leq \frac{M_\infty(3\sqrt{\log n} + 4\sqrt{\log(1/\beta)})}{\sqrt{2N}} + 2\delta. \tag{2.29}$$

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

$$\begin{aligned}\sqrt{\left(3\sqrt{\log n} + 4\sqrt{\log(1/\beta)}\right)^2} &= \sqrt{9\log n + 16\log(1/\beta) + 24\sqrt{\log n}\sqrt{\log(1/\beta)}} \\ &\leq \sqrt{18\log n + 32\log(1/\beta)}.\end{aligned}$$

Using this for (2.29), we get the statement of the theorem

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

□

2.2.2 Penalization in the SAA Approach

In this section, we study the SAA approach for non-strongly convex problem (2.22). We regularize this problem by 1-strongly convex w.r.t. x penalty function $r(x, x^1)$ in the ℓ_2-norm

$$\min_{x \in X \subseteq \mathbb{R}^n} F_\lambda(x) \triangleq \mathbb{E}f(x, \xi) + \lambda r(x, x^1) \tag{2.30}$$

and we prove that the sample sizes in the SA and the SAA approaches will be equal up to logarithmic terms. The empirical counterpart of problem (2.30) is

$$\min_{x \in X} \hat{F}_\lambda(x) \triangleq \frac{1}{m}\sum_{i=1}^m f(x, \xi_i) + \lambda r(x, x^1). \tag{2.31}$$

Let us define $\hat{x}_\lambda = \arg\min_{x \in X} \hat{F}_\lambda(x)$. The next lemma proves the statement from (Shalev-Shwartz et al., 2009) on boundness of the population sub-optimality in terms of the square root of empirical sub-optimality.

Lemma 2.2.2. *Let $f(x, \xi)$ be convex and M-Lipschitz continuous w.r.t. x in the ℓ_2-norm. Let $r(x, x^1)$ be 1-strongly convex and M_r-Lipschitz continuous w.r.t. x in the ℓ_2-norm. Then for any $x \in X$ with probability at least $1 - \beta$ the following holds*

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

where $x^*_\lambda = \arg\min_{x \in X} F_\lambda(x)$, $M_\lambda \triangleq M + \lambda M_r$.

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 $M_\lambda \triangleq M + \lambda M_r$. From 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

$$|F_\lambda(x) - F_\lambda(\hat{x}_\lambda)| \leq M_\lambda \|x - \hat{x}_\lambda\|_2, \qquad \forall x \in X. \tag{2.32}$$

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

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

From this and (2.32) it follows

$$F_\lambda(x) - F_\lambda(\hat{x}_\lambda) \leq \sqrt{\frac{2M_\lambda^2}{\lambda}\left(\hat{F}_\lambda(x) - \hat{F}_\lambda(\hat{x}_\lambda)\right)}. \tag{2.33}$$

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

$$F_\lambda(x) - F_\lambda(x_\lambda^*) = F_\lambda(x) - F_\lambda(\hat{x}_\lambda) + F_\lambda(\hat{x}_\lambda) - F_\lambda(x_\lambda^*). \tag{2.34}$$

From (Shalev-Shwartz et al., 2009, Theorem 6) we have with probability at least $1-\beta$

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

Using this and (2.33) for (2.34) we obtain with probability at least $1-\beta$

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

□

The next theorem proves the eliminating the linear dependence on n in the sample size of the regularized SAA approach for a non-strongly convex objective (see estimate (2.3)), and estimates the auxiliary precision for the regularized SAA problem (2.5).

Theorem 2.2.3. *Let $f(x,\xi)$ be convex and M-Lipschitz continuous w.r.t. x in the ℓ_2-norm and let $r(x,x^1)$ be 1-strongly convex and M_r-Lipschitz continuous w.r.t. x in the ℓ_2-norm. Let $\hat{x}_{\varepsilon'}$ be such that*

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

To satisfy

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

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

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

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 2.2.2 we get for $x = \hat{x}_{\varepsilon'}$,

$$F_\lambda(\hat{x}_{\varepsilon'}) - F_\lambda(x^*_\lambda) \leq \sqrt{\frac{2M_\lambda^2}{\lambda}\left(\hat{F}_\lambda(\hat{x}_{\varepsilon'}) - \hat{F}_\lambda(\hat{x}_\lambda)\right)} + \frac{4M_\lambda^2}{\beta\lambda m} = \sqrt{\frac{2M_\lambda^2}{\lambda}\varepsilon'} + \frac{4M_\lambda^2}{\beta\lambda m}, \tag{2.35}$$

where we used the definition of $\hat{x}_{\varepsilon'}$ from the statement of the this theorem and $M_\lambda \triangleq M + \lambda M_r$. Then we subtract $F(x^*)$ in both sides of (2.35) and get

$$F_\lambda(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{2M_\lambda^2\varepsilon'}{\lambda}} + \frac{4M_\lambda^2}{\beta\lambda m} + F_\lambda(x^*_\lambda) - F(x^*). \tag{2.36}$$

Then we use

$$\begin{aligned} F_\lambda(x^*_\lambda) &\triangleq \min_{x\in X}\left\{F(x) + \lambda r(x, x^1)\right\} \\ &\leq F(x^*) + \lambda r(x^*, x^1) \qquad \text{The inequality holds for any } x \in X, \\ &\triangleq F(x^*) + \lambda\mathcal{R}^2 \end{aligned}$$

where $\mathcal{R}^2 \triangleq r(x^*, x^1)$. Then from this and (2.36) and the definition of $F_\lambda(\hat{x}_{\varepsilon'})$ in (2.30) we get

$$\begin{aligned} F(\hat{x}_{\varepsilon'}) - F(x^*) &\leq \sqrt{\frac{2M_\lambda^2}{\lambda}\varepsilon'} + \frac{4M_\lambda^2}{\beta\lambda m} - \lambda r(\hat{x}_{\varepsilon'}, x^1) + \lambda\mathcal{R}^2 \\ &\leq \sqrt{\frac{2M_\lambda^2\varepsilon'}{\lambda}} + \frac{4M_\lambda^2}{\beta\lambda m} + \lambda\mathcal{R}^2. \end{aligned} \tag{2.37}$$

Let us remind that $M_\lambda \triangleq M + \lambda M_r$. Then assuming $M \gg \lambda M_r$ and choosing $\lambda = \varepsilon/(2\mathcal{R}^2)$ in (2.37), we get the following

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{4M^2\mathcal{R}^2\varepsilon'}{\varepsilon}} + \frac{8M^2\mathcal{R}^2}{\beta m\varepsilon} + \varepsilon/2. \tag{2.38}$$

Equating the first and the second terms in the r.h.s. of (2.38) to $\varepsilon/4$ respectively, we obtain the the rest statements of the theorem including $F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \varepsilon$. □

2.3 Fréchet Mean with respect to Entropy-Regularized Optimal Transport

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

$$\min_{p \in \Delta_n} W_\gamma(p) \triangleq \mathbb{E}_q W_\gamma(p, q), \qquad q \sim \mathbb{P}. \tag{2.39}$$

2.3.1 Properties of Entropy-Regularized Optimal Transport

Entropic regularization of transport distances (Cuturi, 2013) improves their statistical properties (Klatt et al., 2020; Bigot et al., 2019a) and reduces their computational complexity. Entropic regularization has shown good results in generative models (Genevay et al., 2017), multi-label learning (Frogner et al., 2015), dictionary learning (Rolet et al., 2016), image processing (Cuturi and Peyré, 2016; Rabin and Papadakis, 2015), neural imaging (Gramfort et al., 2015).

Let us firstly remind optimal transport problem (introduced in Eq. (1.1)) between histograms $p, q \in \Delta_n$ with cost matrix $C \in \mathbb{R}_+^{n \times n}$

$$W(p, q) \triangleq \min_{\pi \in U(p,q)} \langle C, \pi \rangle, \tag{2.40}$$

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 $\rho \geq 1$, $C_{ij} = \mathtt{d}(x_i, x_j)^\rho$ in (2.40), where $\mathtt{d}(x_i, x_j)$ 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 (2.40) is a more general than the Wasserstein distances.

Following (Cuturi, 2013), we introduce entropy-regularized optimal transport problem

$$W_\gamma(p, q) \triangleq \min_{\pi \in U(p,q)} \{\langle C, \pi \rangle - \gamma E(\pi)\}, \tag{2.41}$$

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^2} in the ℓ_1-norm, the objective in (2.41) is γ-strongly convex with respect to π in the ℓ_1-norm on Δ_{n^2}, and hence problem (2.41) 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 (Bigot et al., 2019b, Theorem 3.4).

One particular advantage of the entropy-regularized optimal transport is a closed-form representation for its dual function (Agueh and Carlier, 2011; Cuturi and

Peyré, 2016) defined by the Fenchel–Legendre transform of $W_\gamma(p,q)$ as a function of p

$$\begin{aligned} W^*_{\gamma,q}(u) &= \max_{p\in\Delta_n}\{\langle u,p\rangle - W_\gamma(p,q)\} = \gamma\left(E(q) + \langle q, \log(K\beta)\rangle\right) \\ &= \gamma\left(-\langle q,\log q\rangle + \sum_{j=1}^n [q]_j \log\left(\sum_{i=1}^n \exp\left(([u]_i - C_{ji})/\gamma\right)\right)\right) \end{aligned} \tag{2.42}$$

where $\beta = \exp(u/\gamma)$, $K = \exp(-C/\gamma)$ and $[q]_j$ is j-th component of vector q. Functions such as log 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 (Cuturi and Peyré, 2016)

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

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

$$\forall l = 1,\ldots,n \qquad [\nabla W^*_{\gamma,q}(u)]_l = \sum_{j=1}^n [q]_j \frac{\exp\left(([u]_l - C_{lj})/\gamma\right)}{\sum_{i=1}^n \exp\left(([u]_i - C_{ji})/\gamma\right)}. \tag{2.43}$$

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

$$\begin{aligned} W_\gamma(p,q) &= \min_{\pi\in U(p,q)} \sum_{i,j=1}^n \left(C_{ij}\pi_{i,j} + \gamma\pi_{i,j}\log\pi_{i,j}\right) \\ &= \max_{u,\nu\in\mathbb{R}^n}\left\{\langle u,p\rangle + \langle \nu,q\rangle - \gamma\sum_{i,j=1}^n \exp\left(([u]_i + [\nu]_j - C_{ij})/\gamma - 1\right)\right\} \\ &= \max_{u\in\mathbb{R}^n}\left\{\langle u,p\rangle - \gamma\sum_{j=1}^n [q]_j \log\left(\frac{1}{[q]_j}\sum_{i=1}^n \exp\left(([u]_i - C_{ij})/\gamma\right)\right)\right\}. \end{aligned} \tag{2.44}$$

Any solution $\begin{pmatrix} u^* \\ \nu^* \end{pmatrix}$ of (2.44) is a subgradient of $W_\gamma(p,q)$ (Peyré et al., 2019, Proposition 4.6)

$$\nabla W_\gamma(p,q) = \begin{pmatrix} u^* \\ \nu^* \end{pmatrix}. \tag{2.45}$$

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

The next theorem (Bigot et al., 2019b) 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 2.3.1. *(Bigot et al., 2019b, Theorem 3.4, Lemma 3.5)*

- *For any* $q \in \Delta_n$, $W_\gamma(p, q)$ *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

$$\begin{aligned} M &\overset{\text{(Bigot et al., 2019b)}}{=} O\left(\sqrt{\sum_{j=1}^{n} \left(\inf_{i \in [n]} \sup_{l \in [n]} |C_{jl} - C_{il}| \right)^2} \right) \\ &= O\left(\sqrt{\sum_{j=1}^{n} \sup_{l \in [n]} C_{jl}^2} \right) = O\left(\sqrt{n} \sup_{j,l \in [n]} C_{jl} \right) \\ &= O\left(\sqrt{n} \sup_{j \in [n]} \sum_{l \in [n]} C_{jl} \right) = O\left(\sqrt{n} \|C\|_\infty \right). \end{aligned}$$

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 Lipshitz 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.

2.3.2 The SA Approach: Stochastic Gradient Descent

For problem (2.39), as a particular case of problem (2.1), stochastic gradient descent method can be used. From Eq. (2.45), it follows that an approximation for the gradient of $W_\gamma(p, q)$ with respect to p can be calculated by Sinkhorn algorithm (Altschuler et al., 2017; Peyré et al., 2019; Dvurechensky et al., 2018b) through the computing dual variable u with δ-precision

$$\|\nabla_p W_\gamma(p, q) - \nabla_p^\delta W_\gamma(p, q)\|_2 \leq \delta, \quad \forall q \in \Delta_n. \tag{2.46}$$

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 (2.10) for $\eta_k = \frac{1}{\gamma k}$ with Sinkhorn algorithm (Algorithm 1) and Algorithm 2 making the projection onto the simplex Δ_n.

For Algorithm 3 and problem (2.39), Theorem 2.1.2 can be specified as follows

Algorithm 1 Sinkhorn's algorithm (Peyré et al., 2019) for calculating $\nabla_p^\delta W_\gamma(p^k, q^k)$

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

Algorithm 2 Euclidean Projection $\Pi_{\Delta_n}(p) = \arg\min_{v \in \Delta_n} \|p - v\|_2$ onto Simplex Δ_n (Duchi et al., 2008)

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

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

Input: starting point $p^1 \in \Delta_n$, realization q^1, δ, γ.
for $k = 1, 2, 3, \ldots$ **do**
  $\eta_k = \frac{1}{\gamma k}$
  $\nabla_p^\delta W_\gamma(p^k, q^k) \leftarrow$ SINKHORN(p^k, q^k, C, γ) or the accelerated Sinkhorn (Guminov et al., 2019)
  $p^{(k+1)/2} \leftarrow p^k - \eta_k \nabla_p^\delta W_\gamma(p^k, q^k)$
  $p^{k+1} \leftarrow$ PROJECTION$(p^{(k+1)/2})$
  Sample q^{k+1}
end for
Output: $p^1, p^2, p^3 ...$

Theorem 2.3.2. *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 $1 - \beta$ the following holds*

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

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 $1-\beta$

$$W_\gamma(\tilde{p}^N) - W_\gamma(p^*_\gamma) \le \varepsilon \quad \textit{and} \quad \|\tilde{p}^N - p^*_\gamma\|_2 \le \sqrt{2\varepsilon/\gamma}.$$

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 $\kappa \triangleq \lambda^+_{\min}\left(\nabla^2 W^*_{\gamma,q}(u^*)\right)$.

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

$$\begin{aligned}\max_{p,q\in\Delta_n} W_\gamma(p,q) &= \max_{p,q\in\Delta_n} \min_{\substack{\pi\in\mathbb{R}^{n\times n}_+,\\ \pi\mathbf{1}=p,\\ \pi^T\mathbf{1}=q}} \sum_{i,j=1}^n (C_{ij}\pi_{ij} + \gamma\pi_{ij}\log\pi_{ij}) \\ &\le \max_{\substack{\pi\in\mathbb{R}^{n\times n}_+,\\ \sum_{i,j=1}^n \pi_{ij}=1}} \sum_{i,j=1}^n (C_{ij}\pi_{ij} + \gamma\pi_{ij}\log\pi_{ij}) \le \|C\|_\infty.\end{aligned}$$

Therefore, $W_\gamma(p,q) : \Delta_n \times \Delta_n \to [-2\gamma\log n, \|C\|_\infty]$. Then we apply Theorem 2.1.2 with $B = \|C\|_\infty$ and $D = \max\limits_{p',p''\in\Delta_n} \|p'-p''\|_2 = \sqrt{2}$, and we sharply get

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

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 $\|\tilde{p}^N - p^*_\gamma\|_2 \le \sqrt{2\varepsilon/\gamma}$ 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 firstly define $\tilde{\delta}$ as the accuracy in function value of the inexact solution u of maximization problem in (2.44). Using this we formulate the number of iteration of the Sinkhorn's (Franklin and Lorenz, 1989; Carlier, 2021; Kroshnin et al., 2019; Stonyakin et al., 2019)

$$\widetilde{O}\left(\exp\left(\frac{\|C\|_\infty}{\gamma}\right)\left(\frac{\|C\|_\infty}{\gamma}+\log\left(\frac{\|C\|_\infty}{\tilde{\delta}}\right)\right)\right). \tag{2.47}$$

The number of iteration for the accelerated Sinkhorn's can be improved (Guminov et al., 2019)

$$\widetilde{O}\left(\sqrt{\frac{n\|C\|_\infty^2}{\gamma\varepsilon'}}\right). \tag{2.48}$$

Here ε' is the accuracy in the function value, which is the expression $\langle u,p\rangle + \langle\nu,q\rangle - \gamma\sum_{i,j=1}^n \exp\left((-C_{ji}+u_i+\nu_j)/\gamma - 1\right)$ under the maximum in (2.44). 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

$$\varepsilon' \geq \frac{\gamma}{2}\|u - u^*\|_2^2 = \frac{\kappa}{2}\delta, \tag{2.49}$$

where $\kappa \triangleq \lambda_{\min}^{+}\left(\nabla^2 W_{\gamma,q}^*(u^*)\right)$. From (Bigot et al., 2019b, Proposition A.2.), for the eigenvalue of $\nabla^2 W_{\gamma,q}^*(u^*)$ it holds that $0 = \lambda_n\left(\nabla^2 W_{\gamma,q}^*(u^*)\right) < \lambda_k\left(\nabla^2 W_{\gamma,q}^*(u^*)\right)$ for all $k = 1, ..., n-1$. Inequality (2.49) holds due to $\nabla_p^0 W_\gamma(p,q) := u$ in Algorithm 3 and $\nabla_p W_\gamma(p,q) \triangleq u^*$ in (2.45). Multiplying both of estimates (2.47) and (2.48) by the complexity of each iteration of the (accelerated) Sinkhorn's algorithm $O(n^2)$ and the number of iterations $N = \widetilde{O}\left(\frac{M^2}{\gamma\varepsilon}\right)$ (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 (Delon and Desolneux, 2020), we study the convergence to the true barycenter of the generated truncated Gaussian measures. Figure 2.1 illustrates the convergence in the 2-Wasserstein distance within 40 seconds.

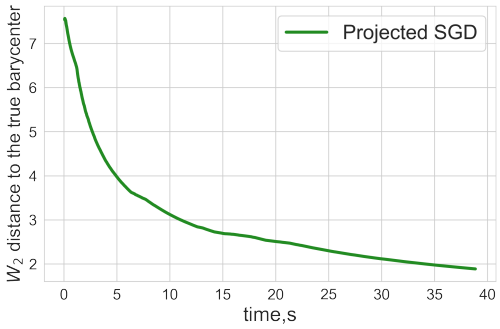


Figure 2.1: Convergence of projected stochastic gradient descent to the true barycenter of 2×10^4 Gaussian measures in the 2-Wasserstein distance.

2.3.3 The SAA Approach

The empirical counterpart of problem (2.39) is the (empirical) Wasserstein barycenter problem

$$\min_{p \in \Delta_n} \frac{1}{m}\sum_{i=1}^{m} W_\gamma(p, q_i), \tag{2.50}$$

where $q_1, q_2, ..., q_m$ are some realizations of random variable with distribution $\mathbb{P}$.

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

$$\frac{1}{m}\sum_{i=1}^{m} W_\gamma(\hat{p}_{\varepsilon'}, q_i) - \frac{1}{m}\sum_{i=1}^{m} W_\gamma(\hat{p}_\gamma^m, q_i) \leq \varepsilon'. \tag{2.51}$$

For instance, $\hat{p}_{\varepsilon'}$ can be calculated by the IBP algorithm (Benamou et al., 2015) or the accelerated IBP algorithm (Guminov et al., 2019). The next theorem specifies Theorem 2.1.3 for the Wassertein barycenter problem (2.50).

Theorem 2.3.3. *Let $\hat{p}_{\varepsilon'}$ satisfies (2.51). Then, with probability at least $1-\beta$*

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

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

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

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

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

Proof. From Theorem 2.1.3 we get the first statement of the theorem

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

From (Guminov et al., 2019) 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 2.1.3

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

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 2.1 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 seconds. For the convergence to the true barycenter w.r.t. the 2-Wasserstein distance in the SAA approach, we refer to (Boissard et al., 2015), however, considering the convergence in the ℓ_2-norm (Theorem 2.3.3) allows to obtain better convergence rate in comparison with the bounds for the 2-Wasserstein distance.

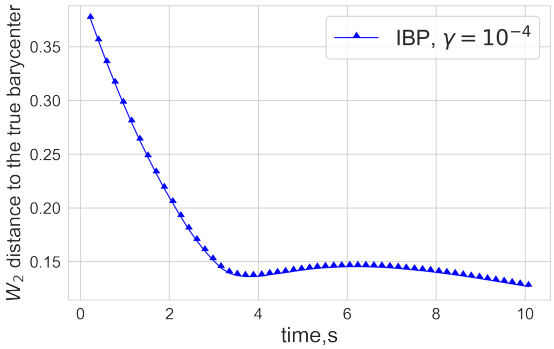


Figure 2.2: Convergence of the Iterative Bregman Projections to the true barycenter of 2×10^4 Gaussian measures in the 2-Wasserstein distance.

2.3.4 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 (2.39). For the brevity, we skip the high probability details since we can fixed β (say $\beta = 0.05$) in the all bounds. Moreover, based on (Shalev-Shwartz et al., 2009), we assume that in fact all bounds of this paper have logarithmic dependence on β which is hidden in $\widetilde{O}(\cdot)$ (Feldman and Vondrák, 2019; Klochkov and Zhivotovskiy, 2021).

Table 2.1: Total complexity of the SA and the SAA implementations for the problem $\min_{p \in \Delta_n} \mathbb{E}_q W_\gamma(p, q)$.

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

Table 2.1 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 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 $\kappa \triangleq \lambda_{\min}^+\left(\nabla^2 W_{\gamma,q}^*(u^*)\right)$. 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.

2.4 Fréchet Mean with respect to Optimal Transport

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

$$\min_{p \in \Delta_n} W(p) \triangleq \mathbb{E}_q W(p, q). \tag{2.52}$$

2.4.1 The SA Approach with Regularization: Stochastic Gradient Descent

The next theorem explains how the solution of strongly convex problem (2.39) approximates a solution of convex problem (2.52) under the proper choice of the regularization parameter γ.

Theorem 2.4.1. *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 $N = \widetilde{O}\left(\frac{n\|C\|_\infty^2}{\gamma\varepsilon}\right)$. Let $\gamma = \varepsilon/(2\mathcal{R}^2)$ with $\mathcal{R}^2 = 2\log n$. Then, with probability at least $1-\beta$ the following holds*

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

where p^* is a solution of (2.52).

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 $\kappa \triangleq \lambda_{\min}^+\left(\nabla^2 W^*_{\gamma,q}(u^*)\right)$.

Proof. The proof of this theorem follows from Theorem 2.3.2 and the following (Gasnikov et al., 2015; Kroshnin et al., 2019; Peyré et al., 2019)

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

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

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

This means that solving problem (2.39) with $\varepsilon/2$ precision, we get a solution of problem (2.52) with ε precision.

When γ is not too large, Algorithm 3 uses the accelerated Sinkhorn's algorithm (instead of Sinkhorn's 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.

□

2.4.2 The SA Approach: Stochastic Mirror Descent

Now we propose an approach to solve problem (2.52) without additional regularization. The approach is based on mirror descent given by the iterative formula (2.23). We use simplex setup which provides a closed form solution for (2.23). Algorithm 4 presents the application of mirror descent to problem (2.52), where the gradient of $W(p^k, q^k)$ can be calculated using dual representation of OT (Peyré et al., 2019) by any LP solver exactly

$$W(p,q) = \max_{\substack{(u,\nu)\in\mathbb{R}^n\times\mathbb{R}^n,\\ u_i+\nu_j\le C_{ij},\forall i,j\in[n]}} \{\langle u,p\rangle + \langle \nu,q\rangle\}. \tag{2.53}$$

Then

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

where u^* is a solution of (2.53) such that $\langle u^*, \mathbf{1}\rangle = 0$.

Algorithm 4 Stochastic Mirror Descent for the Wasserstein Barycenter Problem

Input: 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** $k = 1, \dots, N$ **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**

Output: $\breve{p}^N = \frac{1}{N}\sum_{k=1}^N p^k$

The next theorem estimates the complexity of Algorithm 4

Theorem 2.4.2. *Let $\breve{p}^N$ be the output of Algorithm 4 processing N measures. Then, with probability at least $1-\beta$ we have*

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

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 $1-\beta$

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

The total complexity of Algorithm 4 is

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

Proof. From Theorem 2.2.1 and using $M_\infty = O\left(\|C\|_\infty\right)$, we have

$$W(\check{p}^N) - W(p^*) = O\left(\frac{\|C\|_\infty \sqrt{\log(n/\beta)}}{\sqrt{N}} + 2\delta\right). \tag{2.54}$$

Notice, that $\nabla_{p^k} W(p^k, q^k)$ can be calculated exactly by any LP solver. Thus, we take $\delta = 0$ in (2.54) and get the first statement of the theorem.

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

To get the complexity bounds we notice that the complexity for calculating $\nabla_p W(p^k, q^k)$ is $\tilde{O}(n^3)$ (Ahuja et al., 1993; Dadush and Huiberts, 2018; Dong et al., 2020; Gabow and Tarjan, 1991), multiplying this by $N = O\left(\|C\|_\infty^2 R^2/\varepsilon^2\right)$ with $R^2 \triangleq \mathrm{KL}(p^*, p^1) \leq \log n$, 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 (2.52). 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 2.3 confirms better convergence of stochastic mirror descent than projected stochastic gradient descent as stated in their theoretical complexity (Theorems 2.4.1 and 2.4.2).

2.4.3 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 (2.39) approximates a solution of convex problem (2.52).

Theorem 2.4.3. *Let $\hat{p}_{\varepsilon'}$ satisfy*

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

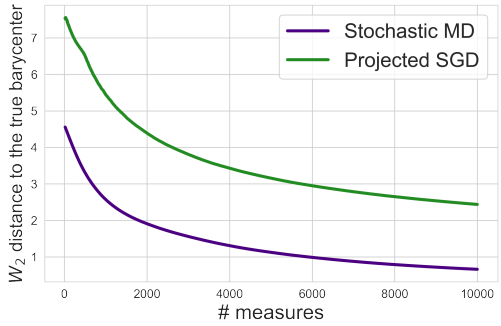


Figure 2.3: Convergence of projected stochastic gradient descent, and stochastic mirror descent to the true barycenter of 2×10^4 Gaussian measures in the 2-Wasserstein distance.

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

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

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

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

Proof. The proof follows from Theorem 2.3.3 and the proof of Theorem 2.4.1 with $\gamma = \varepsilon/(4\log n)$. □

2.4.4 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 (Ben-Tal and Nemirovski, 2001)

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

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 $\widetilde{O}(1)$-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 (Ballu et al., 2020; Bigot et al., 2019c), 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 (Bigot et al., 2019c) 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 $R^2 = \|p^* - p^1\|_2^2 \leq \|p^* - p^1\|_1^2 \leq 2$ and $R_d^2 = B_d(p^*, p^1) = O(\log n)$ are equal up to a logarithmic factor.

The regularized SAA problem is following

$$\min_{p \in \Delta_n} \left\{ \frac{1}{m} \sum_{k=1}^{m} W(p, q^k) + \lambda B_d(p, p^1) \right\}. \tag{2.55}$$

The next theorem is particular case of Theorem (2.2.3) for the population WB problem (2.52) with $r(p, p^1) = B_d(p, p^1)$.

Theorem 2.4.4. *Let $\hat{p}_{\varepsilon'}$ be such that*

$$\frac{1}{m} \sum_{k=1}^{m} W(\hat{p}_{\varepsilon'}, q^k) + \lambda B_d(\hat{p}_{\varepsilon'}, p^1) - \min_{p \in \Delta_n} \left\{ \frac{1}{m} \sum_{k=1}^{m} W(p, q^k) + \lambda B_d(p, p^1) \right\} \leq \varepsilon'. \tag{2.56}$$

To satisfy

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

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

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

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

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

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

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

Proof. The proof is based on saddle-point reformulation of the WB problem. Further, we provide the explanation how to do this (for more details see Chapter 4). Firstly we rewrite the OT as (Jambulapati et al., 2019)

$$W(p, q) = \min_{x \in \Delta_{n^2}} \max_{y \in [-1,1]^{2n}} \{ d^\top x + 2\|d\|_\infty (y^\top A x - b^\top y) \}, \tag{2.57}$$

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 (Dvinskikh and Tiapkin, 2021)

$$\min_{\substack{p \in \Delta^n, \\ \mathbf{x} \in \mathcal{X} \triangleq \underbrace{\Delta_{n^2} \times \ldots \times \Delta_{n^2}}_{m}}} \max_{\mathbf{y} \in [-1,1]^{2mn}} \frac{1}{m} \left\{ \boldsymbol{d}^\top \mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A} \mathbf{x} - \mathbf{b}^\top \mathbf{y} \right) \right\}, \tag{2.58}$$

where $\mathbf{x} = (x_1^\top, \dots, x_m^\top)^\top$, $\mathbf{y} = (y_1^\top, \dots, y_m^\top)^\top$, $\mathbf{b} = (p^\top, q_1^\top, ..., p^\top, q_m^\top)^\top$, $\boldsymbol{d} = (d^\top, \dots, d^\top)^\top$, and $\boldsymbol{A} = \text{diag}\{A, ..., A\} \in \{0,1\}^{2mn \times mn^2}$ is block-diagonal matrix. Similarly to (2.58) we reformulate (2.55) as a saddle-point problem

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

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

$$G(\mathbf{x}, p, \mathbf{y}) = \begin{pmatrix} \nabla_{\mathbf{x}} f \\ \nabla_p f \\ -\nabla_{\mathbf{y}} f \end{pmatrix} = \frac{1}{m} \begin{pmatrix} \boldsymbol{d} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{y} \\ -2\|d\|_\infty \{[y_i]_{1...n}\}_{i=1}^m + \lambda(\nabla d(p) - \nabla d(p^1)) \\ 2\|d\|_\infty (\boldsymbol{A}\mathbf{x} - \mathbf{b}) \end{pmatrix}, \tag{2.59}$$

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 (Dvinskikh and Tiapkin, 2021) with (2.59). The total complexity is

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

Then we use Theorem 2.2.3 and get the exspressions for m, ε' with $\lambda = \varepsilon/(2{R_d}^2)$, where ${R_d}^2 = B_d(p^*, p^1)$. The number of measures is

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

The precision ε' is defined as

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

□

2.4.5 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 (2.52). Table 2.2 presents the total complexity for the numerical algorithms.

Table 2.2: Total complexity of the SA and the SAA implementations for the problem $\min_{p \in \Delta_n} \mathbb{E}_q W(p, q)$.

Algorithm	**Theorem**	**Complexity**
Projected SGD (SA) with $\gamma = \frac{\varepsilon}{4 \log n}$	2.4.1	$\widetilde{O}\left(\frac{n^3 \sqrt{n} \|C\|_\infty^3}{\varepsilon^3 \sqrt{\varepsilon \kappa}}\right)$
Stochastic MD (SA)	2.4.2	$\widetilde{O}\left(\frac{n^3 \|C\|_\infty^2}{\varepsilon^2}\right)$
Accelerated IBP (SAA) with $\gamma = \frac{\varepsilon}{4 \log n}$	2.4.3	$\widetilde{O}\left(\frac{n^4 \|C\|_\infty^4}{\varepsilon^4}\right)$
Mirror Prox with $B_d(p^*, p^1)$ penalization (SAA)	2.4.4	$\widetilde{O}\left(\frac{n^2 \sqrt{n} \|C\|_\infty^5}{\varepsilon^5}\right)$

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 2.3.

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 (Dvinskikh and Tiapkin, 2021) instead of Mirror Prox allows to omit $\sqrt{n}$ in the penalized approach.

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 (Gorbunov et al., 2019). This is the case of the Wasserstein barycenter problem, indeed, the objective is Lipschitz continuous but not Lipschitz smooth.

Chapter 3

Dual Methods for Strongly Convex Optimization

In this Chapter, we firstly present a stochastic dual algorithm for an optimization problem with affine constraints whose objective is strongly convex. Then, for the objective given by the sum of strongly convex functions, we show how to perform this algorithm in a decentralized manner over a network of agents. This algorithm allows us to obtain optimal bounds on the number of communication rounds and oracle calls of dual objective per node. Finally, we show that the results can be naturally applied to the Wasserstein barycenter problem since the dual formulation of entropy-regularized Wasserstein distances and their derivatives have closed-form representations and can be computed for a cheaper price than the primal representations.

3.1 Dual Problem Formulation

We consider a convex optimization problem with affine constraint

$$\min_{Ax=b,\ x\in\mathbb{R}^n} F(x), \tag{3.1}$$

where $F(x)$ is γ_F-strongly convex and possibly presented by the expectation $F(x) = \mathbb{E}F(x,\xi)$ w.r.t. $\xi \in \Xi$.

The dual problem for problem (3.1), written as a maximization problem, is given by the following problem with the Lagrangian dual variable $y \in \mathbb{R}^n$

$$\min_{y\in\mathbb{R}^n} \Psi(y) \triangleq \max_{x\in\mathbb{R}^n} \left\{\langle y, Ax - b\rangle - F(x)\right\}. \tag{3.2}$$

By the Theorem 0.0.6, if $F(x)$ is γ_F-strongly convex, then function $\Psi(y)$ is L_Ψ-Lipschitz smooth with $L_\Psi = \lambda_{\max}(A^\top A)/\gamma_F$, where $\lambda_{\max}(B)$ is the maximum eigenvalue of symmetric matrix B.

When primal function $F(x) = \mathbb{E}F(x,\xi)$, the dual function is also presented by its expectation $\Psi(y) = \mathbb{E}\Psi(y,\xi)$ as well as its gradient. In this case we refer

to stochastic dual oracle. For a deterministic function $F(x)$, we can also refer to stochastic dual oracle when deterministic dual oracle, which returns the gradient of Ψ, is unavailable or very expensive.

3.1.1 Preliminaries on Stochastic Oracle

We make the following assumptions on the stochastic dual oracle which returns the gradient of the dual objective Ψ for all $y \in \mathbb{R}^n$

$$\begin{aligned} &\mathbb{E}\nabla\Psi(y,\xi) = \nabla\Psi(y) \\ &\mathbb{E}\exp\left(\|\nabla\Psi(y,\xi) - \nabla\Psi(y)\|_2^2/\sigma_\Psi^2\right) \le \exp(1). \end{aligned} \tag{3.3}$$

We construct a stochastic approximation for $\nabla\Psi(y)$ by using batches of size r

$$\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r) \triangleq \frac{1}{r}\sum_{i=1}^r \nabla\Psi(y,\xi^i). \tag{3.4}$$

To estimate the variance of minibatch stochastic gradient (3.4), we refer to (Juditsky and Nemirovski, 2008, Theorem 2.1) and (Lan et al., 2012, Lemma 2) on large-deviations theory.

Lemma 3.1.1. *(Juditsky and Nemirovski, 2008, Theorem 2.1) Let $\{D_k\}_{k=1}^N$ be a sequence of random vectors (martingale-difference sequence) such that for all $k = 1, ..., N$, $\mathbb{E}[D_k|D_1, D_2, ..., D_{k-1}] = 0$. Let the sequence $\{D_k\}_{k=1}^N$ satisfies 'light-tail' assumption*

$$\mathbb{E}\left[\exp\left(\frac{\|D_k\|_2^2}{\sigma_k^2}\right)\Big|\eta_1,\dots,\eta_{k-1}\right] \le \exp(1) \quad (a.s.), \quad k = 1, ..., N.$$

Then for all $\Omega \ge 0$

$$\mathbb{P}\left(\left\|\sum_{k=1}^N c_k D_k\right\|_2 \ge (\sqrt{2}+\sqrt{2}\Omega)\sqrt{\sum_{k=1}^N c_k^2\sigma_k^2}\right) \le \exp\left(\frac{-\Omega^2}{3}\right),$$

where $c_1, \dots, c_N$ are positive numbers.

Lemma 3.1.2. *(Lan et al., 2012, Lemma 2) Let for all $k = 1, \dots N$, $D_k = D_k(\{\eta_l\}_{l=1}^k)$ be a deterministic function of i.i.d. realizations $\{\eta_l\}_{l=1}^k$ such that*

$$\mathbb{E}\left[\exp\left(\frac{D_k^2}{\sigma_k^2}\right)\Big|\eta_1,\dots,\eta_{k-1}\right] \le \exp(1) \quad (a.s.), \quad k = 1, ..., N.$$

Then for all $\Omega \ge 0$

$$\mathbb{P}\left(\sum_{k=1}^N c_k D_k^2 \ge (1+\Omega)\sum_{k=1}^N c_k\sigma_k^2\right) \le \exp\left(-\Omega\right),$$

where $c_1, \dots, c_N$ are positive numbers.

Now we use these two lemmas to estimate the variance of the mini-batch stochastic gradient (3.4). The next lemma gives exact constant for the reduced sub-Gaussian variance of the mini-batch gradient.

Lemma 3.1.3 (Sub-Gaussian variance reduction)**.** *Let stochastic gradient* $\nabla\Psi(y,\xi)$ *satisfies the following conditions*

$$\begin{aligned}&\mathbb{E}\nabla\Psi(y,\xi)=\nabla\Psi(y)\\&\mathbb{E}\exp\left(\|\nabla\Psi(y,\xi)-\nabla\Psi(y)\|_2^2/\sigma_\Psi^2\right)\le\exp(1).\end{aligned}$$

Then, for the minibatch gradient $\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)=\frac{1}{r}\sum_{i=1}^r\nabla\Psi(y,\xi^i)$ *with batch size* r*, the following holds with* $\hat{\sigma}_\Psi^2=50\sigma_\Psi^2/r$

$$\begin{aligned}&\mathbb{E}\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)=\nabla\Psi(y)\\&\mathbb{E}\exp\left(\|\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)-\nabla\Psi(y)\|_2^2/\hat{\sigma}_\Psi^2\right)\le\exp(1),\end{aligned}$$

Proof. Lemma 3.1.1 allows us to write for any $\Omega\ge 0$ the following

$$\mathbb{P}\left(\left\|\sum_{i=1}^r\frac{1}{r}\nabla\Psi(y,\xi^i)-\nabla\Psi(y)\right\|_2\ge(\sqrt{2}+\sqrt{2}\Omega)\frac{\sigma_\Psi}{\sqrt{r}}\right)\le\exp\left(\frac{-\Omega^2}{3}\right).$$

Here we used $D_i=\nabla\Psi(y,\xi^i)-\nabla\Psi(y)$ as a martingale difference, $c_i=\frac{1}{r}$, and $\sigma_i^2=\sigma_\Psi^2$ for all $i=1,...,r$.

For $\Omega\ge\frac{1}{3}$ let us rewrite the previous bound as follows

$$\mathbb{P}\left(\left\|\sum_{i=1}^r\frac{1}{r}\nabla\Psi(y,\xi^i)-\nabla\Psi(y)\right\|_2\ge 4\sqrt{2}\Omega\frac{\sigma_\Psi}{\sqrt{r}}\right)\le\exp\left(\frac{-\Omega^2}{3}\right).\tag{3.5}$$

Next we estimate $\hat{\sigma}_\Psi^2$

$$\begin{aligned}&\mathbb{E}\exp\left(\|\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)-\nabla\Psi(\mathbf{y})\|_2^2/\hat{\sigma}_\Psi^2\right)\\&\triangleq\int_0^\infty\mathbb{P}\left(\exp\left(\|\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)-\nabla\Psi(y)\|_2^2/\hat{\sigma}_\Psi^2\right)\ge x\right)dx\\&\le\int_0^{\exp\left(\frac{32\sigma_\Psi^2}{9r\hat{\sigma}_\Psi^2}\right)}\mathbb{P}\left(\exp\left(\|\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)-\nabla\Psi(y)\|_2^2/\hat{\sigma}_\Psi^2\right)\ge x\right)dx\\&+\int_{\frac{4\sqrt{2}\sigma_\Psi}{3\sqrt{r}}}^\infty\mathbb{P}\left(\|\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)-\nabla\Psi(y)\|_2\ge z\right)\frac{2z}{\hat{\sigma}_\Psi^2}\exp\left(\frac{z^2}{\hat{\sigma}_\Psi^2}\right)dz\\&\le\exp\left(\frac{32\sigma_\Psi^2}{9r\hat{\sigma}_\Psi^2}\right)\\&+\int_{\frac{4\sqrt{2}\sigma_\Psi}{3\sqrt{r}}}^\infty\mathbb{P}\left(\|\nabla^r\Psi(y,\{\xi^i\}_{i=1}^r)-\nabla\Psi(y)\|_2\ge z\right)\frac{2z}{\hat{\sigma}_\Psi^2}\exp\left(\frac{z^2}{\hat{\sigma}_\Psi^2}\right)dz,\end{aligned}\tag{3.6}$$

where we used $\mathbb{P}(\cdot) \leq 1$ and the change of variable $z^2 = \hat{\sigma}_\Psi^2 \ln x$ for $x \in \left[\exp\left(\frac{32\sigma_\Psi^2}{9r\hat{\sigma}_\Psi^2}\right), \infty\right)$. Making the following change of variable $z = 4\sqrt{2}\Omega\frac{\sigma_\Psi}{\sqrt{r}}$ in (3.6) from (3.5) we have for any $\Omega \geq \frac{1}{3}$

$$
\begin{aligned}
&\mathbb{E}\exp\left(\|\nabla^r\Psi(y, \{\xi^i\}_{i=1}^r) - \nabla\Psi(y)\|_2^2/\hat{\sigma}_\Psi^2\right) \\
&\leq \exp\left(\frac{32\sigma_\Psi^2}{9r\hat{\sigma}_\Psi^2}\right) + \int_{\frac{1}{3}}^{\infty} \exp\left(\frac{-\Omega^2}{3}\right) 8\sqrt{2}\Omega\frac{\sigma_\Psi^2}{r\hat{\sigma}_\Psi^2}\exp\left(\frac{32\Omega^2\sigma_\Psi^2}{r\hat{\sigma}_\Psi^2}\right) d\Omega \\
&= \exp\left(\frac{32\sigma_\Psi^2}{9r\hat{\sigma}_\Psi^2}\right) + \frac{\sigma_\Psi^2}{r\hat{\sigma}_\Psi^2}\int_{\frac{1}{3}}^{\infty} 8\sqrt{2}\Omega\exp\left(\frac{32\Omega^2\sigma_\Psi^2 - \Omega^2 r\hat{\sigma}_\Psi^2}{3r\hat{\sigma}_\Psi^2}\right) d\Omega \\
&= \exp\left(\frac{32\sigma_\Psi^2}{9r\hat{\sigma}_\Psi^2}\right) + \frac{4\sqrt{2}\sigma_\Psi^2}{r\hat{\sigma}_\Psi^2}\int_{\frac{1}{9}}^{\infty} \exp\left(-\Omega^2\frac{r\hat{\sigma}_\Psi^2 - 32\sigma_\Psi^2}{3r\hat{\sigma}_\Psi^2}\right) d(\Omega^2) \\
&= \exp\left(\frac{32\sigma_\Psi^2}{9r\hat{\sigma}_\Psi^2}\right) + \frac{12\sqrt{2}\sigma_\Psi^2}{r\hat{\sigma}_\Psi^2 - 32\sigma_\Psi^2}\exp\left(-\frac{r\hat{\sigma}_\Psi^2 - 32\sigma_\Psi^2}{27r\hat{\sigma}_\Psi^2}\right),
\end{aligned}
$$

where we used $\int_{1/9}^{\infty} e^{-ax}dx = \frac{1}{a}e^{-a/9}$ for any $a > 0$. If we take $\hat{\sigma}_\Psi^2 = 50\sigma_\Psi^2/r$ in the last inequality we get the statement of the theorem

$$
\begin{aligned}
&\mathbb{E}\exp\left(\|\nabla^r\Psi(y, \{\xi^i\}_{i=1}^{r_k}) - \nabla\Psi(y)\|_2^2/\hat{\sigma}_\Psi^2\right) \\
&\leq \exp(16/(9*25)) + \frac{6\sqrt{2}}{9}\exp(-18/27) \sim 1.6 \leq \exp(1).
\end{aligned}
$$

□

3.1.2 Algorithm and Convergence Rate

Now we propose an algorithm (Algorithm 5) to solve the pair of problems (3.1) and (3.2). The algorithm is an accelerated version of the gradient descent method. The next theorem studies its convergence.

Theorem 3.1.4. *Let $F(x)$ be γ_F-strongly convex. Let R_λ be such that $\|\lambda^*\|_2 \leq R_\lambda$, where λ^* is an exact solution of dual problem (3.2). Then, after $N = O\left(\sqrt{\frac{L_\Psi R_\lambda^2}{\varepsilon}}\right)$ iterations, the output x^N of Algorithm 5 satisfies the following with probability at least $1 - \alpha$*

$$
F(x^N) - F(x^*) \leq \varepsilon, \quad \|Ax^N - b\|_2 \leq \varepsilon/R_\lambda, \tag{3.12}
$$

where $L_\Psi = \lambda_{\max}(A^\top A)/\gamma_F$. The number of dual oracle calls of $\nabla\Psi(\lambda, \xi)$ is

$$
\widetilde{O}\left(\max\left\{\sqrt{\frac{L_\Psi {R_\lambda}^2}{\varepsilon}}, \frac{\sigma_\Psi^2 {R_\lambda}^2}{\varepsilon^2}\right\}\right),
$$

where σ_Ψ^2 is sub-Gaussian variance of $\nabla\Psi(\lambda, \xi)$.

Algorithm 5 Dual Stochastic Accelerated Gradient Algorithm

Input: Starting point $\lambda^0 = \eta^0 = \zeta^0 = 0$, number of iterations N, $A_0 = \alpha_0 = 0$, $L_\Psi = \lambda_{\max}(A^\top A)/\gamma_F$.

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

2:
$$A_{k+1} = A_k + \alpha_{k+1} = 2L_\Psi \alpha_{k+1}^2. \tag{3.7}$$

3:
$$\lambda^{k+1} = (\alpha_{k+1}\zeta^k + A_k\eta^k)/A_{k+1}. \tag{3.8}$$

4: Calculate $\nabla^{r_{k+1}}\Psi(\lambda^{k+1}, \{\xi_i\}_{i=1}^{r_{k+1}})$ according to (3.4) with batch size
$$r_{k+1} = \max\left\{1, 50\sigma_\Psi^2 \alpha_{k+1}\log(N/\alpha)/\varepsilon\right\}. \tag{3.9}$$

5:
$$\zeta^{k+1} = \zeta^k - \alpha_{k+1}\nabla^{r_{k+1}}\Psi(\lambda^{k+1}, \{\xi^\ell\}_{\ell=1}^{r_{k+1}}). \tag{3.10}$$

6:
$$\eta^{k+1} = (\alpha_{k+1}\zeta^{k+1} + A_k\eta^k)/A_{k+1}. \tag{3.11}$$

7: **end for**

Output: $x^N \triangleq \frac{1}{A_N}\sum_{k=0}^N \alpha_k x(\lambda^k, \{\xi^\ell\}_{\ell=1}^{r_k})$, where
$$x(\lambda^k, \{\xi^\ell\}_{\ell=1}^{r_k}) \triangleq \frac{1}{r_k}\sum_{\ell=1}^{r_k} x(\lambda^k, \xi^\ell) = \nabla^{r_k}\Psi(\lambda^k, \{\xi^\ell\}_{\ell=1}^{r_k})$$

Sketch of the Proof. Let us define the set $B_{2R_\lambda}(0) = \{\lambda : \|\lambda\|_2 \le 2R_\lambda\}$. From (Dvurechensky et al., 2018a, Theorem 1) it follows that Algorithm 5 generates the sequences $\{\lambda^N, \zeta^N, y^N, \alpha^N, A^N\}$ satisfying

$$\begin{aligned} A_N\Psi(\eta^N) \le & \min_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^N \alpha_k\left(\Psi(\lambda^k) + \langle\nabla^{r_k}\Psi(\lambda^k, \{\xi^i\}_{i=1}^{r_k}), \lambda - \lambda^k\rangle\right)\right\} + 2R_\lambda^2 \\ & + \sum_{k=0}^{N-1} A_{k+1}\langle\nabla^{r_{k+1}}\Psi(\lambda^{k+1}, \{\xi^i\}_{i=1}^{r_{k+1}}) - \nabla\Psi(\lambda^{k+1}), \eta^k - \lambda^{k+1}\rangle \\ & + \sum_{k=0}^N \frac{A_k}{2L_\Psi}\|\nabla^{r_k}\Psi(\lambda^k, \{\xi^i\}_{i=1}^{r_k}) - \nabla\Psi(\lambda^k)\|_2^2. \end{aligned} \tag{3.13}$$

We denote the stochastic terms in (3.13) as follows

1. $H_1 = \min\limits_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^N \alpha_k\left(\Psi(\boldsymbol{\lambda}^k) + \langle\nabla^{r_k}\Psi(\lambda^k, \{\xi^i\}_{i=1}^{r_k}), \lambda - \lambda^k\rangle\right)\right\}$,
2. $H_2 = \sum_{k=0}^{N-1} A_{k+1}\langle\nabla^{r_k}\Psi(\lambda^{k+1}, \{\xi^i\}_{i=1}^{r_k}) - \nabla\Psi(\lambda^{k+1}), \eta^k - \boldsymbol{\lambda}^{k+1}\rangle$,
3. $H_3 = \sum_{k=0}^N \frac{A_k}{2L_\Psi}\|\nabla^{r_k}\Psi(\lambda^k, \{\xi^i\}_{i=1}^{r_k}) - \nabla\Psi(\lambda^k)\|_2^2$.

By adding and subtracting $\sum_{k=0}^{N-1}\alpha_{k+1}\langle\nabla\Psi(\lambda^{k+1}),\lambda^*-\lambda^{k+1}\rangle$ under the minimum in H_1 we get

$$\begin{aligned}
H_1 = &\min_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^{N}\alpha_k\left(\Psi(\lambda^k)+\langle\nabla^{r_k}\Psi(\lambda^k,\{\xi^i\}_{i=1}^{r_k})-\nabla\Psi(\lambda^k),\lambda-\lambda^k\rangle\right.\right.\\
&\left.\left.+\langle\nabla\Psi(\lambda^k),\lambda-\lambda^k\rangle\right)\right\}\\
&\le \min_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^{N}\alpha_k\left(\Psi(\lambda_k)+\langle\nabla\Psi(\lambda^k),\lambda-\lambda^k\rangle\right)\right\}\\
&+\max_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^{N}\alpha_k\langle\nabla^{r_k}\Psi(\lambda^k,\{\xi^i\}_{i=1}^{r_k})-\nabla\Psi(\lambda^k),\lambda\rangle\right\}\\
&+\sum_{k=0}^{N}\alpha_k\langle\nabla\Psi(\lambda^k)-\nabla^{r_k}\Psi(\lambda^k,\{\xi^i\}_{i=1}^{r_k}),\lambda^k\rangle\\
&\le \min_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^{N}\alpha_k(\Psi(\lambda^k)+\langle\nabla\Psi(\lambda^k),\lambda-\lambda^k\rangle)\right\}\\
&+2R_\lambda\|\sum_{k=0}^{N}\alpha_k(\nabla^r\Psi(\lambda^k,\{\xi^i\}_{i=1}^{r})-\nabla\Psi(\lambda^k))\|_2\\
&+\sum_{k=0}^{N}\alpha_k\langle\nabla\Psi(\lambda^k)-\nabla^{r_k}\Psi(\lambda^k,\{\xi^i\}_{i=1}^{r_k}),\lambda^k\rangle. \qquad (3.14)
\end{aligned}$$

We denote the terms in (3.14) as follows

1. $H_4 = 2R_\lambda\|\sum_{k=0}^{N}\alpha_k(\nabla^{r_k}\Psi(\lambda^k,\{\xi^i\}_{i=1}^{r_k})-\nabla\Psi(\lambda^k))\|_2$,
2. $H_5 = \sum_{k=0}^{N}\alpha_k\langle\nabla\Psi(\lambda^k)-\nabla^{r_k}\Psi(\lambda^k,\{\xi^i\}_{i=1}^{r_k}),\lambda^k\rangle$.

By Cauchy–Schwarz inequality for $H_6 = H_2 + H_5$ we have

$$H_6 \le \sum_{k=0}^{N-1}\|\nabla^{r_{k+1}}\Psi(\lambda^{k+1},\{\xi^i\}_{i=1}^{r_k})-\nabla\Psi(\lambda^{k+1})\|_2\|A_{k+1}\eta^k - A_{k+1}\lambda^{k+1}-\alpha_{k+1}\lambda^{k+1}\|_2 \qquad (3.15)$$

$$\overset{(3.8),(3.7)}{=} \sum_{k=0}^{N-1}\alpha_{k+1}\|\nabla^{r_{k+1}}\Psi(\lambda^{k+1},\{\xi^i\}_{i=1}^{r_k})-\nabla\Psi(\lambda^{k+1})\|_2\|\eta^k-\boldsymbol{\zeta}^k-\lambda^{k+1}\|_2$$

$$\le 3\mathcal{R}\sum_{k=0}^{N-1}\alpha_{k+1}\|\nabla^{r_{k+1}}\Psi(\lambda^{k+1},\{\xi^i\}_{i=1}^{r_k})-\nabla\Psi(\lambda^{k+1})\|_2, \qquad (3.16)$$

where $\mathcal{R}\ge\max\{\|\eta^k\|_2,\|\lambda^{k+1}\|_2,\|\zeta^k\|_2,2R_\lambda\}$ for all $k=1,\dots,N$.

For H_4, H_6 we will use Lemma 3.1.1, and for H_3 we will refer to Lemma 3.1.2. We also will use Lemma 3.1.3 to estimate $\hat{\sigma}_\Psi^2 = 50\sigma_\Psi^2/r$.

Now we use Lemma 3.1.1 for H_6. We take $D_k = \nabla^{r_k}\Psi(\lambda^k, \{\xi^i\}_{i=1}^{r_k}) - \nabla\Psi(\lambda^k)$, $c_k = 3\mathcal{R}\alpha_k$ and $\sigma_k^2 = \hat{\sigma}_\Psi^2 = 50\sigma_\Psi^2/r_k$. Therefore, we get

$$\begin{aligned}&\mathbb{P}\left\{H_6 \geq 3\mathcal{R}(\sqrt{2}+\sqrt{2}\Omega)\sqrt{\sum_{k=1}^N 50\alpha_k^2\sigma_\Psi^2/r_k}\right\}\\ &= \mathbb{P}\left\{H_6 \geq 30\mathcal{R}\sigma_\Psi(1+\Omega)\sqrt{\sum_{k=1}^N \alpha_k^2/r_k}\right\} \leq \exp\left(\frac{-\Omega^2}{3}\right),\end{aligned} \tag{3.17}$$

For H_4 we also use Lemma 3.1.1. We take $D_k = \nabla^{r_k}\Psi(\lambda^k, \{\xi^i\}_{i=1}^{r_k}) - \nabla\Psi(\lambda^k)$, $c_k = 2R_\lambda\alpha_k$ and $\sigma_k^2 = \hat{\sigma}_\Psi^2 = 50\sigma_\Psi^2/r_k$.

$$\begin{aligned}&\mathbb{P}\left\{H_4 \geq 2R_\lambda(\sqrt{2}+\sqrt{2}\Omega)\sqrt{\sum_{k=1}^N 50\alpha_k^2\sigma_\Psi^2/r_k}\right\}\\ &= \mathbb{P}\left\{H_4 \geq 20R_\lambda\sigma_\Psi(1+\Omega)\sqrt{\sum_{k=1}^N \alpha_k^2/r_k}\right\} \leq \exp\left(\frac{-\Omega^2}{3}\right),\end{aligned} \tag{3.18}$$

Now we use Lemma 3.1.2 for H_3. We take $D_k = \|\nabla^{r_k}\Psi(\lambda^k, \{\xi^i\}_{i=1}^{r_k}) - \nabla\Psi(\lambda^k)\|_2$, $c_k = \frac{A_k}{2L_\Psi}$, and $\sigma_k^2 = \hat{\sigma}_\Psi^2 = 50\sigma_\Psi^2/r_k$ for all $k = 1, ..., N$ and get the following

$$\mathbb{P}\left\{H_3 \geq 25(1+\Omega)\sigma_\Psi^2\sum_{k=1}^N \frac{A_k}{L_\Psi r_k}\right\} = \mathbb{P}\left\{H_3 \geq 50\sigma_\Psi^2(1+\Omega)\sum_{k=1}^N \frac{\alpha_k^2}{r_k}\right\} \leq \exp\left(-\Omega\right).$$

We can equivalently rewrite it as follows

$$\mathbb{P}\left\{H_3 \geq 50\sigma_\Psi^2(1+\Omega^2/3)\sum_{k=1}^N \alpha_k^2\right\} \leq \exp\left(-\frac{\Omega^2}{3}\right). \tag{3.19}$$

Next we again consider (3.13)

$$\begin{aligned}A_N\Psi(\eta^N) &\leq H_1 + 2R_\lambda^2 + H_2 + H_3\\ &\leq \min_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^N \alpha_k(\Psi(\lambda^k) + \langle\nabla\Psi(\lambda^k), \lambda - \lambda^k\rangle)\right\} + 2R_\lambda^2\\ &\quad + H_2 + H_3 + H_4 + H_5.\end{aligned} \tag{3.20}$$

Next we estimate the r.h.s of (3.20). We consider

$$\frac{1}{A_N}\min_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^N \alpha_k(\Psi(\lambda^k) + \langle\nabla\Psi(\lambda^k), \lambda - \lambda^k\rangle)\right\}. \tag{3.21}$$

By the definition of the dual function and by the Demyanov–Danskin theorem we have

$$\Psi(\lambda) = \langle \lambda,\ Ax(\lambda) - b\rangle - F(x(\lambda)) \quad \text{and} \quad \nabla\Psi(\lambda) = Ax(\lambda) - b,$$

where $x(\lambda) = \arg\max_{x\in\mathbb{R}^n}\{\langle \lambda,\ Ax - b\rangle - F(x)\}$. Using this in (3.21) we get

$$\begin{aligned}
&\frac{1}{A_N}\min_{\lambda\in B_{2R_\lambda}(0)}\left\{\sum_{k=0}^{N}\alpha_k\left(\langle\lambda^k,\ Ax(\lambda^k) - b\rangle - F(x(\lambda^k)) + \langle Ax(\lambda^k) - b,\ \lambda - \lambda^k\rangle\right)\right\}\\
&= -\frac{1}{A_N}\sum_{k=0}^{N}\alpha_k F(x(\lambda^k)) + \min_{\lambda\in B_{2R_\lambda}(0)}\left\{\frac{1}{A_N}\sum_{k=0}^{N}\alpha_k\langle Ax(\lambda^k) - b,\ \lambda\rangle\right\}\\
&\le -F(\hat{x}^N) - \max_{\lambda\in B_{2R_\lambda}(0)}\left\{\langle A\hat{x}^N - b,\ \lambda\rangle\right\} = -F(\hat{x}^N) - 2R_\lambda\|A\hat{x}^N - b\|_2. \qquad (3.22)
\end{aligned}$$

where we used $\hat{x}^N \triangleq \frac{1}{A_N}\sum_{k=0}^{N}\alpha_k x(\lambda^k)$.

Then we estimate the rest terms of the r.h.s. of (3.20). From the union bound applied for (3.19), and (3.18), (3.17) and making the change $\alpha = \exp\left(-\frac{\Omega^2}{3}\right)$ we have with probability $\ge 1 - 3\alpha$

$$\begin{aligned}
&H_3 + H_4 + H_6 \le\\
&50\sigma_\Psi^2(1+\ln(1/\alpha))\sum_{k=1}^{N}\alpha_k^2/r_k + 20R_\lambda\sigma_\Psi(1+\sqrt{3\ln(1/\alpha)})\sqrt{\sum_{k=1}^{N}\alpha_k^2/r_k}\\
&+ 30\mathcal{R}\sigma_\Psi(1+\sqrt{3\ln(1/\alpha)})\sqrt{\sum_{k=1}^{N}\alpha_k^2/r_k}.
\end{aligned}$$

By the definition of $\mathcal{R} \ge 2R_\lambda$ we have

$$\begin{aligned}
&H_3 + H_4 + H_6 \le\\
&50\sigma_\Psi^2(1+\ln(1/\alpha))\sum_{k=1}^{N}\alpha_k^2/r_k + 40\mathcal{R}\sigma_\Psi(1+\sqrt{3\ln(1/\alpha)})\sqrt{\sum_{k=1}^{N}\alpha_k^2/r_k}.
\end{aligned}$$

By the definition of r (3.9) we get

$$\begin{aligned}
&H_3 + H_4 + H_6\\
&\le 50\sigma_\Psi^2(1+\ln(1/\alpha))\sum_{k=1}^{N}\frac{\varepsilon\alpha_k}{\sigma_\Psi^2\ln(N/\delta)} + 40\mathcal{R}\sigma_\Psi(1+\sqrt{3\ln(1/\alpha)})\sqrt{\sum_{k=1}^{N}\frac{\varepsilon\alpha_k}{\sigma_\Psi^2\ln(N/\delta)}}\\
&= 50(1+\ln(1/\alpha))\frac{\varepsilon A_N}{\ln(N/\delta)} + 40\mathcal{R}(1+\sqrt{3\ln(1/\alpha)})\sqrt{\frac{\varepsilon A_N}{\ln(N/\delta)}}, \qquad (3.23)
\end{aligned}$$

where we used $A_N = \sum_{k=1}^{N} \alpha_k$ from (3.7) in the last equality. We sum up (3.22) and (3.23) we rewrite (3.20) and divide it by A_N. We get with probability $\geq 1 - 3\alpha$ the following

$$\Psi(\eta^N) + F(\hat{x}^N) + 2R_\lambda \|A\hat{x}^N - b\|_2 \leq \frac{2R_\lambda^2}{A_N}$$
$$+50(1 + \ln(1/\alpha))\frac{\varepsilon}{\ln(N/\delta)} + 40\mathcal{R}(1 + \sqrt{3\ln(1/\alpha)})\sqrt{\frac{\varepsilon}{A_N \ln(N/\delta)}} \tag{3.24}$$

The next steps are to prove $\mathcal{R} = O(R_\lambda)$ and transfer from the $\hat{x}^N$ to the output of the Algorithm 5, that is x^N, by using large deviation bounds. This can be found in the paper (Gorbunov et al., 2019),

□

3.2 Decentralized Optimization

Background on Distributed Optimization. A distributed system is a system of computing nodes (agents, machines, processing units), whose interactions are constrained by the system structure. In distributed computing, a problem is divided into many tasks, assigned to different agents. The agents cooperatively solve the global task by solving their local problems and transferring information (usually, a vector) to other nodes.

Distributed optimization has recently gained increased interest due to large-scale problems encountered in machine learning. Usually these problems aggregate enormous data and they need to be solved in a reasonable time with no prohibitive expenses. It can also occur that the data itself is stored or collected in a distributed manner (e.g., sensors in a sensor network obtained the state of the environment from different geographical parts, or micro-satellites collecting local information). In both these settings, distributed systems can be used. They process faster and more data than one computer since the work is divided between many computing nodes. The application of distributed systems includes formation control of unmanned vehicle (Ren, 2006), power system control (Ram et al., 2009), information processing and decision making in sensor networks, distributed averaging, statistical inference and learning (Nedić et al., 2017).

There are two scenarios of distributed optimization: centralized and decentralized. In centralized optimization, there is a central node (master) which coordinates the work of other nodes (slaves). Parallel architecture is a special case of the centralized architecture as it always contains master node. Unfortunately, centralized architecture has a synchronization drawback and a high requirement for the master node (Scaman et al., 2017). To address these disadvantages to some extent, a decentralized distributed architecture should be used (Bertsekas and Tsitsiklis, 1997; Kibardin, 1979). In decentralized scenario, there is no particular node, all agents are equivalent and their communications are constrained only by a network arhcitecture: each agent can communicate only with its immediate neighbors. This decentralized

setting is more robust since decentralized algorithm does not crash when one of computing node fails. Moreover, decentralized computing can be preformed on time-varying (wireless) communication networks.

A large number of distributed algorithms have been developed to minimize an objective given in the form of the average of functions f_i's accessible by different nodes (agents, computers) in a network Thus, we consider the following convex optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{3.25}$$

where $f_i(x)$'s are γ-strongly convex and possibly presented by the expectation $f_i(x) = \mathbb{E}f_i(x,\xi)$ w.r.t. $\xi \in \Xi$.

3.2.1 Decentralized Dual Problem Formulation

To solve (3.25) on a network of agents, a transition to its dual problem is used. For this, we introduce artificial constraint $x_1 = x_2 = \cdots = x_m$ to (3.25) and rewrite it as follows

$$\min_{\substack{x_1=\ldots=x_m,\\ x_1,\ldots,x_m\in\mathbb{R}^n,}} F(\mathbf{x}) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x_i), \tag{3.26}$$

where $\mathbf{x} = (x_1^\top, x_2^\top, ..., x_n^\top)^\top$ is the stack column vector. Further, we will replace the constraint $x_1 = \cdots = x_m$ with affine constraints representing the network structure.

Network system. Let a network of m nodes (agents, computing units) be presented by a fixed connected undirected graph $G = (V, E)$, where V is a set of m nodes, and $E = \{(i,j) : i,j \in V\}$ is a set of edges. The network structure imposes communication constraints: agent i can communicate (exchange information) only with its immediate neighbors (i.e., with agent $j \in V$ such that $(i,j) \in E$.

Let us also define a symmetric and positive semi-definite matrix $W \in \mathbb{R}^{m\times m}$, which will represent a network structure. We define this matrix by the Laplacian matrix of the graph G. The elements of W are presented as

$$[W]_{ij} = \begin{cases} -1, & \text{if } (i,j)\in E,\\ \deg(i), & \text{if } i = j,\\ 0, & \text{otherwise,}\end{cases}$$

where $\deg(i)$ is the degree of vertex i (i.e., the number of neighboring nodes).

Let us further define matrix

$$\mathbf{W} \triangleq W \otimes I_n, \tag{3.27}$$

where $\otimes$ is the Kronecker product and I_n is the identity matrix. Matrix $\mathbf{W}$ inherits the properties of W, including the symmetry and positive semi-definiteness. Furthermore, the vector $\mathbf{1}$ is the unique (up to a scaling factor) eigenvector of $\mathbf{W}$

associated with the eigenvalue $\lambda = 0$. Thus, the equality constraint $x_1 = \cdots = x_m$ is equivalent to affine constraint $\mathbf{W}\mathbf{x} = 0$. Moreover, the following identity holds (Scaman et al., 2017)

$$x_1 = \cdots = x_m \quad \Longleftrightarrow \quad \mathbf{W}\mathbf{x} = 0 \Longleftrightarrow \sqrt{\mathbf{W}}\mathbf{x} = 0.$$

Thus, the problem (3.26) can be rewritten as optimization problem with affine constraints

$$\min_{\substack{\sqrt{\mathbf{W}}\mathbf{x}, \\ x_1,\ldots,x_m \in \mathbb{R}^n}} F(\mathbf{x}) \triangleq \frac{1}{m}\sum_{i=1}^m f_i(x_i). \tag{3.28}$$

The dual problem for problem (3.26) (written as a maximization problem) is given by the following problem with the Lagrangian dual variable $\mathbf{y} \in \mathbb{R}^{mn}$

$$\begin{aligned}
\min_{\mathbf{y}\in\mathbb{R}^{mn}} \Psi(\sqrt{\mathbf{W}}\mathbf{y}) &\triangleq \max_{\mathbf{x}\in\mathbb{R}^{mn}} \left\{\langle \mathbf{y}, \sqrt{\mathbf{W}}\mathbf{x}\rangle - F(\mathbf{x})\right\} \\
&= \max_{\mathbf{x}\in\mathbb{R}^{mn}} \left\{\langle \mathbf{y}, \sqrt{\mathbf{W}}\mathbf{x}\rangle - \frac{1}{m}\sum_{i=1}^m f_i(x_i)\right\} \\
&= \frac{1}{m}\max_{\mathbf{x}\in\mathbb{R}^{mn}} \sum_{i=1}^m \left\{m\langle y_i, [\sqrt{\mathbf{W}}\mathbf{x}]_i\rangle - f_i(x_i)\right\} \\
&= \frac{1}{m}\sum_{i=1}^m \psi_i\left(m[\sqrt{\mathbf{W}}\mathbf{y}]_i\right),
\end{aligned} \tag{3.29}$$

where each $\psi_i(\lambda_i) = \max\limits_{x_i\in\mathbb{R}^n}\{\langle \lambda_i, x_i\rangle - f_i(x_i)\}$ is the Fenchel–Legendre transform of $f_i(x_i)$ and the vector $[\sqrt{\mathbf{W}}\mathbf{x}]_i$ represents the i-th n-dimensional block of $\sqrt{\mathbf{W}}\mathbf{x}$.

By Theorem 0.0.6, if $F(\mathbf{x})$ is γ_F-strongly convex, then $\Psi(\sqrt{\mathbf{W}}\mathbf{y})$ is L_Ψ-Lipschitz smooth with $L_\Psi = \lambda_{\max}(W)/\gamma_F$, where $\gamma_F = \gamma/m$.

By Demyanov–Danskin theorem (Demyanov and Malozemov, 1990; Danskin, 2012) we have

$$\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y}) = \sqrt{\mathbf{W}}\mathbf{x}(\sqrt{\mathbf{W}}\mathbf{y}), \tag{3.30}$$

where $\mathbf{x}(\sqrt{\mathbf{W}}\mathbf{y}) = \arg\max\limits_{\mathbf{x}\in\mathbb{R}^{mn}} \left\{\langle \mathbf{x}, \sqrt{\mathbf{W}}\mathbf{y}\rangle - F(\mathbf{x})\right\}$.

We construct a stochastic approximation for $\nabla\Psi(\mathbf{y})$ by using batches of size r

$$\nabla^r\Psi(\mathbf{y}, \{\boldsymbol{\xi}^\ell\}_{\ell=1}^r) \triangleq \frac{1}{r}\sum_{\ell=1}^r \nabla\Psi(\mathbf{y}, \boldsymbol{\xi}^\ell). \tag{3.31}$$

With the change of variable $\bar{\mathbf{y}} := \sqrt{\mathbf{W}}\mathbf{y}$, this can be rewritten as

$$\nabla^r\Psi(\sqrt{\mathbf{W}}\mathbf{y}, \{\boldsymbol{\xi}^\ell\}_{\ell=1}^r) = \sqrt{\mathbf{W}}\nabla^r\Psi(\bar{\mathbf{y}}, \{\boldsymbol{\xi}^\ell\}_{\ell=1}^r) = \frac{1}{r}\sum_{\ell=1}^r \sqrt{\mathbf{W}}\nabla\Psi(\bar{\mathbf{y}}, \boldsymbol{\xi}^\ell). \tag{3.32}$$

If each $\nabla\psi_i(\bar{y}_i, \xi_i)$ has sub-Gaussian variance

$$\mathbb{E}\exp\left(\|\nabla\psi_i(\bar{y}_i,\xi_i) - \nabla\psi_i(\bar{y}_i)\|_2^2/\sigma_\psi^2\right) \le \exp(1).$$

Then $\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y}, \boldsymbol{\xi})$ has sub-Gaussian variance with $\sigma_\Psi^2 = O\left(\lambda_{\max}(W)m\sigma_\psi^2\right)$ (Lemma 3.2.1).

Lemma 3.2.1. *Let each $\nabla\psi_i(\bar{y}_i, \xi_i)$ $(i = 1, ..., m)$ has σ_ψ^2 sub-Gaussian variance*

$$\mathbb{E}\nabla\psi_i(\bar{y}_i,\xi_i) = \nabla\psi_i(\bar{y}_i),$$
$$\mathbb{E}\exp\left(\|\nabla\psi_i(\bar{y}_i,\xi_i) - \nabla\psi_i(\bar{y}_i)\|_2^2/\sigma_\psi^2\right) \le \exp(1).$$

Then $\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y}, \boldsymbol{\xi})$ has $\sigma_\Psi^2 = O\left(\lambda_{\max}(W)m\sigma_\psi^2\right)$ sub-Gaussian variance, where

$$\Psi(\sqrt{\mathbf{W}}\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m}\psi_i\left(m[\sqrt{\mathbf{W}}\mathbf{y}]_i\right).$$

Sketch of the Proof. We provide the proof of this lemma for variance σ_Ψ^2 (non-sub-Gaussian). Let

$$\mathbb{E}\|\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y},\boldsymbol{\xi}) - \mathbb{E}\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y},\boldsymbol{\xi})\|^2 \le \sigma_\Psi^2.$$

Then we estimate $\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y}, \boldsymbol{\xi})$

$$\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y},\boldsymbol{\xi}) = \sqrt{\mathbf{W}}\nabla\Psi(\bar{\mathbf{y}},\boldsymbol{\xi}) = \sqrt{\mathbf{W}}\cdot\frac{1}{m}\cdot m\begin{pmatrix}\nabla\psi_1(\bar{y}_1,\xi_1)\\ \vdots \\ \nabla\psi_m(\bar{y}_m,\xi_m)\end{pmatrix},$$

where $\bar{\mathbf{y}} = \sqrt{\mathbf{W}}\mathbf{y}$. Then

$$\|\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y},\boldsymbol{\xi}) - \mathbb{E}\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y},\boldsymbol{\xi})\|_2^2 = \|\sqrt{\mathbf{W}}\nabla\Psi(\bar{\mathbf{y}},\boldsymbol{\xi}) - \sqrt{\mathbf{W}}\mathbb{E}\nabla\Psi(\bar{\mathbf{y}},\boldsymbol{\xi})\|_2^2$$
$$= \left\langle\begin{pmatrix}\nabla\psi_1(\bar{y}_1,\xi_1) - \mathbb{E}\nabla\psi_1(\bar{y}_1,\xi_1)\\ \vdots \\ \nabla\psi_m(\bar{y}_m,\xi_m) - \mathbb{E}\nabla\psi_m(\bar{y}_m,\xi_m)\end{pmatrix}, \mathbf{W}\begin{pmatrix}\nabla\psi_1(\bar{y}_1,\xi_1) - \mathbb{E}\nabla\psi_1(\bar{y}_1,\xi_1)\\ \vdots \\ \nabla\psi_m(\bar{y}_m,\xi_m) - \mathbb{E}\nabla\psi_m(\bar{y}_m,\xi_m)\end{pmatrix}\right\rangle$$
$$\le \lambda_{\max}(W)\left\|\begin{pmatrix}\nabla\psi_1(\bar{y}_1,\xi_1) - \mathbb{E}\nabla\psi_1(\bar{y}_1,\xi_1)\\ \vdots \\ \nabla\psi_m(\bar{y}_m,\xi_m) - \mathbb{E}\nabla\psi_m(\bar{y}_m,\xi_m)\end{pmatrix}\right\|_2^2.$$

Taking the expectation we obtain

$$\sigma_\Psi^2 \le \mathbb{E}\|\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y},\boldsymbol{\xi}) - \mathbb{E}\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y},\boldsymbol{\xi})\|_2^2$$
$$\le \lambda_{\max}(W)\mathbb{E}\left\|\begin{pmatrix}\nabla\psi_1(\bar{y}_1,\xi_1) - \mathbb{E}\nabla\psi_1(\bar{y}_1,\xi_1)\\ \vdots \\ \nabla\psi_m(\bar{y}_m,\xi_m) - \mathbb{E}\nabla\psi_m(\bar{y}_m,\xi_m)\end{pmatrix}\right\|_2^2$$
$$\le \lambda_{\max}(W)m\sigma_\psi^2.$$

More precise proof with sub-Gaussian variance can be performed similarly to the proof of Lemma 3.1.3. □

The optimization problem (3.29) is convex unconstrained optimization problem and can be solved by gradient-type algorithms. If the gradient of Ψ is L_Ψ-Lipschitz continuous, then the gradient descent method does not provide optimal estimates in contradistinction to its accelerated version (Nesterov, 2004). However, for the clarity we explain how problem (3.29) can be solved in a decentralized manner using the gradient descent method in the following example.

Example 3.2.2. *The iterative procedure of the gradient descent algorithm for problem (3.29) is presented as follows ($k = 0, 1, 2, ..., N$)*

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L_\Psi}\nabla\Psi(\sqrt{\mathbf{W}}\mathbf{y}^k) \overset{(3.30)}{=} \mathbf{y}^k - \frac{1}{L_\Psi}\sqrt{\mathbf{W}}\mathbf{x}(\sqrt{\mathbf{W}}\mathbf{y}^k). \tag{3.33}$$

Without change of variable, it is unclear how to perform this procedure in a distributed manner. Let $\bar{\mathbf{y}} := \sqrt{W}\mathbf{y}$, then the gradient step (3.33) multiplied by $\sqrt{\mathbf{W}}$ can be rewritten as

$$\bar{\mathbf{y}}^{k+1} = \bar{\mathbf{y}}^k - \frac{1}{L_\Psi}\mathbf{W}\mathbf{x}(\bar{\mathbf{y}}^k).$$

This procedure can be performed in a decentralized manner on a network of agents. Namely each agent $i = 1, ..., m$ calculates

$$\bar{y}_i^{k+1} = \bar{y}_i^k - \frac{1}{L_\Psi}[\mathbf{W}\mathbf{x}(\bar{\mathbf{y}}^k)]_i = \bar{y}_i^k - \frac{1}{L_\Psi}\sum_{j=1}^{n}\mathbf{W}_{ij}x_j(\bar{y}_j^k).$$

Multiplication $\mathbf{W}\mathbf{x}$ naturally defines communications in the network because the elements of matrix $\mathbf{W}_{ij}$

$$\mathbf{W}_{ij} = \begin{cases} -I_{n\times n}, & \text{if } (i,j) \in E, \\ deg(i)I_{n\times n}, & \text{if } i = j, \\ 0_{n\times n}, & \text{otherwise,} \end{cases}$$

are non-zero only for neighboring nodes i, j, and

$$[\mathbf{x}(\bar{\mathbf{y}}^k)]_j = \arg\max_{\mathbf{x}\in\mathbb{R}^{mn}}\left\{\langle x_j, \bar{y}_j^k\rangle - f(x_j)\right\} = x_j(\bar{y}_j^k).$$

Similarly, to the gradient descent method, we can apply its accelerated version (Algorithm 5) in a decentralized manner. The decentralized version of Algorithm 5 with change of variables

$$\bar{\boldsymbol{\eta}} = \sqrt{\mathbf{W}}\boldsymbol{\eta}, \quad \bar{\boldsymbol{\lambda}} = \sqrt{\mathbf{W}}\boldsymbol{\lambda}, \quad \bar{\boldsymbol{\zeta}} = \sqrt{\mathbf{W}}\boldsymbol{\zeta}$$

is presented in Algorithm 6

Algorithm 6 Decentralized Dual Stochastic Accelerated Gradient Algorithm

Input: Starting point $\bar{\boldsymbol{\lambda}}^0 = \bar{\boldsymbol{\eta}}^0 = \bar{\boldsymbol{\zeta}}^0 = 0$, number of iterations N, $A_0 = \alpha_0 = 0$,

1: For each agent $i \in V$ $(i = 1, ..., m)$
2: **for** $k = 0, \ldots, N-1$ **do**
3: $A_{k+1} = A_k + \alpha_{k+1} = 2L_\Psi \alpha_{k+1}^2$.
4: $\bar{\lambda}_i^{k+1} = (\alpha_{k+1}\bar{\zeta}_i^k + A_k\bar{\eta}_i^k)/A_{k+1}$.
5: Calculate $\nabla^{r_{k+1}}\psi_i(\bar{\lambda}_i^{k+1}, \{\xi_i^\ell\}_{\ell=1}^{r_{k+1}})$ from (3.29) according to (3.31) with mini-batch size

$$r_{k+1} = \max\left\{1, 50\sigma_\Psi^2 \alpha_{k+1} \ln(N/\alpha)/\varepsilon\right\},$$

where $\sigma_\Psi^2 = O\left(\lambda_{\max}(W) m \sigma_\psi^2\right)$
6: $\bar{\zeta}_i^{k+1} = \bar{\zeta}_i^k - \alpha_{k+1}\sum_{j=1}^m \mathbf{W}_{ij}\nabla^{r_{k+1}}\psi_j(\bar{\lambda}_j^{k+1}, \{\xi_j^\ell\}_{\ell=1}^{r_{k+1}})$.
7: $\bar{\eta}_i^{k+1} = (\alpha_{k+1}\bar{\zeta}_i^{k+1} + A_k\bar{\eta}_i^k)/A_{k+1}$.
8: **end for**

Output: $\boldsymbol{x}^N = (\tilde{x}_1^\top, \ldots, \tilde{x}_m^\top)^\top$, where $\tilde{x}_i \triangleq \frac{1}{A_N}\sum_{k=0}^N \alpha_k x_i(\bar{\lambda}_i^k, \{\xi_i^\ell\}_{\ell=1}^{r_k})$ for all $i = 1, \ldots, m$ with

$$x_i(\bar{\lambda}_i^k, \{\xi_i^\ell\}_{\ell=1}^{r_k}) \triangleq \frac{1}{r_k}\sum_{\ell=1}^{r_k} x_i(\bar{\lambda}_i^k, \xi_i^\ell) = \nabla^{r_k}\psi_i(\bar{\lambda}_i^k, \{\xi_i^\ell\}_{\ell=1}^{r_k}).$$

3.2.2 Algorithm and Convergence Rate

The next theorem is a decentralized variant of Theorem 3.1.4 for particular case of matrix $A = \sqrt{\mathbf{W}}$ and $b = 0$ together with the fact $\mathcal{R} = O(R_\lambda)$ from (Gorbunov et al., 2019). Let $\chi(W) = \frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)}$ be the condition number of matrix W.

Theorem 3.2.3. *Let $f_i(x)$'s be γ-strongly convex functions. Let R_λ be such that $\|\lambda^*\|_2 \le R_\lambda$, where λ^* is an exact solution of dual problem (3.29). Let for all $i = 1, ..., m$, $\|\nabla f_i(x^*)\|_2 \le M$, where x^* is the solution of (3.28). Then, after $N = O\left(\sqrt{\frac{M^2}{\gamma\varepsilon}}\chi(W)\right)$ iterations, the output $\mathbf{x}^N$ of Algorithm 6 satisfies the following with probability at least $1 - 3\alpha$*

$$F(\mathbf{x}^N) - F(\mathbf{x}^*) \le \varepsilon, \quad \|\sqrt{\mathbf{W}}\mathbf{x}^N\|_2 \le \varepsilon/R_\lambda.$$

The number of dual oracle calls of $\nabla\psi_i(\lambda_i, \xi_i)$ is

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}, \frac{M^2\sigma_\psi^2}{\varepsilon^2}\chi(W)\right\}\right),$$

where σ_ψ^2 is sub-Gaussian variance of $\nabla\psi_i(\lambda_i, \xi_i)$.

Proof. Using the fact $\mathcal{R} = O(R_\lambda)$ proved in (Gorbunov et al., 2019), we improve the number of iterations from Theorem 3.1.4 as follows

$$N = O\left(\sqrt{\frac{L_\Psi R_\lambda^2}{\varepsilon}}\right), \tag{3.34}$$

where $L_\Psi = \lambda_{\max}(W)/\gamma_F$ is the constant of Lipschitz smoothness for $\Psi(\sqrt{\mathbf{Wy}})$, and $\gamma_F = \gamma/m$. Then we use (Lan et al., 2017) to estimate the radius of the dual solution λ^* (corresponding to the minimal Euclidean distance if there are more than one solution)

$$\|\lambda^*\|_2^2 \leq R_\lambda^2 = \frac{\|\nabla F(\mathbf{x}^*)\|_2^2}{\lambda_{\min}^+(W)} \leq \frac{\left\| \frac{1}{m} \begin{pmatrix} \nabla f_1(x^*) \\ \vdots \\ \nabla f_m(x^*) \end{pmatrix} \right\|_2^2}{\lambda_{\min}^+(W)} = \frac{\sum_{i=1}^m \|\nabla f_i(x^*)\|_2^2}{m^2 \lambda_{\min}^+(W)} \leq \frac{M^2}{m\lambda_{\min}^+(W)}, \tag{3.35}$$

where $\lambda_{\min}^+(W)$ is the minimal non-zero eigenvalue of matrix W. Then using $L_\Psi = \lambda_{\max}(W)/\gamma_F$, $\gamma_F = \gamma/m$ and (3.35) in (3.34) we get

$$N = O\left(\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}\right).$$

The number of dual oracle calls of $\nabla\psi_i(\lambda_i, \xi_i)$ is (Theorem 3.1.4)

$$\begin{aligned} &\widetilde{O}\left(\max\left\{N, \frac{\sigma_\Psi^2 R_\lambda^2}{\varepsilon^2}\right\}\right) \\ &\overset{(3.35)}{=} \widetilde{O}\left(\max\left\{\sqrt{\frac{M_F^2}{\gamma\varepsilon}\chi(W)}, \frac{\lambda_{\max}(W) m\sigma_\psi^2}{\varepsilon^2} \cdot \frac{M^2}{m\lambda_{\min}^+(W)}\right\}\right) \\ &= \widetilde{O}\left(\max\left\{\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}, \frac{M^2\sigma_\psi^2}{\varepsilon^2}\chi(W)\right\}\right), \end{aligned}$$

where we used $\sigma_\Psi^2 = O(\lambda_{\max}(W) m\sigma_\psi^2)$ (Lemma 3.2.1) and $\chi(W) = \frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)}$ is the condition number of matrix W.

□

3.3 Wasserstein Barycenter Problem

In this section, we apply the results stated above in a broad sense to the Wasserstein barycenter problem defined with respect to entropy-regularized optimal transport

$$\min_{p \in \Delta_n} \frac{1}{m} \sum_{i=1}^m W_\gamma(p, q_i), \tag{3.36}$$

where $W_\gamma(p, q_i)$ is γ-strongly convex w.r.t p in the ℓ_2-norm.

3.3.1 Decentralized Dual Formulation

To state the Wasserstein barycenter problem (3.36) in a decentralized manner, we rewrite it as follows

$$\min_{\substack{p_1=\ldots=p_m,\\ p_1,\ldots,p_m\in\Delta_n}} \frac{1}{m}\sum_{i=1}^m W_\gamma(p_i,q_i) = \min_{\substack{\sqrt{\mathbf{W}}\mathbf{p}=0,\\ p_1,\ldots,p_m\in\Delta_n}} \frac{1}{m}\sum_{i=1}^m W_\gamma(p_i,q_i), \tag{3.37}$$

where $\mathbf{p} = (p_1^\top, ..., p_m^\top)^\top$ is the column vector and $\mathbf{W}$ is defined in (3.27). The dual problem to (3.37) is

$$\min_{\mathbf{y}\in\mathbb{R}^{nm}} W^*_{\gamma,\mathbf{q}}(\sqrt{\mathbf{W}}\mathbf{y}) \triangleq \frac{1}{m}\sum_{i=1}^m W^*_{\gamma,q_i}(m[\sqrt{\mathbf{W}}\mathbf{y}]_i), \tag{3.38}$$

where $\mathbf{y} = (y_1^\top, ..., y_m^\top)^\top \in \mathbb{R}^{nm}$ is the Lagrangian dual multiplier, $\mathbf{q} = (q_1^\top, ..., q_m^\top)^\top \in \mathbb{R}^{nm}$ and

$$W^*_{\gamma,\mathbf{q}}(\sqrt{\mathbf{W}}\mathbf{y}) \triangleq \max_{p_1,\ldots p_m\in\Delta_n} \left\{ \left\langle \sqrt{\mathbf{W}}\mathbf{y}, \mathbf{p} \right\rangle - \frac{1}{m}\sum_{i=1}^m W_\gamma(p_i,q_i) \right\} \tag{3.39}$$

$$= \frac{1}{m}\sum_{i=1}^m \max_{p_i\in\Delta_n} \left\{ \left\langle m[\sqrt{\mathbf{W}}\mathbf{y}]_i, p_i \right\rangle - W_\gamma(p_i,q_i) \right\} \tag{3.40}$$

$$= \frac{1}{m}\sum_{i=1}^m W^*_{\gamma,q_i}(m[\sqrt{\mathbf{W}}\mathbf{y}]_i), \tag{3.41}$$

Recovery of the Primal Solution. By Demyanov–Danskin theorem (Demyanov and Malozemov, 1990; Danskin, 2012) and from the definition of dual funtion for Wasserstein distances (2.42), we have

$$\nabla W^*_{\gamma,q}(\lambda) = p(\lambda), \tag{3.42}$$

where (2.43)

$$\forall l = 1, ..., n \qquad [p(\lambda)]_l = \sum_{j=1}^n [q]_j \frac{\exp\left(([\lambda]_l - C_{lj})/\gamma\right)}{\sum_{i=1}^n \exp\left(([\lambda]_i - C_{ji})/\gamma\right)}. \tag{3.43}$$

In papers (Uribe et al., 2018; Dvinskikh et al., 2019) a dual distributed algorithm for the Wasserstein barycenter problem was proposed. This algorithm is a deterministic version of Algorithm 6. The next theorem states its convergence.

Theorem 3.3.1. *(Dvinskikh et al., 2019, Corollary 6) After* $N = \widetilde{O}\left(\sqrt{\frac{n\|C\|_\infty^2}{\gamma\varepsilon}}\chi(W)\right)$ *iterations, the output of* $\tilde{\mathbf{p}} = (\tilde{p}_1^T, \cdots, \tilde{p}_m^T)^T$ *of distributed accelerated gradient method with the primal solution recovery* (3.43) *satisfies*

$$\frac{1}{m}\sum_{i=1}^m W_\gamma(\tilde{p}_i, q_i) - \frac{1}{m}\sum_{i=1}^m W_\gamma(p^*, q_i) \le \varepsilon, \qquad \|\sqrt{\mathbf{W}}\tilde{\mathbf{p}}\|_2 \le \varepsilon/R_{\mathbf{y}}.$$

The total per node complexity is

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

3.3.2 Decentralized Dual Stochastic Algorithm

The complexity of dual oracle call for the gradient of the dual function for entropy-regularized optimal transport (3.42) is $O(n^2)$. Using randomize technique, we can reduce it to $O(n)$. To do so, we randomize the true gradient (3.43) by taking component j with probability $[q]_j$

$$[\nabla W^*_{\gamma,q}(\lambda,\xi)]_l = \frac{\exp\left(([\lambda]_l - C_{l\xi})/\gamma\right)}{\sum_{\ell=1}^n \exp\left(([\lambda]_\ell - C_{\ell\xi})/\gamma\right)}, \qquad \forall l = 1,...,n.$$

Recovery of the Primal Solution. We construct a stochastic approximation for $\nabla W^*_{\gamma,\mathbf{q}}(\sqrt{\mathbf{W}}\mathbf{y})$ by using batches of size r and the change of variable $\bar{\mathbf{y}} := \sqrt{\mathbf{W}}\mathbf{y}$

$$\begin{aligned}\nabla^r W^*_{\gamma,\mathbf{q}}(\sqrt{\mathbf{W}}\mathbf{y}, \{\boldsymbol{\xi}^j\}_{j=1}^r) = \sqrt{\mathbf{W}}\nabla^r W^*_{\gamma,\mathbf{q}}(\bar{\mathbf{y}}, \{\boldsymbol{\xi}^j\}_{j=1}^r) &= \frac{1}{r}\sum_{j=1}^r \sqrt{\mathbf{W}}\nabla W^*_{\gamma,\mathbf{q}}(\bar{\mathbf{y}}, \boldsymbol{\xi}^j) \\ &= \frac{1}{r}\sum_{j=1}^r \sqrt{\mathbf{W}}\mathbf{p}(\bar{\mathbf{y}}, \boldsymbol{\xi}^j), \quad (3.45)\end{aligned}$$

where $[\mathbf{p}(\bar{\mathbf{y}}, \boldsymbol{\xi})]_i = p_i(\bar{y}_i, \xi_i)$ is

$$\forall l = 1,...,n \qquad [p_i(\bar{y}_i, \xi_i)]_l = \frac{\exp\left(([\bar{y}_i]_l - C_{l\xi_i})/\gamma\right)}{\sum_{\ell=1}^n \exp\left(([\bar{y}_i]_\ell - C_{\ell\xi_i})/\gamma\right)}. \qquad (3.46)$$

The next theorem presents an application of Theorem 3.2.3 (with changing the constant for the weighted problem) to the Wasserstein barycenter problem.

Theorem 3.3.2. *Let R_λ be such that $\|\lambda^*\| \le R_\lambda$, where λ^* be an exact solution of dual problem (3.38). Let the batch size $r_k = \max\left\{1, \frac{50}{\varepsilon}\lambda_{\max}(W)m\alpha_{k+1}\ln(\frac{N}{\alpha})\right\}$. Then after $N = O\left(\sqrt{\frac{n\|C\|_\infty^2}{\gamma\varepsilon}\chi(W)}\right)$ iterations for the output $\tilde{\mathbf{p}} = (\tilde{p}_1^\top, ..., \tilde{p}_m^\top)^\top$ of Algorithm 7 the following holds with probability at least $1 - 3\alpha$*

$$\frac{1}{m}\sum_{i=1}^m W_\gamma(\tilde{p}_i, q_i) - \frac{1}{m}\sum_{i=1}^m W_\gamma(p^*, q_i) \le \varepsilon, \qquad \|\sqrt{\mathbf{W}}\tilde{\mathbf{p}}\|_2 \le \varepsilon/R_\lambda.$$

Moreover, the per node complexity of Algorithm 7 is

$$\widetilde{O}\left(n \cdot \max\left\{\sqrt{\frac{n\|C\|_\infty^2}{\gamma\varepsilon}\chi(W)}, \frac{n\|C\|_\infty^2}{\varepsilon^2}\chi(W)\right\}\right).$$

Algorithm 7 Decentralized Dual Stochastic Accelerated Gradient Algorithm for WB's

Input: Starting point $\bar{\boldsymbol{\lambda}}^0 = \bar{\boldsymbol{\eta}}^0 = \bar{\boldsymbol{\zeta}}^0 = \mathbf{x}^0 = 0$, number of iterations N, $A_0 = \alpha_0 = 0$,

1: For each agent $i \in V$ $(i = 1, ..., m)$
2: **for** $k = 0, \ldots, N-1$ **do**
3: $A_{k+1} = A_k + \alpha_{k+1} = 2L_\Psi \alpha_{k+1}^2$.
4: $\bar{\lambda}_i^{k+1} = (\alpha_{k+1}\bar{\zeta}_i^k + A_k\bar{\eta}_i^k)/A_{k+1}$.
5: For each $i = 1, ..., m$, calculate $\nabla^{r_{k+1}} W^*_{\gamma,q_i}(\bar{\lambda}_i^{k+1}, \{\xi_i^\ell\}_{\ell=1}^{r_{k+1}})$

$$[\nabla^{r_{k+1}} W^*_{\gamma,q_i}(\bar{\lambda}_i^{k+1}, \{\xi_i^\ell\}_{\ell=1}^{r_{k+1}})]_l = \frac{1}{r_{k+1}} \sum_{\ell=1}^{r_{k+1}} [\nabla W^*_{\gamma,q_i}(\bar{\lambda}_i^{k+1}, \xi_i^\ell)]_l$$
$$= \frac{1}{r_{k+1}} \sum_{\ell=1}^{r_{k+1}} \frac{\exp\left(([\bar{\lambda}_i^{k+1}]_l - C_{l\xi_i^\ell})/\gamma\right)}{\sum_{t=1}^n \exp\left(([\bar{\lambda}_i^{k+1}]_t - C_{t\xi_i^\ell})/\gamma\right)}, \tag{3.44}$$

for all $l = 1, ..., n$ with batch size

$$r_{k+1} = \max\left\{1, 50\lambda_{\max}(W) m \alpha_{k+1} \ln(N/\alpha)/\varepsilon\right\}.$$

6: $\bar{\zeta}_i^{k+1} = \bar{\zeta}_i^k - \alpha_{k+1} \sum_{j=1}^m \mathbf{W}_{ij} \nabla^{r_{k+1}} W^*_{\gamma,q_j}(\bar{\lambda}_j^{k+1}, \{\xi_i^\ell\}_{\ell=1}^{r_{k+1}})$.
7: $\bar{\eta}_i^{k+1} = (\alpha_{k+1}\bar{\zeta}_i^{k+1} + A_k\bar{\eta}_i^k)/A_{k+1}$.
8: **end for**

Output: $\tilde{\mathbf{p}} = (\tilde{p}_1^\top, ..., \tilde{p}_m^\top)^\top$, where $\tilde{p}_i = \frac{1}{A_N} \sum_{k=0}^N \alpha_k p_i(\bar{\lambda}_i^k, \{\xi_i^\ell\}_{\ell=1}^{r_k})$ for all $i = 1, \ldots, m$ with

$$p_i(\bar{\lambda}_i^k, \{\xi_i^\ell\}_{\ell=1}^{r_k}) \triangleq \frac{1}{r_k} \sum_{\ell=1}^{r_k} p_i(\bar{\lambda}_i^k, \xi_i^\ell) = \nabla^{r_k} W^*_{\gamma,q_i}(\bar{\lambda}_i^k, \{\xi_i^\ell\}_{\ell=1}^{r_k}).$$

Proof. The proof of the theorem follows from the Theorem 3.2.3. Thus, we have the following number of iterations

$$N = O\left(\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}\right) \overset{Th.2.3.1}{=} O\left(\sqrt{\frac{n\|C\|_\infty^2}{\gamma\varepsilon}\chi(W)}\right).$$

The number of oracle calls of $\nabla W^*_{\gamma,q_i}(\bar{y}_i, \xi_i)$ is (Theorem 3.2.3)

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}, \frac{M^2\sigma_\psi^2}{\varepsilon^2}\chi(W)\right\}\right), \tag{3.47}$$

where σ_ψ^2 is sub-Gaussian variance of $\nabla W^*_{\gamma,q_i}(\bar{y}_i, \xi_i)$. Now we estimate variance σ_ψ^2

of $\nabla W^*_{\gamma,q_i}(\bar{y}_i, \xi_i)$ (3.46)

$$\begin{aligned}\sigma_\psi^2 &= \max_{\bar{y}_i} \left\{ \mathbb{E}\|p_i(\bar{y}_i, \xi_i)\|_2^2 - \left(\mathbb{E}\|p_i(\bar{y}_i, \xi_i)\|_2\right)^2 \right\} \\ &\leq \max_{\bar{y}_i} \left\{ \mathbb{E}\|p_i(\bar{y}_i, \xi_i)\|_2^2 \right\} \leq \max_{\bar{y}_i} \left\{ \mathbb{E}\|p_i(\bar{y}_i, \xi_i)\|_1^2 \right\} = 1.\end{aligned}$$

Thus, we have $\sigma_\psi^2 \leq 1$. Using this and $M \leq \sqrt{n}\|C\|_\infty$ (Theorem 2.3.1) in (3.47) we get

$$\widetilde{O}\left(\max\left\{ \sqrt{\frac{n\|C\|_\infty^2}{\gamma\varepsilon}\chi(W)}, \frac{n\|C\|_\infty^2}{\varepsilon^2}\chi(W) \right\} \right).$$

Multiplying this by the cost for calculating $\nabla W^*_{\gamma,q_i}(\bar{y}_i, \xi_i)$, which is $O(n)$, we get the per node complexity

$$\widetilde{O}\left(n \cdot \max\left\{ \sqrt{\frac{n\|C\|_\infty^2}{\gamma\varepsilon}\chi(W)}, \frac{n\|C\|_\infty^2}{\varepsilon^2}\chi(W) \right\} \right).$$

□

Chapter 4

Saddle Point Approach for the Wasserstein Barycenter Problem

In this Chapter, we provide a primal algorithm to compute unregularized Wasserstein barycenters with no limitations in contrast to the regularized-based methods, which are numerically unstable under a small value of the regularization parameter. The algorithm is based on the saddle point problem reformulation and the application of mirror prox algorithm with a specific norm. We also show how the algorithm can be executed in a decentralized manner. The complexity of the proposed algorithms meets the best known results in decentralized and non-decentralized setting.

Previous Works. Optimal transport problem (OT) (1.1) is not an easy task. Indeed, to solve this problem between two discrete histograms of size n, one needs to make $\widetilde{O}(n^3)$ arithmetic calculations (Tarjan, 1997; Peyré et al., 2019), 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 $\widetilde{O}(n^2\|C\|_\infty^2/\varepsilon^2)$ convergence rate (Altschuler et al., 2017; Dvurechensky et al., 2018b) 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. Also a faster practice convergence is achieved also by modifications of the Sinkhorn's algorithm, e.g., the Greenkhorn algorithm (Altschuler et al., 2017) of the same convergence rate as the Sinkhorn's algorithm.

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 et al., 2019; 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). Table 4.1, incorporates the most popular algorithms solving OT problem.

Table 4.1: Algorithms for OT problem and their rates of convergence

Paper	**Approach**	**Complexity**
(Dvurechensky et al., 2018b)	Sinkhorn	$\widetilde{O}\left(\frac{n^2\|C\|_\infty^2}{\varepsilon^2}\right)$
(Guminov et al., 2019)	Accelerated Sinkhorn	$\widetilde{O}\left(\frac{n^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
(Jambulapati et al., 2019)	Optimal algorithm based on dual extrapolation with area-convexity	$\widetilde{O}\left(\frac{n^2\|C\|_\infty}{\varepsilon}\right)$

Wasserstein barycenter (WB) problem (1.3) of m measures consists in minimizing the sum of m squared 2-Wasserstein distances (generated by OT metric) to all objects in the set. 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 $\widetilde{O}\left(mn^2\|C\|_\infty^2/\varepsilon^2\right)$ (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 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 $\widetilde{O}\left(mn^2\sqrt[3]{n}\|C\|_\infty^{4/3}/\varepsilon^{4/3}\right)$.

Contribution. We propose a new algorithm, based on mirror prox with specific prox-function, for the WB problem 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. Moreover, this algorithm can be performed in a decentralized manner.

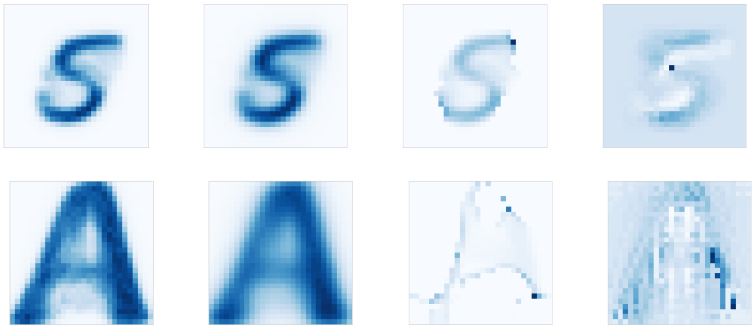
Table 4.2 illustrates the contribution by comparing our new algorithm, called 'Mirror prox with specific norm', with the most popular algorithms for the WB problem. Algorithm 'Dual extrapolation with area-convexity' was proposed in joint paper (Dvinskikh and Tiapkin, 2021) together with 'Mirror prox with specific norm' as an improved version of 'Mirror prox with specific norm' under the weaker convergence requirements of area-convexity. 'Dual extrapolation with area-convexity'

has the best theoretical rate of convergence for the Wasserstein barycenter problem, which is probably optimal. However, it does not have so obvious decentralized interpretation which 'Mirror prox with specific norm' has.

Table 4.2: Algorithms for the WB problem and their rates of convergence

Approach	**Paper**	**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	(Dvinskikh and Tiapkin, 2021)	$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
Dual extrapolation with area-convexity	(Dvinskikh and Tiapkin, 2021)	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty}{\varepsilon}\right)$

Figure 4.1 illustrates numerically instability of the IBP with regularizing parameter γ algorithm when a high-precision ε of calculating Wasserstein barycenters is desired since γ must be selected proportional to ε (Peyré et al., 2019; Kroshnin et al., 2019). 'Dual extrapolation with area-convexity' and 'Mirror prox with specific norm' (Dvinskikh and Tiapkin, 2021) produce good results.



Mirror Prox for WB | Dual Extrapolation | IBP, $\gamma = 10^{-3}$ | IBP, $\gamma = 10^{-5}$

Figure 4.1: Wasserstein barycenters of hand-written digits '5' from the MNIST dataset (first row) and Wasserstein barycenters of letters 'A' from the notMNIST dataset (second row).

Figure 4.2 demonstrates better approximations of the true Gaussian barycenter by 'Dual extrapolation with area-convexity' and 'Mirror prox with specific norm' 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.

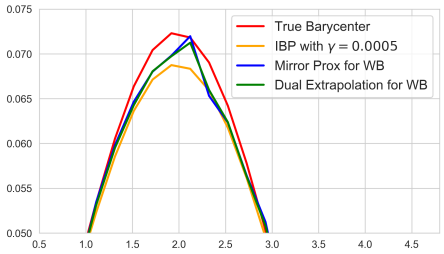


Figure 4.2: Convergence of the barycenters to the true barycenter of Gaussian measures.

The algorithm 'Mirror prox with specific norm' can be also preformed in a decentralized manner and has the same per node complexity as Decentralized FGD (Dvinskikh et al., 2019) up to the dependence on communication matrix. For the star network, we can compare the complexity of decentralized mirror-prox with the complexity of the IBP running in $\widetilde{O}\left(n^2/\varepsilon^2\right)$ time per node (Kroshnin et al., 2019). Decentralized mirror-prox has better dependence on ε, namely $1/\varepsilon$, as well as the accelerated IBP with $\widetilde{O}\left(n^2\sqrt{n}/\varepsilon\right)$ complexity per node of (Guminov et al., 2019). The details of the comparison can be found in Table 4.3

Table 4.3: Distributed algorithms for the WB problem and their per node complexity

Approach	**Paper**	**Architecture**	**Complexity per node**
IBP	(Kroshnin et al., 2019)	star	$\widetilde{O}\left(\frac{n^2\|C\|_\infty^2}{\varepsilon^2}\right)$
Accelerated IBP	(Guminov et al., 2019)	star	$\widetilde{O}\left(\frac{n^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
FastIBP	(Lin et al., 2020)	star	$\widetilde{O}\left(\frac{n^2\sqrt[3]{n}\|C\|_\infty^{4/3}}{\varepsilon\sqrt[3]{\varepsilon}}\right)$
Decentralized FGD	(Dvinskikh et al., 2019)	any	$\widetilde{O}\left(\frac{n^2\sqrt{n\chi(W)}\|C\|_\infty}{\varepsilon}\right)$
Decentralized mirror prox with specific norm	(Rogozin et al., 2021b)	any	$\widetilde{O}\left(\frac{n^2\sqrt{n\chi(W)}\|C\|_\infty^{3/2}}{\varepsilon}\right)$

4.1 Mirror Prox for Wasserstein Barycenters

Our new approach is based on mirror prox algorithm with specific prox-function for the Wasserstein barycenter problem formulated as a saddle-point problem. To present the Wasserstein barycenter problem as a saddle-point problem, we refer to the work (Jambulapati et al., 2019), where the authors obtain saddle-point

representation for optimal transport problem. To so, they vectorize the cost matrix and transport plan.

4.1.1 Saddle Point Formulation

We consider optimal transport problem (1.1) between two discrete measures $\mathtt{p} = \sum_{i=1}^n p_i \delta_{z_i}$ and $\mathtt{q} = \sum_{i=1}^n p_i \delta_{y_i}$ of support size n. The histograms p and q are from the probability simplex Δ_n. Let d be vectorized cost matrix C, and let x be vectorized transport plan $\pi \in U(p,q) \triangleq \{\pi \in \mathbb{R}_+^{n\times n} : \pi \mathbf{1}_n = p, \pi^T \mathbf{1}_n = q\}$. Due to the marginals p, q of transport plan π are from probability simplex Δ_n, it holds that $\sum_{i,j=1}^n \pi_{ij} = 1$. We also introduce $b = \begin{pmatrix} p \\ q \end{pmatrix}$ and incidence matrix $A = \{0,1\}^{2n\times n^2}$. Then the optimal transport problem (1.1) can be rewrutten as

$$\min_{Ax=b,\ x\in\Delta_{n^2}} d^\top x.$$

And then based on the definition of the ℓ_1-norm, this problem can be presented as a saddle-point problem (Jambulapati et al., 2019)

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

Using this representation for optimal transport problem we present the Wasserstein barycenter problem of histograms $q_1, q_2, ..., q_m \in \Delta_n$ as follows

$$\min_{p\in\Delta_n} \frac{1}{m} \sum_{i=1}^m \min_{x_i\in\Delta_{n^2}} \max_{y_i\in[-1,1]^{2n}} \{d^\top x_i + 2\|d\|_\infty \left(y_i^\top A x_i - b_i^\top y_i\right)\}, \tag{4.1}$$

where $b_i = \begin{pmatrix} p \\ q_i \end{pmatrix}$. 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 $\underbrace{\Delta_{n^2} \times \ldots \times \Delta_{n^2}}_{m} \times \Delta_n$, and present (4.1) 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

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{m} \left\{ \boldsymbol{d}^\top \mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A}\mathbf{x} - \boldsymbol{c}^\top \mathbf{y}\right)\right\}, \tag{4.2}$$

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 $\hat{A} = \mathrm{diag}\{A, ..., A\}$ of m blocks, and matrix

$$\mathcal{E}^\top = \begin{pmatrix} \begin{pmatrix} -I_n & 0_{n\times n} \end{pmatrix} \begin{pmatrix} -I_n & 0_{n\times n} \end{pmatrix} \cdots \begin{pmatrix} -I_n & 0_{n\times n} \end{pmatrix} \end{pmatrix}.$$

Since objective $F(\mathbf{x},\mathbf{y})$ in (4.2) is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, problem (4.2) is a saddle-point representation of the Wasserstein barycenter 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

$$\max_{\mathbf{y}\in\mathcal{Y}} F\left(\widetilde{\mathbf{x}}, \mathbf{y}\right) - \min_{\mathbf{x}\in\mathcal{X}} F\left(\mathbf{x}, \widetilde{\mathbf{y}}\right) \leq \varepsilon. \tag{4.3}$$

4.1.2 Algorithm and Convergence Rate

Setups.

- We endow space $\mathcal{Y} \triangleq [-1,1]^{2nm}$ with the standard Euclidean setup: the Euclidean norm $\|\cdot\|_2$, prox-function $d_{\mathcal{Y}}(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2$ and the corresponding Bregman divergence
  $B_{\mathcal{Y}}(\mathbf{y}, \breve{\mathbf{y}}) = \frac{1}{2}\|\mathbf{y} - \breve{\mathbf{y}}\|_2^2$. We define $R_{\mathcal{Y}}^2 = \sup\limits_{\mathbf{y}\in\mathcal{Y}} d_{\mathcal{Y}}(\mathbf{y}) - \min\limits_{\mathbf{y}\in\mathcal{Y}} d_{\mathcal{Y}}(\mathbf{y})$.
- We endow space $\mathcal{X} \triangleq \prod^m \Delta_{n^2} \times \Delta_n$ with 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. We endow $\mathcal{X}$ with prox-function $d_{\mathcal{X}}(\mathbf{x}) = \sum_{i=1}^m \langle x_i, \log x_i\rangle + m\langle p, \log p\rangle$ and corresponding Bregman divergence

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

  We define $R_{\mathcal{X}}^2 = \sup\limits_{\mathbf{x}\in\mathcal{X}} d_{\mathcal{X}}(\mathbf{x}) - \min\limits_{\mathbf{x}\in\mathcal{X}} d_{\mathcal{X}}(\mathbf{x})$.

The next definition clarifies the notion of smoothness for the objective in convex-concave problems.

Definition 4.1.1. *$F(\mathbf{x}, \mathbf{y})$ 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 $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$,*

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

We consider mirror prox (MP) (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 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 $a_1 = \frac{1}{R_{\mathcal{X}}^2}$, $a_2 = \frac{1}{R_{\mathcal{Y}}^2}$

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

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

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

If $F(\mathbf{x}, \mathbf{y})$ is $(L_{\mathbf{xx}}, L_{\mathbf{xy}}, L_{\mathbf{yx}}, L_{\mathbf{yy}})$-smooth, then to satisfy (4.3) 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

$$N = \frac{4}{\varepsilon} \max\{L_{\mathbf{xx}} R_{\mathcal{X}}^2, L_{\mathbf{xy}} R_{\mathcal{X}} R_{\mathcal{Y}}, L_{\mathbf{yx}} R_{\mathcal{Y}} R_{\mathcal{X}}, L_{\mathbf{yy}} R_{\mathcal{Y}}^2\}) \tag{4.4}$$

iterations of the MP (Nemirovski, 2004; Bubeck, 2014) with

$$\eta = 1/(2\max\{L_{\mathbf{xx}} R_{\mathcal{X}}^2, L_{\mathbf{xy}} R_{\mathcal{X}} R_{\mathcal{Y}}, L_{\mathbf{yx}} R_{\mathcal{Y}} R_{\mathcal{X}}, L_{\mathbf{yy}} R_{\mathcal{Y}}^2\}). \tag{4.5}$$

Lemma 4.1.2. *Objective $F(\mathbf{x}, \mathbf{y})$ in (4.2) 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 (4.2) 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 4.1.1. 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 the Setup we have

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

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

$$\|\boldsymbol{A}^\top(\mathbf{y} - \mathbf{y}')\|_{\mathcal{X}^*} \leq L_{\mathbf{xy}} \|\mathbf{y} - \mathbf{y}'\|_2. \tag{4.6}$$

By the definition of dual norm we have

$$\|\boldsymbol{A}^\top(\mathbf{y} - \mathbf{y}')\|_{\mathcal{X}^*} = \max_{\|\mathbf{x}\|_{\mathcal{X}} \leq 1} \langle \mathbf{x}, \boldsymbol{A}^\top(\mathbf{y} - \mathbf{y}')\rangle. \tag{4.7}$$

As $\langle \mathbf{x}, \boldsymbol{A}^\top(\mathbf{y} - \mathbf{y}')\rangle$ is a linear function, (4.6) can be rewritten using (4.7) as

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

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

$$L_{\mathbf{xy}} = \max_{\|\tilde{\mathbf{y}}\|_2 \leq 1} \max_{\|\mathbf{x}\|_{\mathcal{X}} \leq 1} \langle \boldsymbol{A}\mathbf{x}, \tilde{\mathbf{y}}\rangle. \tag{4.8}$$

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 (4.8) as follows

$$L_{\mathbf{xy}} = \max_{\|\mathbf{x}\|_{\mathcal{X}} \leq 1} \|\boldsymbol{A}\mathbf{x}\|_2. \tag{4.9}$$

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

$$\|\boldsymbol{A}\mathbf{x}\|_2^2 = \sum_{i=1}^{m} \left\| Ax_i - \begin{pmatrix} p \\ 0 \end{pmatrix} \right\|_2^2 \leq \sum_{i=1}^{m} \|Ax_i\|_2^2 + m\|p\|_2^2. \tag{4.10}$$

Algorithm 8 Mirror Prox for the Wasserstein Barycenter Problem

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

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

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

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

4:
$$v_i^{k+1} = y_i^k + \alpha\left(Ax_i^k - \begin{pmatrix} p^k \\ q_i \end{pmatrix}\right),$$
Project v_i^{k+1} onto $[-1, 1]^{2n}$

5:
$$u_i^{k+1} = \frac{x_i^k \odot \exp\left\{-\gamma\left(d + 2\|d\|_\infty A^\top y_i^k\right)\right\}}{\sum\limits_{l=1}^{n^2}[x_i^k]_l \exp\left\{-\gamma\left([d]_l + 2\|d\|_\infty [A^\top y_i^k]_l\right)\right\}}$$

6: **end for**

7:
$$s^{k+1} = \frac{p^k \odot \exp\left\{\beta \sum_{i=1}^m [y_i^k]_{1...n}\right\}}{\sum_{l=1}^n [p^k]_l \exp\left\{\beta \sum_{i=1}^m [y_i^k]_l\right\}}$$

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

9:
$$y_i^{k+1} = y_i^k + \alpha\left(Au_i^{k+1} - \begin{pmatrix} s^{k+1} \\ q_i \end{pmatrix}\right)$$
Project y_i^{k+1} onto $[-1, 1]^{2n}$

10:
$$x_i^{k+1} = \frac{x_i^k \odot \exp\left\{-\gamma\left(d + 2\|d\|_\infty A^\top v_i^{k+1}\right)\right\}}{\sum\limits_{l=1}^{n^2}[x_i^k]_l \exp\left\{-\gamma\left([d]_l + 2\|d\|_\infty [A^\top v_i^{k+1}]_l\right)\right\}}$$

11: **end for**

12:
$$p^{k+1} = \frac{p^k \odot \exp\left\{\beta \sum_{i=1}^m [v_i^{k+1}]_{1...n}\right\}}{\sum_{l=1}^n [p^k]_l \exp\left\{\beta \sum_{i=1}^m [v_i^{k+1}]_l\right\}}$$

13: **end for**

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

The last bound holds due to $\langle Ax_i, (p^\top, 0_n^\top)^\top \rangle \geq 0$ since the entries of A, x, p are

non-zero. By the definition of vector $\mathbf{x}$ we have

$$\begin{aligned}\max_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}\|\boldsymbol{A}\mathbf{x}\|_2^2 &= \max_{\|\mathbf{x}\|_{\mathcal{X}}^2\leq 1}\|\boldsymbol{A}\mathbf{x}\|_2^2 = \max_{\sum_{i=1}^m\|x_i\|_1^2+m\|p\|_1^2\leq 1}\|\boldsymbol{A}\mathbf{x}\|_2^2 \\ &\overset{(4.22)}{=} \max_{\alpha\in\Delta_{m+1}}\left(\sum_{i=1}^m\max_{\|x_i\|_1\leq\sqrt{\alpha_i}}\|Ax_i\|_2^2 + \max_{\|p\|_1\leq\sqrt{\frac{\alpha_{m+1}}{m}}} m\|p\|_2^2\right) \\ &= \max_{\alpha\in\Delta_{m+1}}\left(\sum_{i=1}^m\alpha_i\max_{\|x_i\|_1\leq 1}\|Ax_i\|_2^2 + \max_{\|p\|_1\leq 1}\alpha_{m+1}\|p\|_2^2\right). \end{aligned} \tag{4.11}$$

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,

$$\|Ax_i\|_2^2 = \|h_1\|_2^2 + \|h_2\|_2^2 \leq \|h_1\|_1^2 + \|h_2\|_1^2 = 2. \tag{4.12}$$

For the second term in the r.h.s. of (4.11) we have

$$\max_{\|p\|_1\leq 1}\alpha_{m+1}\|p\|_2^2 \leq \max_{\|p\|_1\leq 1}\alpha_{m+1}\|p\|_1^2 = \alpha_{m+1}. \tag{4.13}$$

Using (4.12) and (4.13) in (4.11) we get

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

Using this for (4.9) 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 of the MP algorithm for the Wasserstein barycenter problem with prox-function $d_{\mathcal{Z}}(\mathbf{z})$. For this particular problem, formulated as a saddle-point problem (4.2), the MP algorithm has closed-form solutions presented in Algorithm 8.

Theorem 4.1.3. *Assume that $F(\mathbf{x}, \mathbf{y})$ in (4.2) is $(0, 2\sqrt{2}\|d\|_\infty/m, 2\sqrt{2}\|d\|_\infty/m, 0)$-smooth and $R_{\mathcal{X}} = \sqrt{3m\log n}$, $R_{\mathcal{Y}} = \sqrt{mn}$. Then after $N = 8\|d\|_\infty\sqrt{6n\log n}/\varepsilon$ iterations, Algorithm 8 with $\eta = \frac{1}{4\|d\|_\infty\sqrt{6n\log n}}$ outputs a pair $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}}) \in (\mathcal{X}, \mathcal{Y})$ such that*

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

The total complexity of Algorithm 8 is

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

Proof. By Lemma 4.1.2, $F(\mathbf{x}, \mathbf{y})$ 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 (4.4) and (4.5).

The complexity of one iteration of Algorithm 8 is $O(mn^2)$ 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 4.1.3 with respect to C as $O\left(mn^2\sqrt{n \log n}\|C\|_\infty \varepsilon^{-1}\right)$.

Moreover, the complexity results may be improved by $\sqrt{n}$ term (Dvinskikh and Tiapkin, 2021).

Theorem 4.1.4. *(Dvinskikh and Tiapkin, 2021) Dual Extrapolation algorithm with area-convexity after*

$$N = 8\|d\|_\infty(60 \log n + 9\|d\|_\infty)/\varepsilon$$

iterations outputs a pair $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}}) \in (\mathcal{X}, \mathcal{Y})$ *such that*

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

It can be done in wall-clock time $\widetilde{O}(mn^2\|d\|_\infty\varepsilon^{-1})$.

4.2 Decentralized Mirror Prox for Wasserstein Berycenters

4.2.1 Decentralized Saddle-Point Formulation

To present the Mirror Prox algorithm for the Wasserstein Barycenter problem in a decentralized manner, we rewrite problem (4.1) by introducing artificial constraints $p_1 = p_2 = ... = p_m$ as follows

$$\frac{1}{m}\sum_{i=1}^{m} \min_{\substack{p\in\Delta_n,\\ p_1=...=p_m}} \min_{x_i\in\Delta_{n^2}} \max_{y_i\in[-1,1]^{2n}} \{d^\top x_i + 2\|d\|_\infty \left(y_i^\top A x_i - b_i^\top y_i\right)\}. \tag{4.14}$$

Next we rewrite this problem for the stacked column vectors $\mathbf{p} = (p_1^\top \in \Delta_n, \cdots, p_m^\top \in \Delta_n)^\top \in \mathcal{P} \triangleq \prod^m \Delta_n$, $\mathbf{x} = (x_1^\top \in \Delta_{n^2}, \ldots, x_m^\top \in \Delta_{n^2})^\top \in \mathcal{X} \triangleq \prod^m \Delta_{n^2}$ (where $\prod^m \Delta_{n^2}$ is the Cartesian product of m simplices), and $\mathbf{y} = (y_1^\top, \ldots, y_m^\top)^\top \in \mathcal{Y} \triangleq [-1, 1]^{2mn}$. Then we rewrite the objective in (4.14) without normalizing factor $1/m$. We intend to minimize this objective with accuracy $m\varepsilon$.

$$\min_{\substack{\mathbf{x}\in\mathcal{X},\\ \mathbf{p}\in\mathcal{P},\\ p_1=...=p_m}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{p}, \mathbf{y}) \triangleq \boldsymbol{d}^\top \mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A}\mathbf{x} - \mathbf{b}^\top \mathbf{y}\right), \tag{4.15}$$

where $\mathbf{b} = (p_1^\top, q_1^\top, ..., p_m^\top, q_m^\top)^\top$, $\boldsymbol{d} = (d^\top, \dots, d^\top)^\top$ and $\boldsymbol{A} = \mathrm{diag}\{A, ..., A\} \in \{0,1\}^{2mn \times mn^2}$ is block-diagonal matrix. To enable distributed computation of this problem, the constraint $p_1 = \dots = p_m$ is replaced by $\mathbf{W}\mathbf{p} = 0$ (matrix W is defined in (3.27)). Finally, we introduce Lagrangian dual variable $\mathbf{z} = (z_1^\top, ..., z_m^\top) \in \mathcal{Z} \triangleq \mathbb{R}^{nm}$, scaled by γ, to constraint $\mathbf{W}\mathbf{p} = 0$ for the problem (4.15) and rewrite it as follows

$$\min_{\substack{\mathbf{x}\in\mathcal{X},\ \mathbf{y}\in\mathcal{Y},\\ \mathbf{p}\in\mathcal{P}\ \mathbf{z}\in\mathbb{R}^{nm}}} \max F(\mathbf{x},\mathbf{p},\mathbf{y},\mathbf{z}) \triangleq \boldsymbol{d}^\top\mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A}\mathbf{x} - \mathbf{b}^\top\mathbf{y}\right) + \gamma\langle \mathbf{z}, \mathbf{W}\mathbf{p}\rangle. \tag{4.16}$$

4.2.2 Algorithm and Convergence Rate

Setup.

- We endow space $\mathcal{V} \triangleq \mathcal{Y} \times \mathcal{Z} \triangleq [-1,1]^{2nm} \times \mathbb{R}^{nm}$ with the standard Euclidean setup: the Euclidean norm $\|\cdot\|_2$, prox-function $d_{\mathbf{v}}(\mathbf{v}) = \frac{1}{2}\|\mathbf{v}\|_2^2$, and the corresponding Bregman divergence $B_{\mathbf{v}} = \frac{1}{2}\|\mathbf{v} - \breve{\mathbf{v}}\|_2^2$. We define $R_{\mathbf{v}}^2 = \max\limits_{\mathbf{v}\in\mathcal{V}\cap B_R(0)} d_{\mathbf{v}}(\mathbf{v}) - \min\limits_{\mathbf{v}\in\mathcal{V}\cap B_R(0)} d_{\mathbf{v}}(\mathbf{v})$. Here $B_R(0)$ is a ball of radius R centered in 0.

- We endow space $\mathcal{U} \triangleq \mathcal{X} \times \mathcal{P} \triangleq \prod^m \Delta_{n^2} \times \prod^m \Delta_n$ with the folllowing norm $\|\mathbf{u}\|_{\mathbf{u}} = \sqrt{\sum_{i=1}^m \|x_i\|_1^2 + \sum_{i=1}^m \|p_i\|_1^2}$, where $\|\cdot\|_1$ is the ℓ_1-norm, prox-function $d_{\mathbf{u}}(\mathbf{u}) = \sum_{i=1}^m \langle x_i, \log x_i\rangle + \sum_{i=1}^m \langle p_i, \log p_i\rangle$, and the corresponding Bregman divergence

$$\begin{aligned} B_{\mathbf{u}}(\mathbf{u}, \breve{\mathbf{u}}) = &\sum_{i=1}^m \langle x_i, \log(x_i/\breve{x}_i)\rangle - \sum_{i=1}^m \mathbf{1}_{n^2}^\top (x_i - \breve{x}_i) \\ &+ \sum_{i=1}^m \langle p_i, \log(p_i/\breve{p}_i)\rangle - \sum_{i=1}^m \mathbf{1}_n^\top (p_i - \breve{p}_i). \end{aligned}$$

  We define $R_{\mathbf{u}}^2 = \max\limits_{\mathbf{u}\in\mathcal{U}} d_{\mathbf{u}}(\mathbf{u}) - \min\limits_{\mathbf{u}\in\mathcal{U}} d_{\mathbf{u}}(\mathbf{u})$,

We consider mirror prox algorithm on space $\mathcal{U} \times \mathcal{V}$ with the prox-function $ad_{\mathbf{u}}(\mathbf{u}) + bd_{\mathbf{v}}(\mathbf{v})$ and the corresponding Bregman divergence $aB_{\mathbf{u}}(\mathbf{u}, \breve{\mathbf{u}}) + bB_{\mathbf{v}}(\mathbf{v}, \breve{\mathbf{v}})$, where $a = \frac{1}{R_{\mathbf{u}}^2}$, $b = \frac{1}{R_{\mathbf{v}}^2}$.

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

$$G(\mathbf{x},\mathbf{p},\mathbf{y},\mathbf{z}) = \begin{pmatrix} \nabla_{\mathbf{x}} F(\mathbf{x},\mathbf{p},\mathbf{y},\mathbf{z}) \\ \nabla_{\mathbf{p}} F(\mathbf{x},\mathbf{p},\mathbf{y},\mathbf{z}) \\ -\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{p},\mathbf{y},\mathbf{z}) \\ -\nabla_{\mathbf{z}} F(\mathbf{x},\mathbf{p},\mathbf{y},\mathbf{z}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{d} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{y} \\ \gamma \mathbf{W}^\top \mathbf{z} - 2\|d\|_\infty \{[y_i]_{1\dots n}\}_{i=1}^m \\ -2\|d\|_\infty (\boldsymbol{A}\mathbf{x} - \mathbf{b}) \\ -\gamma \mathbf{W}\mathbf{p} \end{pmatrix}.$$

Here $[y_i]_{1\dots n}$ is the first n component of vector $y_i \in [-1,1]^{2n}$, and $\{[y_i]_{1\dots n}\}_{i=1}^m$ is a short form of $([y_1]_{1\dots n}, [y_2]_{1\dots n}, ..., [y_m]_{1\dots n})$.

Lemma 4.2.1. *Objective $F(\mathbf{u},\mathbf{v})$ in (4.16) is $(L_{\mathbf{uu}}, L_{\mathbf{uv}}, L_{\mathbf{vu}}, L_{\mathbf{vv}})$-smooth with $L_{\mathbf{uu}} = L_{\mathbf{vv}} = 0$ and $L_{\mathbf{uv}} = L_{\mathbf{vu}} = \sqrt{8\|d\|_\infty^2 + \gamma\lambda_{\max}(W)^2}$.*

Proof of Lemma (4.2.1). As $F(\mathbf{u}, \mathbf{v})$ is bilinear, $L_{\mathbf{uu}} = L_{\mathbf{vv}} = 0$. Next, we estimate $L_{\mathbf{uv}}$ and $L_{\mathbf{vu}}$. By the definition of $L_{\mathbf{uv}}$ and the spaces $\mathcal{U}, \mathcal{V}$ we have

$$\|\nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}) - \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}')\|_{\mathcal{U}^*} \leq L_{\mathbf{uv}} \|\mathbf{v} - \mathbf{v}'\|_2. \tag{4.17}$$

From the definition of dual norm, it follows

$$\|\nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}) - \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}')\|_{\mathcal{U}^*} = \max_{\|\mathbf{u}\|_{\mathcal{U}} \leq 1} \langle \mathbf{u}, \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}) - \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}') \rangle.$$

From this and (4.17) we get

$$\max_{\|\mathbf{u}\|_{\mathcal{U}} \leq 1} \langle \mathbf{u}, \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}) - \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}') \rangle \leq L_{\mathbf{uv}} \|\mathbf{v} - \mathbf{v}'\|_2. \tag{4.18}$$

By the definition of $F(\cdot)$ and $\mathcal{U} = \mathcal{X} \times \mathcal{P}$ we have

$$\nabla_{\mathbf{u}} F = \begin{pmatrix} \nabla_{\mathbf{x}} F \\ \nabla_{\mathbf{p}} F \end{pmatrix} = \begin{pmatrix} \boldsymbol{d} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{y} \\ \gamma \mathbf{W}^\top \mathbf{z} - 2\|d\|_\infty \{[y_i]_{1\ldots n}\}_{i=1}^m \end{pmatrix}.$$

From this and $\mathcal{V} \triangleq \mathcal{Y} \times \mathcal{Z}$,

$$\begin{aligned} \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}) - \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}') &= \begin{pmatrix} 2\|d\|_\infty \boldsymbol{A}^\top (\mathbf{y} - \mathbf{y}') \\ \gamma \mathbf{W}^\top (\mathbf{z} - \mathbf{z}') - 2\|d\|_\infty (\{[y_i - y_i']_{1\ldots n}\}_{i=1}^m) \end{pmatrix} \\ &= \begin{pmatrix} 2\|d\|_\infty \boldsymbol{A} & -2\|d\|_\infty \mathcal{E} \\ 0_{mn \times mn^2} & \gamma \mathbf{W} \end{pmatrix}^\top \begin{pmatrix} \mathbf{y} - \mathbf{y}' \\ \mathbf{z} - \mathbf{z}' \end{pmatrix}, \end{aligned}$$

where $\mathcal{E} \in \{1, 0\}^{2mn \times mn}$ is block-diagonal matrix

$$\mathcal{E} = \begin{pmatrix} \begin{pmatrix} I_n \\ 0_{n \times n} \end{pmatrix} & \cdots & 0_{2n \times n} \\ \vdots & \ddots & \vdots \\ 0_{2n \times n} & \cdots & \begin{pmatrix} I_n \\ 0_{n \times n} \end{pmatrix} \end{pmatrix}.$$

From this it follows that $\nabla_{\mathbf{u}} F(\cdot)$ is linear function in $\mathbf{v} - \mathbf{v}'$, then (4.18) can be rewritten as

$$L_{\mathbf{uv}} = \max_{\|\mathbf{v} - \mathbf{v}'\|_2 \leq 1} \max_{\|\mathbf{u}\|_{\mathcal{U}} \leq 1} \left\langle \mathbf{u}, \begin{pmatrix} 2\|d\|_\infty \boldsymbol{A} & -2\|d\|_\infty \mathcal{E} \\ 0_{mn \times mn^2} & \gamma \mathbf{W} \end{pmatrix}^\top (\mathbf{v} - \mathbf{v}') \right\rangle. \tag{4.19}$$

By the same arguments we can get the same expression for $L_{\mathbf{vu}}$ up to rearrangement of maximums. Next, we use the fact that the ℓ_2-norm is the conjugate norm for the ℓ_2-norm. From this and (4.19) it follows

$$L_{\mathbf{uv}} = \max_{\|\mathbf{u}\|_{\mathcal{U}} \leq 1} \left\| \begin{pmatrix} 2\|d\|_\infty \boldsymbol{A} & -2\|d\|_\infty \mathcal{E} \\ 0_{mn \times mn^2} & \gamma \mathbf{W} \end{pmatrix} \mathbf{u} \right\|_2. \tag{4.20}$$

After that, we write

$$\begin{aligned}\max_{\|\mathbf{u}\|_{\mathcal{U}}\leq 1} & \left\| \begin{pmatrix} 2\|d\|_\infty \boldsymbol{A} & -2\|d\|_\infty \mathcal{E} \\ 0_{mn\times mn^2} & \gamma \mathbf{W} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{p} \end{pmatrix} \right\|_2^2 \\ &= \max_{\|\mathbf{u}\|_{\mathcal{U}}^2\leq 1} \left\| \begin{pmatrix} 2\|d\|_\infty (\boldsymbol{A}\mathbf{x} - \mathcal{E}\mathbf{p}) \\ \gamma \mathbf{W}\mathbf{p} \end{pmatrix} \right\|_2^2 \\ &= \max_{\|\mathbf{x}\|_{\mathcal{X}}^2+\|\mathbf{p}\|_{\mathcal{P}}^2\leq 1} \left(4\|d\|_\infty^2 \|\boldsymbol{A}\mathbf{x} - \mathcal{E}\mathbf{p}\|_2^2 + \gamma^2 \|\mathbf{W}\mathbf{p}\|_2^2\right) \\ &\leq 4\|d\|_\infty^2 \max_{\|\mathbf{x}\|_{\mathcal{X}}^2+\|\mathbf{p}\|_{\mathcal{P}}^2\leq 1} \|\boldsymbol{A}\mathbf{x} - \mathcal{E}\mathbf{p}\|_2^2 + \gamma^2 \max_{\|\mathbf{p}\|_{\mathcal{P}}^2\leq 1} \|\mathbf{W}\mathbf{p}\|_2^2. \end{aligned} \tag{4.21}$$

We consider the first term of the r.h.s. of (4.21) under the minimum

$$\|\boldsymbol{A}\mathbf{x} - \mathcal{E}\mathbf{p}\|_2^2 = \sum_{i=1}^m \left\| Ax_i - \begin{pmatrix} p_i \\ 0_n \end{pmatrix} \right\|_2^2 \leq \sum_{i=1}^m \|Ax_i\|_2^2 + \sum_{i=1}^m \|p_i\|_2^2. \tag{4.22}$$

The last bound holds due to $\langle Ax_i, (p_i^\top, 0_n^\top)\rangle \geq 0$ as the entries of $A, \mathbf{x}, \mathbf{p}$ are non-negative. Next we take the minimum in (4.22)

$$\begin{aligned}\max_{\|\mathbf{x}\|_{\mathcal{X}}^2+\|\mathbf{p}\|_{\mathcal{P}}^2\leq 1} \|\boldsymbol{A}\mathbf{x} - \mathcal{E}\mathbf{p}\|_2^2 &= \max_{\sum_{i=1}^m \left(\|x_i\|_1^2+\|p_i\|_1^2\right)\leq 1} \|\boldsymbol{A}\mathbf{x} - \mathcal{E}\mathbf{p}\|_2^2 \\ &\overset{(4.22)}{\leq} \max_{\alpha\in\Delta_{2m}} \left(\sum_{i=1}^m \max_{\|x_i\|_1\leq\sqrt{\alpha_i}} \|Ax_i\|_2^2 + \sum_{i=1}^m \max_{\|p_i\|_1\leq\sqrt{\alpha_{i+m}}} \|p_i\|_2^2 \right) \\ &= \max_{\alpha\in\Delta_{2m}} \left(\sum_{i=1}^m \alpha_i \max_{\|x_i\|_1\leq 1} \|Ax_i\|_2^2 + \sum_{i=1}^m \alpha_{i+m} \max_{\|p_i\|_1\leq 1} \|p_i\|_2^2 \right). \end{aligned} \tag{4.23}$$

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

$$\|Ax_i\|_2^2 = \|h_1\|_2^2 + \|h_2\|_2^2 \leq \|h_1\|_1^2 + \|h_2\|_1^2 = 2. \tag{4.24}$$

As $p_i \in \Delta_n, \forall i = 1, ..., m$ we have

$$\max_{\|p_i\|_1\leq 1} \|p_i\|_2^2 \leq \max_{\|p_i\|_1\leq 1} \|p_i\|_1^2 = 1. \tag{4.25}$$

Using (4.24) and (4.25) in (4.23) we get

$$\max_{\|\mathbf{x}\|_{\mathcal{X}}^2+\|\mathbf{p}\|_{\mathcal{P}}^2\leq 1} \|\boldsymbol{A}\mathbf{x} - \mathcal{E}\mathbf{p}\|_2^2 \leq \max_{\alpha\in\Delta_{2m}} \left(2\sum_{i=1}^m \alpha_i + \sum_{i=m+1}^{2m} \alpha_i \right) \leq \max_{\alpha\in\Delta_{2m}} 2\sum_{i=1}^{2m} \alpha_i = 2. \tag{4.26}$$

Now we consider the second term of the r.h.s. of (4.21).

$$\max_{\|\mathbf{p}\|_{\mathcal{P}}^2\leq 1} \|\mathbf{W}\mathbf{p}\|_2^2 = \max_{\sum_{i=1}^m \|p_i\|_1^2\leq 1} \|\mathbf{W}\mathbf{p}\|_2^2. \tag{4.27}$$

The set $\sum_{i=1}^m \|p_i\|_1^2 \leq 1$ is contained in the set $\sum_{j=1}^n \sum_{i=1}^m [p_i]_j^2 \leq 1$ as cros-product terms of $\|p_i\|_1^2$ are non-negative. Thus, we can change the constraint in the minimum in (4.27) as follows

$$\max_{\sum_{i=1}^m \|p_i\|_1^2 \leq 1} \|\mathbf{W}\mathbf{p}\|_2^2 \leq \max_{\sum_{j=1}^n \sum_{i=1}^m [p_i]_j^2 \leq 1} \|\mathbf{W}\mathbf{p}\|_2^2 = \max_{\|\mathbf{p}\|_2^2 \leq 1} \|\mathbf{W}\mathbf{p}\|_2^2 \tag{4.28}$$

$$= \max_{\|\mathbf{p}\|_2 \leq 1} \|\mathbf{W}\mathbf{p}\|_2^2 \triangleq \lambda_{\max}(\mathbf{W})^2 = \lambda_{\max}(W)^2. \tag{4.29}$$

The last inequality holds due to $\mathbf{W} \triangleq W \otimes I_n$ and the properties of the Kronecker product for eigenvalues. Using(4.26) and (4.28) in (4.21) for the estimation of $L_{\mathbf{uv}}$ from (4.20), we get

$$L_{\mathbf{uv}} = L_{\mathbf{vu}} = \sqrt{8\|d\|_\infty^2 + \gamma^2 \lambda_{\max}(W)^2}.$$

□

This lemma allows us to obtain the following convergence result

Theorem 4.2.2. *Let $\|\mathbf{z}\|_2^2 \leq R^2$, then $R_{\mathbf{u}} = \sqrt{3m \log n}$ and $R_{\mathbf{v}} = \sqrt{mn + R^2/2}$ with*

$$R^2 = \frac{\|\nabla_{\mathbf{p}} f(\mathbf{x}, \mathbf{p}^*, \mathbf{y})\|_2^2}{\gamma \lambda_{\min}^+(W)} \leq \frac{4mn\|d\|_\infty^2}{\gamma \lambda_{\min}^+(W)},$$

where $\lambda_{\min}^+(W)$ is the minimal positive eigenvalue of W. Then after $N = \frac{4L_{\mathbf{uv}} R_{\mathbf{u}} R_{\mathbf{v}}}{m\varepsilon}$ iterations, Algorithm 9 with $\eta = \frac{1}{2L_{\mathbf{uv}} R_{\mathbf{u}} R_{\mathbf{u}}}$ outputs a pair $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$ such that

$$\max_{\substack{\mathbf{y} \in \mathcal{Y}, \\ \|\mathbf{z}\|_2 \leq R}} F(\widetilde{\mathbf{u}}, \mathbf{y}, \mathbf{z}) - \min_{\substack{\mathbf{x} \in \mathcal{X}, \\ \mathbf{p} \in \mathcal{P}}} F(\mathbf{x}, \mathbf{p}, \widetilde{\mathbf{v}}) \leq \varepsilon.$$

The total complexity of Algorithm 9 per node is

$$O\left(\frac{n^2}{\varepsilon}\sqrt{n \log n}\sqrt{\chi}\|d\|_\infty^{3/2}\right).$$

Proof of Theorem 4.2.2. The constants of smoothness for $F(\mathbf{u}, \mathbf{v})$ follows from Lemma 4.2.1. The bound on duality gap follows from the theory of Mirror-Prox with proper $R_{\mathcal{U}}$, $R_{\mathcal{V}}$ and $L_{\mathbf{uv}}$, $L_{\mathbf{uv}}$, $L_{\mathbf{vu}}$, $L_{\mathbf{vv}}$.

To estimate R, we calculate the ℓ_2-norm of the objective in (4.15)

$$\|\nabla_{\mathbf{p}} f(\mathbf{x}, \mathbf{p}^*, \mathbf{y})\|_2^2 = \|2\|d\|_\infty \{[y_i]_{1\ldots n}\}_{i=1}^m\|_2^2 = \sum_{i=1}^m 4\|d\|_\infty^2 \|[y_i]_{1\ldots n}\|_2^2 \leq 4mn\|d\|_\infty^2.$$

Thus, we get

$$R^2 = \frac{\|\nabla_{\mathbf{p}} f(\mathbf{x}, \mathbf{p}^*, \mathbf{y})\|_2^2}{\gamma \lambda_{\min}^+(W)} \leq \frac{4mn\|d\|_\infty^2}{\gamma \lambda_{\min}^+(W)}.$$

To simplify the expression for $R_{\mathbf{u}}$, $R_{\mathbf{v}}$ and $L_{\mathbf{uv}}$ we use that for any a, b, $a + b \leq 2\max\{a, b\}$ and $\sqrt{a^2 + b^2} \leq \sqrt{2}\max\{a, b\}$:

$$R_{\mathbf{v}} = mn + R^2/2 \leq \sqrt{2mn}\max\left\{1, \frac{2\|d\|_\infty}{\sqrt{\gamma\lambda^+_{\min}(W)}}\right\}$$
$$L_{\mathbf{uv}} \leq \sqrt{2}\max\{\sqrt{8}\|d\|_\infty, \gamma\lambda_{\max}(W)\}$$

The complexity of one iteration of Alg. 9 per node is $O(n^2)$ as the number of non-zero elements in matrix A is $2n^2$. Multiplying this by the number of iterations N we get

$$\begin{aligned} O(n^2 N) &= O\left(n^2 L_{\mathbf{uv}} R_u R_v/(m\varepsilon)\right) \\ &= O\left(\frac{n^2}{m\varepsilon}\sqrt{m\log n}\sqrt{mn}\max\left\{1, \frac{2\|d\|_\infty}{\sqrt{\gamma\lambda^+_{\min}(W)}}\right\}\max\{\sqrt{8}\|d\|_\infty, \gamma\lambda_{\max}(W)\}\right) \\ &= O\left(\frac{n^2}{\varepsilon}\sqrt{n\log n}\max\left\{1, \frac{2\|d\|_\infty}{\sqrt{\gamma\lambda^+_{\min}(W)}}\right\}\max\{\sqrt{8}\|d\|_\infty, \gamma\lambda_{\max}(W)\}\right). \end{aligned}$$

We can minimize this expression over γ to get the minimal total complexity. We take $\gamma = \frac{\sqrt{8}\|d\|_\infty}{\lambda_{\max(W)}}$ and we get the final statement

$$O\left(\frac{n^2}{\varepsilon}\sqrt{n\log n}\sqrt{\chi}\|d\|_\infty^{3/2}\right), \tag{4.30}$$

where we used the notation of the condition number for matrix W: $\chi = \frac{\lambda_{\max}(W)}{\lambda^+_{\min}(W)}$. □

4.2.3 Experiments

Next, we illustrate the work of Algorithm 9. 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) on the Erdős-Rényi random graph with probability of edge creation $p = 0.5$. Figure 4.3 shows the the convergence of Algorithm 9 with respect to the function optimality gap and consensus gap. The slope ration -1 on logarithmic scale fits theoretical dependence of the desired accuracy ε on number of iterations ($N \sim \varepsilon^{-1}$, Theorem 4.2.2).

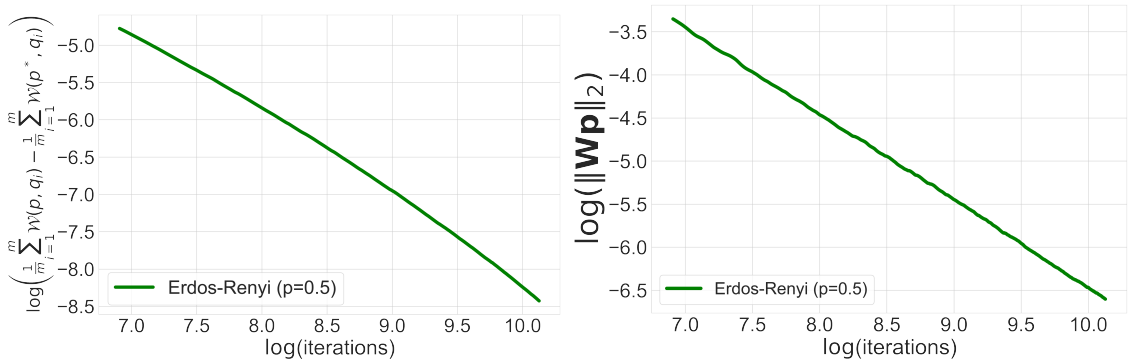


Figure 4.3: Convergence of Decentralized Mirror-Prox for Wasserstein Barycenters

Algorithm 9 Decentralized Mirror-Prox for Wasserstein Barycenters

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

$\alpha = 2\|d\|_\infty \eta(mn + R^2/2)/m$, $\beta = 6\|d\|_\infty \eta \log n$, $\kappa = 3\eta \log n$, $\theta = \eta(mn + R^2/2)/m$

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

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

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

$$s_i^{k+1} = \frac{p_i^k \odot \exp\left\{\beta[y_i^k]_{1...n} - 3\eta\log n \sum_{j=1}^m \gamma W_{ij} z_j^k\right\}}{\sum_{l=1}^n [p_i^k]_l \exp\left\{\beta[y_i^k]_l - 3\eta\log n \left[\sum_{j=1}^m \gamma W_{ij} z_j^k\right]_l\right\}}$$

$$v_i^{k+1} = y_i^k + \alpha\left(Ax_i^k - \begin{pmatrix} p_i^k \\ q_i \end{pmatrix}\right), \quad \text{project } v_i^{k+1} \text{ onto } [-1,1]^{2n}.$$

$$\lambda_i^{k+1} = z_i^k + \theta \sum_{j=1}^m \gamma W_{ij} p_j^k$$

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

$$p_i^{k+1} = \frac{p_i^k \odot \exp\left\{\beta[v_i^{k+1}]_{1...n} - 3\eta\log n \sum_{j=1}^m \gamma W_{ij} \lambda_j^{k+1}\right\}}{\sum_{l=1}^n [p_i^k]_l \exp\left\{\beta[v_i^{k+1}]_l - 3\eta\log n \left[\sum_{j=1}^m \gamma W_{ij} \lambda_j^{k+1}\right]_l\right\}}$$

$$y_i^{k+1} = y_i^k + \alpha\left(Au_i^{k+1} - \begin{pmatrix} s_i^{k+1} \\ q_i \end{pmatrix}\right), \quad \text{project } y_i^{k+1} \text{ onto } [-1,1]^{2n}.$$

$$z_i^{k+1} = z_i^k + \theta \sum_{j=1}^m \gamma W_{ij} s_j^{k+1}$$

  end for

end for

Output: $\widetilde{\mathbf{u}} = \frac{1}{N}\sum\limits_{k=1}^N \begin{pmatrix} u_1^k \\ \vdots \\ u_m^k \\ s_1^k \\ \vdots, s_m^k \end{pmatrix}$, $\widetilde{\mathbf{v}} = \frac{1}{N}\sum\limits_{k=1}^N \begin{pmatrix} v_1^k \\ \vdots \\ v_m^k \\ \lambda_1^k \\ \vdots \\ \lambda_m^k \end{pmatrix}$

Chapter 5

Decentralized Algorithms for Stochastic Optimization

This Chapter has interests other than Wasserstein barycenters, it presents the optimal bounds on the number of communication rounds and dual oracle calls of the gradient of the dual objective per node in the problem of minimizing the sum of strongly convex and Lipschitz smooth functions. This chapter complements Chapter 3 for the case of additionally Lipschitz smooth (stochastic) objectives.

We consider minimizing the average of functions in a distributed manner

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} f_i(x), \tag{5.1}$$

where $f_i(x)$'s are γ-strongly convex and L-Lipschitz smooth. We assume that each $f_i(x)$ has the Fenchel–Legendre representation

$$f_i(x) = \max_{y \in \mathbb{R}^n} \{\langle x, y \rangle - \psi_i(y)\}$$

with convex $\psi_i(y)$. The case when f_i's are dual-friendly (have the Fenchel–Legendre representation) is the case of the Wasserstein barycenter problem (see Chapter 3).

5.1 Dual Approach for Optimization Problem with Affine Constraints

Similarly to Chapter 3, we firstly derive (stochastic) dual algorithms for a general minimization problem with affine constrains where the objective is strongly convex and Lipschitz smooth, and then we show how to execute these algorithms in a decentralized setting for problem (5.1).

We consider a general minimization problem with affine constrains

$$\min_{Ax=b,\ x \in \mathbb{R}^n} F(x), \tag{5.2}$$

where $F(x)$ has L_F-Lipschitz continuous gradient and γ_F-strongly convex in the ℓ_2-norm, $\mathrm{Ker}A \neq \varnothing$. Let $x^* = \arg\min\limits_{\substack{Ax=b,\\ x\in\mathbb{R}^n}} F(x)$.

Remark 2. *We notice that turning to the dual problems does not oblige us using the dual oracle. Instead, we can use the primal oracle and the Moreau theorem (Rockafellar, 2015) with the Fenchel–Legendre representation. The corresponding maximization problem can be solved using the first-order oracle for the primal objective. However, such approach will not contribute to obtaining the optimal bounds on the number of primal first-order oracle calls.*

The dual problem (up to a sign) to (5.2) is the following

$$\min_{y\in\mathbb{R}^n} \Psi(y) \triangleq \max_{x\in\mathbb{R}^n} \{\langle y, Ax - b\rangle - F(x)\}, \tag{5.3}$$

where Ψ is L_Ψ–Lipschitz smooth with $L_\Psi = \frac{\lambda_{\max}(A^TA)}{\gamma_F}$ and γ_Ψ–strongly convex with $\gamma_\Psi = \frac{\lambda^+_{\min}(A^TA)}{L_F}$ in the ℓ_2-norm in $y^0 + (\mathrm{Ker}A^T)^\perp$.[1] The Lipschitz smoothness of dual objective Ψ follows from strong convexity of the primal objective $F(x)$ (Theorem 0.0.6), the strong convexity of Ψ follows from Lipschitz smoothness of $f(x)$ (Kakade et al., 2009; Rockafellar, 2015).

By Demyanov–Danskin theorem we have

$$\nabla\Psi(y) = Ax(A^\top y) - b, \tag{5.4}$$

where

$$x(A^Ty) = \arg\max_{x\in\mathbb{R}^n} \{\langle y, Ax - b\rangle - F(x)\}. \tag{5.5}$$

Let us define $y^* = \arg\min\limits_{y\in\mathbb{R}^n} \Psi(y)$.

For strongly convex objective, fast gradient method (see Chapter 3 for its stochastic version) is not a primal-dual method (Nesterov, 2009; Nemirovski et al., 2010), hence, it cannot be used to solve the primal-dual pair of problems (5.2) and (5.3). The restart technique (Juditsky and Nesterov, 2014; Nemirovskij and Yudin, 1983; Gasnikov, 2017) cannot be also used because the radius of a solution is not a distance from a starting point: $R_y \triangleq \|y^0\|_2 + \|y^0 - y^*\|_2$ (usually we take $y^0 = 0$). The next theorem provides a method to solve the primal-dual pair of problems (5.2) and (5.3).

Theorem 5.1.1. *Let the objective $F(x)$ in (5.2) be L_F-Lipschitz smooth and γ_F-strongly convex in the ℓ_2-norm. Let y^N be an output of the OGM-G algorithm (Kim and Fessler, 2021). Let R_y be such that $\|y^*\|_2 \leq R_y$, where y^* is the solution of (5.3). Then after*

$$\widetilde{O}\left(\sqrt{\frac{L_F}{\gamma_F}\chi(A^\top A)}\right) \tag{5.6}$$

[1] Since $\mathrm{Im}A = (\mathrm{Ker}A^T)^\perp$ we will have that all the points $\tilde{y}^k, z^k, y^k$, generated by fast gradient method and methods based on fast gradient method, belong to $y^0 + (\mathrm{Ker}A^T)^\perp$. That is, from the point of view of estimates this means, that we can consider Ψ to be γ_Ψ-strongly convex everywhere.

iterations (number of oracle calls of $\nabla\Psi(y)$), the following holds for $x^N = x(A^T y^N)$

$$F(x^N) - F(x^*) \leq \varepsilon, \qquad \|Ax^N - b\|_2 \leq \varepsilon / R_y. \tag{5.7}$$

Proof. Let $\Psi(y)$ be dual function for $F(x)$ defined in (5.3). Ψ is L_Ψ–Lipschitz smooth with $L_\Psi = \frac{\lambda_{\max}(A^T A)}{\gamma_F}$ and γ_Ψ–strongly convex with $\gamma_\Psi = \frac{\lambda^+_{\min}(A^T A)}{L_F}$ in the ℓ_2-norm in $y^0 + (\mathrm{Ker} A^T)^\perp$. Then we have (Allen-Zhu, 2018; Anikin et al., 2017; Nesterov, 2012)

$$F(x(A^T y)) - F(x^*) \leq \langle \nabla\Psi(y), y\rangle = \langle Ax(A^\top y) - b, y\rangle. \tag{5.8}$$

Let x^N and y^N be outputs of an algorithm solving the pair of primal-dual problems (5.2) and (5.3) and let R_y be such that $\|y^*\|_2 \leq R_y$. Since the dual objective is strongly convex, the following relation for x^N and y^N holds

$$x^N = x(A^T y^N). \tag{5.9}$$

We have

$$\begin{aligned} F(x^N) - F(x^*) = F(x(A^T y^N)) - F(x^*) \overset{(5.8)}{\leq} \langle \nabla\Psi(y^N), y^N\rangle &\leq \|\nabla\Psi(y^N)\|_2 \|y^N\|_2 \\ &\leq 2R_y \|\nabla\Psi(y^N)\|_2, \end{aligned}$$

where we used the Cauchy—Schwarz inequality and ($\|y^N\|_2 \leq 2R_y$). Hence, to get $F(x^N) - F(x^*) \leq \varepsilon$ from (5.7) we need to prove

$$\|\nabla\Psi(y^N)\|_2 \leq \varepsilon / (2R_y). \tag{5.10}$$

Moreover, from (5.4) and (5.9) it follows that $\nabla\Psi(y^N) = Ax^N - b$. Thus if we get (5.10), we prove (5.7).

In order to prove this, we refer to a method which converges in term of the norm of the gradient, for instance, OGM-G (Kim and Fessler, 2021). It has the following convergence rate

$$\|\nabla\Psi(y^N)\|_2 = O\left(\frac{L_\Psi \|y^0 - y^*\|_2}{N^2}\right) = O\left(\frac{L_\Psi \|\nabla\Psi(y^0)\|_2}{\gamma_\Psi N^2}\right),$$

where we used

$$\frac{\gamma}{2}\|y^0 - y^*\|_2^2 \leq \Psi(y^0) - \Psi(y^*) \leq \frac{1}{2\gamma_\Psi}\|\nabla\Psi(y^0)\|_2^2.$$

Thus, after $\bar{N} = \widetilde{O}\left(\sqrt{\frac{L_\Psi}{\gamma_\Psi}}\right)$ iterations of OGM-G we will have

$$\|\nabla\Psi(y^{\bar{N}})\|_2 \leq \frac{1}{2}\|\nabla\Psi(y^0)\|_2.$$

So after $l = \log_2 \left(\|\nabla \Psi(y^0)\|_2 \frac{R_y}{\varepsilon} \right)$ restarts ($y^0 := y^{\bar{N}}$) we will obtain (5.10). This approach requires

$$O\left(\sqrt{\frac{L_\Psi}{\gamma_\Psi}} \log \left(\|\nabla \Psi(y^0)\|_2 \frac{R_y}{\varepsilon} \right) \right)$$

number of oracle calls of $\nabla \Psi(y)$ (that is $Ax(A^T y) - b$). Using $L_\Psi = \frac{\lambda_{\max}(A^T A)}{L}$ and $\gamma_\Psi = \frac{\lambda^+_{\min}(A^T A)}{L}$, we obtain

$$\widetilde{O}\left(\sqrt{\frac{L_\Psi}{\gamma_\Psi}} \right) = \widetilde{O}\left(\sqrt{\frac{\lambda_{\max}(A^\top A)/\gamma}{\lambda_{\min}(A^\top A)/L}} \right) = \widetilde{O}\left(\sqrt{\frac{L_F}{\gamma_F} \chi(A^\top A)} \right).$$

□

The same result with the replacement

$$\sqrt{\frac{L_\Psi}{\gamma_\Psi}} \log \left(\|\nabla \Psi(y^0)\|_2 \frac{R_y}{\varepsilon} \right) \to \sqrt{\frac{L_\Psi}{\gamma_\Psi}} \log \left(2 L_\Psi^2 \frac{R_y^4}{\varepsilon^2} \right)$$

can be obtained by using fast gradient method for Lipschitz smooth dual objective (but not strongly convex) with bound $\sqrt{\frac{L_\Psi}{\gamma_\Psi}} \log \left(\frac{L_\Psi R_y^2}{\varepsilon'} \right)$ (Nesterov, 2010) and desired accuracy $\varepsilon' = \frac{\varepsilon^2}{2 L_\Psi R_y^2}$. This follows from

$$\frac{1}{2L_\Psi} \|\nabla \Psi(y^N)\|_2^2 \leq \Psi(y^N) - \Psi(y^*) \leq \varepsilon'.$$

5.2 Stochastic Dual Approach for Optimization Problem with Affine Constraints

Now we assume that we are given stochastic oracle $\nabla \Psi(y, \xi)$ with sub-Gaussian variance σ_Ψ^2 (Jin et al., 2019).

$$\begin{aligned} &\mathbb{E} \nabla \Psi(y, \xi) = \nabla \Psi(y) \\ &\mathbb{E} \exp\left(\|\nabla \Psi(y, \xi) - \nabla \Psi(y)\|_2^2 / \sigma_\Psi^2 \right) \leq \exp(1). \end{aligned}$$

Now we consider a method form (Foster et al., 2019) called RRMA+AC-SA^2 (see also (Allen-Zhu, 2018) in the non-accelerated but composite case). This algorithm converges as follows (for simplicity we skip polylogarithmic factors and high probability terminology)

$$\|\nabla \Psi(y^N)\|_2^2 = \widetilde{O}\left(\frac{L_\Psi^2 \|y^0 - y^*\|_2^2}{N^4} + \frac{\sigma_\Psi^2}{N} \right) = \widetilde{O}\left(\frac{L_\Psi^2 \|\nabla \Psi(y^0)\|_2^2}{\gamma_\Psi^2 N^4} + \frac{\sigma_\Psi^2}{N} \right).$$

If we use restart technique of size $\bar{N} = \tilde{O}\left(\sqrt{\frac{L_\Psi}{\gamma_\Psi}}\right)$ and batched gradient with batch size

$$r_{k+1} = \tilde{O}\left(\frac{\sigma_\Psi^2}{\bar{N}\|\nabla\Psi(\bar{y}^{k+1})\|_2^2}\right),$$

where $\bar{y}^k$ is the output from the previous restart, then after $l = O\left(\log_2\left(\|\nabla\Psi(y^0)\|_2\frac{R_y}{\varepsilon}\right)\right)$ restarts we will get

$$\|\nabla\Psi(\bar{y}^l)\|_2 \leq \varepsilon/R_y.$$

Therefore, the total number of stochastic dual oracle calls will be

$$\widetilde{O}\left(\frac{\sigma_\Psi^2 R_y^2}{\varepsilon^2}\right). \tag{5.11}$$

Note that the same bound takes place in the non-strongly convex case ($\gamma_\Psi = 0$). From (Allen-Zhu, 2018; Jin et al., 2019) it is known that this bound cannot be improved.

5.3 Decentralized Optimization

Next we apply the results for minimizing the average of the functions in a distributed setting

$$\min_{x\in\mathbb{R}^n} F(x) \triangleq \frac{1}{m}\sum_{i=1}^m f_i(x), \tag{P1}$$

where f_i's are L-Lipschitz smooth and γ-strongly convex. We seek to solve (P1) on a network of agents in a decentralized manner. To do so, we similarly to Chapter 3 equivalently rewrite (P1) using communication matrix $\mathbf{W}$ defined in (3.27) as follows

$$\min_{\substack{\sqrt{\mathbf{W}}\mathbf{x}=0,\\ x_1,\ldots,x_m\in\mathbb{R}^n}} F(\mathbf{x}) \triangleq \frac{1}{m}\sum_{i=1}^m f_i(x_i), \tag{P2}$$

$\mathbf{x} = (x_1^\top, x_2^\top, , ..., x_n^\top)^\top$ is the stack column vector. We also consider a stochastic version of problem (P2), whose objectives f_i's are given by their expectations: $f_i(x_i) = \mathbb{E}f_i(x_i, \xi_i)$. If f_i's are dual-friendly then we can construct the dual problem to problem (P2) with dual Lagrangian variable $\mathbf{y} = [y_1^T \in \mathbb{R}^n, \cdots, y_m^T \in \mathbb{R}^n]^T \in \mathbb{R}^{mn}$

$$\min_{\mathbf{y}\in\mathbb{R}^{mn}} \Psi(\mathbf{y}) \triangleq \frac{1}{m}\sum_{i=1}^m \psi_i(m[\sqrt{\mathbf{W}}\mathbf{y}]_i), \tag{D2}$$

where $\psi_i(\lambda_i) = \max\limits_{x_i\in\mathbb{R}^n}\{\langle\lambda_i, x_i\rangle - f_i(x_i)\}$ is the Fenchel–Legendre transform of $f_i(x_i)$ and the vector $[\sqrt{\mathbf{W}}\mathbf{x}]_i$ represents the i-th n-dimensional block of $\sqrt{\mathbf{W}}\mathbf{x}$. From the fact that $F(\mathbf{x})$ is L_F–Lipschitz smooth and γ_F–strongly convex it follows that $\Psi(\mathbf{y})$ is L_Ψ–Lipschitz smooth with $L_\Psi = \frac{\lambda_{\max}(W)}{\gamma_F}$ and γ_Ψ–strongly convex with

$\gamma_\Psi = \frac{\lambda^+_{\min}(W)}{L_F}$ in the ℓ_2-norm in $y^0 + (\mathrm{Ker} A^T)^\perp$. Here $L_F = L/m$, $\gamma_F = \gamma/m$. We also consider the stochastic version of problem (D2), whose objectives ψ_i's are given by their expectations $\psi_i(\lambda_i) = \mathbb{E}[\psi_i(\lambda_i,\ \xi_i)]$.

We consider the unbiased stochastic dual oracle returns $\nabla\psi_i(\lambda_i, \xi_i)$ under the following σ^2_ψ-sub-Gaussian variance condition (for all $i = 1, ..., m$)

$$\mathbb{E}\exp\left(\|\nabla\psi_i(\lambda_i, \xi_i) - \nabla\psi_i(\lambda_i)\|_2^2/\sigma^2_\psi\right) \le \exp(1).$$

Problem (D2) can be considered as a particular case of problem (5.2) with with $A = \sqrt{W}$, $b = 0$ and $\sigma^2_\Psi = \lambda_{\max}(W)m\sigma^2_\psi$ (Lemma 3.2.1).

Similarly to Chapter 3 we make the following change of variables

$$\tilde{\mathbf{y}} := \sqrt{\mathbf{W}}\tilde{\mathbf{y}}, \quad \mathbf{z} := \sqrt{\mathbf{W}}\mathbf{z}, \quad \mathbf{y} := \sqrt{\mathbf{W}}\mathbf{y}$$

to present the algorithms of this Chapter solving the pair of primal-dual problems (P2) and (D2) in a decentralized manner. We also need to multiply the corresponding steps in the algorithm by $\sqrt{\mathbf{W}}$.

The bound (5.6) for the pair of decentralized primal-dual problems (P2) and (D2) will change as follows

$$\widetilde{O}\left(\sqrt{\frac{L}{\gamma}\chi(W)}\right),$$

where we used $A = \sqrt{W}$ and the symmetry of $\sqrt{W}$, $L_F = L/m$, $\gamma_F = \gamma/m$ for (P2).

The bound (5.11) for the pair of decentralized primal-dual problems (P2) and (D2) will change as follows

$$\widetilde{O}\left(\max\left\{\frac{\sigma^2_\Psi R_y^2}{\varepsilon^2}, \sqrt{\frac{L}{\gamma}\chi(W)}\right\}\right) = \widetilde{O}\left(\max\left\{\frac{M^2\sigma^2_\psi}{\varepsilon^2}\chi(W), \sqrt{\frac{L}{\gamma}\chi(W)}\right\}\right),$$

where we used $\sigma^2_\Psi = \lambda_{\max}(W)m\sigma^2_\psi$ (Lemma 3.2.1) and (Lan et al., 2017)

$$\|\lambda^*\|_2^2 \le R^2_\lambda = \frac{\|\nabla F(\mathbf{x}^*)\|_2^2}{\lambda^+_{\min}(W)} \le \frac{\left\|\frac{1}{m}\begin{pmatrix}\nabla f_1(x^*)\\ \vdots \\ \nabla f_m(x^*)\end{pmatrix}\right\|_2^2}{\lambda^+_{\min}(W)} = \frac{\sum_{i=1}^m \|\nabla f_i(x^*)\|_2^2}{m^2\lambda^+_{\min}(W)}$$
$$\le \frac{M^2}{m\lambda^+_{\min}(W)}.$$

Tables 5.1 and 5.2 summarize these bounds together with bounds from Chapter 5. Note that the bounds on communication steps (rounds) are optimal (up to a logarithmic factor) due to (Arjevani and Shamir, 2015; Scaman et al., 2017, 2018). Bounds for the oracle calls per node are probably optimal in the class of methods with optimal number of communication steps (up to a logarithmic factor) in the

Table 5.1: The optimal bounds for dual deterministic oracle

Property of f_i	γ-strongly convex, L-smooth	γ-strongly convex, $\|\nabla f_i(x^*)\|_2 \leq M$
The number of communication rounds	$\widetilde{O}\left(\sqrt{\frac{L}{\gamma}\chi(W)}\right)$	$O\left(\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}\right)$
The number of oracle calls of $\nabla\psi_i(\lambda_i)$ per node i	$\widetilde{O}\left(\sqrt{\frac{L}{\gamma}\chi(W)}\right)$	$O\left(\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}\right)$

Table 5.2: The optimal bounds for dual stochastic (unbiased) oracle

Property of f_i	γ-strongly convex, L-smooth	γ-strongly convex, $\|\nabla f_i(x^*)\|_2 \leq M$
The number of communication rounds	$\widetilde{O}\left(\sqrt{\frac{L}{\gamma}\chi(W)}\right)$	$O\left(\sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}\right)$
The number of oracle calls of $\nabla\psi_i(\lambda_i, \xi_i)$ per node i	$\widetilde{O}\left(\max\left\{\frac{M^2\sigma_\psi^2}{\varepsilon^2}\chi(W), \sqrt{\frac{L}{\gamma}\chi(W)}\right\}\right)$	$O\left(\max\left\{\frac{M^2\sigma_\psi^2}{\varepsilon^2}\chi(W), \sqrt{\frac{M^2}{\gamma\varepsilon}\chi(W)}\right\}\right)$

deterministic case (Allen-Zhu, 2018; Foster et al., 2019; Woodworth et al., 2018) and optimal for the non-smooth stochastic primal oracle and stochastic dual oracle for parallel architecture.[2] For stochastic oracle the bounds hold in terms of high probability deviations (we skip the corresponding logarithmic factor).

The detailed proofs of the statements of this Chapter can be found in the arXiv preprint (Gorbunov et al., 2019).

[2]In parallel architecture the bounds on stochastic oracle calls per node of type $\max\{B, D\}$ can be parallel up to B/D processors.

Bibliography

Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). Network flows: theory, algorithms and applications.

Allen-Zhu, Z. (2018). How to make the gradients small stochastically: Even faster convex and nonconvex sgd. In *Advances in Neural Information Processing Systems*, pages 1157–1167.

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.

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

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

Arjevani, Y. and Shamir, O. (2015). Communication complexity of distributed convex learning and optimization. In *Advances in neural information processing systems*, pages 1756–1764.

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

Ballu, M., Berthet, Q., and Bach, F. (2020). Stochastic optimization for regularized Wasserstein estimators. In *International Conference on Machine Learning*, pages 602–612. PMLR.

Ben-Tal, A. and Nemirovski, A. (2001). *Lectures on Modern Convex Optimization*. Society for Industrial and Applied Mathematics.

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.

Bertsekas, D. P. and Tsitsiklis, J. N. (1997). *Parallel and distributed computation: numerical methods*. Athena Scientific, Belmont, MA.

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

Bigot, J., Cazelles, E., and Papadakis, N. (2019b). Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755.

Bigot, J., Cazelles, E., and Papadakis, N. (2019c). Penalization of barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 51(3):2261–2285.

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*.

Boissard, E., Le Gouic, T., Loubes, J.-M., et al. (2015). Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759.

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

Bures, D. (1969). An extension of kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212.

Carlier, G. (2021). On the linear convergence of the multi-marginal sinkhorn algorithm.

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 Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc.

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

Dadush, D. and Huiberts, S. (2018). A friendly smoothed analysis of the simplex method. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 390–403. ACM.

Danskin, J. M. (2012). *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media.

Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Mayo-Íscar, A. (2019). Robust clustering tools based on optimal transportation. *Statistics and Computing*, 29(1):139–160.

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.

Demyanov, V. F. and Malozemov, V. N. (1990). *Introduction to minimax*. Courier Corporation.

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

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279.

Dvinskikh, D. (2020). Stochastic Approximation versus Sample Average Approximation for population Wasserstein barycenters. *arXiv preprint arXiv:2001.07697*.

Dvinskikh, D. and Gasnikov, A. (2021). Decentralized and parallel primal and dual accelerated methods for stochastic convex programming problems. *Journal of Inverse and Ill-posed Problems*.

Dvinskikh, D., Gorbunov, E., Gasnikov, A., Dvurechensky, P., and Uribe, C. A. (2019). On primal and dual approaches for distributed stochastic convex optimization over networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7435–7440. IEEE.

Dvinskikh, D. and Tiapkin, D. (2021). Improved complexity bounds in wasserstein barycenter problem. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1738–1746. PMLR.

Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedić, A. (2018a). Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *Advances in Neural Information Processing Systems 31*, pages 10783–10793. arXiv:1806.03915.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018b). 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*.

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

Foster, D. J., Sekhari, A., Shamir, O., Srebro, N., Sridharan, K., and Woodworth, B. (2019). The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR.

Franklin, J. and Lorenz, J. (1989). On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717 – 735. Special Issue Dedicated to Alan J. Hoffman.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061.

Gabow, H. N. and Tarjan, R. E. (1991). Faster scaling algorithms for general graph matching problems. *Journal of the ACM (JACM)*, 38(4):815–853.

Gasnikov, A. (2017). Universal gradient descent. *arXiv preprint arXiv:1711.00394*.

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

Gasnikov, A. V., Lagunovskaya, A. A., Usmanova, I. N., and Fedorenko, F. A. (2016). Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77(11):2018–2034. arXiv:1412.3890.

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

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

Gramfort, A., Peyré, G., and Cuturi, M. (2015). Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer.

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

Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

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

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A short note on concentration inequalities for random vectors with SubGaussian norm. *arXiv preprint arXiv:1902.03736*.

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

Juditsky, A. and Nemirovski, A. (2012). First order methods for non-smooth convex large-scale optimization, i: General purpose methods. In Suvrit Sra, Sebastian Nowozin, S. W., editor, *Optimization for Machine Learning*, pages 121–184. Cambridge, MA: MIT Press.

Juditsky, A. and Nemirovski, A. S. (2008). Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*.

Juditsky, A. and Nesterov, Y. (2014). Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80.

Juditsky, A., Rigollet, P., Tsybakov, A. B., et al. (2008). Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206.

Kakade, S., Shalev-Shwartz, S., and Tewari, A. (2009). On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09.pdf*, 2(1).

Kakade, S. M. and Tewari, A. (2009). On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808.

Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422.

Kibardin, V. (1979). Decomposition into functions in the minimization problem. *Avtomatika i Telemekhanika*, pages 66–79.

Kim, D. and Fessler, J. A. (2021). Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219.

Klatt, M., Tameling, C., and Munk, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443.

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

Kroshnin, A., Dvinskikh, D., Tupitsa, N., 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.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Lan, G., Lee, S., and Zhou, Y. (2017). Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, pages 1–48.

Lan, G., Nemirovski, A., and Shapiro, A. (2012). Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458.

Li, H., Fang, C., Yin, W., and Lin, Z. (2018). A sharp convergence rate analysis for distributed accelerated gradient methods. *arXiv preprint arXiv:1810.01053*.

Lin, T., Ho, N., Chen, X., Cuturi, M., and Jordan, M. I. (2020). Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. *Advances in Neural Information Processing Systems*, 33.

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

Nedić, A., Olshevsky, A., and Uribe, C. A. (2017). Fast convergence rates for distributed non-bayesian learning. *IEEE Transactions on Automatic Control*, 62(11):5538–5553.

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.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nemirovski, A., Onn, S., and Rothblum, U. G. (2010). Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78.

Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152.

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

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259. First appeared in 2005 as CORE discussion paper 2005/67.

Nesterov, Y. (2010). *Introduction to Convex Optimization*. Moscow, MCCME.

Nesterov, Y. (2012). How to make the gradients small. *Optima*, 88:10–11.

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

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rabin, J. and Papadakis, N. (2015). Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269. Springer.

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

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

Ram, S. S., Veeravalli, V. V., and Nedic, A. (2009). Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM 2009*, pages 3001–3005. IEEE.

Ren, W. (2006). Consensus based formation control strategies for multi-vehicle systems. In *2006 American Control Conference*, pages 6–pp. IEEE.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

Rogozin, A., Bochko, M., Dvurechensky, P., Gasnikov, A., and Lukoshkin, V. (2021a). An accelerated method for decentralized distributed stochastic optimization over time-varying graphs.

Rogozin, A., Dvurechensky, P., Dvinkikh, D., Beznosikov, A., Kovalev, D., and Gasnikov, A. (2021b). Decentralized distributed optimization for saddle point problems. *arXiv preprint arXiv:2102.07758*.

Rolet, A., Cuturi, M., and Peyré, G. (2016). Fast dictionary learning with a smoothed Wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638.

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

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. (2017). Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3027–3036.

Scaman, K., Bach, F., Bubeck, S., Massoulié, L., and Lee, Y. T. (2018). Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems*, pages 2740–2749.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009). Stochastic convex optimization. In *COLT*.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on Stochastic Programming*. Society for Industrial and Applied Mathematics.

Shapiro, A. and Nemirovski, A. (2005). On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer.

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.

Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238.

Spokoiny, V. et al. (2012). Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909.

Srivastava, S., Cevher, V., Dinh, Q., and Dunson, D. (2015). Wasp: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920. PMLR.

Stonyakin, F. S., Dvinskikh, D., Dvurechensky, P., Kroshnin, A., Kuznetsova, O., Agafonov, A., Gasnikov, A., Tyurin, A., Uribe, C. A., Pasechnyuk, D., and Artamonov, S. (2019). Gradient methods for problems with inexact model of the objective. In Khachay, M., Kochetov, Y., and Pardalos, P., editors, *Mathematical Optimization Theory and Operations Research*, pages 97–114, Cham. Springer International Publishing. arXiv:1902.09001.

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

Thorpe, M., Park, S., Kolouri, S., Rohde, G. K., and Slepvcev, D. (2017). A transportation L^p distance for signal analysis. *Journal of mathematical imaging and vision*, 59(2):187–210.

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

Uribe, C. A., Lee, S., Gasnikov, A., and Nedić, A. (2017). Optimal algorithms for distributed optimization. *arXiv preprint arXiv:1712.00232*.

Uribe, C. A., Lee, S., Gasnikov, A., and Nedić, A. (2020). A dual approach for optimal algorithms in distributed optimization over networks. *Optimization Methods and Software*, pages 1–40.

Wang, W., Ozolek, J. A., Slepcev, D., Lee, A. B., Chen, C., and Rohde, G. K. (2010). An optimal transportation approach for nuclear structure-based pathology. *IEEE transactions on medical imaging*, 30(3):621–631.

Weed, J., Bach, F., et al. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648.

Woodworth, B. E., Wang, J., Smith, A., McMahan, B., and Srebro, N. (2018). Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 8505–8515.

Declaration

I declare that I have completed the thesis independently. I have not applied for a doctor's degree in the doctoral subject elsewhere and do not hold a corresponding doctor's degree.

Berlin, 03.05.2021 Darina Dvinskikh