Distributed algorithms with variance reduction for calculating Wasserstein barycenters

Arzhantsev Andrei
Faculty of Computer Science
HSE University
Moscow, Russia
aiarzhantsev_1@edu.hse.ru

***Abstract*—this work is focused on Wasserstein barycenters (WB) problem, which is an important application of optimal transport (OT) problem. WB can be considered as a saddle point optimization problem, for which many methods are developed and studied. Recent articles related to this topic had improvements in complexity bounds for the problem, obtained through the use of optimization methods such as MirrorProx or Dual extrapolation. Other studies in the field of saddle point optimization problems demonstrate that algorithms can be improved due to the proper utilization of variance reduction techniques. Investigation of the non-Euclidean setup of WB problem and generalization of the newest algorithms for it is the main goal of the work. The methodology of constructing a distributed algorithm and a comparison of methods is considered as well.**

***Index Terms*—optimal transport, Wasserstein barycenters, variational inequalities, variance reduction, saddle point optimization, Bregman divergence, distributed optimization**

I. Introduction

“Wasserstein barycenters” is a frequently employed tool in optimal transport problems [1]. For instance, it can be utilized in Probabilistic Multilevel Clustering [2] or representation learning [3] across various tasks. Wasserstein barycenters problem entails seeking a probability distribution that serves as the optimal relative to given probability distributions within the Wasserstein space. The problem is NP-hard [4], thus predominantly iterative methods and their various optimizations are investigated for its solution.

The WB problem can be approached from various configurations and perspectives. Recent research [5] indicates that treating the WB problem as a saddle-point problem and applying Dual Extrapolation has led to improvements over previous state-of-the-art methods for the WB, FastIBP [6] and Accelerated IBP [7]. Aim of this work is to explore algorithms for solving the problem in a non-Euclidean setup, leveraging insights from researchers tackling this problem in Euclidean spaces.

The outcome of this endeavor is a distributed algorithm employing variance reduction techniques for solving the problem, demonstrating superior convergence compared to previous algorithm versions. Beyond the practical utility of the result, due to the accelerated algorithm for solving the WB problem, valuable theoretical knowledge applicable to related tasks is expected as well. Additionally, numerical experiments on relevant datasets will be presented.

To enhance the complexity estimation of the algorithm, in addition to analyzing existing algorithms, exploration of optimization methods for similar problems in general is required in order to identify relevant ideas, correct and adapt them to the specific task. These ideas are mostly connected to various variance reduction techniques and methods of distributed optimization.

The subsequent part of the article is structured as follows. Section 2 will consist of a formal description of the WB problem and its necessary interpretations for subsequent analysis. In Section 3 related research will be reviewed. In section 4 methodology will be explained, along with mentioning plans and current results. Section 5 will be the conclusion.

II. Problem Statement

In this section WB problem will be defined and described. Assume we have two categorical distributions p and q with n variables, and the cost matrix C. Let U be a set of matrices X, where $X\mathbb{1} = p, X^T\mathbb{1} = q$ ($\mathbb{1}$ is $1 \times n$ vector of ones). Then OT problem can be written as an optimization task:

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

where $W(p, q)$ describes distances between two distributions p and q. Wasserstein barycenters use this OT distance to find optimal representatives of distributions in terms of their geometry. Having set of m different categorical distributions $P = \{p_1, p_2, ..., p_m\}$, WB of them can be written as an optimization problem:

$$WB(P) = \operatorname*{argmin}_{p \in P} \frac{1}{m} \sum_{p_i \in P} W(p_i, p)$$

Using techniques provided in literature [8], we can rewrite this optimization problem as follows:

$$WB(P) = \min_{x \in X} \max_{y \in Y} \frac{1}{m} (d^T x + 2\|d\|_\infty (y^T A x c^T y),$$

where X denotes a artesian product of $m + 1$ spaces $(\Delta(n^2),..., \Delta(n2), \Delta(n))$, $Y = [-1, 1]^{2mn}$, $d = (vec(C)^T, ..., vec(C)^T, 0_n^T)$, $c = (0_n^T, q_1^T ..., 0_n^T, q_m^T)$ and A is a block matrix with incidence matrices on the diagonal and $((-I_n, 0_{n\times n}), ..., ((-I_n, 0_{n\times n}))$ in the last row. Full formula and its derivation you can find in article [5].

This problem reformulation shows that WB is a is a saddle-point problem, generally written as

$$\min_{x\in X}\max_{y\in Y} f(x,y) + g1(x) - g2(y)),$$

which gives us the opportunity to use specific methods and techniques for optimization in variational inequalities.

Another necessary point about such problems, is that they can be formulated in different kinds of setups. In a Euclidean setup the domain area, which is in our case (X, Y), is endowed with a Euclidean structure. In Bregman setup Bregman distance is used, which was described in [9]

III. Literature review

As far as we are working with saddle point problem, Variational Inequality Methods can be used for its solution [12]. Variatinal Inequality in general is optimization problems in such form:

$$\text{find } x \in X \text{ such that}$$

$$\forall x_0 \in X \ \langle F(x), x - x_0\rangle + g(x) - g(x_0)) \geq 0,$$

where F is monotone operator and g in convex semicontinuous function. Since many algorithms are developed, we will take a closer look at the most interesting for us methods.

Extragrdient - method for convex optimization and variational inequalities, that is similar to default gradient method, but can find saddle points [11]. It initially was presented for Euclidean space problems, but can be extended to Bregman distances as well. General idea is to make iterations with combination of gradient step and projection on the corresponding space.

MirrorProx - method for variational inequalities, that is strongly connected with dual theory [13].It utilizes dual function and dual space derived from the initial problem setup and can be applied in non-Euclidean setup. The algorithm makes two types of iterations: first is Mirror Descent Step, which minimizes the dual variable and a Bregman divergence term, and Proximal Step, where the objective function and a Bregman divergence term are minimized.

There are definitely some other methods and techniques, that can be used for our task, such as Dual Exterpolation method [10] or forward-backward-forward method.

Since this work is focused on methods with implementation of variance reduction techniques, it's worth pointing out that some necessary methods have already been developed and thus are strongly related to this article.

Variance reduction in general is a technique to reduce the variability of estimates or gradients [15]. It can be done in different ways, including *mini-batching*, where gradient is computed on batch of indices rather than all indices as in casual gradient descent or just one index as in stochastic gradient descent, and "momentum," where the current gradient is updated by incorporating the previous step's gradient estimate. However, it would be more accurate to explore contemporary methods further.

SAGA - a variance reduction technique in which, instead of approximating the gradient itself, a specicial unbiased estimate of this gradient is utilized:

$$g = \nabla f_i(x_i) - \nabla f_i(\hat{x}_i) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\hat{x}_j),$$

here $\hat{x}$ is point where gradient is already calculated.

For the iterative descent method employing this technique, the value of x is not stored; instead, the gradient value at this point v is preserved and updated concurrently, yielding the following algorithm step:

$$g_k = g_{k-1} + \frac{v_i - v_{old}}{n}$$

$$x_{k+1} = x_k - \varepsilon(g_k + v_i - v_{old}),$$

where v_{old} is each step initialized with previos v_i, v_i is i-th component of gradint in correspondning x, and i is chosen randomly each step.

This method has already been successfully applied to the saddle point problem [16]. In addition to the standard method, mini-batching and resampling for coordinates undergoing recalculation were employed.

SVRG - method, which is considered less memory-demanding, then SAGA, because it stores values of referred x instead of gradient values for this point. The downside of this feature is that extra iteration loop is required inside the main loop, and its length is a hyperparameter that should be tuned. In main loop x_0 is initialized with x from previous step, full gradient v in x_0 is calculated and stored . Then internal loop looks like this:

$$g_k = \nabla_i(x_k) - v_i + v$$

$$x_{k+1} = x_k - \varepsilon g_k,$$

where i is chosen randomly for each k. This technique appeared in solutions for saddle point problems as well [16].

Returning to Variatinal Inequality in general, it is advisable to refer to an article that addresses the approach to this problem in a non-Euclidean setup [9]. In this paper, variance reduction techniques have been applied to both Euclidean and Bregman setups. In Euclidean setup, correct implementation of SVRG was tuned with proper sampling and used in gradient step that appears as a part of the whole extragradient method step.

Now Brggman the setup of the problem should be explained. Since the space, in which the variables from the task are located, is not endowed with a Euclidean structure, we will need another distance metric, which is in this case Bregman divergence:

$$D(u, v) = h(u) - h(v) - \langle \nabla h(v), u - v\rangle$$

Here h is a chosen function, with some assumptions needed for the presented method. In general, this setup is similar to Euclidean, but for proposed extragradient method convergence rate wasn't found due to calculations complexity. MirrorProx, as a more useful tool for problems in non-Eucledan setup, provided an easier way to implement SVRG. The whole algorithm is similar to a casual MirrorProx, but it has an additional loop for both Mirror Descent and Proximal Steps with as in casual SVRG. For a better convergence rate a proper sampling was used.

Finally, the article presenting an enhanced distributed algorithm for the Wasserstein barycenters problem will be reviewed [5]. With a reformulation as a saddle point point problem already mentioned above, two new methods were suggested for the problem in Euclidean setup. First method is varianton of MirrorProx, which gave the complexity bound $O(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon})$. Second method is variation of Dual Extrapolation with proven complexity bound $O(\frac{mn^2\|C\|_\infty}{\varepsilon})$. Both solutions reduced the necessary conditions for convergence compared to the past state-of-the-art algorithms, while second reduced convergence rate.

IV. Methodology

The article will explore fundamental methods for solving the Wasserstein barycenters problem or broader tasks, such as MirrorProx, extragradients, and DualExterpolation, and their variations necessary for solving problems in a non-Euclidean setup and improving convergence estimates. The core of the method involves implementing variance reduction techniques and enhancing convergence rates through their implementation. Baseline algorithms for comparison include implementations of Mirror prox with specific norms and Dual extrapolation with area-convexity, described in [5] for the Euclidean setup, as well as methods like Accelerated IBP [7] and Fast IBP [6], which were once state-of-the-art solutions to the problem. Additionally, a method for making the algorithm distributed will be proposed.

The presented algorithm will be examined as follows. Firstly, the formulation of the algorithm, theoretical convergence analysis, and convergence rate estimates will be provided with all necessary theoretical derivations or references to literature.

Secondary, methods comparison in numerical experiments is also considered. While datasets for validation are not finalized yet, experiments can include finding barycenter of Gaussian measures, validation on policeman and burglar matrix or other available matrix optimization problem's data (in entropic setup for non-Euclidean setup) or image datasets such as MNIST or COVTYPE.

The current progress in the work involves researching relevant literature, including the examination of existing optimization methods in the context of the problem, existing variance reduction methods, and their applications in related fields. Valuable insights into solving the problem have been found, such as reformulating it as a saddle point problem, and for utilizing variance reduction methods, such as employing enhanced sampling in SGVR. Currently, hypotheses are being formulated based on this information and will continue to be tested.

V. Conclusion

The Wasserstein barycenters problem, being an application of optimal transport, interests many researchers in the field of optimization methods. The variety of generalizations and potential setups allow this problem to be viewed from different perspectives, enabling the application of various methods and the discovery of improvements in algorithms for its solution. Despite being a highly theoretical research area, it is not devoid of practical applications.

Although this paper primarily elaborates on the theme of the presentation and discusses previous research rather than presenting its own results, within the context of a larger theoretical project, this is valuable, and the project plans are formulated and clear. The enhanced solution to the problem, which will be proposed in the final report, can be considered a relatively significant contribution to the field in general, while the preceding research, along with theoretical derivations and results, may help identify new ideas and methods for future improvements.

Word Count: 1980

References

[1] Martial Agueh, Guillaume Carlier, Barycenters in the Wasserstein Space, January 2011, SIAM Journal on Mathematical Analysis

[2] Nhat Ho, Viet Huynh, Dinh Phung, Michael I. Jordan, Probabilistic Multilevel Clustering via Composite Transportation Distance, October 30, 2018

[3] Sidak Pal Singh, Andreas Hug, Aymeric Dieuleveut, Martin Jaggi, Context Mover's Distance & Barycenters: Optimal Transport of Contexts for Building Representations, February, 2020

[4] Jason M. Altschuler, Enric Boix-Adser'a, Wasserstein barycenters are NP-hard to compute, December, 2020

[5] Darina Dvinskikh, Daniil Tiapkin, Improved Complexity Bounds in Wasserstein Barycenter Problem, October, 2020

[6] Sergey Guminov, Pavel Dvurechensky, Alexander Gasnikov, Accelerated Alternating Minimization, June, 2019

[7] Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, Michael I. Jordan, June 2022

[8] Arun Jambulapati, Aaron Sidford, Kevin Tian, A Direct $\tilde{O}(1/\varepsilon)$ Iteration Parallel Algorithm for Optimal Transport, June, 2019

[9] Ahmet Alacaoglu, Yura Malitsky, Stochastic Variance Reduction for Variational Inequality Methods, June, 2022

[10] Yu. Nesterov, Dual extrapolation and its applications for solving variational inequalities and related problems, August, 2003

[11] G. Korpelevich, The extragradient method for findingsaddle points and other problems, 1976

[12] Anatoli Juditsky, Arkadi Nemirovski, Claire Tauvel, Solving variational inequalities with Stochastic Mirror-Prox, October, 2018

[13] Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, Alexander Gasnikov, Optimal Algorithms for Decentralized Stochastic Variational Inequalities, NeurIPS 2022

[14] Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, Alexander Gasnikov, Optimal Algorithms for Decentralized Stochastic Variational Inequalities, NeurIPS 2022

[15] Robert M. Gowera, Mark Schmidt, Francis Bach, and Peter Richtárikd, Variance-Reduced Methods for Machine Learning, October, 2020

[16] P. Balamurugan, Francis Bach, Stochastic Variance Reduction Methods for Saddle-Point Problems, November, 2016