

Multimarginal Optimal Transport by Accelerated Alternating Minimization

Nazarii Tupitsa, Pavel Dvurechensky, Alexander Gasnikov, César A. Uribe

Abstract— We study multimarginal optimal transport (MOT) problems, which include, as a particular case, the Wasserstein barycenter problem. In MOT problems, one has to find an optimal coupling between m probability measures, which amounts to finding a tensor of order m . We propose a method based on accelerated alternating minimization and estimate the complexity to find an approximate solution. We use entropic regularization with a sufficiently small regularization parameter and apply accelerated alternating minimization to the dual problem. A novel primal-dual analysis is used to reconstruct the approximately optimal coupling tensor. Our algorithm exhibits a better computational complexity than the state-of-the-art methods for some regimes of the problem parameters.

I. INTRODUCTION

Optimal transport (OT) has gained increasing interest in recent years from its broad range of applications ranging from medical image processing [1], machine learning [2], graph-theory [3], control theory [4], among many others. Fundamentally, many of these applications require the comparison and quantification of distances between probability distributions [5]. In Kantorovich formulation, the OT problem seeks to minimize

$$\int_{M_1 \times \dots \times M_m} c(x_1, \dots, x_m) d\pi(x_1, \dots, x_m),$$

over the set $\Pi(p_1, \dots, p_m)$ of positive joint measures π on the product space $M_1 \times \dots \times M_m$ whose marginals are the p_k 's, where p_1, \dots, p_m (marginals) is a set of probability measures on smooth manifolds M_1, \dots, M_m , and $c(x_1, \dots, x_m)$ is a cost function [6].

The work of P. Dvurechensky, A. Gasnikov, and N. Tupitsa in part III-A – III-C, IV was funded by Russian Science Foundation (project 18-71-10108). The work of C.A. Uribe and A. Gasnikov in part III-D was partially funded by the Yahoo! Faculty Engagement Program. The work of P. Dvurechensky in part II was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy [Pleaseinsert"PrerenderUnicode~intopreamble] The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689). The research of N. Tupitsa in part V 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.

N.T. is with the Moscow Institute of Physics and Technology, Institute for Information Transmission Problems and National Research University Higher School of Economics, Russia (tupitsa@phystech.edu). P.D. is with the Weierstrass Institute for Applied Analysis and Stochastics, Germany, and the Institute for Information Transmission Problems, Russia (pavel.dvurechensky@wias-berlin.de). A.G. is with the Moscow Institute of Physics and Technology, Institute for Information Transmission Problems, Russia, and National Research University Higher School of Economics, Russia, and Sirius University of Science and Technology, Russia, and Caucasus Mathematical Center, Adyge State University, Russia. (gasnikov@yandex.ru). C.A.U. is with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, USA (cauribe@mit.edu).

Although the OT problem formulation is mathematically precise, see, for example, the seminal monograph by Villani [7], and references therein, its translation to practical applications heavily depends on the availability of computationally attractive methods. Many of the OT related problems are computationally intense, and much effort has been put into analyzing the underlying complexity of such problems [8]–[11].

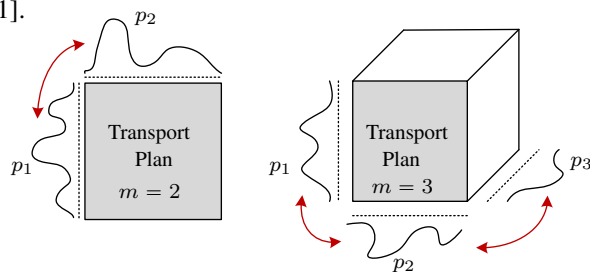


Fig. 1. A visual representation of the multimarginal optimal transport problem for $m = 2$ and $m = 3$. When $m = 2$, the transport plan defines the optimal cost of moving p_1 into p_2 . For discrete distributions this corresponds to a matrix with marginals p_1 and p_2 . When $m = 3$, in the discrete case is, the transport plan is a three dimensional tensor, whose marginals are p_1 , p_2 , and p_3 .

Classically, OT has been studied for quantifying distances between *two* probability distributions (i.e., $m = 2$) for which theory is fairly well understood [7], [12], [13]. However, for $m \geq 3$, i.e., the multimarginal optimal transport (MOT) problem, much less is known, even though such regime has been recently shown useful for many applications, like tomographic image reconstruction [14], generative adversarial networks [15], economics [16], and density functional theory [17]. Figure 1 shows a visual representation of the MOT problem for $m = 3$. See [6] for a recent survey of fundamental theoretical formulations and applications of the MOT problem.

Computational aspects of the MOT problem were studied in [18], where an Iterative Bregman Projections algorithm was proposed for this problem, yet without complexity analysis. It was also pointed out that the MOT problem can be applied to calculate the barycenter of m measures without fixing the barycenter's support. In [19], the authors propose and analyze the complexity of two algorithms for the MOT problem. We follow [19] by using the entropy regularization approach as well [20].

In this paper, we develop an algorithm for the computation of approximate solutions for the MOT problem using recently developed methods of alternating minimization. Our contributions are three-fold:

- We develop a novel algorithm for the approximate computation of MOT maps based on accelerated alternating

minimization algorithm.

- We formally prove the computational complexity of the proposed algorithm. We show that the proposed algorithm has an iteration complexity $\tilde{O}(m^2 n^{1/2}/\varepsilon)$, and a computational complexity of $\tilde{O}(m^3 n^{m+1/2}/\varepsilon)$ arithmetic operations. Our result indicates an upper exponential bound for the Wasserstein barycenter problem's complexity with free support, which is known to be a non-convex optimization problem.
- We show that in some regimes of the MOT problem parameters m (number of distributions), n (dimension of the distributions), and ε (desired accuracy), the proposed algorithm has better iteration complexity in comparison with existing methods.

This paper is organized as follows. Section II presents the problem formulation and the dual aspects of the OT problem. Section III contains the algorithm design methodology and the theoretical primal-dual analysis required for the establishment of the algorithmic complexity. Section V shows some preliminary experiments. Section IV discusses the specific computational complexity results. Finally, Section VI presents the conclusions and future work.

II. THE ENTROPY REGULARIZED MOT PROBLEM

In what follows, Δ^n denotes the probability simplex in \mathbb{R}_+^n : $\Delta^n = \{u \in \mathbb{R}_+^n : \mathbf{1}_n^\top u = 1\}$. For a tensor $A = (A_{i_1, \dots, i_m}) \in \mathbb{R}^{n_1 \times \dots \times n_m}$, we write $\|A\|_\infty = \max_{1 \leq i_k \leq n_j, \forall k \in \{1, \dots, m\}} |A_{i_1, \dots, i_m}|$ and $\|A\|_1 = \sum_{1 \leq i_k \leq n_j, \forall k \in \{1, \dots, m\}} |A_{i_1, \dots, i_m}|$, and denote by $p_k(A) \in \mathbb{R}^{n_k}$ its k -th marginal for $k \in \{1, \dots, m\}$ where each component is defined as

$$[p_k(A)]_j = \sum_{1 \leq i_l \leq n_l, \forall l \neq k} A_{i_1, \dots, i_{k-1}, j, i_{k+1}, \dots, i_m}.$$

For two tensors of the same dimension, we denote the Frobenius inner product of A and B by

$$\langle A, B \rangle = \sum_{1 \leq i_k \leq n_k, \forall k \in \{1, \dots, m\}} A_{i_1, \dots, i_m} B_{i_1, \dots, i_m}.$$

The MOT problem between $m \geq 2$ discrete probability distributions with n support points¹ has the following form:

$$\min_{X \in \mathbb{R}_+^{n \times \dots \times n}, p_k(X) = p_k, \forall k \in \{1, \dots, m\}} \langle C, X \rangle, \quad (1)$$

where X denotes a multimarginal transportation plan and $C \in \mathbb{R}_+^{n \times \dots \times n}$ is a cost tensor. For all $k \in \{1, \dots, m\}$, a vector $p_k = (p_{kj})$ is given as a probability vector in Δ^n .

The MOT problem is a linear program with mn equality constraints, and n^m variables and inequality constraints. When $m = 2$, the MOT problem reduces to the classical OT problem [7].

In the general case of m measures, one of the applications of MOT is grid-free Wasserstein barycenter computation [18]. Despite the linear programming (LP) formulation being in its standard form, the problem's dimension, which is exponential

¹For simplicity we consider same cardinality of the support set for each distribution. This can be extended for general case.

in m , does not allow the use of standard LP solvers such as interior-point methods [21], [22]. Next, we describe how to apply the entropic regularization approach so ameliorate such computational requirements.

Following [18], [20], we consider a regularized version of (1), in which we add an entropic penalty to the multimarginal transportation plan. The resulting problem has the following form:

$$\min_{\substack{X \in \mathbb{R}_+^{n \times \dots \times n}, \\ p_k(X) = p_k, \quad \forall k \in \{1, \dots, m\} \\ \sum_{i_1, \dots, i_m} X_{i_1, \dots, i_m} = 1, \quad 1 \leq i_j \leq n}} F(X) := \langle C, X \rangle - \gamma H(X), \quad (2)$$

where $\gamma > 0$ is the regularization parameter, and $H(X)$ is the entropic regularization term: $H(X) := -\langle X, \log(X) \rangle$. Here logarithm of a tensor should be understood as component-wise. We underline that we add a constraint that X belongs to probability simplex of the size n^m . This constraint is a corollary of the fact that all the vectors p_k , $k = 1, \dots, m$ belong to Δ^n . Adding this constraint does not change the problem's solution, but it is crucial to obtain a dual optimization problem to have a Lipschitz-continuous gradient. The reason for the latter is that entropy is strongly convex on the probability simplex w.r.t. the 1-norm.

The next lemma shows that the entropy regularized MOT problem has a closed-form dual representation that we can exploit for developing computationally efficient approaches.

Lemma 1. *The dual problem formulation of the entropy regularized MOT problem (2) is defined as $\max_{\Lambda} \phi(\Lambda)$, where*

$$\phi(\Lambda) := -\gamma \left[\ln \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} \exp \left\{ -\sum_{k=1}^m \frac{[\lambda_k]_{i_k}}{\gamma} - \frac{C_{i_1 \dots i_m}}{\gamma} - 1 \right\} + 1 + \frac{1}{\gamma} \sum_{k=1}^m \lambda_k^T p_k \right]. \quad (3)$$

Moreover, the primal variable can computed as

$$X_{i_1 \dots i_m}(\Lambda) = \frac{\exp \left(-\sum_{k=1}^m \frac{[\lambda_k]_{i_k}}{\gamma} - \frac{C_{i_1 \dots i_m}}{\gamma} \right)}{\sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} \exp \left\{ -\sum_{k=1}^m \frac{[\lambda_k]_{i_k}}{\gamma} - \frac{C_{i_1 \dots i_m}}{\gamma} \right\}} \quad (4)$$

Finally, with the change of variable $u_k = -\frac{\lambda_k}{\gamma} - \frac{1}{m}$ the dual problem becomes

$$\min_U \phi(U) \equiv \phi(u_1, \dots, u_m). \quad (5)$$

where $U = (u_1^T, \dots, u_m^T)^T \in \mathbb{R}^{mn}$.

All the proofs of this paper can be found in [23].

Proof. We introduce dual variables $\lambda_i \in \mathbb{R}^n$ for $i \in \{1, \dots, m\}$ and define the Lagrangian function as follows:

$$L(X, \Lambda, \mu) = \langle C, X \rangle + \gamma \langle X, \log(X) \rangle + \sum_{k=1}^m \lambda_k^T (p_k(X) - p_k) + \mu \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} X_{i_1 \dots i_m} - \mu, \quad (6)$$

where $\Lambda = (\lambda_1^T, \dots, \lambda_m^T)^T \in \mathbb{R}^{mn}$, and formulate the dual unconstrained problem

$$\max_{\Lambda \in \mathbb{R}^{mn}} \max_{\mu \in \mathbb{R}} \min_{X \in \mathbb{R}_+^{n^m}} L(X, \Lambda, \mu).$$

Taking the derivative with respect to $X_{i_1 \dots i_m}$ and setting it to zero yields

$$\frac{\partial L}{\partial X_{i_1 \dots i_m}}(X, \Lambda, \mu) = C_{i_1 \dots i_m} + \gamma + \gamma \log(X_{i_1 \dots i_m}) + \sum_{k=1}^m [\lambda_k]_{i_k} + \mu = 0. \quad (7)$$

the solution of the above problem is

$$X_{i_1 \dots i_m}(\Lambda, \mu) = \exp\left(\frac{-\sum_{k=1}^m [\lambda_k]_{i_k} - C_{i_1 \dots i_m} - \gamma - \mu}{\gamma}\right).$$

Therefore, we have

$$L(\Lambda, \mu) = -\gamma \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} X_{i_1, \dots, i_m}(\Lambda, \mu) - \sum_{k=1}^m \lambda_k^T p_k - \mu.$$

By taking a derivative w.r.t μ and setting it to zero we have

$$\sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} X_{i_1, \dots, i_m}(\Lambda, \mu(\Lambda)) - 1 = 0.$$

From where we can express $\mu(\Lambda)$ as

$$\exp\left\{-\frac{\mu}{\gamma}\right\} \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} \exp\left\{-\sum_{k=1}^m \frac{\lambda_k i_k}{\gamma} - \frac{C_{i_1 \dots i_m}}{\gamma} - 1\right\} = 1,$$

yielding the theorem's statements.

As it is known [24], the objective in (3) has Lipschitz continuous gradient. This follows from the fact that entropy is strongly convex on the probability simplex. Since the dual objective has Lipschitz gradient, we can use gradient-type of methods to solve the dual problem and obtain the corresponding complexity.

Finally, with the change of variable $u_k = -\frac{\lambda_k}{\gamma} - \frac{1}{m}$ the dual objective becomes

$$\phi(U) \equiv \phi(u_1, \dots, u_m) \equiv \gamma \left[\ln \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} \exp\left\{\sum_{k=1}^m [u_k]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma}\right\} - \sum_{k=1}^m u_k^T p_k \right], \quad (8)$$

where $U = (u_1^T, \dots, u_m^T)^T \in \mathbb{R}^{mn}$.

III. ALGORITHM DESIGN BASED ON THE ALTERNATING MINIMIZATION APPROACH

In this section, we describe the proposed approach for designing an algorithm to approximately solve the MOT problem, based on an alternating minimization approach.

First, we introduce the tensor $B(U) \in \mathbb{R}_+^{n^m}$ with elements given as

$$B_{i_1, \dots, i_m}(u_1, \dots, u_m) = \exp\left\{\sum_{k=1}^m [u_k]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma}\right\},$$

and element-wise sum given as

$$\Sigma(U) = \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n, 1 \leq j \leq m}} B_{i_1, \dots, i_m}(u_1, \dots, u_m).$$

Moreover, it follows that the partial derivatives of the dual function ϕ are

$$\frac{1}{\gamma} \left[\frac{\partial \phi}{\partial u_\xi} \right]_\eta = \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m \\ i_\xi = \eta}} \frac{\exp\left\{\sum_{k=1}^m [u_k]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma}\right\}}{\Sigma(U)} - [p_\xi]_\eta = \frac{[p_\xi(B(U))]_\eta}{\Sigma(U)} - [p_\xi]_\eta. \quad (9)$$

Therefore, as shown in the next lemma, we obtain a closed-form solution for alternating minimization of the dual problem.

Lemma 2. *The iterations*

$$u_k^{t+1} \in \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \phi(u_1^t, \dots, u_{k-1}^t, u, u_{k+1}^t, \dots, u_m^t),$$

can be written explicitly as

$$u_k^{t+1} = u_k^t + \ln p_k - \ln p_k(B(U^t)),$$

or entry-wise as

$$[u_k^{t+1}]_\eta = [u_k^t]_\eta + \ln [p_k]_\eta - \ln [p_k(B(U^t))]_\eta. \quad (10)$$

Proof. Consider the following tensor

$$\begin{aligned} B_{i_1, \dots, i_m}(u_1^t, \dots, u_{\xi-1}^t, u_\xi^{t+1}, u_{\xi+1}^t, \dots, u_m^t) &= \exp\left\{[u_\xi^{t+1}]_{i_\xi} + \sum_{k \neq \xi} [u_k^t]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma}\right\} \\ &= \frac{\exp[u_\xi^{t+1}]_{i_\xi}}{\exp[u_\xi^t]_{i_\xi}} \exp\left\{[u_\xi^t]_{i_\xi} + \sum_{k \neq \xi} [u_k^t]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma}\right\} \\ &= \frac{\exp[u_\xi^{t+1}]_{i_\xi}}{\exp[u_\xi^t]_{i_\xi}} B(U^t), \end{aligned}$$

□

and plug in the expression (10) from the lemma statement

$$\begin{aligned}
& \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} B_{i_1, \dots, i_m}(u_1^t, \dots, u_{\xi-1}^t, u_{\xi}^{t+1}, u_{\xi+1}^t, \dots, u_m^t) \\
&= \sum_{\eta} \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m \\ i_{\xi} = \eta}} B_{i_1, \dots, i_m}(u_1^t, \dots, u_{\xi-1}^t, u_{\xi}^{t+1}, u_{\xi+1}^t, \dots, u_m^t) \\
&= \sum_{\eta} \frac{\exp[u_{\xi}^{t+1}]_{\eta}}{\exp[u_{\xi}^t]_{\eta}} [p_{\xi}(B(U^t))]_{\eta} \\
&\stackrel{(10)}{=} \sum_{\eta} \frac{[p_{\xi}]_{\eta}}{[p_{\xi}(B(U^t))]_{\eta}} [p_{\xi}(B(U^t))]_{\eta} = 1.
\end{aligned}$$

Next, we plug (10) in the optimality conditions $\frac{\partial \phi}{\partial [u_{\xi}]_{\eta}} = 0$ and show that the conditions are satisfied

$$\begin{aligned}
& [p_{\xi}]_{\eta} = \\
& \frac{\exp([u_{\xi}^{t+1}]_{\eta}) \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m \\ i_{\xi} = \eta}} \exp \left\{ \sum_{k \neq \xi} [u_k^t]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma} \right\}}{\sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m}} B_{i_1, \dots, i_m}(u_1^t, \dots, u_{\xi-1}^t, u_{\xi}^{t+1}, u_{\xi+1}^t, \dots, u_m^t)} \\
&= \exp([u_{\xi}^{t+1}]_{\eta}) \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m \\ i_{\xi} = \eta}} \exp \left\{ \sum_{k \neq \xi} [u_k^t]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma} \right\} \\
&= \frac{e^{[u_{\xi}^{t+1}]_{\eta}}}{e^{[u_{\xi}^t]_{\eta}}} \sum_{\substack{\dots \\ i_{\xi} = \eta}} B(U^t) \stackrel{(10)}{=} \frac{[p_{\xi}]_{\eta}}{[p_{\xi}(B(U^t))]_{\eta}} [p_{\xi}(B(U^t))]_{\eta}.
\end{aligned}$$

□

Lemma 2 implies that the dual objective ϕ can be explicitly minimized in each of the m blocks of variables u_k , $k = 1, \dots, m$, suggesting to use alternating minimization algorithms for the dual problem. Note that the nature of the Iterative Bregman Projections algorithm [18] is different since it is an alternating projection algorithm for the primal problem.

A. General Primal-Dual Accelerated Alternating Minimization

In order to analyze the proposed algorithm, first we develop a general framework for primal-dual accelerated alternating minimization. We consider a general minimization problem

$$(P_1) \quad \min_{x \in Q \subseteq E} \{f(x) : \mathcal{A}x = b\},$$

where E is a finite-dimensional real vector space, Q is a simple closed convex set, \mathcal{A} is a given linear operator from E to some finite-dimensional real vector space H , $b \in H$ is given. This problem template, in particular, covers Problem (2). The Lagrange dual problem to Problem (P₁) is

$$(D_1) \quad \max_{\lambda \in \Lambda} \left\{ -\langle \lambda, b \rangle + \min_{x \in Q} (f(x) + \langle \mathcal{A}^T \lambda, x \rangle) \right\}.$$

Here, we define $\Lambda = H^*$. Note also that Problem (3) is a particular case of this general dual template. It is convenient to rewrite Problem (D₁) in the equivalent form of a minimization problem

$$(P_2) \quad \min_{\lambda \in \Lambda} \left\{ \varphi(\lambda) = \langle \lambda, b \rangle + \max_{x \in Q} (-f(x) - \langle \mathcal{A}^T \lambda, x \rangle) \right\}.$$

Since f is convex, $\varphi(\lambda)$ is a convex function. Thus, by Danskin's theorem (see e.g. [24]), its subgradient is

$$\nabla \varphi(\lambda) = b - \mathcal{A}x(\lambda), \quad (11)$$

where $x(\lambda)$ is some solution of the convex problem

$$\max_{x \in Q} (-f(x) - \langle \mathcal{A}^T \lambda, x \rangle). \quad (12)$$

In what follows, we assume that $\varphi(\lambda)$ is L -smooth and that the dual problem (D₁) has a solution λ^* and there exist some $R > 0$ such that $\|\lambda^*\|_2 \leq R$. We underline that the quantity R will be used only in the convergence analysis, but not in the algorithm itself.

To describe our algorithm we also need the following notation. The set $\{1, \dots, N\}$ of indices of the orthonormal basis vectors $\{e_i\}_{i=1}^N$ is divided into m disjoint subsets (blocks) I_k , $k \in \{1, \dots, m\}$. Let $S_k(x) = x + \text{span}\{e_i : i \in I_k\}$, i.e. the affine subspace containing x and all the points differing from x only over the block k .

The idea of the Algorithm 1 is to use greedy alternating minimization steps in the dual and combine them with momentum, as in Nesterov's accelerated methods. This allows us to obtain an accelerated convergence rate for the dual problem. Further, we add a step which updates the primal variable, which is our actual objective, since it corresponds to the multimarginal transportation tensor.

Algorithm 1 Primal-Dual Accelerated Alternating Minimization (PD-AAM)

- 1: $A_0 = \alpha_0 = 0$, $\eta^0 = \zeta^0 = \theta^0 = \mathbf{0}_{mn}$
- 2: **for** $t \geq 0$ **do**
- 3: Set $\beta_t = \text{argmin}_{\beta \in [0,1]} \varphi(\eta^t + \beta(\zeta^t - \eta^t))$
- 4: Set $\theta^t = \eta^t + \beta(\zeta^t - \eta^t)$
- 5: Choose $i_t = \text{argmax}_{i \in \{1, \dots, n\}} \|\nabla_i \varphi(\theta^t)\|_2^2$
- 6: Set $\eta^{t+1} = \text{argmin}_{\eta \in S_{i_t}(\theta^t)} \varphi(\eta)$
- 7: Find largest a_{t+1} from the quadratic equation

$$\varphi(\theta^t) - \frac{a_{t+1}^2}{2(A_t + a_{t+1})} \|\nabla \varphi(\theta^t)\|_2^2 = \varphi(\eta^{t+1})$$

- 8: Set $A_{t+1} = A_t + a_{t+1}$
- 9: Set $\zeta^{t+1} = \zeta^t - a_{t+1} \nabla \varphi(\theta^t)$
- 10: Set $\hat{x}^{t+1} = \frac{a_{t+1} x(\theta^t) + A_t \hat{x}^t}{A_{t+1}}$, where $x(\theta^t)$ is the primal variable reconstruction (Eq. (4) in the case of MOT)
- 11: **end for**

Output: The points \hat{x}^{t+1} , η^{t+1} .

The key result for this method is that it guarantees convergence in terms of the constraints and the duality gap

for the primal problem, provided that the dual is smooth, in the spirit of [25]–[30].

Theorem 3 ([31], Theorem 3). *Let the objective φ in the problem (P_2) be L -smooth and the solution of this problem be bounded, i.e. $\|\lambda^*\|_2 \leq R$. Then, for the sequences $\hat{x}_{t+1}, \eta_{t+1}$, $t \geq 0$, generated by Algorithm 1, we have*

$$f(\hat{x}^t) - f^* \leq f(\hat{x}^t) + \varphi(\eta^t) \leq \frac{2mLR^2}{t^2}, \quad (13)$$

$$\|\mathcal{A}\hat{x}^t - b\|_2 \leq \frac{8mLR}{t^2}. \quad (14)$$

To apply this result we need to estimate the Lipschitz constant L of the gradient of the dual objective and provide a bound R for an optimal solution.

Later, we will see the application of Theorem 3 to the MOT problem based on the following change of variables.

$$x \rightleftharpoons X, \quad f(x) \rightleftharpoons F(X), \quad \varphi(\Lambda) \rightleftharpoons \phi(U) \rightleftharpoons \phi(\Lambda)$$

$$\{x : \mathcal{A}x = b\} \rightleftharpoons \{X : p_k(X) = p_k, \quad \forall k \in \{1, \dots, m\}\}$$

$$Q \rightleftharpoons \{X \in \mathbb{R}_+^{n \times \dots \times n} : \sum_{i_1, \dots, i_m} X_{i_1, \dots, i_m} = 1, \quad 1 \leq i_j \leq n\}$$

The primal variable X is reconstructed from the dual variable U or Λ using (4).

B. Bound for L

We endow the space of transportation tensors with 1-norm, which leads to the primal objective in (2) being strongly convex on the feasible set of this problem with parameter γ . Further, we use the 2-norm for the dual space of Lagrange multipliers Λ in (3). Hence, the dual objective in (3) is L -smooth with the parameter $L \leq \|\mathcal{A}\|_{1 \rightarrow 2}^2 / \gamma$ [24]. Here $\mathcal{A} : \mathbb{R}^{n^m} \rightarrow \mathbb{R}^{mn}$ is the linear operator defining the linear constraints of the problem, which, in the case of the multimarginal optimal transport problem, is defined by $\mathcal{A} \text{vec}(X) = (p_1(X)^T, \dots, p_m(X)^T)^T$. Thus, each column of the matrix \mathcal{A} contains no more than m non-zero elements, which are equal to one. Hence, since $\|\mathcal{A}\|_{1 \rightarrow 2}$ is equal to maximum 2-norm of the column of this matrix, we have that $\|\mathcal{A}\|_{1 \rightarrow 2} = \sqrt{m}$. Finally, we have that $L \leq \frac{m}{\gamma}$.

C. Bound for R

We return to the particular dual problem (5) for the MOT problem to estimate the norm of an optimal dual solution in this particular case.

Lemma 4. *For every u_ξ^* entry of $U^* = ([u_1^*]^T, \dots, [u_m^*]^T)^T$ the following holds*

$$\max_{\eta} [u_\xi^*]_{\eta} - \min_{\eta} [u_\xi^*]_{\eta} \leq -\ln \nu \min_{\eta} [p_\xi]_{\eta}.$$

Proof. By the optimality condition (9)

$$0 = \frac{\partial \phi}{\partial [u_\xi]_{\eta}} = -[p_\xi]_{\eta} + \frac{\exp([u_\xi]_{\eta})}{\Sigma(U)} \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 1 \leq j \leq m \\ i_\xi = \eta}} \exp \left\{ \sum_{k \neq \xi} [u_k]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma} \right\},$$

where $\nu = \exp \frac{-\|C\|_{\infty}}{\gamma}$. Since $p_\xi \in \Delta_n$, we obtain the bound for the the solution of the above optimality conditions

$$\begin{aligned} 1 &\geq [p_\xi]_{\eta} \\ &= \frac{\exp([u_\xi^*]_{\eta})}{\Sigma(U^*)} \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 0 \leq j \leq m \\ i_\xi = \eta}} \exp \left\{ \sum_{k \neq \xi} [u_k^*]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma} \right\} \\ &\geq \nu \exp([u_\xi^*]_{\eta}) \Sigma(U^*)^{-1} \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 0 \leq j \leq m \\ i_\xi = \eta}} \exp \left\{ \sum_{k \neq \xi} [u_k^*]_{i_k} \right\} \\ &= \nu \exp([u_\xi^*]_{\eta}) \Sigma(U^*)^{-1} \sum_{\substack{k=1 \\ k \neq \xi}}^m \langle \mathbf{1}, e^{u_k^*} \rangle. \end{aligned} \quad (15)$$

From the above inequality we have

$$[u_\xi^*]_{\eta} \leq \ln \Sigma(U^*) - \ln \nu - \ln \sum_{\substack{k=1 \\ k \neq \xi}}^m \langle \mathbf{1}, e^{u_k^*} \rangle. \quad (16)$$

On the other hand,

$$\begin{aligned} [p_\xi]_{\eta} &= \frac{\exp([u_\xi^*]_{\eta})}{\Sigma(U^*)} \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 0 \leq j \leq m \\ i_\xi = \eta}} \exp \left\{ \sum_{k \neq \xi} [u_k^*]_{i_k} - \frac{C_{i_1 \dots i_m}}{\gamma} \right\} \\ &\leq \exp([u_\xi^*]_{\eta}) \Sigma(U^*)^{-1} \sum_{\substack{i_1, \dots, i_m \\ 1 \leq i_j \leq n \\ 0 \leq j \leq m \\ i_\xi = \eta}} \exp \left\{ \sum_{k \neq \xi} [u_k^*]_{i_k} \right\}, \end{aligned} \quad (17)$$

leads to

$$[u_\xi^*]_{\eta} \geq \ln [p_\xi]_{\eta} + \ln \Sigma(U^*) - \ln \sum_{\substack{k=1 \\ k \neq \xi}}^m \langle \mathbf{1}, e^{u_k^*} \rangle. \quad (18)$$

Combining (18) and (16) we have, for all $\xi = 1, \dots, m$,

$$\max_{\eta} [u_\xi^*]_{\eta} - \min_{\eta} [u_\xi^*]_{\eta} \leq -\ln \nu \min_{\eta} [p_\xi]_{\eta}.$$

□

Lemma 5. *Defining $\Lambda^0 = -\frac{\gamma}{m} \mathbf{1}_{mn}$, there exists a solution Λ^* of the dual problem (3) such that*

$$R = \|\Lambda^* - \Lambda^0\|_2 \leq \frac{\sqrt{mn}}{2} \left(\|C\|_{\infty} - \frac{\gamma}{2} \ln \min_{i,j} \{ [p_i]_j \} \right).$$

Proof. We begin by deriving an upper bound on $\|(u_1^{*T}, \dots, u_m^{*T})^T\|_2$. Using the results of the previous lemma, it remains to notice that the objective $\phi(U)$ is invariant under transformations $u_i \rightarrow u_i + t_i \mathbf{1}$, $t_i \in \mathbb{R}$ for $i \in \{1, \dots, m\}$, so there must exist some solution with $\max_{\eta} [u_i^*]_{\eta} = -\min_{\eta} [u_i^*]_{\eta} = \|u_i^*\|_{\infty}$, so

$$\|u_i^*\|_{\infty} \leq -\frac{1}{2} \ln \nu \min_{\eta} [p_i]_{\eta}.$$

As a consequence,

$$\begin{aligned} \|u_i^*\|_2 &\leq \sqrt{n} \|U^*\|_\infty \leq \\ &\leq -\frac{\sqrt{n}}{2} \ln \nu \min_{i,j} \{[p_i]_j\} \\ &\leq \frac{\sqrt{n}}{2} \left(\frac{\|C\|_\infty}{\gamma} - \frac{1}{2} \ln \min_{i,j} \{[p_i]_j\} \right). \end{aligned}$$

and

$$\|U^*\|_2 = \sqrt{\sum_i^m \|u_i^*\|_2^2} \leq \frac{\sqrt{mn}}{2} \left(\frac{\|C\|_\infty}{\gamma} - \frac{1}{2} \ln \min_{i,j} \{[p_i]_j\} \right)$$

By definition, $u_i = -\frac{1}{\gamma} \lambda_i - \frac{1}{m} \mathbf{1}$, so we have the inverse transformation $\lambda_i = -\gamma u_i - \frac{\gamma}{m} \mathbf{1}$. Finally, with $\Lambda^0 = -\frac{\gamma}{m} \mathbf{1}_{mn}$

$$\begin{aligned} R &= \|\Lambda^* - \Lambda^0\|_2 = \\ &= \left\| \left(-\gamma u_1^* - \frac{\gamma}{m} \mathbf{1}, \dots, -\gamma u_m^* - \frac{\gamma}{m} \mathbf{1} \right) \right. \\ &\quad \left. - \left(-\frac{\gamma}{m} \mathbf{1}, \dots, -\frac{\gamma}{m} \mathbf{1} \right) \right\|_2 = \left\| -\gamma (u_1^*, \dots, u_m^*) \right\|_2 \\ &= \gamma \|U^*\|_2 \leq \frac{\sqrt{mn}}{2} \left(\|C\|_\infty - \frac{\gamma}{2} \ln \min_{i,j} \{[p_i]_j\} \right). \end{aligned}$$

□

D. Projection on the feasible set

The Algorithm 1 may return a point in the primal space which does not satisfy the equality constraints. In this subsection, we provide a procedure to project approximate transport tensor to obtain a feasible point for the primal problem, i.e. find such $\hat{X} \approx \hat{X}^t$ that $p_i(\hat{X}) = p_i$. To do this we formulate Algorithm 2, which is a generalization of rounding procedure in [32], see also [19].

Algorithm 2 Multimarginal Rounding

```

1:  $V_1 = U$ 
2: for  $r = 1, \dots, m-1$  do
3:    $[X_r]_i = \min \{ [p_r]_i / [p_r(V_r)]_i, 1 \}$ 
4:    $X_r = \text{DiagTensor}(x_r)$ 
5:    $V_{r+1} = \text{ProdTensor}_r(V_r, X_r)$ 
6: end for
7: for  $r = 1, \dots, m$  do
8:    $\text{err}_r = p_r - p_r(V_m)$ 
9: end for
Output:  $\hat{V} = V_m + \bigotimes_{r=1}^m \text{err}_r / \|\text{err}_m\|_1^{m-1}$ 

```

Note that in Algorithm 2 the function $\text{DiagTensor}(\cdot)$ takes a vector as input and outputs a m -dimensional tensor with the input as its diagonal. Moreover, $\text{ProdTensor}_r(A, B)$ takes two m -dimensional tensors as input, and multiplies them in the direction r . We use \bigotimes to denote the tensor product of the input factors. The next lemma shows that the output of Algorithm 2 is in the desired space with the corresponding marginals, and bounds the error induced by the projection.

Lemma 6. *Let $\{p_k\}_{k=1}^m \in \Delta_n$, and $U \in \mathbb{R}_+^{n \times \dots \times n}$, then Algorithm 2 outputs a matrix \hat{F} with marginals $\{p_k\}_{k=1}^m$ satisfying $\|U - \hat{V}\|_1 \leq 2 \sum_{r=1}^m \|p_r - p_r(U)\|_1$.*

Algorithm 3 Approximate MOT by PD-AAM

Input: Accuracy ε .

- 1: Set $\gamma = \frac{\varepsilon}{2m \ln n}$, $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
- 2: Define $\tilde{p}_k = \left(1 - \frac{\varepsilon'}{4m}\right) p_k + \frac{\varepsilon'}{4mn} \mathbf{1}_n$, $k = 1, \dots, m$.
- 3: Apply PD-AAM to the dual problem (5) with marginals \tilde{p}_k , $k = 1, \dots, m$ until the stopping criterion $2 \sum_{k=1}^m \|p_k(\hat{X}^t) - \tilde{p}_k\|_1 + F(\hat{X}^t) + \phi(\eta^t) \leq \varepsilon/2$.
- 4: Find \hat{X} as the projection of \hat{X}^t on $\{X \in \mathbb{R}_+^{n \times \dots \times n}, p_k(X) = p_k, \forall k = 1, \dots, m\}$ by the Algorithm 2.

Output: \hat{X} .

Proof. Initially, note that for all $k = 1, \dots, m$, we have

$$\begin{aligned} p_k(\hat{V}) &= p_k(V_m) + p_k \left(\bigotimes_{r=1}^m \text{err}_r / \|\text{err}_m\|_1^{m-1} \right) \\ &= p_k(V_m) + \text{err}_k = p_k. \end{aligned}$$

Thus, the output of the \hat{U} has the desired marginals. Now, define $I = \|U\|_1 - \|V_m\|_1$, thus,

$$I = \sum_{r=1}^m \sum_{i=1}^n ([p_r(V_r)]_i - [p_r]_i)_+$$

Moreover, we have

$$\|\hat{V} - U\| \leq I + \left\| \bigotimes_{r=1}^m \text{err}_r / \|\text{err}_m\|_1^{m-1} \right\|$$

$$I + 1 - \|V_m\|_1 = 2I + 1 - \|U\|_1 = 2 \sum_{r=1}^m \|p_r - p_r(U)\|_1,$$

where the last line follows the same arguments as the proof of Lemma 7 in [32]. □

IV. COMPLEXITY OF MULTIMARGINAL OT

In this section, we prove the computational complexity of finding a ε -solution for the original *non-regularized* MOT problem (1), i.e. we estimate the complexity to find \hat{X} satisfying all the constraints in (1) and also satisfying

$$\langle C, \hat{X} \rangle \leq \langle C, X^* \rangle + \varepsilon, \quad (19)$$

where X^* is an optimal solution for (1). The approximation is produced by Algorithm 3 below.

To obtain its complexity, we combine all the above building blocks, i.e., analysis of the PD-AAM algorithm and estimates for R and L , and the rounding procedure.

To adapt Algorithm 1 to Problem (5), one should replace *Step 4* with: Choose $I = \arg\max_{i \in \{1, \dots, m\}} \left\| \frac{\partial \phi}{\partial u_i}(\theta^t) \right\|_2$, and *Step 5* with: Set

$$\eta_i^{t+1} = \begin{cases} \theta_i^t + \ln p_i - \ln p_i(B(\theta^t)), & i = I \\ \theta_i^t, & \text{otherwise.} \end{cases}$$

Theorem 7. *The output \hat{X} of Algorithm 3 is an ε -solution for the original non-regularized MOT problem (1), e.g.*

$$\langle C, \hat{X} \rangle \leq \langle C, X^* \rangle + \varepsilon. \quad (20)$$

Proof. By Lemma 6, \hat{X} is a feasible point for Problem (1). Let us estimate the objective residual. We have

$$\begin{aligned} \langle C, \hat{X} \rangle &= \langle C, X^* \rangle + \langle C, X_\gamma^* - X^* \rangle + \langle C, \hat{X}^t - X_\gamma^* \rangle \\ &+ \langle C, \hat{X} - \hat{X}^t \rangle \leq \langle C, X^* \rangle + \gamma m \ln n + F(\hat{X}^t) + \phi(\eta^t) \\ &\quad + 2 \sum_{k=1}^m \|p_k(\hat{X}^t) - p_k\|_1 \|C\|_\infty, \end{aligned} \quad (21)$$

where \hat{X}^t is the output of Algorithm 1, \hat{X} is a projection of \hat{X}^t by Algorithm 2 on the feasible set, X^* is a solution to the non-regularized multimarginal OT problem (1), X_γ^* is a solution to the entropy-regularized multimarginal OT problem (2). To obtain the last inequality we used the fact that the Entropy on the standard simplex in the dimension n^m belongs to the interval $-H(X) \in [-m \ln n, 0]$, and, hence, $\langle C, X_\gamma^* - X^* \rangle \leq 0$ and

$$\begin{aligned} \langle C, \hat{X}^t - X_\gamma^* \rangle &= (\langle C, \hat{X}^t \rangle - \gamma H(\hat{X}^t)) \\ &\quad - (\langle C, X_\gamma^* \rangle - \gamma H(X_\gamma^*)) + \gamma (H(\hat{X}^t) - H(X_\gamma^*)) \\ &\stackrel{(13)}{\leq} F(\hat{X}^t) + \phi(\eta^t) + \gamma m \ln n. \end{aligned} \quad (22)$$

Finally, by the Hölder inequality and Lemma 6,

$$\langle C, \hat{X} - \hat{X}^t \rangle \leq \|C\|_\infty \|\hat{X} - \hat{X}^t\|_1 \leq 2 \|C\|_\infty \sum_{k=1}^m \|p_k(\hat{X}^t) - p_k\|_1.$$

This finishes the proof of inequality (21).

Further, we have

$$\sum_{k=1}^m \|p_k(\hat{X}^t) - p_k\|_1 \leq \sum_{k=1}^m \left(\|p_k(\hat{X}^t) - \tilde{p}_k\|_1 + \|\tilde{p}_k - p_k\|_1 \right) \leq \varepsilon',$$

by the construction of \tilde{p}_k and the stopping criterion in step 3 of Algorithm 3. Combining this, (21), the choice of γ and ε' as well as the stopping criterion in step 3 of Algorithm 3, we obtain that (20) holds. \square

It remains to estimate the complexity of the algorithm. By Theorem 3, we obtain that

$$\begin{aligned} \sum_{k=1}^m \|p_k(\hat{X}^t) - \tilde{p}_k\|_1 &\leq \sqrt{mn} \|\mathcal{A}\hat{X}^t - b\|_2 \leq \frac{8m^{\frac{3}{2}}n^{\frac{1}{2}}LR}{t^2} \\ &\leq \frac{8m^{\frac{3}{2}}n^{\frac{1}{2}}}{t^2} \cdot \frac{m \cdot 2m \ln n}{\varepsilon} \cdot \frac{\sqrt{mn} \left(\|C\|_\infty + \frac{\varepsilon}{4m \ln n} \ln \frac{4mn \cdot 8 \|C\|_\infty}{\varepsilon} \right)}{2} \\ &= \frac{8m^4 n \|C\|_\infty \ln n}{\varepsilon t^2} \left(1 + \frac{\varepsilon}{4m \|C\|_\infty \ln n} \ln \frac{32mn \|C\|_\infty}{\varepsilon} \right), \end{aligned}$$

where the operator \mathcal{A} is defined in Sect III-B and we used that by the choice of \tilde{p}_k , $\min_{i,j} \{p_i\}_j \geq \frac{\varepsilon'}{4mn}$. At the same time,

$$\begin{aligned} F(\hat{X}^t) + \phi(\eta^t) &\leq \frac{2mLR^2}{t^2} \\ &\leq \frac{2m}{t^2} \cdot \frac{m \cdot 2m \ln n}{\varepsilon} \cdot \frac{mn}{4} \left(\|C\|_\infty + \frac{\varepsilon}{4m \ln n} \ln \frac{32mn \|C\|_\infty}{\varepsilon} \right)^2 \\ &= \frac{m^4 n \|C\|_\infty^2 \ln n}{\varepsilon t^2} \left(1 + \frac{\varepsilon}{4m \|C\|_\infty \ln n} \ln \frac{32mn \|C\|_\infty}{\varepsilon} \right)^2. \end{aligned}$$

Let us denote $\delta_\varepsilon = 1 + \frac{\varepsilon}{4m \|C\|_\infty \ln n} \ln \frac{32mn \|C\|_\infty}{\varepsilon}$. Since ε is small and m, n are large, we can think of this quantity as

$\delta_\varepsilon = O(1)$. Then, to satisfy the stopping criterion in step 3 of Algorithm 3 we need to take

$$\begin{aligned} t &\geq \sqrt{\frac{128m^4 n \|C\|_\infty^2 \delta_\varepsilon \ln n}{\varepsilon^2}} = \tilde{O} \left(\frac{m^2 n^{1/2} \|C\|_\infty}{\varepsilon} \right), \text{ and} \\ t &\geq \sqrt{\frac{4m^4 n \|C\|_\infty^2 \delta_\varepsilon^2 \ln n}{\varepsilon^2}} = \tilde{O} \left(\frac{m^2 n^{1/2} \|C\|_\infty}{\varepsilon} \right). \end{aligned}$$

Since in each iteration we need to calculate the full gradient of the dual objective, which amounts to calculating m marginals $p_k(B(U))$, $k = 1, \dots, m$ of the m -dimensional tensor $B(U)$, the cost of this operation is $O(mn^m)$ and it dominates the complexity of other operations in each iteration. This gives the following theorem and the main result of the paper.

Theorem 8. *The computational complexity of finding an ε -approximate solution for the non-regularized MOT problem using Algorithm 3 is*

$$\tilde{O} \left(\frac{m^3 n^{m+1/2} \|C\|_\infty}{\varepsilon} \right).$$

We now discuss the scalability of the proposed algorithm. As already mentioned, the most expensive operation on each iteration is the calculation of m marginals $p_k(B(U))$ of the m -dimensional tensor $B(U)$. This operation can be organized in parallel if we store this tensor in shared memory and allow m workers to access it. Then, they can independently calculate all the marginals. The total amount of arithmetic operations remains the same, but the work time is now proportional to n^m rather than mn^m .

Next, we compare our complexity results with the estimates in the preprint [19]. By inspecting their Algorithm 2 and Algorithm 5, we see that similarly to our algorithm, in each iteration, they need to calculate all the marginals (which they denote by $r_i(B(\beta))$) to choose the block I , which will be updated. The complexity of this operation dominates the complexity of other operations in each step. Thus, since each iteration in their algorithms and our algorithm is the same, we compare the iteration complexity of the algorithms. The iteration complexity of our algorithm is $\tilde{O}(m^2 n^{1/2} \|C\|_\infty / \varepsilon)$. The iteration complexity of the multimarginal Sinkhorn's algorithm [19] is $\tilde{O}(m^3 \|C\|_\infty^2 / \varepsilon^2)$, which has worse dependence on ε and m than our bound. The claimed iteration complexity of multimarginal RANDKHORN algorithm in [19] is $\tilde{O}(m^{8/3} n^{1/3} \|C\|_\infty^{4/3} / \varepsilon)$, which has worse dependence on m and $\|C\|_\infty$ than our bound. Moreover, the multimarginal RANDKHORN is a randomized algorithm, and its complexity is estimated on average, whereas our algorithm and complexity are deterministic.

V. EXPERIMENTS

This section provides a numerical comparison of multimarginal Sinkhorn's algorithm from [19] with our AAM method. We performed experiments using randomly chosen vectors $p_i \in \Delta_n$ and tensor $C \in \mathbb{R}_+^{n^m}$. We slightly modified the smaller values of p_i as described above to lower bound

their minimal value. We choose several values of accuracy $\varepsilon \in [0.25, 0.0125]$, and run the methods until the stopping criterion was reached. One can see that our AAM algorithm outperforms multimarginal Sinkhorn's algorithm from [19].² Unfortunately, we were not able to implement the multimarginal RANDKHORN algorithm since its stopping criterion $\bar{E}_t > \varepsilon'$ depends on *expected* residual in the constraints given in [19, Eq. (28)], which is unavailable in practice.

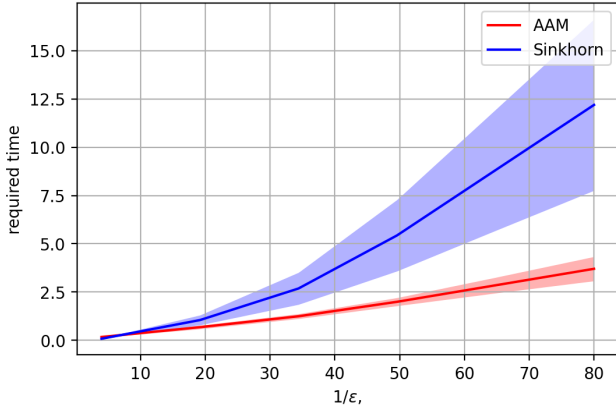


Fig. 2. Performance comparison between multimarginal Sinkhorn's algorithm and Algorithm 3 ($n = 15$, $m = 4$).

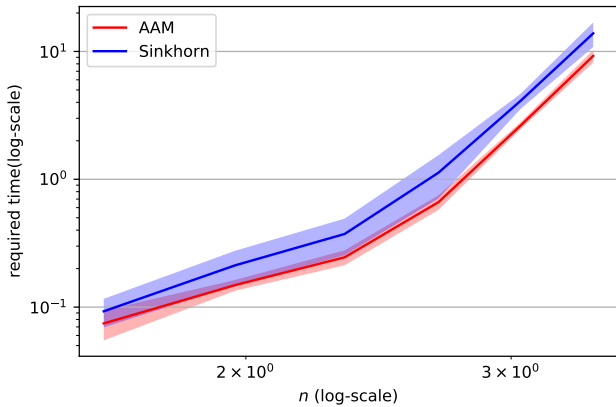


Fig. 3. Performance comparison between multimarginal Sinkhorn's algorithm and Algorithm 3 ($m = 4$, $\varepsilon = 0.05$).

VI. CONCLUSIONS

We provide a novel algorithm for the computation of approximate solutions to the multimarginal optimal transport problem. Our results are based on a new primal-dual analysis of the entropy regularized optimal transport problem. We show that the iteration complexity of our algorithm is better than the state-of-the-art methods in a large set of problem regimes to the number of distributions, dimension of the distributions, and desired accuracy.

As a byproduct of our analysis, given that the Wasserstein barycenter of a set of distributions can be recovered from the optimal multimarginal transport plan [18], we provide some evidence of an exponential complexity bound for the

computation of the free-support barycenter which is known to be a non-convex problem.

Future work will include the study of fully decentralized approaches and extensive experimental results for applications related to signal processing.

REFERENCES

- [1] A. Gramfort, G. Peyré, and M. Cuturi, "Fast optimal transport averaging of neuroimaging data," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 261–272.
- [2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv:1701.07875*, 2017.
- [3] S. Asoodeh, T. Gao, and J. Evans, "Curvature of hypergraphs via multi-marginal optimal transport," in *CDC 2018*, 2018, pp. 1180–1185.
- [4] Y. Chen, T. T. Georgiou, and M. Pavon, "Optimal transport over a linear dynamical system," *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2137–2152, 2016.
- [5] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Multi-marginal optimal mass transport with partial information," *arXiv:1905.03847*, 2019.
- [6] B. Pass, "Multi-marginal optimal transport: theory and applications," *Math. Model. & Num. Anal.*, vol. 49, no. 6, pp. 1771–1790, 2015.
- [7] C. Villani, *Topics in Optimal Transportation*, ser. Graduate studies in mathematics. American Mathematical Society, 2003.
- [8] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, "Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 1367–1376.
- [9] A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe, "On the complexity of approximating Wasserstein barycenters," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 2019, pp. 3530–3540.
- [10] T. Lin, N. Ho, and M. Jordan, "On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 2019, pp. 3982–3991.
- [11] A. Jambulapati, A. Sidford, and K. Tian, "A direct $O(1/\varepsilon)$ iteration parallel algorithm for optimal transport," in *NeurIPS 2019*, 2019, pp. 11 359–11 370.
- [12] L. Ambrosio and N. Gigli, "A user's guide to optimal transport," in *Modelling and optimisation of flows on networks*. Springer, 2013.
- [13] R. J. McCann, "A glimpse into the differential topology and geometry of optimal transport," *arXiv:1207.1867*, 2012.
- [14] I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier, "Tomographic reconstruction from a few views: a multi-marginal optimal transport approach," *Appl. Math. & Opt.*, vol. 75, no. 1, pp. 55–73, 2017.
- [15] J. Cao, L. Mo, Y. Zhang, K. Jia, C. Shen, and M. Tan, "Multi-marginal wasserstein gan," in *NeurIPS*, 2019, pp. 1774–1784.
- [16] I. Ekeland, "An optimal matching problem," *ESAIM: Control, Optimization and Calculus of Variations*, vol. 11, no. 1, pp. 57–71, 2005.
- [17] R. G. Parr, "Density functional theory of atoms and molecules," in *Horizons of Quantum Chemistry*. Springer, 1980, pp. 5–15.
- [18] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative bregman projections for regularized transportation problems," *SIAM Journal on Scientific Computing*, vol. 37, no. 2, pp. A1111–A1138, 2015.
- [19] T. Lin, N. Ho, M. Cuturi, and M. I. Jordan, "On the Complexity of Approximating Multimarginal Optimal Transport," *arXiv e-prints*, 2019, arXiv:1910.00152.
- [20] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2292–2300.
- [21] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.
- [22] Y. T. Lee and A. Sidford, "Path finding methods for linear programming: Solving linear programs in $\tilde{O}(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow," *FOCS 2014*, pp. 424–433, 2014.
- [23] N. Tupitsa, P. Dvurechensky, A. Gasnikov, and C. A. Uribe, "Multimarginal optimal transport by accelerated gradient descent," *arXiv:2004.02294*, 2020.
- [24] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

² The code available <https://rb.gy/siirke>

- [25] P. Dvurechensky, A. Gasnikov, E. Gasnikova, S. Matsievsky, A. Rodomanov, and I. Usik, "Primal-dual method for searching equilibrium in hierarchical congestion population games," in *Supplementary Proceedings of Conference on Discrete Optimization and Operations Research (DOOR 2016)*, 2016, pp. 584–595.
- [26] P. Dvurechensky, D. Dvinskikh, A. Gasnikov, C. A. Uribe, and A. Nedić, "Decentralize and randomize: Faster algorithm for Wasserstein barycenters," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 10783–10793.
- [27] C. A. Uribe, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and A. Nedić, "Distributed computation of Wasserstein barycenters over networks," in *IEEE Conference on Decision and Control*, 2018, pp. 6544–6549.
- [28] S. V. Guminov, Y. E. Nesterov, P. E. Dvurechensky, and A. V. Gasnikov, "Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems," *Doklady Mathematics*, vol. 99, no. 2, pp. 125–128, 2019.
- [29] P. Dvurechensky, A. Gasnikov, S. Omelchenko, and A. Tiurin, "A stable alternative to Sinkhorn's algorithm for regularized optimal transport," in *Mathematical Optimization Theory and Operations Research*, A. Kononov, M. Khachay, V. A. Kalyagin, and P. Pardalos, Eds. Cham: Springer International Publishing, 2020, pp. 406–423.
- [30] Y. Nesterov, A. Gasnikov, S. Guminov, and P. Dvurechensky, "Primal-dual accelerated gradient methods with small-dimensional relaxation oracle," *Optimization Methods and Software*, 2020. [Online]. Available: <https://doi.org/10.1080/10556788.2020.1731747>
- [31] S. Guminov, P. Dvurechensky, N. Tupitsa, and A. Gasnikov, "Accelerated Alternating Minimization, Accelerated Sinkhorn's Algorithm and Accelerated Iterative Bregman Projections," *arXiv e-prints*, 2019, arXiv:1906.03622.
- [32] J. Altschuler, J. Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via sinkhorn iteration," in *NeurIPS 2017*. Curran Associates, Inc., 2017, pp. 1961–1971.