Multimarginal Optimal Transport by Accelerated Alternating Minimization

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

Abstract—We consider a multimarginal optimal transport, which includes as a particular case the Wasserstein barycenter problem. In this problem one has to find an optimal coupling between \( m \) probability measures, which amounts to finding a tensor of the order \( m \). We propose an accelerated method based on accelerated alternating minimization and estimate its complexity to find the approximate solution to the problem. We use entropic regularization with 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 the Kantorovich formulation, the optimal transport problem seeks to minimize

\[
\int_{M_1 \times \cdots \times M_m} c(x_1, \ldots, x_m) d\pi(x_1, \ldots, x_m),
\]

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

Although the optimal transport 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 have been shown to be computationally intense, and a lot of effort has been put into analyzing the underlying complexity of such problems [8]–[10].

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, optimal transport has been studied for the case of comparing and quantifying distances between two probability distributions (i.e., \( m = 2 \)) for which theory is fairly well understood [7], [11], [12]. However, for \( m \geq 3 \), so called multimarginal optimal transport problem, much less is known, even though such regime has been recently shown useful for many applications, like tomographic image reconstruction [13], generative adversarial networks [14], economics [15], and density functional theory [16]. See [6] for a recent survey of fundamental theoretical formulations and applications of the multimarginal optimal transport problem. Computational aspects of the multimarginal problem were studied in [17], where an Iterative Bregman Projections algorithm was proposed for this problem, yet without complexity analysis. It was also pointed that multimarginal optimal transport problem can be applied to calculate the barycenter of \( m \) measures without fixing the support of the barycenter. In the preprint [18] the authors propose and analyze complexity of two algorithms for the multimarginal OT problem. Both papers, as well as ours, follow the entropy regularization approach [19].

Our objective in this paper is to develop a novel algorithm for the computation of approximate solutions for the multimarginal optimal transport problem using recently developed methods of alternating minimization. Our contributions are three fold:

- We develop a novel algorithm for the approximate computation of multimarginal optimal transport maps based on accelerated alternating minimization algorithm.
- We formally prove the computational complexity of

The work of P. Dvurechensky, A. Gasnikov, and N. Tupitsa in part III-A – III-C, IV and V 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.

N.T. is with the Moscow Institute of Physics and Technology, Institute for Information Transmission Problems ( 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 ( gasnikov@yandex.ru ). C.A.U. is with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, USA ( cauribe @mit.edu ).
the proposed algorithm. Explicitly, we show that the proposed algorithm has an iteration complexity $O \left( m^2/n^{1/2}/\epsilon \right)$ and computational complexity of $O \left( m^3 n^{m+1/2}/\epsilon \right)$ arithmetic operations. Our result indicates an upper exponential bound for the complexity of Wasserstein barycenter problem with free support, which is known to be a non-convex optimization problem.

- We show that under some regimes of the problem parameters $m$ (number of distributions), $n$ (dimension of the distributions), and $\epsilon$ (desired accuracy) the proposed algorithm has better iteration complexity in comparison with existing methods.

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

II. PROBLEM STATEMENT

A. The Multimarginal Optimal Transport Problem

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

$$[p_k(A)]_j = \sum_{1 \leq i_j \leq n_j, \forall k \neq j} A_{i_1,\ldots,i_{j-1},j,i_{j+1},\ldots,i_m}.$$

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

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

The problem of computing the multimarginal OT problem between $m \geq 2$ discrete probability distributions with $n$ support points has the following form:

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

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

The multimarginal OT problem is a linear program with $mn$ equality constraints and $n^m$ variables and inequality constraints. When $m = 2$, the multimarginal OT problem reduces to the classical optimal transport problem [7]. In the general case of $m$ measures one of the applications of multimarginal OT is grid-free Wasserstein barycenter computation [17]. Despite LP formulation is in standard form, the dimension of the problem, which is exponential in $m$ does not allow to apply standard LP solvers such as interior point methods [20], [21]. Thus, in the next section we apply entropic regularization approach.

B. Entropy Regularized Multimarginal Optimal Transport Problem

Following [17], [19], we consider a regularized version of problem (1) in which we add an entropic penalty for the multimarginal transportation plan. The resulting problem has the following form:

$$\min_{X \in \mathbb{R}_+^{n \times \cdots \times n}, \ p_k(X) = p_k, \ \forall k \in \{1,\ldots,m\}} \langle C, X \rangle - \gamma H(X), \tag{2}$$

where $\gamma > 0$ is the regularization parameter, and $H(X)$ denotes the entropic regularization term:

$$H(X) := -\langle X, \log(X) \rangle.$$

Here logarithm of a tensor is taken 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 $p_k$, $k = 1,\ldots,m$ belong to $\Delta^n$. Adding this constraint does not change the solution of the problem, but it is crucial to obtain a dual optimization problem with the objective having Lipschitz-continuous gradient. The reason for the latter is that entropy is strongly convex on the probability simplex w.r.t. the 1-norm.

C. Dual Problem Formulation

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

$$L(X, \Lambda, \mu) = \langle C, X \rangle + \gamma \langle X, \log(X) \rangle + \sum_{k=1}^m \lambda_k^T (p_k(X) - p_k) + \mu \sum_{i_1 \leq \cdots \leq i_m \in \{1,\ldots,n\}} X_{i_1,\ldots,i_m} - \mu, \tag{3}$$

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

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

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

$$\frac{\partial L}{\partial X_{i_1,\ldots,i_m}} (X, \Lambda, \mu) = C_{i_1,\ldots,i_m} + \gamma + \gamma \log(X_{i_1,\ldots,i_m}) + \sum_{k=1}^m [\lambda_k]_{i_k} + \mu = 0. \tag{4}$$

The solution of the above problem is

$$X_{i_1,\ldots,i_m}(\Lambda, \mu) = \exp \left( \frac{\sum_{k=1}^m [\lambda_k]_{i_k} - C_{i_1,\ldots,i_m} - \gamma - \mu}{\gamma} \right).$$
Therefore, we have
\[ L(\Lambda, \mu) = -\gamma \sum_{i_1 \leq j \leq m, 1 \leq i \leq n} X_{i_1, \ldots, i_m}(\Lambda, \mu) - m \lambda^T p_k - \mu. \]

By taking a derivative w.r.t. \( \mu \) and setting it to zero we have
\[ \sum_{i_1 \leq j \leq m, 1 \leq i \leq n} X_{i_1, \ldots, i_m}(\Lambda, \mu(\Lambda)) + \gamma = 0. \]

From where we can express \( \mu(\Lambda) \) as
\[ \exp \left\{ -\frac{\mu}{\gamma} \right\} \sum_{i_1 \leq j \leq m, 1 \leq i \leq n} \exp \left\{ -\frac{1}{\gamma} \sum_{k=1}^{m} \frac{\lambda k_i}{\gamma} - \frac{C_{1i, \ldots, i_m}}{\gamma} - 1 \right\} = 1, \]
yielding the following form of the dual problem:
\[ \max_{\lambda} -\gamma \left[ \ln \sum_{i_1 \leq j \leq m, 1 \leq i \leq n} \exp \left\{ -\sum_{k=1}^{m} \frac{[\lambda k_i]}{\gamma} - \frac{C_{1i, \ldots, i_m}}{\gamma} - 1 \right\} \right] \]
\[ + 1 + \frac{1}{\gamma} \sum_{k=1}^{m} \lambda^T p_k \]. \hspace{1cm} (5)

As it is known [22], the objective in [5] 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 methods to solve the dual problem and obtain the corresponding complexity.

Finally, with the change of variable \( u_k = -\frac{\lambda k_i}{\gamma} - \frac{1}{m} \) the dual objective becomes
\[ \varphi(U) = \varphi(u_1, \ldots, u_m) = \gamma \left[ \ln \sum_{1 \leq i \leq n} \exp \left\{ -\sum_{k=1}^{m} \frac{[u k_i]}{\gamma} - \frac{C_{1i, \ldots, i_m}}{\gamma} \right\} - \sum_{k=1}^{m} u_k^T p_k \right], \hspace{1cm} (6) \]
where \( U = (u_1^T, \ldots, u_m^T)^T \in \mathbb{R}^{mn} \), and we have the dual problem
\[ \min_U \varphi(U) = \varphi(u_1, \ldots, u_m). \hspace{1cm} (7) \]

### III. Algorithm Design Based on the Alternating Minimization Approach

Introducing the notation for the tensor \( B(U) \in \mathbb{R}^{mn} \) with elements given as
\[ B_{i_1, \ldots, i_m}(u_1, \ldots, u_m) = \exp \left\{ \sum_{k=1}^{m} [u k_i] \lambda k_i - \frac{C_{1i, \ldots, i_m}}{\gamma} \right\} \]
and for its element-wise sum
\[ \Sigma(U) = \sum_{i_1 \leq j \leq m, 1 \leq i \leq n} B_{i_1, \ldots, i_m}(u_1, \ldots, u_m). \]

we obtain partial derivatives
\[ \left[ \frac{\partial \varphi}{\partial \xi} \right]_{\eta} = \sum_{i_1 \leq j \leq m, 1 \leq i \leq n} \exp \left\{ \sum_{k=1}^{m} [u k_i] \lambda k_i - \frac{C_{1i, \ldots, i_m}}{\gamma} \right\} \frac{[p \xi(B(U))]_{\eta}}{\Sigma(U)} - [p \xi]_{\eta} \]
\[ = [p \xi(B(U))]_{\eta} - [p \xi]_{\eta}. \hspace{1cm} (8) \]

**Lemma 1.** The iterations
\[ u_{k+1}^t \in \arg \min_{u \in \mathbb{R}^n} \varphi(u_1^t, \ldots, u_{k-1}^t, u, u_{k+1}^t, \ldots, u_m^t), \]
can be written explicitly as
\[ u_{k+1}^t = u_k^t + \ln p_k - \ln p_k(B(U^t)), \]
or entry-wise as
\[ [u_{k+1}^t]_{\eta} = [u_k^t]_{\eta} + \ln [p_k]_{\eta} - \ln [p_k(B(U^t))]_{\eta}. \hspace{1cm} (9) \]

**Proof.** Consider the following tensor
\[ B_{i_1, \ldots, i_m}(u_1^t, \ldots, u_{k-1}^t, u_{k+1}^t, \ldots, u_m^t) \]
\[ = \exp \left\{ [u_{k+1}^t]_\eta \sum_{k \neq \xi} [u_k^t]_\xi - \frac{C_{1i, \ldots, i_m}}{\gamma} \right\} \]
\[ = \frac{\exp[u_{k+1}^t]_\eta}{\exp[u_k^t]_\xi} \exp \left\{ [u_{k+1}^t]_\eta \sum_{k \neq \xi} [u_k^t]_\xi - \frac{C_{1i, \ldots, i_m}}{\gamma} \right\} \]
\[ = -\frac{\exp[u_{k+1}^t]_\eta}{\exp[u_k^t]_\xi} B(U^t), \]
and plug in the expression (9) from the lemma statement
\[ \sum_{i_1 \leq j \leq m} B_{i_1, \ldots, i_m}(u_1^t, \ldots, u_{\xi-1}^t, u_{\xi+1}^t, \ldots, u_m^t) \]
\[ = \sum_{\eta} \sum_{i_1 \leq j \leq m} B_{i_1, \ldots, i_m}(u_1^t, \ldots, u_{\xi-1}^t, u_{\xi+1}^t, \ldots, u_m^t) \]
\[ = \sum_{\eta} \frac{\exp[u_{k+1}^t]_\eta}{\exp[u_k^t]_\xi} [p \xi(B(U^t))]_{\eta} \]
\[ \sum_{\eta} [p \xi(B(U^t))]_{\eta} [p \xi(B(U^t))]_{\eta} = 1. \]

Next, we plug (9) in the optimality conditions \[ \frac{\partial \varphi}{\partial \xi} \eta = 0 \]
and show that the conditions are satisfied

\[
[p_\xi]_\eta = 
\exp\left(\sum_{\xi \notin \eta} \exp\left\{ \sum_{k \neq \xi} u_k^\eta_i - \frac{C_{i_1 \ldots i_m}}{\gamma} \right\} \right)
\]

\[
= \sum_{1 \leq i_1 \leq n} \ldots \sum_{1 \leq i_m \leq n} B_{i_1 \ldots i_m}(u_1^\eta, \ldots, u_m^\eta, u_{\xi_1}^\eta, \ldots, u_{\xi_m}^\eta)
\]

\[
= \frac{e^{[u^\eta_{\xi^\eta}]}_{\eta}}{e^{[u^\eta]}_{\eta}} \sum_{i=\eta} B(U^\eta) \left[ [p_\xi]_\eta \left[ p_\xi(B(U^\eta)) \right]_\eta \right].
\]

Lemma \[\text{1}\] implies that the dual objective \(\varphi\) can be explicitly minimized in each of the \(m\) blocks of variables \(u_k, k = 1, \ldots, m\), suggesting to use alternating minimization algorithms for the dual problem. Note that the nature of the Iterative Bregman Projections algorithm \[\text{17}\] is different since it is an alternating projection algorithm for the primal problem.

A. General Primal-Dual Accelerated Alternating Minimization

In this section, we consider a general minimization problem

\[ (P_1) \quad \min_{x \in Q \subseteq E} \{ f(x) : Ax = b \}, \]

where \(E\) is a finite-dimensional real vector space, \(Q\) is a simple closed convex set, \(A\) 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 the problem \[\text{2}\]. The Lagrange dual problem to Problem \(P_1\) is

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

Here we denote \(\Lambda = H^*. \) Note also that the dual entropy multimarginal OT problem \[\text{5}\] is a particular case of this general dual problem. It is convenient to rewrite Problem \(D_1\) in the equivalent form of a minimization problem

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

Since \(f\) is convex, \(\varphi(\lambda)\) is a convex function and, by Danskin’s theorem, its subgradient is equal to

\[ \nabla \varphi(\lambda) = b - Ax(\lambda), \]

\[ \text{(10)} \]

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

\[ \max_{x \in Q} \{ -f(x) - \langle A^T \lambda, x \rangle \}. \]

\[ \text{(11)} \]

In what follows, we assume that \(\varphi(\lambda)\) is \(L\)-smooth and that the dual problem \(D_1\) has a solution \(\lambda^*\) 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 \(\{1, \ldots, N\}\) of indices of the orthonormal basis vectors \((e_i)_{i=1}^N\) is divided into \(m\) disjoint subsets (blocks) \(I_k, k \in \{1, \ldots, m\}\). 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 \[\text{1}\] is to use greedy alternating minimization steps in the dual and combine them with momentum step \(9\), which is responsible for Nesterov’s acceleration and allows to obtain accelerated convergence rate for the dual problem. Further, we add a step \(10\) which updates the primal variable, which is our actual goal since it corresponds to the multimarginal transportation tensor.

\[ \text{Algorithm 1 Primal-Dual Accelerated Alternating Minimization (PD-AAM)} \]

1: \(A_0 = \alpha_0 = 0, \eta^0 = \zeta^0 = 0^m, \)
2: for \(t \geq 0\) do
3: \(\beta_t = \arg \min_{\beta \in [0,1]} \varphi(\eta^t + \beta(\zeta^t - \eta^t)) \)
4: \(\theta^t = \eta^t + \beta(\zeta^t - \eta^t) \)
5: \(\gamma_t = \arg \max_{i \in \{1, \ldots, n\}} \|\nabla_i \varphi(\theta^t)\|_2 \)
6: \(\gamma_t = \arg \min_{\eta \in S_t(\theta^t)} \varphi(\eta) \)
7: \(\text{Find largest } \alpha_{t+1} \text{ from the quadratic equation} \)
8: \(\varphi(\theta^t) - \frac{\alpha_{t+1}^2}{2(A_t + \alpha_{t+1})} \|\nabla \varphi(\theta^t)\|_2^2 = \varphi(\eta^{t+1}) \)
9: \(\eta_{t+1} = \eta_t + \alpha_{t+1} \nabla \varphi(\theta^t) \)
10: \(\zeta_{t+1} = \zeta_t - \alpha_{t+1} \nabla \varphi(\theta^t) \)
11: end for

Output: The points \(\hat{x}^{t+1}, \eta^{t+1}\).

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.

**Theorem 2** ([23], Theorem 3). Let the objective \(\varphi\) 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 \[\text{2}\] we have

\[ f(\hat{x}^k) - f^* \leq f(\hat{x}^t) + \varphi(\eta^t) \leq \frac{2mL^2R^2}{t^2}, \]

\[ \|A_x t - b\|_2 \leq \frac{8mL^2R}{t^2}. \]

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.

B. Bound for \(R\)

Now we return to the particular dual problem \[\text{7}\] for the multimarginal optimal transport problem to estimate the norm
of an optimal dual solution in this particular case. By the optimality condition \( \mathbf{8} \)

\[
0 = \frac{\partial \varphi}{\partial [u_\xi]} = -[p_\xi]_\eta \\
+ \exp([u_\xi]_\eta) \sum_{i_1, \ldots, i_m \leq n} \exp \left( \sum_{k \neq \xi} [u_k]_{i_k} - \frac{C_{i_1 \ldots i_m}}{\gamma} \right),
\]

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

\[
1 \geq [p_\xi]_\eta = \frac{\exp([u_\xi]_\eta) \Sigma(U^*)^{-1}}{\Sigma(U^*)} \sum_{i_1, \ldots, i_m \leq n} \exp \left( \sum_{k \neq \xi} [u_k]_{i_k} - \frac{C_{i_1 \ldots i_m}}{\gamma} \right) \\
\geq \nu \exp([u_\xi]_\eta) \Sigma(U^*)^{-1} \sum_{i_1, \ldots, i_m \leq n} \exp \left( \sum_{k \neq \xi} [u_k]_{i_k} \right) \\
= \nu \exp([u_\xi]_\eta) \Sigma(U^*)^{-1} \sum_{k=1}^{m} \langle 1, u_k^\# \rangle. \tag{14}
\]

From the above inequality we have

\[
[u_\xi]_\eta \leq \ln \Sigma(U^*) - \ln \nu - \ln \sum_{k=1}^{m} \langle 1, u_k^\# \rangle. \tag{15}
\]

On the other hand,

\[
[p_\xi]_\eta = \frac{\exp([u_\xi]_\eta) \Sigma(U^*)^{-1}}{\Sigma(U^*)} \sum_{i_1, \ldots, i_m \leq n} \exp \left( \sum_{k \neq \xi} [u_k]_{i_k} - \frac{C_{i_1 \ldots i_m}}{\gamma} \right) \\
\leq \exp([u_\xi]_\eta) \Sigma(U^*)^{-1} \sum_{i_1, \ldots, i_m \leq n} \exp \left( \sum_{k \neq \xi} [u_k]_{i_k} \right), \tag{16}
\]

leads to

\[
[u_\xi]_\eta \geq \ln[p_\xi]_\eta + \ln \Sigma(U^*) - \ln \sum_{k=1}^{m} \langle 1, u_k^\# \rangle. \tag{17}
\]

Combining (17) and (15) we have, for all \( \xi = 1, \ldots, m \),

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

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

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

**Proof.** We begin by deriving an upper bound on \( \|u_1^T, \ldots, u_m^T\|_2^2 \). Using the results of the previous lemma, it remains to notice that the objective \( \varphi(U) \) is invariant under transformations \( u_i \to u_i + t_i \mathbf{1}, t_i \in \mathbb{R} \) for \( i \in \{1, \ldots, m\} \), so there must exist some solution with \( \max_{\eta} [u_\xi]_\eta = -\min_{\eta} [u_\xi]_\eta = \|u_\xi\|_\infty \), so

\[
\|u_\xi\|_\infty \leq -\frac{1}{2} \ln \nu \min_{\eta} [r_i]_\eta.
\]

As a consequence,

\[
\|U^*\|_2 \leq \sqrt{mn}\|U^*\|_\infty \leq \\
\leq \frac{\sqrt{mn}}{2} \ln \nu \min_{i,j} \{[r_i]_j\} \\
\leq \frac{\sqrt{mn}}{2} \left( \|C\|_\infty - \frac{\gamma}{2} \ln \min_{i,j} \{[r_i]_j\} \right).
\]

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

\[
R = \|\Lambda^* - \Lambda^0\|_2 = \\
= \left\| (-\gamma u_1^T - \frac{1}{m} \mathbf{1}, \ldots, -\gamma u_m^T - \frac{1}{m} \mathbf{1}) \\
- (-\frac{\gamma}{m} \mathbf{1}, \ldots, -\frac{\gamma}{m} \mathbf{1}) \right\|_2 = \| -\gamma (u_1^*, \ldots, u_m^*) \|_2 \\
= \gamma \|U^*\|_2 \leq \sqrt{mn} \left( \|C\|_\infty - \frac{\gamma}{2} \ln \min_{i,j} \{[r_i]_j\} \right).
\]

\[\square\]

**C. 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 \( \gamma \). Further, we use the 2-norm for the dual space of Lagrange multipliers \( \Lambda \) in (3). Hence, the dual objective in (5) is \( L \)-smooth with the parameter \( L \leq \|A\|_1 \gamma^{-1} / 2 \) [22]. Here \( A : \mathbb{R}^{nm} \to \mathbb{R}^{mn} \) is the linear operator defining the linear constraints of the problem, which, in the case of the multilmarginal optimal transport problem, is defined by \( A \mathbf{vec}(X) = (p_1(X)^2, \ldots, p_m(X)^2)^T \). Thus, each column of the matrix \( A \) contains no more than \( m \) non-zero elements, which are equal to one. Hence, \( \|A\|_1 \gamma^{-1} \) is equal to maximum 2-norm of the column of this matrix, we have that \( \|A\|_1 \gamma^{-1} = \sqrt{m} \). We finally have that \( L \leq \frac{m \gamma}{2} \).

**D. Projection on the feasible set**

It may happen that the Algorithm \[\ref{1}\] returns 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 \( \tilde{X} \approx \hat{x}^T \) that \( p_i(\tilde{X}) = p_i \). To do this we formulate Algorithm \[\ref{2}\] which is a generalization of rounding procedure in [24], see also [18].

Note that in Algorithm \[\ref{2}\] the function \( \text{DiagTensor}() \) takes a vector as input and outputs a \( m \)-dimensional tensor with the input as its diagonal. Moreover, \( \text{ProdTensor}(A, B) \) takes two \( m \)-dimensional tensors as input, and multiplies them in the direction \( r \). We use \( \otimes \) to denote the tensor
Algorithm 2 Multimarginal Rounding

1: $F_1 = U$
2: for $r = 1, \ldots, m - 1$ do
3: \[ x_r = \min \{ [p_r]/[p_r(F_r)], 1 \} \]
4: \[ X_r = \text{DiagTensor}(x_r) \]
5: \[ F_{r+1} = \text{ProdTensor}(F_r, X_r) \]
6: end for
7: for $r = 1, \ldots, m - 1$ do
8: \[ \text{err}_r = p_r - p_r(F_m) \]
9: end for

Output: $F = F_m + \bigotimes_{r=1}^{m} \text{err}_r/\|\text{err}_m\|_1^{m-1}$

Algorithm 3 Approximate MOT by PD-AAM

Input: Accuracy $\varepsilon$.
1: Set $\gamma = \frac{\varepsilon}{2m \ln n}$, $\varepsilon' = \frac{\varepsilon}{8\sqrt{C \ln n}}$
2: Define $\tilde{p}_k = \left(1 - \frac{\varepsilon'}{4m}\right)(p_k + \frac{\varepsilon'}{4(4m - \varepsilon)})^k$, $k = 1, \ldots, m$.
3: Apply PD-AAM to the dual problem $\{7\}$ with marginals $\tilde{p}_k$, $k = 1, \ldots, m$ until the stopping criterion $\sum_{k=1}^{m} \| p_k(\hat{X}^t) - \tilde{p}_k \|_1 \leq \varepsilon'/2$ and $f(\hat{x}^t) + \varphi(\eta^t) \leq \varepsilon/4$.
4: Find $\hat{X}$ as the projection of $\hat{X}^t$ on $\{X \in \mathbb{R}_{+}^{n \times \cdots \times n}, p_k(X) = p_k, \forall k = 1, \ldots, m \}$ by the Algorithm 2.

Output: $\hat{X}$. 

The PD-AAM algorithm and estimates for $R$ and $L$, and the rounding procedure.

To adapt Algorithm $[1]$ to the problem $\{7\}$ one should replace step 4 with Choose $I = \arg\max_{i \in \{1, \ldots, m\}} \left\{ \| x_i^{\text{opt}}(\theta^t) \|_2 \right\}$ and step 5 with
\[ \eta_i^{t+1} = \left\{ \begin{array}{ll} \theta_i^t + \ln p_i - \ln p_i(B(\theta^t)), & i = I \\ \theta_i^t, & \text{otherwise}. \end{array} \right. \]

By Lemma $[4]$ $\hat{X}$ is a feasible point for Problem $\{1\}$. Let us estimate the objective residual. We have
\begin{align*}
\langle C, \hat{X} \rangle &= \langle C, X^* \rangle + \langle C, X^* - X^* \rangle + \langle C, \hat{X} - X^* \rangle \\
&\leq \langle C, X^* \rangle + \gamma m \ln n + f(\hat{x}^t) + \varphi(\eta^t) \\
&\quad + 2\sum_{k=1}^{m} \| p_k(\hat{X}^t) - p_k \|_1 \| C \|_2, \quad (19)
\end{align*}
where $\hat{X}^t$ is the output of Algorithm $[1]$, $\hat{X}$ is a projection of $X^t$ by Algorithm $[2]$ on the feasible set, $X^*$ is a solution to the non-regularized multimarginal OT problem $\{1\}$, $X^*$ 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, $\langle C, X^* - X^* \rangle \leq 0$ and
\begin{align*}
\langle C, \hat{X} - X^* \rangle &= \langle C, \hat{X}^t \rangle - \gamma H(\hat{X}^t) \\
&\quad - \langle C, X^* - \gamma H(X^*) \rangle + H(\hat{X}^t) - H(X^*) \\
&\leq f(\hat{x}^t) + \varphi(\eta^t) + \gamma m \ln n. \quad (20)
\end{align*}
Finally, by the Hölder inequality and Lemma $[4]$
\begin{align*}
\langle C, \hat{X} - \hat{X}^t \rangle &\leq \| C \|_2 \| \hat{X} - \hat{X}^t \|_1 \leq 2\| C \|_2 \sum_{k=1}^{m} \| p_k(\hat{X}^t) - p_k \|_1.
\end{align*}
This finishes the proof of inequality $\{19\}$.

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, $\{19\}$, the choice of $\gamma$ and $\varepsilon'$
as well as the stopping criterion in step 3 of Algorithm [3] we obtain that (18) holds.

It remains to estimate the complexity of the algorithm. By Theorem [2] we obtain that

\[
\frac{1}{l^2} \sum_{k=1}^{m} \left| p_k(\hat{X}^t) - \hat{p}_k \right|_1 \leq \sqrt{\frac{m}{\varepsilon}} \left| A \hat{X}^t - b \right|_2 \leq \frac{8m^3 n^2 \frac{1}{\varepsilon}}{l^2}.
\]

\[
= 8m^3 n \varepsilon \frac{1}{\varepsilon} \ln n \ln \frac{4m n}{\varepsilon} \left( 2 \frac{1}{\varepsilon} \right) \frac{\varepsilon}{4m n \ln \frac{32mn}{\varepsilon}},
\]

where the operator \( A \) is defined in Sect III-C and we used that by the choice of \( \hat{p}_k \), \( m_{i,j} \geq \frac{1}{4mn} \). At the same time,

\[
f(\hat{x}^t) + \varphi(\eta^t) \leq \frac{2m LR^2}{l^2} \frac{1}{\varepsilon} \ln \frac{1}{\varepsilon} \left( \frac{2}{\varepsilon} \right) \frac{\varepsilon}{4m \ln \frac{32mn}{\varepsilon}}.
\]

Let us denote \( \delta = 1 + \frac{\varepsilon}{4m \ln \frac{32mn}{\varepsilon}} \). Since \( \varepsilon \) is small and \( m, n \) are large, we can think of this quantity as \( \delta \approx O(1) \). Then, to satisfy the stopping criterion in step 3 of Algorithm [3] we need to take

\[
t \geq \sqrt{\frac{128m^4 n \varepsilon^2}{\varepsilon^2}} \frac{1}{\varepsilon} \frac{\varepsilon}{4m \ln \frac{32mn}{\varepsilon}} = O \left( \frac{m^2 n^{1/2} \varepsilon^2}{\varepsilon} \right),
\]

and

\[
t \geq \sqrt{\frac{4m^4 n \varepsilon^2}{\varepsilon^2}} \frac{1}{\varepsilon} \frac{\varepsilon}{4m \ln \frac{32mn}{\varepsilon}} = O \left( \frac{m^2 n^{1/2} \varepsilon^2}{\varepsilon} \right).
\]

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

**Theorem 5.** The complexity to find \( \varepsilon \)-approximate solution to the non-regularized multimarginal OT problem by Algorithm [3] is

\[
\hat{O} \left( \frac{m^3 n^{m+1/2} \varepsilon}{\varepsilon^2} \right).
\]

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 a shared memory and allow \( m \) workers to access this memory. Then, they can independently calculate all the marginals. The total amount of arithmetic operations remains the same, but the time of work is now proportional to \( m^m \) rather than to \( mn^m \).

Next, we compare our complexity results with the estimates in the preprint [18]. 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 the complexity of each iteration in their algorithms and in our algorithm are the same, we compare the iteration complexity of the algorithms. The iteration complexity of our algorithm is \( \tilde{O} \left( m^2 n^{1/2} \| C \|_x / \varepsilon \right) \). The iteration complexity of the multimarginal Sinkhorn’s algorithm [18] is \( \tilde{O} \left( m^3 \| C \|_x^2 / \varepsilon^2 \right) \), which has worse dependence on \( \varepsilon \) and \( m \) than our bound. The claimed iteration complexity of multimarginal RANDKHORN algorithm in [18] is \( \tilde{O} \left( m^{8/3} n^{1/3} \| C \|_x^{4/3} / \varepsilon \right) \), which has worse dependence on \( m \) and \( \| C \|_x \) than our bound. Moreover, the multimarginal RANDKHORN is a randomized algorithm and its complexity is estimated in average, whereas our algorithm and complexity is deterministic.

**V. PRELIMINARY EXPERIMENTS**

In this section, we provide a numerical comparison of multimarginal Sinkhorn’s algorithm from [18] with our AAM method. We performed experiments using randomly chosen vectors \( p_i \in \Delta_n \) and tensor \( C \in \mathbb{R}_+^m \). We slightly modified the smaller values of \( p_i \) as described above to lower bound their minimal value. We choose several values of accuracy \( \varepsilon \in [0.01, 0.25] \), and run the methods until the stopping criterion was reached. One can see, that our AAM algorithm outperforms multimarginal Sinkhorn’s algorithm from [18]. But in our experiment better performance of our method was shown with slightly different constants: a) we modified vectors \( p_i \) with \( \varepsilon' = \varepsilon / (32 \| C \|_x) \), b) for stopping criteria we fulfilled \( \sum_{k=1}^{m} \| p_k(\hat{X}^t) - \hat{p}_k \|_1 \leq \varepsilon / 8 \) and \( f(\hat{x}^t) + \varphi(\eta^t) \leq 3\varepsilon / 16 \). Unfortunately, we were not able to implement the multimarginal RANDKHORN algorithm since its stopping criterion \( E_t > \varepsilon' \) depends on expected residual in the constraints given in [18, Eq. (28)], which is unavailable in practice.
VI. CONCLUSIONS

We provide a novel algorithm for the computation of approximate solutions to the multimarginal optimal transport problem. Our results are based on 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 with respect to 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 [17], 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, as well as extensive experimental results for applications related to signal processing.

REFERENCES

[1] A. Gramfort, P. 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 preprint arXiv:1701.07875, 2017.
[3] S. Asoodeh, T. Gao, and J. Evans, “Curvature of hypergraphs via multimarginal optimal transport,” in 2018 IEEE Conference on Decision and Control (CDC). IEEE, 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 preprint arXiv:1905.03847, 2019.
[6] B. Pass, “Multi-marginal optimal transport: theory and applications,” ESAIM: Mathematical Modelling and Numerical Analysis, vol. 49, no. 6, pp. 1771–1790, 2015.
[7] C. Villani, Topics in Optimal Transportation, ser. Graduate studies in mathematics. American Mathematical Society, 2003. [Online]. Available: https://books.google.ru/books?id=idyFAwAAQBAJ
[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, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80, 2018, pp. 1367–1376, arXiv:1802.04367.
[9] A. Kroshnin, N. Tupitsa, D. Vdvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe, “On the complexity of approximating Wasserstein barycenters,” in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 3530–3540, arXiv:1901.08686.
[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, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 3982–3991.
[11] L. Ambrosio and N. Gigli, “A user’s guide to optimal transport,” in Modelling and optimisation of flows on networks. Springer, 2013, pp. 1–155.
[12] R. J. McCann, “A glimpse into the differential topology and geometry of optimal transport,” arXiv preprint arXiv:1207.1867, 2012.
[13] I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier, “Tomographic reconstruction from a few views: a multi-marginal optimal transport approach,” Applied Mathematics & Optimization, vol. 75, no. 1, pp. 55–73, 2017.