Journal of Machine Learning Research 25 (2024) 1-33 Submitted 5/22; Revised 7/24; Published 10/24

On the Convergence of Projected Alternating Maximization for Equitable and Optimal Transport

Minhui Huang MHHUANG@UCDAVIS.EDU
Department of Electrical and Computer Engineering
University of California
Davis, CA 95616-5270, USA

Shiqian Ma SQMA@RICE.EDU
Department of Computational Applied Mathematics and Operations Research
Rice University
Houston, TX 77005-1827, USA

Lifeng Lai LFLAI@UCDAVIS.EDU
Department of Electrical and Computer Engineering
University of California
Davis, CA 95616-5270, USA

Editor: Suvrit Sra

Abstract

This paper studies the equitable and optimal transport (EOT) problem, which has many applications such as fair division problems and optimal transport with multiple agents etc. In the discrete distributions case, the EOT problem can be formulated as a linear program (LP). Since this LP is prohibitively large for general LP solvers, (Scetbon et al., 2021) suggests to perturb the problem by adding an entropy regularization. They proposed a projected alternating maximization algorithm (PAM) to solve the dual of the entropy regularized EOT. In this paper, we provide the first convergence analysis of PAM. A novel rounding procedure is proposed to help construct the primal solution for the original EOT problem. We also propose a variant of PAM by incorporating the extrapolation technique that can numerically improve the performance of PAM. Results in this paper may shed lights on block coordinate (gradient) descent methods for general optimization problems.

Keywords: Equitable and Optimal Transport, Fairness, Saddle Point Problem, Projected Alternating Maximization, Block Coordinate Descent, Acceleration, Rounding.

1. Introduction

Optimal transport (OT) is a classical problem that recently finds many emerging applications in machine learning and artificial intelligence, including generative models (Arjovsky et al., 2017), representation learning (Ozair et al., 2019), reinforcement learning (Bellemare et al., 2017) and word embeddings (Alvarez-Melis et al., 2019) etc. More recently, (Scetbon et al., 2021) proposed an equitable and optimal transport (EOT) problem that targets to fairly distribute the workload of OT when there are multiple agents. In this problem formulation, there are multiple agents working together to move mass from measures μ to ν and each agent has its unique cost function. A very important issue that needs to be considered here is the fairness, which aims at finding transportation plans such that the workloads among all the agents are equal to each other. This can be achieved by minimizing the largest transportation cost among all agents, which leads to a convex-concave saddle point problem. The EOT problem has wide applications in economics and machine learning, such as fair division or the cake-cutting problem (Moulin, 2003; Brandt et al., 2016), multi-type resource allocation (Mackin and Xia, 2015), internet minimal transportation time and sequential optimal transport (Scetbon et al., 2021).

We now describe the EOT problem formally. Given two discrete probability measures $\mu_n = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu_n = \sum_{i=1}^n b_i \delta_{y_i}$, the EOT studies the problem of transporting mass from μ to ν by N agents. Here, $\{x_1, x_2, ..., x_n\} \subset \mathbb{R}^d$ and $\{y_1, y_2, ..., y_n\} \subset \mathbb{R}^d$ are the support points of each measure and $a = [a_1, a_2, ..., a_n]^\top \in \Delta_+^n$, $b = [b_1, b_2, ..., b_n]^\top \in \Delta_+^n$ are corresponding weights for each measure, where Δ_+^n denotes the probability simplex in $\mathbb{R}^n$. Moreover, throughout this paper, we assume vector $b > 0$. For each agent k, we denote its unique cost function as $c^k(x, y), k \in [N] = \{1, \ldots, N\}$ and its cost matrix as C^k, where $C^k_{i,j} = c^k(x_i, y_j)$. Moreover, we define the following coupling decomposition set

$$\Pi_{a,b}^N := \left\{ \boldsymbol{\pi} = (\pi^k)_{k\in[N]} \,\middle|\, r\left(\sum_k \pi^k\right) = a,\ c\left(\sum_k \pi^k\right) = b, \quad \pi^k_{ij} \geq 0, \forall i, j \in [n] \right\},$$

where $r(\pi) = \pi\mathbf{1}, c(\pi) = \pi^\top \mathbf{1}$ are the row sum and column sum of matrix π respectively. Mathematically, the EOT problem can be formulated as

$$\min_{\boldsymbol{\pi}\in\Pi_{a,b}^N} \max_{1\leq k\leq N} \langle \pi^k, C^k \rangle. \tag{1}$$

When $N = 1$, (1) reduces to the standard OT problem. Note that (1) minimizes the point-wise maximum of a finite collection of functions. It is easy to see that (1) is equivalent to the following constrained problem:

$$\min_{\boldsymbol{\pi}\in\Pi_{a,b}^N} \max_{\lambda\in\Delta_+^N} \ell(\boldsymbol{\pi}, \lambda) := \sum_{k=1}^N \lambda_k \langle \pi^k, C^k \rangle. \tag{2}$$

The following proposition shows an important property of EOT: at the optimum of the minimax EOT formulation (2), the transportation costs of the agents are equal to each other.

Proposition 1 *(Scetbon et al., 2021)[Proposition 1] Assume that all entries of all cost matrices $C^k, k \in [N]$ have the same sign. Let $\boldsymbol{\pi}^* \in \Pi_{a,b}^N$ be the optimal solution of* (2)*. It holds that*

$$\langle (\pi^*)^i, C^i \rangle = \langle (\pi^*)^j, C^j \rangle, \quad \forall i, j \in [N]. \tag{3}$$

Note that Proposition 1 requires all entries of all cost matrices to have the same sign. When the cost matrices are all non-negative, (2) solves the transportation problem with multiple agents. When the cost matrices are all non-positive, the cost matrices are interpreted as the utility functions and (2) solves the fair division problem (Moulin, 2003).

The discrete OT is a linear programming (LP) problem (in fact, an assignment problem) with a complexity of $O(n^3 \log n)$ (Tarjan, 1997). Due to this cubic dependence on the dimension n, it is challenging to solve large-scale OT in practice. A widely adopted compromise is to add an entropy regularizer to the OT problem (Cuturi, 2013). The resulting problem is strongly convex and smooth, and its dual problem can be efficiently solved by the celebrated Sinkhorn's algorithm (Sinkhorn and Knopp, 1967; Cuturi, 2013). This strategy is now widely used in the OT community due to its computational advantages as well as improved sample complexity (Genevay et al., 2019). Similar ideas were also used for computing the Wasserstein barycenter (Benamou et al., 2015), projection robust Wasserstein distance (Paty and Cuturi, 2019; Lin et al., 2020; Huang et al., 2021a), projection robust Wasserstein barycenter (Huang et al., 2021b). Motivated by these previous works, (Scetbon et al., 2021) proposed to add an entropy regularizer to (2), and designed a projected alternating maximization algorithm (PAM) to solve its dual problem. However, the convergence of PAM has not been studied. (Scetbon et al., 2021) also proposed an accelerated projected gradient ascent algorithm (APGA) for solving a different form of the dual problem of the entropy regularized EOT. Since the objective function of this new dual form has Lipschitz continuous gradient, APGA is essentially the Nesterov's accelerated gradient method and thus its convergence rate is known. However, numerical experiments conducted in (Scetbon et al., 2021) indicate that APGA performs worse than PAM. We will discuss the reasons in details later.

Our Contributions. There are mainly three issues with the PAM and APGA algorithms in (Scetbon et al., 2021), and we will address all of them in this paper. Our results may shed lights on designing new block coordinate descent algorithms. Our main contributions are given below.

- The PAM algorithm in (Scetbon et al., 2021) only returns the dual variables. How to find the primal solution of (2), i.e., the optimal transport plans $\boldsymbol{\pi}$, was not discussed in (Scetbon et al., 2021). In this paper, we propose a novel rounding procedure to find the primal solution. Our rounding procedure is different from the one widely used in the literature (Altschuler et al., 2017).

- We provide the first convergence analysis of the PAM algorithm, and analyze its iteration complexity for finding an ϵ-optimal solution to the EOT problem (2). In particular, we show that it takes at most $O(Nn^2\epsilon^{-2})$ arithmetic operations to find an ϵ-optimal solution to (2). This matches the rate of the Sinkhorn's algorithm for computing the Wasserstein distance (Dvurechensky et al., 2018).

- We propose a variant of PAM that incorporates the extrapolation technique as used in Nesterov's accelerated gradient method. We name this variant as PAM with Extrapolation (PAME). The iteration complexity of PAME is also analyzed. Though we are not able to prove a better complexity over PAM at this moment, we find that PAME performs much better than PAM numerically.

Notation. For vectors a and b with the same dimension, $a./b$ denotes their entry-wise division. We denote $c_\infty := \max_k \|C^k\|_\infty$. Throughout this paper, we assume vector $b > 0$, and we denote $\iota := \min_j \log(b_j)$. We use $\mathbf{1}_n$ to denote the n-dimensional vector whose entries are all equal to one. We use $\mathbb{I}_{\mathcal{X}}(x)$ to denote the indicator function of set $\mathcal{X}$, i.e., $\mathbb{I}_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$, and $\mathbb{I}_{\mathcal{X}}(x) = \infty$ otherwise. We denote $c^t = c(\sum_{k=1}^N \pi^k(f^{t+1}, g^t, \lambda^t))$. For integer $N > 0$, we denote $[N] := \{1, \ldots, N\}$. We also denote $\boldsymbol{\pi}(f, g, \lambda) = [\pi^k(f, g, \lambda)]_{k\in[N]}$.

2. Projected Alternating Maximization Algorithm

The PAM algorithm proposed in (Scetbon et al., 2021) aims to solve the entropy regularized EOT problem, which is given by

$$\min_{\boldsymbol{\pi}\in\Pi_{a,b}^N} \max_{\lambda\in\Delta_+^N} \ \ell_\eta(\boldsymbol{\pi},\lambda) := \sum_{k=1}^N p_\eta^k(\pi^k,\lambda) \tag{4}$$

where $\eta > 0$ is a regularization parameter, $p_\eta^k(\pi^k,\lambda) := \lambda_k\langle\pi^k, C^k\rangle - \eta H(\pi^k)$, and the entropy function H is defined as $H(\pi) = -\sum_{i,j}\pi_{i,j}(\log\pi_{i,j} - 1)$. The entropy regularization was first introduced into the OT problem by (Cuturi, 2013) and is now widely used in the OT community. By adding an entropy regularizer, the primal problem becomes strongly convex and the dual problem is unconstrained and is suitable for alternating maximization. This leads to the Sinkhorn's algorithm which has low per-iteration complexity and thus is scalable. The PAM algorithm proposed by (Scetbon et al., 2021) used the same idea for the EOT problem. Note that (4) is a strongly-convex-concave minimax problem whose constraint sets are convex and bounded, and thus the Sion's minimax theorem (Sion, 1958) guarantees that

$$\min_{\boldsymbol{\pi}\in\Pi_{a,b}^N} \max_{\lambda\in\Delta_+^N} \ \ell_\eta(\boldsymbol{\pi},\lambda) = \max_{\lambda\in\Delta_+^N} \min_{\boldsymbol{\pi}\in\Pi_{a,b}^N} \ \ell_\eta(\boldsymbol{\pi},\lambda). \tag{5}$$

Now we consider the dual problem of $\min_{\boldsymbol{\pi}\in\Pi_{a,b}^N} \ell_\eta(\boldsymbol{\pi},\lambda)$. First, we add a redundant constraint $\sum_{k,i,j}\pi_{i,j}^k = 1$ and consider the dual of

$$\min_{\boldsymbol{\pi}\in\Pi_{a,b}^N, \sum_{k,i,j}\pi_{i,j}^k=1} \ell_\eta(\boldsymbol{\pi},\lambda). \tag{6}$$

The reason for adding this redundant constraint is to guarantee that the dual objective function is Lipschitz smooth. It is easy to verify that the dual problem of (6) is given by

$$\max_{f,g} \min_{\substack{\sum_{k,i,j}\pi_{i,j}^k=1,\\ \boldsymbol{\pi}\in\left(\mathbb{R}_+^{n\times n}\right)^N}} \sum_{k=1}^N \lambda_k\langle\pi^k, C^k\rangle - \eta H(\boldsymbol{\pi}) + f^\top\left(a - r\left(\sum_k \pi^k\right)\right) + g^\top\left(b - c\left(\sum_k (\pi^k)^\top\right)\right), \tag{7}$$

where f and g are the dual variables and $H(\boldsymbol{\pi}) = \sum_k H(\pi^k)$. It is noted that problem (7) admits the following solution:

$$\pi^k(f,g,\lambda) = \frac{\zeta^k(f,g,\lambda)}{\sum_k\|\zeta^k(f,g,\lambda)\|_1}, \quad \forall k\in[N], \tag{8}$$

where

$$\zeta^k(f,g,\lambda) = \exp\left(\frac{f\mathbf{1}_n^\top + \mathbf{1}_n g^\top - \lambda_k C^k}{\eta}\right), \quad \forall k\in[N]. \tag{9}$$

By plugging (8) into (7), we obtain the following dual problem of (6):

$$\max_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^n} \langle f,a\rangle + \langle g,b\rangle - \eta\log\left(\sum_{k=1}^N\|\zeta^k(f,g,\lambda)\|_1\right) - \eta. \tag{10}$$

Plugging (10) into (5), we know that the entropy regularized EOT problem (4) is equivalent to a pure maximization problem:

$$\max_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^n,\ \lambda\in\Delta_+^N} F(f,g,\lambda) := \langle f, a\rangle + \langle g, b\rangle - \eta\log\left(\sum_{k=1}^N \|\zeta^k(f,g,\lambda)\|_1\right) - \eta. \tag{11}$$

Function $F(f,g,\lambda)$ is a smooth concave function with three block variables (f,g,λ). We use (f^*,g^*,λ^*) to denote an optimal solution of (11), and we denote $F^* = F(f^*,g^*,\lambda^*)$. The PAM algorithm proposed in (Scetbon et al., 2021) is essentially a block coordinate descent (BCD) algorithm for solving (11). More specifically, the PAM updates the three block variables by the following scheme:

$$f^{t+1} \in \operatorname*{argmax}_f F(f, g^t, \lambda^t), \tag{12a}$$

$$g^{t+1} \in \operatorname*{argmax}_g F(f^{t+1}, g, \lambda^t), \tag{12b}$$

$$\lambda^{t+1} := \mathrm{Proj}_{\Delta_+^N}\left(\lambda^t + \tau\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t)\right). \tag{12c}$$

Each iteration of PAM consists of two exact maximization steps followed by one projected gradient step. Importantly, the two exact maximization problems (12a)-(12b) have numerous optimal solutions, and we choose to use the following ones:

$$f^{t+1} = f^t + \eta\log\left(\frac{a}{r\left(\sum_{k=1}^N \zeta^k(f^t,g^t,\lambda^t)\right)}\right), \tag{13}$$

$$g^{t+1} = g^t + \eta\log\left(\frac{b}{c\left(\sum_{k=1}^N \zeta^k(f^{t+1},g^t,\lambda^t)\right)}\right). \tag{14}$$

Furthermore, the optimality conditions of (12a)-(12b) imply that

$$\begin{aligned} &a - \frac{r\left(\sum_{k=1}^N \zeta^k(f^{t+1},g^t,\lambda^t)\right)}{\sum_k \|\zeta^k(f^{t+1},g^t,\lambda^t)\|_1} = 0,\\ &b - \frac{c\left(\sum_{k=1}^N \zeta^k(f^{t+1},g^{t+1},\lambda^t)\right)}{\sum_k \|\zeta^k(f^{t+1},g^{t+1},\lambda^t)\|_1} = 0, \quad \forall t. \end{aligned} \tag{15}$$

However, we need to point out that the PAM (12) only returns the dual variables (f^t,g^t,λ^t). One can compute the primal variable $\boldsymbol{\pi}$ using (8), but it is not necessarily a feasible solution. That is, $\boldsymbol{\pi}$ computed from (8) does not satisfy $\boldsymbol{\pi}\in\Pi_{a,b}^N$. How to obtain an optimal primal solution from the dual variables was not discussed in (Scetbon et al., 2021). For the OT problem, i.e., $N=1$, a rounding procedure for returning a feasible primal solution has been proposed in (Altschuler et al., 2017). However, this rounding procedure cannot be applied to the EOT problem directly. In the next section, we propose a new rounding procedure for returning a primal solution based on the dual solution (f^t,g^t,λ^t). This new rounding procedure involves a dedicated way to compute the margins.

2.1 The Rounding Procedure and the Margins

Given $a \in \Delta_+^n$, $b \in \Delta_+^n$, and $\boldsymbol{\pi} = \{\pi^k\}_{k\in[N]}$ satisfying $r(\sum_k \pi^k) = a$, we construct vectors $a^k, b^k \in \mathbb{R}^n, k \in [N]$ from the procedure

$$(a^k, b^k)_{k\in[N]} = \text{Margins}(\boldsymbol{\pi}, a, b). \tag{16}$$

The details of this procedure are given below. First, we set $a^k = r(\pi^k)$, which immediately implies $\sum_{k=1}^N a^k = a$. We then construct b^k such that the following properties hold (these properties are required in our convergence analysis later):

(i) $b^k \geq 0$;

(ii) $\sum_{k=1}^N b^k = b$;

(iii) $\sum_{i=1}^n a_i^k = \sum_{j=1}^n b_j^k, \quad \forall k \in [N]$;

(iv) For any fixed $j \in [n]$, the quantities $b_j^k - [c(\pi^k)]_j$ have the same sign for all $k \in [N]$. That is, for any k and k', we have

$$(b_j^k - [c(\pi^k)]_j) \cdot (b_j^{k'} - [c(\pi^{k'})]_j) \geq 0, \tag{17}$$

which provides the following identity that is useful in our convergence analysis later:

$$\begin{aligned}
&\sum_{k=1}^N \|b^k - c(\pi^k)\|_1 \\
=&\sum_{k=1}^N \sum_{j=1}^n |b_j^k - [c(\pi^k)]_j| = \sum_{j=1}^n \left|\sum_{k=1}^N (b_j^k - [c(\pi^k)]_j)\right| \\
=&\sum_{j=1}^n \left| b_j - \left[c\left(\sum_{k=1}^N \pi^k\right)\right]_j \right| = \left\| b - c\left(\sum_{k=1}^N \pi^k\right)\right\|_1 .
\end{aligned} \tag{18}$$

The procedure on constructing $(b^k)_{k\in[N]}$ satisfying these four properties is provided in Appendix **??**.

After $(a^k, b^k)_{k\in[N]}$ are constructed from (16) with $\boldsymbol{\pi} = \boldsymbol{\pi}(f^T, g^{T-1}, \lambda^{T-1})$, we adopt the rounding procedure proposed in (Altschuler et al., 2017) to output a primal feasible solution $(\hat{\pi}^k)_{k\in[N]}$. The rounding procedure is described in Algorithm 2.

With this new procedure for rounding and computing the margins a^k, b^k, we now formally describe our PAM algorithm in Algorithm 1. Note that the algorithm is terminated when the following criteria are met:

$$\|c^{t-1} - b\|_1 \leq \epsilon/(6(6c_\infty - \eta\iota)), \tag{19a}$$

$$\left\|\lambda^t - \lambda^{t-1}\right\|_2 \leq \eta\epsilon/(18c_\infty^2), \tag{19b}$$

$$\tilde{F}(f^t, g^{t-1}, \lambda^{t-1}) \leq \epsilon/6, \tag{19c}$$

where $\tilde{F}(f, g, \lambda)$ is the suboptimality defined as: $\tilde{F}(f, g, \lambda) = F(f^*, g^*, \lambda^*) - F(f, g, \lambda)$.

Algorithm 1 Projected Alternating Maximization Algorithm

1: **Input:** Cost matrices $\{C^k\}_{1\le k\le N}$, vectors $a, b \in \Delta_+^n$ with $b > 0$, accuracy ϵ. $f^0 = g^0 = [1, ..., 1]^\top$, $\lambda^0 = [1/N, ..., 1/N]^\top \in \Delta_+^N$. $t = 0$
2: Choose parameters as

$$\eta = \min\left\{\frac{\epsilon}{3(\log(n^2N)+1)}, c_\infty\right\}, \quad \tau = \eta/c_\infty^2. \tag{20}$$

3: **while** (19) is not met **do**
4: Compute f^{t+1} by (13)
5: Compute g^{t+1} by (14)
6: Compute λ^{t+1} by (12c)
7: $t \leftarrow t+1$
8: **end while**
9: Assume stopping condition (19) is satisfied at the T-th iteration. Compute $(a^k, b^k)_{k\in[N]} = \text{Margins}(\boldsymbol{\pi}(f^T, g^{T-1}, \lambda^{T-1}), a, b)$ as in Section 2.1.
10: **Output:** $(\hat\pi, \hat\lambda)$ where $\hat\pi^k = \text{Round}(\pi^k(f^T, g^{T-1}, \lambda^{T-1}), a^k, b^k)$, $\forall k \in [N]$, $\hat\lambda = \lambda^{T-1}$.

Algorithm 2 Round(π, a, b)

1: **Input:** $\pi \in \mathbb{R}^{n\times n}$, $a \in \mathbb{R}_+^n$, $b \in \mathbb{R}_+^n$.
2: $X = \text{Diag}\,(x)$ with $x_i = \frac{a_i}{r(\pi)_i} \wedge 1$
3: $\pi' = X\pi$
4: $Y = \text{Diag}\,(y)$ with $y_j = \frac{b_j}{c(\pi')_j} \wedge 1$
5: $\pi'' = \pi' Y$
6: $err_a = a - r(\pi''), err_b = b - c(\pi'')$
7: **Output:** $\pi'' + err_a err_b^\top / \|err_a\|_1$.

2.2 Connections with BCD and BCGD Methods

The block coordinate descent (BCD) algorithm minimizes a multivariate objective function by iteratively updating subsets of variables while keeping the rest fixed. The convergence of BCD has been studied in many literatures (Beck and Tetruashvili, 2013; Diakonikolas and Orecchia, 2018; Sun and Hong, 2015). We now discuss the connections between PAM and the block coordinate descent method and the block coordinate gradient descent (BCGD) method. For the ease of presentation, we now assume that we are dealing with the following general convex optimization problem with m block variables:

$$\min_{x_i\in\mathcal{X}_i, i=1,\ldots,m} J(x_1, x_2, \ldots, x_m), \tag{21}$$

where $\mathcal{X}_i \subset \mathbb{R}^{d_i}$ and J is convex and differentiable. The BCD method for solving (21) iterates as follows:

$$x_i^{t+1} = \operatorname*{argmin}_{x_i\in\mathcal{X}_i} J(x_1^{t+1}, x_2^{t+1}, \ldots, x_{i-1}^{t+1}, x_i, x_{i+1}^t, \ldots, x_m^t), \tag{22}$$

and it assumes that these subproblems are easy to solve. The BCGD method for solving (21) iterates as follows:

$$x_i^{t+1} = \operatorname*{argmin}_{x_i \in \mathcal{X}_i} \langle \nabla_{x_i} J(x_1^{t+1}, \ldots, x_{i-1}^{t+1}, x_i, x_{i+1}^t, \ldots, x_m^t), x_i - x_i^t \rangle + \frac{1}{2\tau} \|x_i - x_i^t\|_2^2, \tag{23}$$

where $\tau > 0$ is the step size. The PAM (12) is a hybrid of BCD (22) and BCGD (23), in the sense that some block variables are updated by exactly solving a maximization problem (the f and g steps), and some other block variables are updated by taking a gradient step (the λ step). Though this hybrid idea has been studied in the literature (Hong et al., 2017; Xu and Yin, 2013), their convergence analysis requires the blocks corresponding to exact minimization to be strongly convex. However, in our problem (11), the negative of the objective function is merely convex. Hence we need to develop new convergence proofs to analyze the convergence of PAM (Algorithm 1). How to extend our convergence results of PAM (Algorithm 1) to more general settings is a very interesting topic for future study.

3. Convergence Analysis of PAM

In this section, we analyze the iteration complexity of Algorithm 1 for obtaining an ϵ-optimal solution to the original EOT problem (2). The ϵ-optimal solution to (2) is defined as follows.

Definition 2 (see, e.g., (Nemirovski, 2005)) *We call $(\hat{\boldsymbol{\pi}}, \hat{\lambda}) \in \Pi_{a,b}^N \times \Delta_+^N$ an ϵ-optimal solution to the EOT problem* (2) *if the following inequality holds:*

$$\max_{\lambda \in \Delta_+^N} \ell(\hat{\boldsymbol{\pi}}, \lambda) - \min_{\boldsymbol{\pi} \in \Pi_{a,b}^N} \ell(\boldsymbol{\pi}, \hat{\lambda}) \leq \epsilon.$$

Note that the left hand side of the inequality is the duality gap of (2).

3.1 Technical Preparations

We first give the partial gradients of F.

$$[\nabla_f F(f, g, \lambda)]_i = a_i - \frac{\sum_{k,j} \exp((f_i + g_j - \lambda_k C_{ij}^k)/\eta)}{\sum_k \|\zeta^k(f, g, \lambda)\|_1} = a_i - \left[r \left(\sum_k \pi^k(f, g, \lambda) \right) \right]_i, \tag{24a}$$

$$[\nabla_g F(f, g, \lambda)]_j = b_j - \frac{\sum_{k,i} \exp((f_i + g_j - \lambda_k C_{ij}^k)/\eta)}{\sum_k \|\zeta^k(f, g, \lambda)\|_1} = b_j - \left[c \left(\sum_k \pi^k(f, g, \lambda) \right) \right]_j, \tag{24b}$$

$$[\nabla_\lambda F(f, g, \lambda)]_k = \frac{\sum_{i,j} C_{ij}^k \exp((f_i + g_j - \lambda_k C_{ij}^k)/\eta)}{\sum_k \|\zeta^k(f, g, \lambda)\|_1} = \langle \pi^k(f, g, \lambda), C^k \rangle. \tag{24c}$$

Since (13) and (14) renormalize the row sum and column sum of $\sum_k \zeta^k(f, g, \lambda)$ to be a and b, we immediately have

$$\sum_{k=1}^N \|\zeta^k(f^{t+1}, g^t, \lambda^t)\|_1 = 1, \quad \sum_{k=1}^N \|\zeta^k(f^{t+1}, g^{t+1}, \lambda^t)\|_1 = 1, \forall t, \tag{25}$$

which, combined with (8), yields

$$\pi^k(f^{t+1}, g^t, \lambda^t) = \zeta^k(f^{t+1}, g^t, \lambda^t), \quad \pi^k(f^{t+1}, g^{t+1}, \lambda^t) = \zeta^k(f^{t+1}, g^{t+1}, \lambda^t), \forall t. \tag{26}$$

The following lemma gives an error bound for Algorithm 2 (see (Altschuler et al., 2017)).

Lemma 3 (Rounding Error) *Let $a, b \in \mathbb{R}^n_+$ with $\sum_{i=1}^n a_i = \sum_{j=1}^n b_j = q$, $\pi \in \mathbb{R}^{n\times n}_+$, and $\hat{\pi} = Round(\pi, a, b)$. The following inequality holds:*

$$\|\hat{\pi} - \pi\|_1 \leq 2(\|r(\pi) - a\|_1 + \|c(\pi) - b\|_1).$$

Proof. This a directly result of (Altschuler et al., 2017)[Lemma 7]. □

The following corollary follows (Scetbon et al., 2021)[Proposition 12] and shows that $\nabla_\lambda F$ is Lipschitz continuous.

Corollary 4 *For any $f, g \in \mathbb{R}^n$ and $\lambda^1, \lambda^2 \in \Delta^N_+$, the following inequality holds*

$$\|\nabla_\lambda F(f, g, \lambda^1) - \nabla_\lambda F(f, g, \lambda^2)\|_2 \leq c_\infty^2 \|\lambda^1 - \lambda^2\|_2/\eta, \tag{27}$$

which immediately implies

$$F(f, g, \lambda^1) \geq F(f, g, \lambda^2) + \langle \nabla_\lambda F(f, g, \lambda^2), \lambda^1 - \lambda^2 \rangle - \frac{c_\infty^2}{2\eta}\|\lambda^1 - \lambda^2\|_2^2. \tag{28}$$

The next corollary is an extension of Dvurechensky et al. (2018)[Lemma 1] and gives a bound for g.

Corollary 5 *Let (f^t, g^t, λ^t) be the sequence generated by Algorithm 1. For any $t \geq 0$, it holds that*

$$\max_j g_j^t - \min_j g_j^t \leq c_\infty - \eta\iota, \tag{29a}$$

$$\max_j g_j^* - \min_j g_j^* \leq c_\infty - \eta\iota. \tag{29b}$$

Lemma 6 *Let $\{f^t, g^t, \lambda^t\}$ be generated by PAM (Algorithm 1). The following equality holds.*

$$\sum_k^N \|\pi^k(f^{t+1}, g^{t+1}, \lambda^t) - \pi^k(f^{t+1}, g^t, \lambda^t)\|_1 = \left\|c^t - b\right\|_1, \forall t.$$

Proof. By (26), we have

$$\begin{aligned}
&\sum_k^N \|\pi^k(f^{t+1}, g^{t+1}, \lambda^t) - \pi^k(f^{t+1}, g^t, \lambda^t)\|_1 \\
=&\sum_k^N \sum_{i,j} |e^{\left(f_i^{t+1} + g_j^{t+1} - \lambda_k^t C_{i,j}^k\right)/\eta} - e^{\left(f_i^{t+1} + g_j^{t} - \lambda_k^t C_{i,j}^k\right)/\eta}| \\
=&\sum_k^N \sum_{i,j} [\pi^k(f^{t+1}, g^t, \lambda^t)]_{i,j} \left|b_j/c_j^t - 1\right| = \sum_j c_j^t |b_j/c_j^t - 1| = \left\|c^t - b\right\|_1.
\end{aligned}$$

□

3.2 Key Lemmas

In this subsection, we provide a few useful lemmas that will lead to our main theorem on the iteration complexity of PAM (Algorithm 1). These lemmas yield the following results: the function F is monotonically increasing (Lemmas 7), the suboptimality of the dual problem can be upper bounded (Lemma 8-10), and the PAM returns an ϵ-optimal solution under conditions (43) (Lemma 11). In Theorem 12 we will show that these conditions can indeed be satisfied.

Lemma 7 *[Increase of F] Let $\{f^t, g^t, \lambda^t\}$ be generated by PAM (Algorithm 1). The following inequalities hold:*

$$\begin{aligned} F(f^{t+1}, g^t, \lambda^t) - F(f^t, g^t, \lambda^t) &\geq 0 & \text{(30a)}\\ F(f^{t+1}, g^{t+1}, \lambda^t) - F(f^{t+1}, g^t, \lambda^t) &\geq \frac{\eta}{2}\left\|c^t - b\right\|_1^2 & \text{(30b)}\\ F(f^{t+1}, g^{t+1}, \lambda^{t+1}) - F(f^{t+1}, g^{t+1}, \lambda^t) &\geq c_\infty^2\|\lambda^{t+1} - \lambda^t\|^2/(2\eta). & \text{(30c)} \end{aligned}$$

Proof.

First, (30a) is a direct consequence of (12a).

Next, we prove (30b). We have

$$\begin{aligned} &F(f^{t+1}, g^{t+1}, \lambda^t) - F(f^{t+1}, g^t, \lambda^t)\\ =& \langle g^{t+1} - g^t, b\rangle - \eta\log\left(\sum_{k=1}^N \|\zeta^k(f^{t+1}, g^{t+1}, \lambda^t)\|_1\right) + \eta\log\left(\sum_{k=1}^N \|\zeta^k(f^{t+1}, g^t, \lambda^t)\|_1\right)\\ =& \langle g^{t+1} - g^t, b\rangle = \eta\sum_{j=1}^n b_j\log(b_j/c_j^t) = \eta\mathcal{K}(b||c^t) \geq \frac{\eta}{2}\|c^t - b\|_1^2, \end{aligned}$$

where $\mathcal{K}(x||y)$ denotes the KL divergence of x and y, the second equality is due to (25), the third equality is due to (14), and the last inequality follows the Pinsker's inequality.

Finally, we prove (30c). From the optimality condition of (12c), we know that there exists

$$h(\lambda^{t+1}) \in \partial\mathbb{I}_{\Delta_N^+}(\lambda^{t+1}) \tag{31}$$

such that

$$\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) - \frac{1}{\tau}(\lambda^{t+1} - \lambda^t) - h(\lambda^{t+1}) = 0. \tag{32}$$

From (28) we have

$$\begin{aligned} &F(f^{t+1}, g^{t+1}, \lambda^{t+1}) - F(f^{t+1}, g^{t+1}, \lambda^t)\\ \geq&\langle\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t), \lambda^{t+1} - \lambda^t\rangle - \frac{c_\infty^2}{2\eta}\|\lambda^{t+1} - \lambda^t\|^2\\ =&\langle\frac{1}{\tau}(\lambda^{t+1} - \lambda^t) + h(\lambda^{t+1}), \lambda^{t+1} - \lambda^t\rangle - \frac{c_\infty^2}{2\eta}\|\lambda^{t+1} - \lambda^t\|^2\\ \geq&\langle\frac{1}{\tau}(\lambda^{t+1} - \lambda^t), \lambda^{t+1} - \lambda^t\rangle - \frac{c_\infty^2}{2\eta}\|\lambda^{t+1} - \lambda^t\|^2\\ =&c_\infty^2\|\lambda^{t+1} - \lambda^t\|^2/(2\eta), \end{aligned}$$

where the first equality is due to (32), the second inequality is due to (31), and the last equality is due to the definition of τ in (20).

$\square$

Before we bound the suboptimality gap, we need the following lemma.

Lemma 8 *Let $\{f^t, g^t, \lambda^t\}$ be generated by PAM (Algorithm 1). For any $\lambda \in \Delta_N^+$, the following inequality holds:*

$$\left\langle \lambda - \lambda^t, \nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \right\rangle \leq 3c_\infty^2 \|\lambda^{t+1} - \lambda^t\|_2/\eta + c_\infty \left\|c^t - b\right\|_1. \tag{33}$$

Proof. The optimality condition of (12c) is given by:

$$\langle \lambda - \lambda^{t+1}, \frac{1}{\tau}(\lambda^{t+1} - \lambda^t) - \nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) \rangle \geq 0, \quad \forall \lambda \in \Delta_N^+, \tag{34}$$

which implies that

$$\begin{aligned} &\langle \lambda^{t+1} - \lambda, -\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) \rangle \\ \leq &\langle \lambda - \lambda^{t+1}, \frac{1}{\tau}(\lambda^{t+1} - \lambda^t) \rangle \leq \frac{1}{\tau}\|\lambda - \lambda^{t+1}\|_2 \|\lambda^{t+1} - \lambda^t\|_2 \leq 2c_\infty^2 \|\lambda^{t+1} - \lambda^t\|_2/\eta, \end{aligned} \tag{35}$$

where the last inequality is due to the fact that the diameter of Δ_N^+ is bounded by $\sqrt{2} \leq 2$. Moreover, we have

$$\begin{aligned} &\langle \lambda^t - \lambda, \nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) - \nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \rangle \\ = &\sum_k^N (\lambda_k^t - \lambda_k) \cdot \langle \pi^k(f^{t+1}, g^{t+1}, \lambda^t) - \pi^k(f^{t+1}, g^t, \lambda^t), C^k \rangle \\ \leq &\sum_k^N \|\pi^k(f^{t+1}, g^{t+1}, \lambda^t) - \pi^k(f^{t+1}, g^t, \lambda^t)\|_1 \|C^k\|_\infty \\ \leq &\, c_\infty \|c^t - b\|_1, \end{aligned} \tag{36}$$

where the equality is due to (24c), and the last inequality is due to Lemma 6. Finally, we have

$$\begin{aligned} &\left\langle \lambda^t - \lambda, -\nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \right\rangle \\ = &\left\langle \lambda^t - \lambda^{t+1}, -\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) \right\rangle + \left\langle \lambda^{t+1} - \lambda, -\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) \right\rangle + \\ &\left\langle \lambda^t - \lambda, \nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) - \nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \right\rangle \\ \leq &\|\lambda^t - \lambda^{t+1}\|_2 \cdot \|\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t)\|_2 + 2c_\infty^2 \|\lambda^{t+1} - \lambda^t\|_2/\eta + c_\infty \|c^t - b\|_1, \end{aligned} \tag{37}$$

where the first inequality is due to (35) and (36). From (24c) we have $\|\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t)\|_2 \leq c_\infty$, which, combined with (37) and the fact that $\eta \leq c_\infty$, yields the desired result. $\square$

The suboptimality of (11) is defined as: $\tilde{F}(f, g, \lambda) = F(f^*, g^*, \lambda^*) - F(f, g, \lambda)$. Note that $\tilde{F}(f, g, \lambda) \geq 0, \forall f, g, \lambda \in \Delta_N^+$.

Lemma 9 *Let (f^t, g^t, λ^t) be generated by PAM (Algorithm 1). The following inequality holds:*

$$\tilde{F}(f^{t+1}, g^t, \lambda^t) \leq (2c_\infty - \eta\iota)\|c^t - b\|_1 + 3c_\infty^2\|\lambda^{t+1} - \lambda^t\|_2/\eta.$$

Proof. Denote $u^t = (\max_j g_j^t + \min_j g_j^t)/2, u^* = (\max_j g_j^* + \min_j g_j^*)/2$. From (26) we get

$$\langle \mathbf{1}, c^t - b\rangle = \sum_{i=1}^n a_i - \sum_{j=1}^n b_j = 0,$$

which further implies

$$\begin{aligned}\left\langle g^t - g^*, c^t - b\right\rangle &= \left\langle (g^t - u^t\mathbf{1}) - (g^* - u^*\mathbf{1}), c^t - b\right\rangle \\ &\leq \left(\|g^t - u^t\mathbf{1}\|_\infty + \|g^* - u^*\mathbf{1}\|_\infty\right)\left\|c^t - b\right\|_1 \leq (c_\infty - \eta\iota)\left\|c^t - b\right\|_1,\end{aligned} \tag{38}$$

where the last inequality is due to Corollary 5. Now we set $\lambda = \lambda^*$ in (33), and we obtain

$$\left\langle \lambda^t - \lambda^*, -\nabla_\lambda F(f^{t+1}, g^t, \lambda^t)\right\rangle \leq 3c_\infty^2\|\lambda^{t+1} - \lambda^t\|_2/\eta + c_\infty\left\|c^t - b\right\|_1. \tag{39}$$

Since $F(f, g, \lambda)$ is a concave function, we have

$$F(f^*, g^*, \lambda^*) \leq F(f^{t+1}, g^t, \lambda^t) + \langle \nabla F(f^{t+1}, g^t, \lambda^t), (f^*, g^*, \lambda^*) - (f^{t+1}, g^t, \lambda^t)\rangle,$$

which, combining with (24) yields

$$\begin{aligned}\tilde{F}(f^{t+1}, g^t, \lambda^t) &= F(f^*, g^*, \lambda^*) - F(f^{t+1}, g^t, \lambda^t) \\ &\leq \langle f^{t+1} - f^*, r(\textstyle\sum_{k=1}^N \pi^k(f^{t+1}, g^t, \lambda^t)) - a\rangle + \left\langle g^t - g^*, c^t - b\right\rangle \\ &\quad + \left\langle \lambda^t - \lambda^*, -\nabla_\lambda F(f^{t+1}, g^t, \lambda^t)\right\rangle \\ &\leq (2c_\infty - \eta\iota)\|c^t - b\|_1 + 3c_\infty^2\|\lambda^{t+1} - \lambda^t\|_2/\eta,\end{aligned}$$

where the last inequality follows from (15), (26), (38) and (39). $\square$

The next lemma shows that the suboptimality gap $\tilde{F}(f, g, \lambda)$ can be bounded by $O(1/t)$.

Lemma 10 *Let (f^t, g^t, λ^t) be generated by PAM (Algorithm 1). The following inequality holds:*

$$\tilde{F}(f^{t+1}, g^{t+1}, \lambda^{t+1}) \leq \frac{4/(\eta\gamma_0)}{t + 1 + 4/(\eta\gamma_0\tilde{F}(f^0, g^0, \lambda^0))},$$

where $\gamma_0 = \min\left\{\frac{1}{(2c_\infty - \eta\iota)^2}, \frac{1}{9c_\infty^2}\right\}$ is a constant.

Proof. Combining (30b) and (30c), we have

$$F(f^{t+1}, g^{t+1}, \lambda^{t+1}) - F(f^{t+1}, g^t, \lambda^t) \geq \frac{\eta}{2}\left\|c^t - b\right\|_1^2 + c_\infty^2\left\|\lambda^{t+1} - \lambda^t\right\|_2^2/(2\eta). \tag{40}$$

Therefore, we have

$$
\begin{aligned}
&\tilde{F}(f^{t+1}, g^t, \lambda^t) - \tilde{F}(f^{t+1}, g^{t+1}, \lambda^{t+1}) \\
\leq & -\frac{\eta}{2}\left\|c^t - b\right\|_1^2 - c_\infty^2\left\|\lambda^{t+1} - \lambda^t\right\|_2^2/(2\eta) \\
\leq & -\frac{\eta}{2}\gamma_0 \cdot \big(((2c_\infty - \eta\iota)\|c^t - b\|_1)^2 + (3c_\infty^2\|\lambda^{t+1} - \lambda^t\|_2/\eta)^2\big) \\
\leq & -\frac{\eta}{4}\gamma_0\big((2c_\infty - \eta\iota)\|c^t - b\|_1 + 3c_\infty^2\|\lambda^{t+1} - \lambda^t\|_2/\eta\big)^2 \\
\leq & -\frac{\eta}{4}\gamma_0\tilde{F}(f^{t+1}, g^t, \lambda^t)^2,
\end{aligned}
\tag{41}
$$

where the last inequality is from Lemma 9. Dividing both sides of (41) by $\tilde{F}(f^{t+1}, g^{t+1}, \lambda^{t+1})\cdot \tilde{F}(f^{t+1}, g^t, \lambda^t)$, we have

$$
\begin{aligned}
\frac{1}{\tilde{F}(f^{t+1}, g^{t+1}, \lambda^{t+1})} &\geq \frac{1}{\tilde{F}(f^{t+1}, g^t, \lambda^t)} + \frac{\eta}{4}\gamma_0 \cdot \frac{\tilde{F}(f^{t+1}, g^t, \lambda^t)}{\tilde{F}(f^{t+1}, g^{t+1}, \lambda^{t+1})} \\
&\geq \frac{1}{\tilde{F}(f^{t+1}, g^t, \lambda^t)} + \frac{\eta}{4}\gamma_0 \geq \frac{1}{\tilde{F}(f^t, g^t, \lambda^t)} + \frac{\eta}{4}\gamma_0,
\end{aligned}
\tag{42}
$$

where the second inequality is due to (41) and the last inequality is from (30a). Summing (42) from 0 to t leads to

$$
\frac{1}{\tilde{F}(f^{t+1}, g^{t+1}, \lambda^{t+1})} \geq \frac{1}{\tilde{F}(f^0, g^0, \lambda^0)} + \frac{\eta(t+1)}{4}\gamma_0,
$$

which implies the desired result. □

The next lemma gives sufficient conditions for the PAM algorithm to return an ϵ-optimal solution to the original EOT problem (2).

Lemma 11 *Assume PAM terminates at the T-iteration, i.e.,*

$$
\|c^{T-1} - b\|_1 \leq \epsilon/(6(6c_\infty - \eta\iota)), \tag{43a}
$$

$$
\left\|\lambda^T - \lambda^{T-1}\right\|_2 \leq \eta\epsilon/(18c_\infty^2), \tag{43b}
$$

$$
\tilde{F}(f^T, g^{T-1}, \lambda^{T-1}) \leq \epsilon/6. \tag{43c}
$$

Then the output $(\hat{\pi}, \hat{\lambda})$ of PAM (Algorithm 1), i.e.,

$$
\hat{\pi}^k = \mathit{Round}(\pi^k(f^T, g^{T-1}, \lambda^{T-1}), a^k, b^k),\ \forall k \in [N], \quad \hat{\lambda} = \lambda^{T-1},
$$

is an ϵ-optimal solution of the original EOT problem (2).

Proof. According to Definition 2, it is sufficient to show that the output $(\hat{\boldsymbol{\pi}}, \hat{\lambda}) \in \Pi_{a,b}^N \times \Delta_N^+$ satisfies the following two inequalities:

$$
\max_{\lambda \in \Delta_N^+} \ell\left(\hat{\boldsymbol{\pi}}, \lambda\right) - \ell(\hat{\boldsymbol{\pi}}, \hat{\lambda}) \leq \frac{\epsilon}{2}, \tag{44a}
$$

$$
\ell(\hat{\boldsymbol{\pi}}, \hat{\lambda}) - \min_{\boldsymbol{\pi} \in \Pi_{a,b}^N} \ell(\boldsymbol{\pi}, \hat{\lambda}) \leq \frac{\epsilon}{2}. \tag{44b}
$$

We prove (44a) first. For ease of presentation, we denote $\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi}(f^T, g^{T-1}, \lambda^{T-1})$, $\boldsymbol{\pi}^* = \boldsymbol{\pi}(f^*, g^*, \lambda^*)$. Note that $\hat{\pi}^k = Round(\tilde{\pi}^k, a^k, b^k), \forall k \in [N]$. We also denote

$$\bar{\lambda}(\boldsymbol{\pi}) := \underset{\lambda \in \Delta_N^+}{\operatorname{argmax}} \left\{ \ell(\boldsymbol{\pi}, \lambda) = \sum_{k=1}^{N} \lambda_k \langle \pi^k, C^k \rangle \right\}. \tag{45}$$

Note that the term on the left hand side of (44a) can be rewritten as

$$\begin{aligned} &\ell\left(\hat{\boldsymbol{\pi}}, \bar{\lambda}(\hat{\boldsymbol{\pi}})\right) - \ell\left(\hat{\boldsymbol{\pi}}, \hat{\lambda}\right) \\ =& \underbrace{(\ell(\hat{\boldsymbol{\pi}}, \bar{\lambda}(\hat{\boldsymbol{\pi}})) - \ell(\tilde{\boldsymbol{\pi}}, \bar{\lambda}(\tilde{\boldsymbol{\pi}})))}_{(I)} + \underbrace{([\ell(\tilde{\boldsymbol{\pi}}, \bar{\lambda}(\tilde{\boldsymbol{\pi}})) - \eta H(\tilde{\boldsymbol{\pi}})] - [\ell(\boldsymbol{\pi}^*, \lambda^*) - \eta H(\boldsymbol{\pi}^*)])}_{(II)} \\ &+ \underbrace{([\ell(\boldsymbol{\pi}^*, \lambda^*) - \eta H(\boldsymbol{\pi}^*)] - [\ell(\tilde{\boldsymbol{\pi}}, \hat{\lambda}) - \eta H(\tilde{\boldsymbol{\pi}})])}_{(III)} + \underbrace{(\ell(\tilde{\boldsymbol{\pi}}, \hat{\lambda}) - \ell(\hat{\boldsymbol{\pi}}, \hat{\lambda}))}_{(IV)}. \end{aligned} \tag{46}$$

We now provide upper bounds for these four terms. Denote

$$\hat{k}^* = \underset{k \in [N]}{\operatorname{argmax}} \langle \hat{\pi}^k, C^k \rangle, \quad \tilde{k}^* = \underset{k \in [N]}{\operatorname{argmax}} \langle \tilde{\pi}^k, C^k \rangle. \tag{47}$$

Since (1) and (2) are equivalent, we have the following for the term (I):

$$\begin{aligned} (I) =& \sum_k [\bar{\lambda}(\hat{\boldsymbol{\pi}})]_k \langle \hat{\pi}^k, C^k \rangle - \sum_k [\bar{\lambda}(\tilde{\boldsymbol{\pi}})]_k \langle \tilde{\pi}^k, C^k \rangle = \langle \hat{\pi}^{\hat{k}^*}, C^{\hat{k}^*} \rangle - \langle \tilde{\pi}^{\tilde{k}^*}, C^{\tilde{k}^*} \rangle \\ \leq& \langle \hat{\pi}^{\hat{k}^*}, C^{\hat{k}^*} \rangle - \langle \tilde{\pi}^{\hat{k}^*}, C^{\hat{k}^*} \rangle \leq \|\hat{\pi}^{\hat{k}^*} - \tilde{\pi}^{\hat{k}^*}\|_1 \|C^k\|_\infty \leq c_\infty \sum_k \|\hat{\pi}^k - \tilde{\pi}^k\|_1 \\ \leq& 2c_\infty \sum_k (\|r(\tilde{\pi}^k) - a^k\|_1 + \|c(\tilde{\pi}^k) - b^k\|_1) = 2c_\infty \|c^{T-1} - b\|_1, \end{aligned} \tag{48}$$

where the first inequality follows from the definition of $\tilde{k}^*$ in (47), the fourth inequality is from Lemma 3, and the last equality follows from (15) and (17).

For the term (II), recall that

$$H(\boldsymbol{\pi}) = -\sum_{k,i,j} \pi_{i,j}^k (\log \pi_{i,j}^k - 1) \text{ and } \tilde{\pi}^k = \exp\left(\frac{f^T \mathbf{1}^\top + \mathbf{1}(g^{T-1})^\top - \lambda_k^{T-1} C^k}{\eta} \right)$$

due to (26), and define $u^{T-1} = \frac{\max_j g_j^{T-1} + \min_j g_j^{T-1}}{2}$. We have

$$
\begin{aligned}
(II) =& \sum_k \bar{\lambda}(\tilde{\boldsymbol{\pi}})_k \langle \tilde{\pi}^k, C^k \rangle + \eta \sum_{k,i,j} \tilde{\pi}_{i,j}^k \left(\frac{f_i^T + g_j^{T-1} - \lambda_k^{T-1} C_{ij}^k}{\eta} - 1 \right) - F^* \\
=& \sum_k (\bar{\lambda}(\tilde{\boldsymbol{\pi}})_k - \hat{\lambda}_k) \langle \tilde{\pi}^k, C^k \rangle + \sum_{k,i,j} \tilde{\pi}_{i,j}^k \left(f_i^T + g_j^{T-1} - \eta \right) - F^* \\
=& \left\langle \bar{\lambda}(\tilde{\boldsymbol{\pi}}) - \hat{\lambda}, \nabla_\lambda F(f^T, g^{T-1}, \lambda^{T-1}) \right\rangle + \left\langle f^T, a \right\rangle + \left\langle g^{T-1}, c^{T-1} \right\rangle - \eta \textstyle\sum_{i,j,k} \tilde{\pi}_{i,j}^k - F^* \\
=& \left\langle \bar{\lambda}(\tilde{\boldsymbol{\pi}}) - \hat{\lambda}, \nabla_\lambda F(f^T, g^{T-1}, \lambda^{T-1}) \right\rangle + \left\langle f^T, a \right\rangle + \left\langle g^{T-1}, c^{T-1} \right\rangle - \log \left(\sum_k \|\tilde{\pi}^k\|_1 \right) - \eta - F^* \\
=& \left\langle \bar{\lambda}(\tilde{\boldsymbol{\pi}}) - \hat{\lambda}, \nabla_\lambda F(f^T, g^{T-1}, \lambda^{T-1}) \right\rangle + \langle g^{T-1}, c^{T-1} - b \rangle + F(f^T, g^{T-1}, \lambda^{T-1}) - F^* \\
\leq& \left\langle \bar{\lambda}(\tilde{\boldsymbol{\pi}}) - \hat{\lambda}, \nabla_\lambda F(f^T, g^{T-1}, \lambda^{T-1}) \right\rangle + \langle g^{T-1}, c^{T-1} - b \rangle \\
\leq& c_\infty \|c^{T-1} - b\|_1 + 3c_\infty^2 \|\lambda^T - \lambda^{T-1}\|_2 / \eta + \langle g^{T-1} - u^{T-1}\mathbf{1}, c^{T-1} - b \rangle \\
\leq& c_\infty \|c^{T-1} - b\|_1 + 3c_\infty^2 \|\lambda^T - \lambda^{T-1}\|_2 / \eta + \|g^{T-1} - u^{T-1}\mathbf{1}\|_\infty \|c^{T-1} - b\|_1 \\
\leq& (3c_\infty/2 - \eta\iota/2) \|c^{T-1} - b\|_1 + 3c_\infty^2 \|\lambda^T - \lambda^{T-1}\|_2 / \eta,
\end{aligned}
\tag{49}
$$

where the third equality uses (26), (24c) and (15), the second inequality follows from Lemma 8 by setting $\lambda = \bar{\lambda}(\tilde{\boldsymbol{\pi}})$ and $t = T - 1$, and the last inequality uses Corollary 5.

For the term (III), we have

$$
\begin{aligned}
(III) \leq& \left| \sum_k \hat{\lambda}_k \langle \tilde{\pi}^k, C^k \rangle + \eta \sum_{i,j} \tilde{\pi}_{i,j}^k \left(\frac{f_i^T + g_j^{T-1} - \lambda_k^{T-1} C^k}{\eta} - 1 \right) - F^* \right| \\
=& |\langle g^{T-1}, c^{T-1} - b \rangle + F(f^T, g^{T-1}, \lambda^{T-1}) - F^*| \\
\leq& (c_\infty/2 - \eta\iota/2) \|c^{T-1} - b\|_1 + |F(f^T, g^{T-1}, \lambda^{T-1}) - F^*|,
\end{aligned}
\tag{50}
$$

where the last inequality follows from Corollary 5.

Finally, for the term (IV), we have

$$
\begin{aligned}
(IV) =& \sum_k \langle \tilde{\pi}^k - \hat{\pi}^k, \hat{\lambda}_k C^k \rangle \leq \sum_k \|\tilde{\pi}^k - \hat{\pi}^k\|_1 \|C^k\|_\infty \\
\leq& 2c_\infty \sum_k (\|r(\tilde{\pi}^k) - a^k\|_1 + \|c(\tilde{\pi}^k) - b^k\|_1) = 2c_\infty \|c^{T-1} - b\|_1,
\end{aligned}
\tag{51}
$$

where the first inequality uses $|\hat{\lambda}_k| \leq 1$, the second inequality uses Lemma 3 and (17). Plugging (48) - (51) into (46), and using (43), we obtain (44a).

Now we prove (44b). For ease of presentation, we denote

$$
\bar{\boldsymbol{\pi}}(\lambda) := \operatorname*{argmin}_{\boldsymbol{\pi} \in \Pi_{a,b}^N} \ell(\boldsymbol{\pi}, \lambda). \tag{52}
$$

We also denote $\tilde{b} = c^{T-1} = \sum_k c\left(\tilde{\pi}^k\right)$ and $\pi'^k = \mathrm{Round}(\bar{\pi}(\hat{\lambda})^k, \tilde{a}^k, \tilde{b}^k)$, where

$$
(\tilde{a}^k, \tilde{b}^k)_{k \in [N]} := \mathrm{Margins}(\bar{\boldsymbol{\pi}}(\hat{\lambda}), a, \tilde{b}),
$$

as defined in (16). From (18) we know that

$$\sum_k \left\| c\left((\bar{\pi}(\hat{\lambda}))^k \right) - \tilde{b}^k \right\|_1 = \left\| \sum_k c\left((\bar{\pi}(\hat{\lambda}))^k \right) - \sum_k \tilde{b}^k \right\|_1 = \|b - \tilde{b}\|_1 = \|b - c^{T-1}\|_1, \tag{53}$$

where the second equality is due to $\bar{\boldsymbol{\pi}}(\hat{\lambda}) \in \Pi^N_{a,b}$ and thus $c(\sum_k (\bar{\boldsymbol{\pi}}(\hat{\lambda}))^k) = b$, and the fact that $\sum_k \tilde{b}^k = \tilde{b}$ due to Property (ii) of the Margins procedure in Section 2.1. By the Sinkhorn's theorem (Sinkhorn, 1967), $\tilde{\boldsymbol{\pi}}$ is the unique optimal solution of $\min_{\boldsymbol{\pi} \in \Pi^N_{a,\tilde{b}}} \ell_\eta(\boldsymbol{\pi}, \hat{\lambda})$. Therefore

$$\sum_k \hat{\lambda}_k \langle \tilde{\pi}^k, C^k \rangle - \eta H(\tilde{\boldsymbol{\pi}}) \le \sum_k \hat{\lambda}_k \langle \pi'^k, C^k \rangle - \eta H(\boldsymbol{\pi}'). \tag{54}$$

Now, note that the left hand side of (44b) can be arranged into three parts:

$$\begin{aligned} &\ell(\hat{\boldsymbol{\pi}}, \hat{\lambda}) - \ell(\bar{\boldsymbol{\pi}}(\hat{\lambda}), \hat{\lambda}) \\ =& \underbrace{\left(\sum_k \hat{\lambda}_k \langle \hat{\pi}^k, C^k \rangle - \sum_k \hat{\lambda}_k \langle \tilde{\pi}^k, C^k \rangle \right)}_{(V)} + \underbrace{\left(\sum_k \hat{\lambda}_k \langle \tilde{\pi}^k, C^k \rangle - \sum_k \hat{\lambda}_k \langle \pi'^k, C^k \rangle \right)}_{(VI)} \\ &+ \underbrace{\left(\sum_k \hat{\lambda}_k \langle \pi'^k, C^k \rangle - \sum_k \hat{\lambda}_k \langle (\bar{\pi}(\hat{\lambda}))^k, C^k \rangle \right)}_{(VII)}. \end{aligned} \tag{55}$$

We now upper bound these three terms. First note that the term (V) is the same as the term (IV) and thus has the same upper bound in (51). Since $0 \le H(\boldsymbol{\pi}) \le \log(n^2 N) + 1$, from (54) we have that

$$(VI) = \sum_k \hat{\lambda}_k \langle \tilde{\pi}^k, C^k \rangle - \sum_k \hat{\lambda}_k \langle \pi'^k, C^k \rangle \le \eta \left| H(\tilde{\boldsymbol{\pi}}) - H(\boldsymbol{\pi}') \right| \le \frac{1}{3}\epsilon, \tag{56}$$

where the last step uses the definition of η in (20).

For the term (VII), we have

$$\begin{aligned} (VII) =& \sum_k \hat{\lambda}_k \langle \pi'^k, C^k \rangle - \sum_k \hat{\lambda}_k \langle (\bar{\pi}(\hat{\lambda}))^k, C^k \rangle \le \sum_k \|\pi'^k - (\bar{\pi}(\hat{\lambda}))^k\|_1 \|C^k\|_\infty \\ \le& 2c_\infty \sum_k (\|r((\bar{\pi}(\hat{\lambda}))^k) - \tilde{a}^k\|_1 + \|c((\bar{\pi}(\hat{\lambda}))^k) - \tilde{b}^k\|_1) = 2c_\infty \|c^{T-1} - b\|_1, \end{aligned} \tag{57}$$

where the second inequality follows from Lemma 3, the second equality uses (53) and the fact that $r((\bar{\pi}(\hat{\lambda}))^k) = \tilde{a}^k$ due to the property of the Margins procedure in (16).

Finally, plugging (51) (note (V)=(IV)), (56) and (57) into (55), and using (43a) and noting $\iota < 0$, we obtain (44b). This completes the proof.

$\square$

3.3 Main Result

We now present our main theorem, which gives the iteration complexity of PAM such that (43) is satisfied, and as a result of Lemma 11, an ϵ-optimal solution to the original EOT problem (2) is obtained.

Theorem 12 *Define $\epsilon' = \epsilon/(6c_\infty - \eta\iota)$, and set T as*

$$T = 5 + \frac{36}{\eta\sqrt{\gamma_0}\epsilon'} + \frac{648c_\infty^2}{\eta\epsilon} + \frac{28}{\eta\gamma_0\epsilon} = O\left(c_\infty^2\epsilon^{-2}\right), \tag{58}$$

where $\gamma_0 = \min\left\{\frac{1}{(2c_\infty-\eta\iota)^2}, \frac{1}{9c_\infty^2}\right\}$ is a constant and we know $\gamma_0 = O(c_\infty^{-2})$. The output pair of Algorithm 1 is an ϵ-optimal solution of the EOT problem (2).

Proof. According to Lemma 11, we only need to show that (43) holds after T iterations as defined in (58). To guarantee (43a) and (43b), we follow the ideas of Dvurechensky *et al.* (Dvurechensky et al., 2018) and construct a switching process. We first reduce $\tilde{F}$ from $\tilde{F}(f^0, g^0, \lambda^0)$ to a constant s by running t_1 steps. In this process, Lemma 10 indicates

$$t_1 \leq 1 + \frac{4}{\eta\gamma_0 s} - \frac{4}{\eta\gamma_0\tilde{F}(f^0, g^0, \lambda^0)}. \tag{59}$$

Secondly, starting from s, we continue running the algorithm, and assume that there are t_2 iterations in which (43a) fails. By (30b) we have

$$t_2 \leq 1 + \frac{72s}{\eta\epsilon'^2}.$$

Therefore, we know that the total iteration number that (43a) fails is upper bounded by

$$T_1 = t_1 + t_2 \leq 2 + \frac{72s}{\eta\epsilon'^2} + \frac{4}{\eta\gamma_0 s} - \frac{4}{\eta\gamma_0\tilde{F}(f^0, g^0, \lambda^0)}$$

iterations. By choosing $s = \epsilon'/(6\sqrt{\gamma_0})$, we know that

$$T_1 \leq \begin{cases} 2 + \frac{12}{\eta\sqrt{\gamma_0}\epsilon'} + \frac{24}{\eta\sqrt{\gamma_0}\epsilon'} - \frac{4}{\eta\gamma_0\tilde{F}(f^0,g^0,\lambda^0)} \leq 2 + \frac{36}{\eta\sqrt{\gamma_0}\epsilon'} & \text{if } \tilde{F}(f^0, g^0, \lambda^0) \geq \frac{\epsilon'}{6\sqrt{\gamma_0}} \\ 2 + \frac{12}{\eta\sqrt{\gamma_0}\epsilon'} + \frac{24}{\eta\sqrt{\gamma_0}\epsilon'} - \frac{4}{\eta\gamma_0\tilde{F}(f^0,g^0,\lambda^0)} \leq 2 + \frac{12}{\eta\sqrt{\gamma_0}\epsilon'} & \text{otherwise.} \end{cases}$$

Therefore, we have $T_1 \leq 2 + \frac{36}{\eta\sqrt{\gamma_0}\epsilon'}$. Similarly, starting from s, the number of iterations that (43b) fails can be bounded by

$$t_3 \leq 1 + \frac{648sc_\infty^2}{\eta\epsilon^2},$$

where we apply (30b). By choosing $s = \epsilon$, we know that the total iteration number that (43b) fails is upper bounded by

$$T_2 = t_1 + t_3 \leq 2 + \frac{648c_\infty^2}{\eta\epsilon} + \frac{4}{\eta\gamma_0\epsilon} - \frac{4}{\eta\gamma_0\tilde{F}(f^0, g^0, \lambda^0)}$$

iterations. Finally, by letting $s = \epsilon/6$ in (59), we know that

$$\tilde{F}(f^{T_3-1}, g^{T_3-1}, \lambda^{T_3-1}) \leq \epsilon/6$$

after

$$T_3 = 1 + \frac{24}{\eta\gamma_0\epsilon}$$

iterations. From (30a) we know that after T_3 iterations, we have

$$\tilde{F}(f^{T_3}, g^{T_3-1}, \lambda^{T_3-1}) \leq \tilde{F}(f^{T_3-1}, g^{T_3-1}, \lambda^{T_3-1}) \leq \epsilon/6,$$

i.e., (43c) holds. Combining the above discussions, we know that after $T = T_1 + T_2 + T_3 + 1$ iterations, there must exist at least one iteration such that (43) holds, and thus the output of PAM is an ϵ-optimal solution to the original EOT problem (2).

□

Remark 13 *Though our complexity result matches the rate of the Sinkhorn's algorithm in terms of the dependence on ϵ, we argue that EOT is a more difficult problem than the entropic regularized OT, and thus our results are promising. First, EOT is a saddle-point problem while entropic regularized OT is a minimization problem. Second, the extra variable λ in EOT requires a gradient projection step in the PAM algorithm, which introduces significant difficulty to the analysis of the convergence behavior. While for Sinkhorn's algorithm it is much easier to analyze, because the dual is unconstrained. Third, since there are multiple agents in EOT, it is more difficult to design the rounding procedure to obtain the primal solution. We also note that the dependence of c_∞ in our result and in the result of Sinkhorn's algorithm (Dvurechensky et al., 2018) are both c_∞^2.*

4. Projected Alternating Maximization with Extrapolation

In this section, we discuss how to accelerate the PAM algorithm (Algorithm 1). It can be shown that the gradient of F in (11) is Lipschitz continuous[1]. Therefore, Scetbon *et al.* (Scetbon et al., 2021) proposed to adopt Nesterov's accelerated gradient method (Nesterov, 2004) to solve (11). Their algorithm, named APGA (Accelerated Projected Gradient Ascent algorithm), iterates as follows:

$$(v, w, z)^\top \leftarrow (f^{t-1}, g^{t-1}, \lambda^{t-1})^\top + \left(\frac{t-2}{t+1}\right) \cdot \left((f^{t-1}, g^{t-1}, \lambda^{t-1})^\top - (f^{t-2}, g^{t-2}, \lambda^{t-2})^\top\right)$$

$$(f^t, g^t)^\top \leftarrow (v, w)^\top + \frac{1}{L}\nabla_{(f,g)}F(v, w, z) \tag{60a}$$

$$(\lambda^t)^\top \leftarrow \mathrm{Proj}_{\Delta_N^+}\left(z + \frac{1}{L}\nabla_\lambda F(v, w, z)\right), \tag{60b}$$

where L is the Lipschitz constant of ∇F. Note that APGA treats the problem (11) as a generic convex and smooth problem, and does not take advantage of the special structures of (11). In particular, f and g are updated using gradient ascent steps. This is in contrast

1. In Lemma 4 we proved that $\nabla_\lambda F$ is Lipschitz continuous. The Lipschitz continuity of $\nabla_f F$ and $\nabla_g F$ can be proved similarly.

Algorithm 3 Projected Alternating Maximization with Extrapolation Algorithm

---

Input: Cost matrices $\{C^k\}_{1\le k\le N}$, accuracy ϵ, $\theta \in (0,1)$.
Initialization: $f^0 = g^0 = [1,...,1]^\top,\ \lambda^0 = [1/N,...,1/N]^\top \in \Delta_N^+$.
Choose parameters as

$$\eta = \min\left\{\frac{\epsilon}{3(\log(n^2N)+1)}, c_\infty\right\}, \quad \tau = \frac{\eta}{2c_\infty^2}. \tag{61}$$

while (63) is not met **do**
  Compute f^{t+1} by (62a)
  Compute g^{t+1} by (62b)
  Compute y^{t+1} by (62c)
  Compute λ^{t+1} by (62d)
  $t \leftarrow t+1$
end while
Assume the stop condition (63) is satisfied at the T-th iteration. Compute $(a^k, b^k)_{k\in[N]} = \text{Margins}(\boldsymbol{\pi}(f^T, g^{T-1}, \lambda^{T-1}), a, b)$ as in Section 2.1.
Output: $(\hat{\pi}, \hat{\lambda})$ where $\hat{\pi}^k = \text{Round}(\pi^k(f^T, g^{T-1}, \lambda^{T-1}), a^k, b^k),\ \forall k \in [N],\ \hat{\lambda} = \lambda^{T-1}$.

---

to PAM in which f and g are obtained by exact maximizations, which is expected to improve the function value of F more significantly. In the following, we will design an accelerated algorithm that utilizes this property. Our method is called PAME (PAM with Extrapolation) and it incorporates the extrapolation technique to the gradient step for updating λ, and f and g are still updated using exact maximizations. We note that currently we are not able to prove a better complexity for PAME. Our iteration complexity result in Theorem 19 is in the same order as that of PAM, but numerically we have observed great improvement of PAME over PAM. It is an interesting future topic to study other accelerations to PAM that can provably achieve improved complexity.

A typical iteration of our PAME algorithm is given below:

$$f^{t+1} = f^t + \eta \log\left(\frac{a}{r\left(\sum_k \zeta^k(f^t, g^t, \lambda^t)\right)}\right), \tag{62a}$$

$$g^{t+1} = g^t + \eta \log\left(\frac{b}{c\left(\sum_k \zeta^k(f^{t+1}, g^t, \lambda^t)\right)}\right), \tag{62b}$$

$$y^{t+1} = \text{Proj}_{\Delta_N^+}\left(\lambda^t + (1-\theta)(\lambda^t - \lambda^{t-1})\right), \tag{62c}$$

$$\lambda^{t+1} = \text{Proj}_{\Delta_N^+}\left(y^{t+1} + \tau\nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1})\right). \tag{62d}$$

Here $\theta \in (0,1)$ is a given parameter for the extrapolation step. We see that steps (62a)-(62b) are the same as (13)-(14) and they are solutions to the exact maximizations (12a)-(12b). Steps (62c)-(62c) give extrapolation to the gradient step for λ, similar to Nesterov's accelerated gradient method. Note that PAME (62) solves the dual entropy-regularized EOT problem (11). We use the same rounding procedure in Section 2.1 to generate a primal solution to the original EOT problem (1). The complete PAME algorithm is described in

Algorithm 3. Note that the algorithm is terminated when the following criteria are met:

$$\|c^{t-1} - b\|_1 \le \epsilon/(6(6c_\infty - \eta\iota)), \tag{63a}$$
$$\|\lambda^{t-1} - \lambda^{t-2}\|_2 \le \eta\epsilon/(60(1-\theta)c_\infty^2), \tag{63b}$$
$$\|\lambda^t - y^t\|_2 \le \eta\epsilon/(42c_\infty^2), \tag{63c}$$
$$\tilde{F}(f^t, g^{t-1}, \lambda^{t-1}) \le \epsilon/6. \tag{63d}$$

5. Convergence Analysis of PAME Algorithm

In this section, we analyze the iteration complexity of PAME (Algorithm 3) for obtaining an ϵ-optimal solution to the original EOT problem (1). The proof for PAME is different from that of PAM, and here we need to analyze the behavior of the following Hamiltonian, inspired by (Jin et al., 2018).

$$E(f, g, \lambda^1, \lambda^2) = F(f, g, \lambda^1) - \frac{1}{2\tau}\|\lambda^1 - \lambda^2\|_2^2. \tag{64}$$

5.1 Technical Preparation

The following simple fact is useful for our analysis later.

$$\|y^{t+1} - \lambda^t\|_2 = \|\mathrm{Proj}_{\Delta_N^+}\left(\lambda^t + (1-\theta)(\lambda^t - \lambda^{t-1})\right) - \mathrm{Proj}_{\Delta_N^+}\left(\lambda^t\right)\|_2 \le (1-\theta)\|\lambda^t - \lambda^{t-1}\|_2, \tag{65}$$

where the equality follows from the definition of y^{t+1} in (62c), and the inequality is due to the non-expansiveness of the projection operator.

The following lemma shows that the Hamiltonian $E(f^t, g^t, \lambda^t, \lambda^{t-1})$ is monotonically increasing when updating λ in Algorithm (3).

Lemma 14 *[Sufficient increase in λ] Let $\{f^t, g^t, y^t, \lambda^t\}$ be generated by PAME (Algorithm 3). The following inequality holds:*

$$E(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t) - E(f^{t+1}, g^{t+1}, \lambda^t, \lambda^{t-1}) \ge \frac{2\theta - \theta^2}{2\tau}\|\lambda^t - \lambda^{t-1}\|_2^2 + \frac{1}{4\tau}\|\lambda^{t+1} - y^{t+1}\|_2^2. \tag{66}$$

Note that since $\theta \in (0, 1)$, the right hand side of (66) is always nonnegative.

Proof. From the optimality condition of (62d) we know that, there exists $h(\lambda^{t+1}) \in \partial\mathbb{I}_{\Delta_N^+}(\lambda^{t+1})$ such that

$$\nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) - \frac{1}{\tau}(\lambda^{t+1} - y^{t+1}) - h(\lambda^{t+1}) = 0. \tag{67}$$

By the convexity of the indicator function $\mathbb{I}_{\Delta_N^+}(\lambda^{t+1})$, we have

$$\langle y^{t+1} - \lambda^{t+1}, h(\lambda^{t+1})\rangle \le 0, \quad \langle \lambda^t - \lambda^{t+1}, h(\lambda^{t+1})\rangle \le 0. \tag{68}$$

Moreover, we have the following inequality:

$$
\begin{aligned}
&\|\lambda^{t+1} - \lambda^t\|_2^2 = \|\lambda^{t+1} - y^{t+1} + y^{t+1} - \lambda^t\|_2^2 \\
=& \|y^{t+1} - \lambda^t\|_2^2 + 2\langle \lambda^{t+1} - y^{t+1}, y^{t+1} - \lambda^t\rangle + \|\lambda^{t+1} - y^{t+1}\|_2^2 \\
\leq& (1-\theta)^2\|\lambda^t - \lambda^{t-1}\|_2^2 + 2\langle \lambda^{t+1} - y^{t+1}, y^{t+1} - \lambda^t\rangle + \|\lambda^{t+1} - y^{t+1}\|_2^2,
\end{aligned} \tag{69}
$$

where the inequality is from (65).

We then have the following inequality:

$$
\begin{aligned}
&F(f^{t+1}, g^{t+1}, \lambda^t) - F(f^{t+1}, g^{t+1}, \lambda^{t+1}) \\
\leq& \big(F(f^{t+1}, g^{t+1}, y^{t+1}) + \langle \nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}), \lambda^t - y^{t+1}\rangle\big) - \\
& \big(F(f^{t+1}, g^{t+1}, y^{t+1}) + \langle \nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}), \lambda^{t+1} - y^{t+1}\rangle - c_\infty^2\|\lambda^{t+1} - y^{t+1}\|_2^2/(2\eta)\big) \\
=& \langle \nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}), \lambda^t - \lambda^{t+1}\rangle + c_\infty^2\|\lambda^{t+1} - y^{t+1}\|_2^2/(2\eta) \\
\leq& \langle \nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) - h(\lambda^{t+1}), \lambda^t - \lambda^{t+1}\rangle + c_\infty^2\|\lambda^{t+1} - y^{t+1}\|_2^2/(2\eta) \\
=& \frac{1}{\tau}\langle \lambda^{t+1} - y^{t+1}, \lambda^t - \lambda^{t+1}\rangle + \frac{1}{4\tau}\|\lambda^{t+1} - y^{t+1}\|^2 \\
=& \frac{1}{\tau}\langle \lambda^{t+1} - y^{t+1}, \lambda^t - y^{t+1} + y^{t+1} - \lambda^{t+1}\rangle + \frac{1}{4\tau}\|\lambda^{t+1} - y^{t+1}\|^2 \\
=& -\frac{1}{\tau}\langle \lambda^{t+1} - y^{t+1}, y^{t+1} - \lambda^t\rangle - \frac{3}{4\tau}\|\lambda^{t+1} - y^{t+1}\|^2,
\end{aligned} \tag{70}
$$

where the first inequality is from the concavity of F with respect to λ and (28), the second inequality is due to (68), the second equality is due to (67). Combining (69) and (70) leads to

$$
\begin{aligned}
&E(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t) = F(f^{t+1}, g^{t+1}, \lambda^{t+1}) - \frac{1}{2\tau}\|\lambda^{t+1} - \lambda^t\|_2^2 \\
\geq& F(f^{t+1}, g^{t+1}, \lambda^t) + \frac{1}{\tau}\langle \lambda^{t+1} - y^{t+1}, y^{t+1} - \lambda^t\rangle + \frac{3}{4\tau}\|\lambda^{t+1} - y^{t+1}\|^2 \\
& - \frac{(1-\theta)^2}{2\tau}\|\lambda^t - \lambda^{t-1}\|_2^2 - \frac{1}{\tau}\langle \lambda^{t+1} - y^{t+1}, y^{t+1} - \lambda^t\rangle - \frac{1}{2\tau}\|\lambda^{t+1} - y^{t+1}\|_2^2 \\
=& F(f^{t+1}, g^{t+1}, \lambda^t) - \frac{1}{2\tau}\|\lambda^t - \lambda^{t-1}\|_2^2 + \frac{2\theta - \theta^2}{2\tau}\|\lambda^t - \lambda^{t-1}\|_2^2 + \frac{1}{4\tau}\|\lambda^{t+1} - y^{t+1}\|^2 \\
=& E(f^{t+1}, g^{t+1}, \lambda^t, \lambda^{t-1}) + \frac{2\theta - \theta^2}{2\tau}\|\lambda^t - \lambda^{t-1}\|_2^2 + \frac{1}{4\tau}\|\lambda^{t+1} - y^{t+1}\|^2,
\end{aligned}
$$

which completes the proof. □

Now we define the following function $\tilde{E}$, and later we will prove that $\tilde{E}(f^t, g^t, \lambda^t, \lambda^{t-1})$ can be upper bounded by $O(1/t)$.

$$
\tilde{E}(f, g, \lambda^1, \lambda^2) = F(f^*, g^*, \lambda^*) - E(f, g, \lambda^1, \lambda^2).
$$

The next lemma is useful for obtaining the upper bound for $\tilde{E}(f^t, g^t, \lambda^t, \lambda^{t-1})$. Moreover, it is noted that $\tilde{E}(f, g, \lambda^1, \lambda^2) \geq 0, \forall f, g, \lambda^1, \lambda^2$, and $\tilde{E}(f, g, \lambda, \lambda) = \tilde{F}(f, g, \lambda), \forall f, g, \lambda$.

Lemma 15 *Let $\{f^t, g^t, y^t, \lambda^t\}$ be generated by PAME (Algorithm 3). For any $\lambda \in \Delta_N^+$, the following inequality holds*

$$\left\langle \lambda - \lambda^t, \nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \right\rangle \leq c_\infty \|c^t - b\|_1 + 7c_\infty^2 \|\lambda^{t+1} - y^{t+1}\|_2/\eta + 5(1-\theta)c_\infty^2 \|\lambda^t - \lambda^{t-1}\|_2/\eta. \tag{71}$$

Proof. From the optimality condition of (62d), we have the following inequality:

$$\left\langle \lambda - \lambda^{t+1}, \frac{1}{\tau}(\lambda^{t+1} - y^{t+1}) - \nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) \right\rangle \geq 0, \quad \forall \lambda \in \Delta_N^+. \tag{72}$$

The left hand side of (71) can be rearranged to three terms.

$$\begin{aligned}
&\langle \lambda - \lambda^t, \nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \rangle \\
=&\underbrace{\langle \lambda^t - \lambda, -\nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) \rangle}_{(I)} + \underbrace{\langle \lambda^t - \lambda, \nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) - \nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) \rangle}_{(II)} \\
&+ \underbrace{\langle \lambda^t - \lambda, \nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) - \nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \rangle}_{(III)}.
\end{aligned} \tag{73}$$

We now bound these three terms one by one. To bound the term (I), we first note that from (24c) and (15), we have

$$\|\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t)\|_2 \leq c_\infty \leq c_\infty^2/\eta, \tag{74}$$

where the second inequality is due to the definition of η (61). Now we can bound the term (I) as follows:

$$\begin{aligned}
(I) =& \left\langle \lambda^t - \lambda^{t+1}, -\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) \right\rangle + \left\langle \lambda^t - \lambda^{t+1}, \nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t) - \nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) \right\rangle \\
&+ \left\langle \lambda^{t+1} - \lambda, -\nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) \right\rangle \\
\leq& \left\|\lambda^t - \lambda^{t+1}\right\|_2 \cdot \left\|\nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t)\right\|_2 + \frac{c_\infty^2}{\eta} \left\|\lambda^t - \lambda^{t+1}\right\|_2 \cdot \left\|\lambda^t - y^{t+1}\right\|_2 \\
&+ \frac{1}{\tau} \left\|\lambda^{t+1} - \lambda\right\|_2 \cdot \left\|\lambda^{t+1} - y^{t+1}\right\|_2 \\
\leq& 3c_\infty^2 \|\lambda^t - \lambda^{t+1}\|_2/\eta + 4c_\infty^2 \|\lambda^{t+1} - y^{t+1}\|_2/\eta,
\end{aligned} \tag{75}$$

where the first inequality uses Lemma 4 and (72), the second inequality uses (74) and the facts that $\left\|\lambda^t - y^{t+1}\right\|_2 \leq 2$ and $\left\|\lambda^t - \lambda\right\|_2 \leq 2$.

For the term (II), Lemma 4 yields:

$$(II) \leq 2 \left\|\nabla_\lambda F(f^{t+1}, g^{t+1}, y^{t+1}) - \nabla_\lambda F(f^{t+1}, g^{t+1}, \lambda^t)\right\|_2 \leq 2c_\infty^2 \left\|y^{t+1} - \lambda^t\right\|_2/\eta. \tag{76}$$

For the term (III), it can be bounded as:

$$\begin{aligned}
(III) =& \sum_{k=1}^N (\lambda_k^t - \lambda_k) \cdot \langle \pi^k(f^{t+1}, g^{t+1}, \lambda^t) - \pi^k(f^{t+1}, g^t, \lambda^t), C^k \rangle \\
\leq& \sum_{k=1}^N \|\pi^k(f^{t+1}, g^{t+1}, \lambda^t) - \pi^k(f^{t+1}, g^t, \lambda^t)\|_1 \|C^k\|_\infty \leq c_\infty \|c^t - b\|_1,
\end{aligned} \tag{77}$$

where the last inequality is due to Lemma (6). Plugging (75) - (77) into (73) and applying the triangle inequality, we obtain

$$\left\langle \lambda - \lambda^t, \nabla_\lambda F(f^{t+1}, g^t, \lambda^t) \right\rangle \leq c_\infty \|c^t - b\|_1 + 7c_\infty^2 \|\lambda^{t+1} - y^{t+1}\|_2/\eta + 5c_\infty^2 \|y^{t+1} - \lambda^t\|_2/\eta,$$

which immediately implies (71) by noting (65).

□

Lemma 16 *Let $(f^t, g^t, y^t, \lambda^t)$ be generated by PAME (Algorithm 3). The following inequality holds:*

$$\tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1}) \leq (2c_\infty - \eta\iota)\|c^t - b\|_1 + 7c_\infty^2 \|\lambda^{t+1} - y^{t+1}\|_2/\eta + (7-5\theta)c_\infty^2 \|\lambda^t - \lambda^{t-1}\|_2/\eta.$$

Proof. Since $F(f, g, \lambda)$ is a concave function, we have

$$F(f^*, g^*, \lambda^*) \leq F(f^{t+1}, g^t, \lambda^t) + \left\langle \nabla F(f^{t+1}, g^t, \lambda^t), (f^*, g^*, \lambda^*) - (f^{t+1}, g^t, \lambda^t) \right\rangle,$$

which implies that

$$\begin{aligned}
\tilde{F}(f^{t+1}, g^t, \lambda^t) \leq & \langle g^t - g^*, c^t - b\rangle + \langle \lambda^t - \lambda^*, -\nabla_\lambda F(f^{t+1}, g^t, \lambda^t)\rangle \\
\leq & (c_\infty - \eta\iota)\|c^t - b\|_1 + c_\infty \|c^t - b\|_1 + 7c_\infty^2 \|\lambda^{t+1} - y^{t+1}\|_2/\eta \\
& + 5(1-\theta)c_\infty^2 \|\lambda^t - \lambda^{t-1}\|_2/\eta.
\end{aligned} \tag{78}$$

where in the first inequality we have used (26), and the second inequality follows from (38) and setting $\lambda = \lambda^*$ in (71). From (78) we immediately get

$$\begin{aligned}
&\tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1}) = \tilde{F}(f^{t+1}, g^t, \lambda^t) + \frac{1}{2\tau}\|\lambda^t - \lambda^{t-1}\|_2^2 \\
&\leq c_\infty^2 \|\lambda^t - \lambda^{t-1}\|_2^2/\eta + (2c_\infty - \eta\iota)\|c^t - b\|_1 + 7c_\infty^2 \|\lambda^{t+1} - y^{t+1}\|_2/\eta + 5(1-\theta)c_\infty^2 \|\lambda^t - \lambda^{t-1}\|_2/\eta \\
&\leq 2c_\infty^2 \|\lambda^t - \lambda^{t-1}\|_2/\eta + (2c_\infty - \eta\iota)\|c^t - b\|_1 + 7c_\infty^2 \|\lambda^{t+1} - y^{t+1}\|_2/\eta + 5(1-\theta)c_\infty^2 \|\lambda^t - \lambda^{t-1}\|_2/\eta,
\end{aligned}$$

where the second inequality is due to $\|\lambda^t - \lambda^{t-1}\|_2 \leq 2$. This completes the proof. □

The following lemma bounds $\tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t)$ by $O(1/t)$.

Lemma 17 *Let $\{f^t, g^t, y^t \lambda^t\}$ be generated by PAME (Algorithm 3). The following inequality holds:*

$$\tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t) \leq \frac{6/(\eta\gamma_1)}{t + 1 + 6/(\eta\gamma_1 \tilde{F}(f^0, g^0, \lambda^0))}, \forall t \geq 0,$$

where we assume $\lambda^{-1} = \lambda^0$, and

$$\gamma_1 = \min\left\{\frac{1}{(2c_\infty - \eta\iota)^2}, \frac{2(2\theta - \theta^2)}{(7-5\theta)^2 c_\infty^2}, \frac{1}{49c_\infty^2}\right\} \tag{79}$$

is a constant.

Proof. Combining (30b) and Lemma 14, we have

$$
\begin{aligned}
& E(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t) - E(f^{t+1}, g^t, \lambda^t, \lambda^{t-1}) \\
= & \left(E(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t) - E(f^{t+1}, g^{t+1}, \lambda^t, \lambda^{t-1})\right) + \left(F(f^{t+1}, g^{t+1}, \lambda^t) - F(f^{t+1}, g^t, \lambda^t)\right) \\
\geq & \frac{\eta}{2}\|c^t - b\|_1^2 + \frac{2\theta - \theta^2}{2\tau}\left\|\lambda^t - \lambda^{t-1}\right\|_2^2 + \frac{1}{4\tau}\left\|\lambda^{t+1} - y^{t+1}\right\|_2^2,
\end{aligned}
$$

which implies that

$$
\begin{aligned}
& \tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t) - \tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1}) \\
\leq & -\frac{\eta}{2}\|c^t - b\|_1^2 - (2\theta - \theta^2)\frac{c_\infty^2}{\eta}\|\lambda^t - \lambda^{t-1}\|_2^2 - \frac{c_\infty^2}{2\eta}\|\lambda^{t+1} - y^{t+1}\|_2^2 \\
\leq & -\frac{\eta}{2}\gamma_1\left[\left((2c_\infty - \eta\iota)\left\|c^t - b\right\|_1\right)^2 + \left((7-5\theta)c_\infty^2\|\lambda^t - \lambda^{t-1}\|_2/\eta\right)^2 + \left(7c_\infty^2\|\lambda^{t+1} - y^{t+1}\|_2/\eta\right)^2\right] \\
\leq & -\frac{\eta}{6}\gamma_1\left[(2c_\infty - \eta\iota)\left\|c^t - b\right\|_1 + (7-5\theta)c_\infty^2\|\lambda^t - \lambda^{t-1}\|_2/\eta + 7c_\infty^2\|\lambda^{t+1} - y^{t+1}\|_2/\eta\right]^2 \\
\leq & -\frac{\eta}{6}\gamma_1\tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1})^2,
\end{aligned}
\tag{80}
$$

where the last inequality applies Lemma 16. We then divide both sides of (80) by

$$
\tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t) \cdot \tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1}),
$$

and we obtain

$$
\begin{aligned}
\frac{1}{\tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t)} \geq & \frac{1}{\tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1})} + \frac{\eta}{6}\gamma_1 \cdot \frac{\tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1})}{\tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t)} \\
\geq & \frac{1}{\tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1})} + \frac{\eta}{6}\gamma_1 \geq \frac{1}{\tilde{E}(f^t, g^t, \lambda^t, \lambda^{t-1})} + \frac{\eta}{6}\gamma_1,
\end{aligned}
\tag{81}
$$

where the second inequality holds because (80) implies that

$$
\tilde{E}(f^{t+1}, g^t, \lambda^t, \lambda^{t-1}) \geq \tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t),
$$

and the last inequality follows from (30a). Summing (81) from 0 to t leads to

$$
\frac{1}{\tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t)} \geq \frac{1}{\tilde{E}(f^0, g^0, \lambda^0, \lambda^{-1})} + \frac{\eta(t+1)}{6}\gamma_1 = \frac{1}{\tilde{F}(f^0, g^0, \lambda^0)} + \frac{\eta(t+1)}{6}\gamma_1,
$$

which immediately leads to the desired result. □

Similar to Lemma 11, the following lemma provides some sufficient conditions for the PAME algorithm to return an ϵ-optimal solution to the original EOT problem (2).

Lemma 18 *Assume PAME terminates at the T-iteration, i.e.,*

$$
\|c^{T-1} - b\|_1 \leq \epsilon/(6(6c_\infty - \eta\iota)), \tag{82a}
$$

$$
\|\lambda^{T-1} - \lambda^{T-2}\|_2 \leq \eta\epsilon/(60(1-\theta)c_\infty^2), \tag{82b}
$$

$$
\|\lambda^T - y^T\|_2 \leq \eta\epsilon/(42c_\infty^2), \tag{82c}
$$

$$
\tilde{F}(f^T, g^{T-1}, \lambda^{T-1}) \leq \epsilon/6. \tag{82d}
$$

Then the output $(\hat{\pi}, \hat{\lambda})$ *of PAME (Algorithm 3), i.e.,*

$$\hat{\pi}^k = Round(\pi^k(f^T, g^{T-1}, \lambda^{T-1}), a^k, b^k), \ \forall k \in [N], \quad \hat{\lambda} = \lambda^{T-1},$$

is an ϵ*-optimal solution of the original EOT problem* (2).

Proof. The proof is essentially the same as that of Lemma 11. More specifically, we again need to show that the output of PAME $(\hat{\boldsymbol{\pi}}, \hat{\lambda})$ satisfies (44). The proof of (44b) is exactly the same as the proof of Lemma 11. The proof of (44a) only requires to develop a new bound for

$$\left\langle \bar{\lambda}(\tilde{\boldsymbol{\pi}}) - \hat{\lambda}, \nabla_\lambda F(f^T, g^{T-1}, \lambda^{T-1}) \right\rangle \tag{83}$$

that is used in (49). Other parts are again exactly the same as the ones in Lemma 11. The new bound of (83) can be obtained by applying Lemma 15 with $\lambda = \bar{\lambda}(\tilde{\boldsymbol{\pi}})$ and $t = T-1$, which yields

$$\begin{aligned} &\langle \bar{\lambda}(\tilde{\boldsymbol{\pi}}) - \hat{\lambda}, \nabla_\lambda F(f^T, g^{T-1}, \lambda^{T-1}) \rangle \\ \le\ & c_\infty \|c^{T-1} - b\|_1 + 5(1-\theta)c_\infty^2 \|\lambda^{T-1} - \lambda^{T-2}\|_2/\eta + 7c_\infty^2 \|\lambda^T - y^T\|_2/\eta. \end{aligned} \tag{84}$$

By combining (84) with (48)-(51), we can bound the left hand side of (44a) by

$$\begin{aligned} &\ell\left(\hat{\boldsymbol{\pi}}, \bar{\lambda}(\hat{\boldsymbol{\pi}})\right) - \ell(\hat{\boldsymbol{\pi}}, \hat{\lambda}) \\ \le\ & (6c_\infty - \eta\iota) \left\|c^{T-1} - b\right\|_1 + 5(1-\theta)c_\infty^2 \left\|\lambda^{T-1} - \lambda^{T-2}\right\|_2/\eta + 7c_\infty^2 \|\lambda^T - y^T\|_2/\eta \\ &+ \left|F(f^T, g^{T-1}, \lambda^{T-1}) - F^*\right| \\ \le\ & \left(\frac{1}{6} + \frac{1}{12} + \frac{1}{12} + \frac{1}{6}\right)\epsilon = \frac{1}{2}\epsilon, \end{aligned} \tag{85}$$

where in the last inequality we have used all the sufficient conditions (82a)-(82d). ☐

Theorem 19 *Define* $\epsilon' = \epsilon/(6c_\infty - \eta\iota)$*, and set* T *to be*

$$\begin{aligned} T =& 8 + \frac{48}{\eta\sqrt{\gamma_1}\epsilon'} + \frac{\left(3600(1-\theta)^2 + 882\right)c_\infty^2 s}{\eta\epsilon^2} + \frac{48}{\eta\gamma_1\epsilon} \\ =& O\left(c_\infty^2 \epsilon^{-2}\right), \end{aligned} \tag{86}$$

where $\gamma_1 = \min\left\{\frac{1}{(2c_\infty - \eta\iota)^2}, \frac{2(2\theta-\theta^2)}{(7-5\theta)^2 c_\infty^2}, \frac{1}{49c_\infty^2}\right\}$ *and we know* $\gamma_1 = O(c_\infty^{-2})$*. At least one of the iterations in Algorithm 3, after rounding, is an* ϵ*-saddle point of the EOT problem* (2).

Proof. According to Lemma 18, we only need to show that (82) holds after T iterations as defined in (86).

We follow the same idea as the proof of Theorem 12. First we reduce $\tilde{E}(f^{t+1}, g^{t+1}, \lambda^{t+1}, \lambda^t)$ from $\tilde{E}(f^0, g^0, \lambda^0, \lambda^{-1}) = \tilde{F}(f^0, g^0, \lambda^0)$ to a constant s by running t_1 steps. By Lemma 17, we have

$$t_1 \le 1 + \frac{6}{\eta\gamma_1 s} - \frac{6}{\eta\gamma_1 \tilde{F}(f^0, g^0, \lambda^0)}. \tag{87}$$

Secondly, starting from s, we continue running the algorithm, and assume that there are t_2 iteration in which (82a) fails. By (30b) we have

$$t_2 \leq 1 + \frac{72s}{\eta\epsilon'^2}.$$

Therefore, we know that the total iteration number that (82a) fails can be upper bounded by

$$T_1 = t_1 + t_2 \leq 2 + \frac{72s}{\eta\epsilon'^2} + \frac{6}{\eta\gamma_1 s} - \frac{6}{\eta\gamma_1 \tilde{F}(f^0, g^0, \lambda^0)}$$

iterations. By choosing $s = \frac{\epsilon'}{6\sqrt{\gamma_1}}$, we know that

$$T_1 \leq \begin{cases} 2 + \frac{12}{\eta\sqrt{\gamma_1}\epsilon'} + \frac{36}{\eta\sqrt{\gamma_1}\epsilon'} - \frac{6}{\eta\gamma_1\tilde{F}(f^0,g^0,\lambda^0)} \leq 2 + \frac{48}{\eta\sqrt{\gamma_1}\epsilon'} & \text{if } \tilde{F}(f^0, g^0, \lambda^0) \geq \frac{\epsilon'}{6\sqrt{\gamma_1}}, \\ 2 + \frac{12}{\eta\sqrt{\gamma_1}\epsilon'} + \frac{36}{\eta\sqrt{\gamma_1}\epsilon'} - \frac{6}{\eta\gamma_1\tilde{F}(f^0,g^0,\lambda^0)} \leq 2 + \frac{12}{\eta\sqrt{\gamma_1}\epsilon'} & \text{otherwise.} \end{cases}$$

Therefore, we have $T_1 \leq 2 + \frac{48}{\eta\sqrt{\gamma_1}\epsilon'}$. Similarly, from Lemma 14 we know that, starting from s, the number of iterations that (82b) and (82c) fail can be respectively bounded by

$$t_3 \leq 1 + \frac{3600(1-\theta)^2 c_\infty^2 s}{\eta\epsilon^2(2\theta - \theta^2)}, \quad \text{and} \quad t_4 \leq 1 + \frac{3528 c_\infty^2 s}{\eta\epsilon^2}.$$

By choosing $s = \epsilon$, we have the total iteration numbers that (82b) and (82c) fail can be respectively bounded by

$$T_2 = t_1 + t_3 \leq 2 + \frac{3600(1-\theta)^2 c_\infty^2}{\eta\epsilon(2\theta - \theta^2)} + \frac{6}{\eta\gamma_1\epsilon} - \frac{6}{\eta\gamma_1\tilde{F}(f^0, g^0, \lambda^0)} \leq 2 + \frac{3600(1-\theta)^2 c_\infty^2}{\eta\epsilon(2\theta - \theta^2)} + \frac{6}{\eta\gamma_1\epsilon}$$

and

$$T_3 = t_1 + t_4 \leq 2 + \frac{3528 c_\infty^2}{\eta\epsilon} + \frac{6}{\eta\gamma_1\epsilon} - \frac{6}{\eta\gamma_1\tilde{F}(f^0, g^0, \lambda^0)} \leq 2 + \frac{3528 c_\infty^2}{\eta\epsilon} + \frac{6}{\eta\gamma_1\epsilon}.$$

Finally, by letting $s = \epsilon/6$ in (87), we know that

$$\tilde{E}(f^{T_4-1}, g^{T_4-1}, \lambda^{T_4-1}, \lambda^{T_4-2}) \leq \epsilon/6 \tag{88}$$

after

$$T_4 = 1 + \frac{36}{\eta\gamma_1\epsilon}$$

iterations. From (88) we know that

$$\tilde{F}(f^{T_4-1}, g^{T_4-1}, \lambda^{T_4-1}) \leq \epsilon/6,$$

which implies that (82d) holds with $T = T_4$ by noting (30a).

Combining the above discussions, we know that after $T = T_1 + T_2 + T_3 + T_4 + 1$ iterations, there must exist at least one iteration such that the sufficient condition (82) holds, and thus the output of PAME is an ϵ-optimal solution to the original EOT problem (2). □

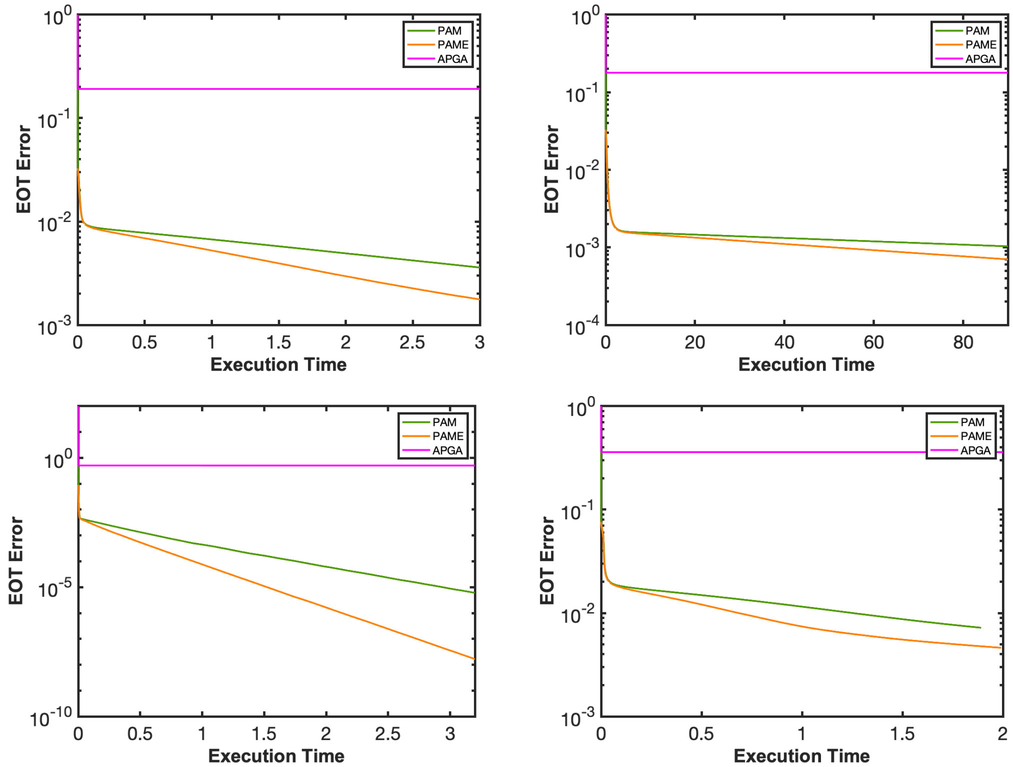

Figure 1: Computational time comparison between PAM, PAME and APGA algorithms on Gaussian distributions. **Upper Left**: $N = 10, n = 100, \eta = 0.1$, **Upper Right**: $N = 10, n = 500, \eta = 0.1$, **Bottom Left**: $N = 10, n = 100, \eta = 0.5$, **Bottom Right**: $N = 5, n = 100, \eta = 0.1$.

Remark 20 *We are not able to analytically prove that PAME has an improved complexity bound at this moment yet. The APGA proposed in (Scetbon et al., 2021) in fact has better complexity than PAM and PAME. However, as demonstrated in (Scetbon et al., 2021) and in our numerical experiments (Sections 6), APGA performs worse than PAM. We believe the reason is that APGA takes gradient step for the variables f and g, while PAM exactly minimizes the subproblems corresponding to these two variables. It is the exact minimization step that led to the improvement. Developing a provably better algorithm is definitely an important and interesting future direction.*

6. Numerical Experiments

We compare the performance of PAME with PAM and APGA (60) (Scetbon et al., 2021) on a synthetic dataset: the Gaussian distributions. We also conduct numerical comparison on another synthetic dataset: the fragmented hypercube dataset. The results are included in the following sections.

6.1 Numerical Results on Gaussian Distribution

Gaussian Distribution: Consider the case when two sets of discrete support $\{x_i\}_{i\in[n]}, \{y_j\}_{j\in[n]}$ are independently sampled from Gaussian distributions

$$\mathcal{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}\right) \tag{89}$$

and

$$\mathcal{N}\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}\right) \tag{90}$$

respectively. The base cost matrix C^{base} is computed by $C^{base}_{i,j} = \|x_i - y_j\|_2^2$. Assume we have N agents. The cost matrix of each agent can be obtained by adding Gaussian noise sampled from $\mathcal{N}(0, 10)$ to each element of the base cost. For instance, for the k-th agent with a cost matrix C^k, we have $C^k_{i,j} = |C^{base}_{i,j} + \mathcal{N}(0, 10)|$.

We then set $a = b = [1/n, ..., 1/n]$ for all experiments. For all algorithms, we set $\tau = \frac{5\eta}{c_\infty^2}$ and we set $\theta = 0.1$ for the PAME algorithm. We consider the EOT error as a measure of optimality. The EOT error at iteration t is defined by

$$Error = |\ell(\pi(f^t, g^t, \lambda^t), \lambda^t) - \ell^*|, \tag{91}$$

where ℓ^* is the approximated optimal value of EOT (2) obtained by running the PAM algorithm for 20000 iterations. Figure 1 plots the EOT error against the execution time for Gaussian distributions. We run each algorithm for 2000 iterations for different parameter settings. In all cases, the PAME and PAM perform significantly better than APGA, and PAME also shows significant improvement over PAM.

Figure 2 shows the optimal couplings obtained from the standard OT and EOT of two Gaussian distributions under three different metrics: the Euclidean cost ($\|\cdot\|_2$), the square Euclidean cost ($\|\cdot\|_2^2$) and the $L_1^{1.5}$ norm ($\|\cdot\|_1^{1.5}$) respectively. We set $n = 4, \eta = 0.05$ and generate samples independently according to (89). For the EOT problem, we consider three agents with cost matrices computed by the three metrics mentioned above. Note that the entropy regularized models lead to a dense transportation plan and Figure 2 only plots the couplings with a probability larger than 10^{-3}. We see that all the agents have the same total cost in the EOT model, and as expected, the cost is smaller than the other three OT costs obtained by using the same metric. The sub figures in the first row imply that if we split the workload to three parts evenly, then the three agents will each have costs 5.935/3, 2.158/3 and 5.030/3, which is not fair because they have different costs. But the EOT model can indeed guarantee the fairness.

We further compare the computational time for Gaussian distributions. We generate the data as previously and set the parameters as $\eta = 0.5, \tau = 5\eta/c_\infty^2$. We stop all the algorithms when the EOT error (91) is less than 10^{-4}. Tables 1 and 2 show the CPU time (in seconds) for different (n, N) pairs. The reported computational time is averaged over 5 runs. In Table 1, the APGA algorithm fails to reach an error of 10^{-4} in 500000 iterations when $n = 100$ and $n = 500$. We conclude that the APGA algorithm converges much slower than PAM and PAME algorithms. The PAME algorithm performs the best among all three algorithms.

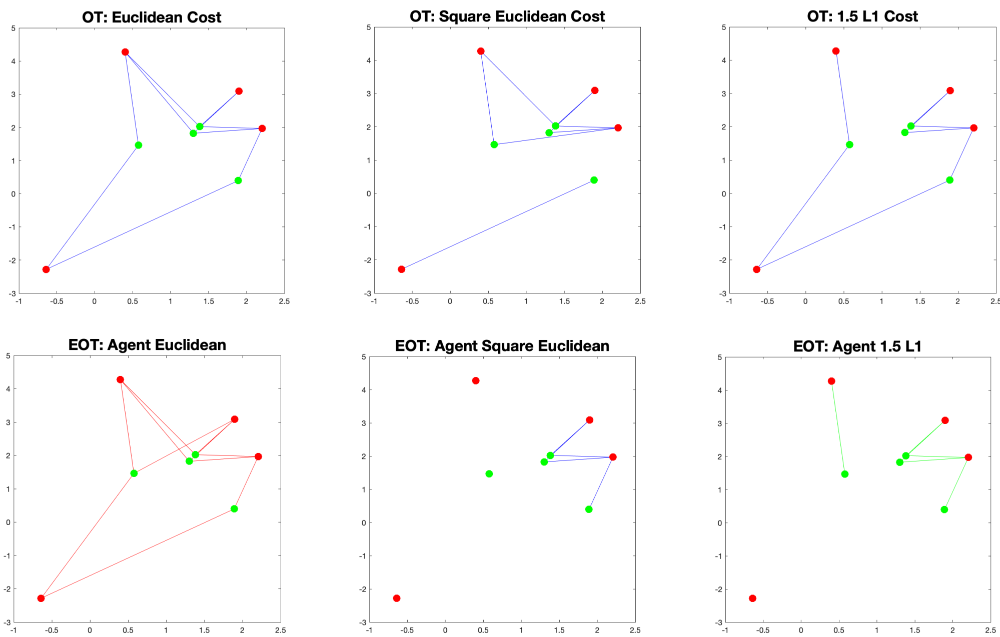

Figure 2: Optimal couplings of standard OT (first row) and EOT (second row). OT Square Euclidean Cost: 5.935; OT Euclidean Cost: 2.158; OT $L_1^{1.5}$ Cost: 5.030; EOT Cost: 0.906.

Table 1: CPU time (in seconds) comparison for Gaussian Distributions. Fixed $N = 3$.

Algorithms	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 500$
PAM	0.038283	0.096039	0.177209	1.038785	1.768560
PAME	0.025210	0.065552	0.091593	0.564618	1.340104
APGA	1.125673	12.364775	106.768840	-	-

6.2 Numerical Results on Fragmented Hypercube Dataset

In this section, we compare the performance of PAME with PAM and APGA (60) (Scetbon et al., 2021) on the fragmented hypercube dataset.

Fragmented Hypercube: We now consider transferring mass between a uniform distribution over a hypercube $\mu = \mathcal{U}([-1,1]^d)$ and a distribution ν obtained by a pushforward $\nu = T_\sharp \mu$ defined by $T(x) = x + 2\text{sign}(x) \odot \left(\sum_{m=1}^{m^*} e_m\right)$. Here $\text{sign}(\cdot)$ is taken elementwisely, $m^* \in [d]$ and $e_i,\ i \in [d]$ is the canonical basis of $\mathbb{R}^d$. In our experiments, we set $d = 10, m^* = 2$ and sample two base support sets $\{x_i^{base}\}_{i\in[n]}, \{y_j^{base}\}_{j\in[n]}$ independently from μ, ν. To obtain the cost matrix for one agent, we first add Gaussian noise sampled from $\mathcal{N}(0,1)$ to the base support sets to get $\{x_i^{noisy}\}_{i\in[n]}, \{y_j^{noisy}\}_{j\in[n]}$ and compute the cost using the noisy support sets. For instance, for the k-th agent, we have $(x_i^{noisy})^k = x_i^{base} + \mathcal{N}(0,1), (y_j^{noisy})^k = y_j^{base} + \mathcal{N}(0,1)$ and $C_{i,j}^k = \|(x_i^{noisy})^k - (y_j^{noisy})^k\|_2^2$.

Table 2: CPU time (in seconds) comparison for Gaussian Distributions. Fixed $n = 50$.

Algorithms	$N=2$	$N=3$	$N=5$	$N=10$	$N=20$
PAM	0.180343	0.177209	1.021598	0.719909	0.903429
PAME	0.105775	0.091593	0.560785	0.385495	0.618381
APGA	70.959300	106.768840	169.213121	178.637889	212.697866

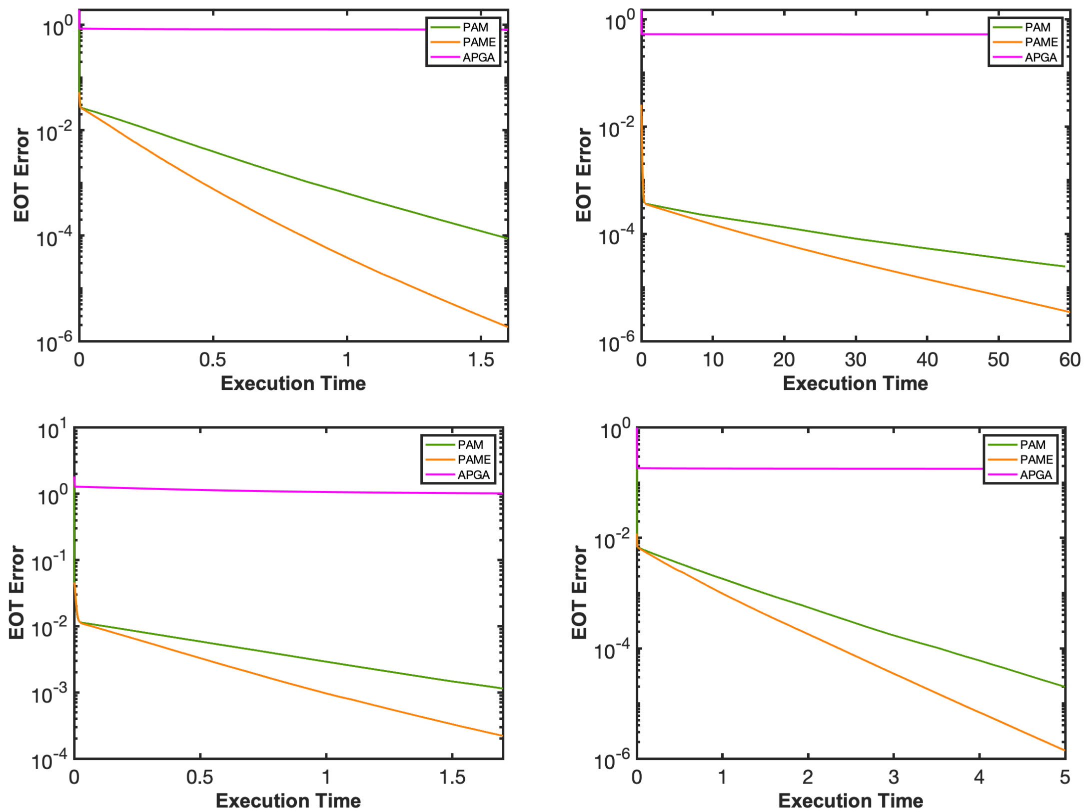

Figure 3: CPU time comparison between PAM, PAME and APGA algorithms on the Fragmented Hypercube dataset. **Upper Left**: $N = 5, n = 100, \eta = 0.2$, **Upper Right**: $N = 5, n = 500, \eta = 0.2$, **Bottom Left**: $N = 5, n = 100, \eta = 0.1$, **Bottom Right**: $N = 10, n = 100, \eta = 0.2$.

Figures 3 plots the EOT error versus the CPU time for Fragmented Hypercube dataset. We run PAM for 20000 iterations to get an approximate optimal ℓ^* and run all algorithms for 2000 iterations for different parameter settings. In all cases, the PAME and PAM perform significantly better than APGA, and PAME also shows significant improvement over PAM.

We then compare the computational time for Fragmented Hypercube dataset. We set the parameters as $\eta = 0.2, \tau = 5\eta/c_\infty^2$. We stop all the algorithms when the EOT error (91) is less than 10^{-4}. Tables 3 and 4 show the CPU time (averaged over 5 runs) for different (n, N) pairs. We see that the PAME algorithm still performs the best among all three algorithms. Note that in Table 3 the APGA algorithm fails to reach an error of 10^{-4} in 500000 iterations when $n = 50$, $n = 100$ and $n = 500$.

Table 3: CPU time (in seconds) comparison for Fragmented Hypercube. Fixed $N = 3$.

Algorithms	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 500$
PAM	0.165165	0.101363	0.177209	1.154193	3.840804
PAME	0.112527	0.068529	0.091593	0.588553	2.007653
APGA	13.771911	22.017804	-	-	-

Table 4: CPU time (in seconds) comparison for Fragmented Hypercube. Fixed $n = 20$.

Algorithms	$N = 2$	$N = 3$	$N = 5$	$N = 10$	$N = 20$
PAM	0.003180	0.101363	0.172696	0.253926	0.231646
PAME	0.040302	0.068529	0.110080	0.150513	0.129156
APGA	1.007166	22.017804	13.801154	6.304324	3.749495

7. Conclusion

In this paper, we have provided the first convergence analysis of the PAM algorithm for solving the EOT problem. Specifically, we have shown that it takes at most $O(\epsilon^{-2})$ iterations for the PAM algorithm to find an ϵ-saddle point. We have proposed a PAME algorithm which incorporates the extrapolation technique to PAM. The PAME shows significant numerical improvement over PAM. Results in this paper might shed lights on designing new BCD type algorithms.

Acknowledgments

Research of Shiqian Ma was supported in part by Office of Naval Research (ONR) grant N00014-24-1-2705, National Science Foundation (NSF) grants DMS-2243650, CCF-2308597, CCF-2311275 and ECCS-2326591, UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program, and a startup fund from Rice University. Research of Lifeng Lai was supported by National Science Foundation under grants CCF-2112504 and CCF-2232907.

References

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in neural information processing systems*, pages 1964–1974, 2017.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458, 2017.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Jelena Diakonikolas and Lorenzo Orecchia. Alternating randomized block coordinate descent. In *International Conference on Machine Learning*, pages 1224–1232. PMLR, 2018.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning*, pages 1367–1376. PMLR, 2018.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.

M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo. Iteration complexity analysis of block coordinate descent method. *Mathematical Programming Series A*, 163(1):85–114, 2017.

Minhui Huang, Shiqian Ma, and Lifeng Lai. A Riemannian block coordinate descent method for computing the projection robust Wasserstein distance. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4446–4455. PMLR, 2021a.

Minhui Huang, Shiqian Ma, and Lifeng Lai. Projection robust Wasserstein barycenters. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4456–4465. PMLR, 2021b.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust Wasserstein distance and Riemannian optimization. In *NeurIPS*, volume 33, 2020.

Erika Mackin and Lirong Xia. Allocating indivisible items in categorized domains. *arXiv preprint arXiv:1504.05932*, 2015.

Hervé Moulin. *Fair division and collective welfare*. MIT press, 2003.

A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2005.

Y. E. Nesterov. *Introductory lectures on convex optimization: A basic course*. Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7553-7.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. In *Advances in Neural Information Processing Systems*, pages 15604–15614, 2019.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *International Conference on Machine Learning*, pages 5072–5081, 2019.

Meyer Scetbon, Laurent Meunier, Jamal Atif, and Marco Cuturi. Equitable and optimal transport with multiple agents. In *International Conference on Artificial Intelligence and Statistics*, pages 2035–2043. PMLR, 2021.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Ruoyu Sun and Mingyi Hong. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. *Advances in Neural Information Processing Systems*, 28, 2015.

Robert E Tarjan. Dynamic trees as search trees via Euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177, 1997.

Y. Xu and W. Yin. A block coordinate descent method for regularized multi-convex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 63(3):1758–1789, 2013.