ON THE CONVERGENCE OF GRADIENT EXTRAPOLATION
METHODS FOR UNBALANCED OPTIMAL TRANSPORT

Quang Minh Nguyen
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA
nmquang@mit.edu

Hoang H. Nguyen
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA, USA
hnguyen455@gatech.edu

Yi Zhou
IBM Research
Almaden Research Center
San Jose, CA, USA
yi.zhou@ibm.com

Lam M. Nguyen
IBM Research
Thomas J. Watson Research Center
Yorktown Heights, NY, USA
LamNguyen.MLTD@ibm.com

ABSTRACT

We study the Unbalanced Optimal Transport (UOT) between two measures of possibly different masses with at most \( n \) components, where marginal constraints of the standard Optimal Transport (OT) are relaxed via Kullback-Leibler divergence with regularization factor \( \tau \). We propose a novel algorithm based on Gradient Extrapolation Method (GEM-UOT) to find an \( \varepsilon \)-approximate solution to the UOT problem in \( O(\kappa n^2 \log \left( \frac{\tau n}{\varepsilon} \right)) \), where \( \kappa \) is the condition number depending on only the two input measures. Compared to the only known complexity \( O\left( \frac{\tau n^2 \log n}{\varepsilon} \log \left( \frac{\log n}{\varepsilon} \right) \right) \) for solving the UOT problem via the Sinkhorn algorithm, ours is better in \( \varepsilon \) and lifts Sinkhorn’s linear dependence on \( \tau \), which hindered its practicality to approximate the standard OT via UOT. Our proof technique is based on a novel dual formulation of the squared \( \ell_2 \)-norm regularized UOT objective, which is of independent interest and also leads to a new characterization of approximation error between UOT and OT in terms of both the transportation plan and transport distance. To this end, we further present an algorithm, based on GEM-UOT with fine tuned \( \tau \) and a post-process projection step, to find an \( \varepsilon \)-approximate solution to the standard OT problem in \( O(\kappa n^2 \log \left( \frac{n}{\varepsilon} \right)) \), which is a new complexity in the literature of OT. Extensive experiments on synthetic and real datasets validate our theories and demonstrate the favorable performance of our methods in practice.

1 Introduction

The optimal transport (OT) problem was originated from the need to find the optimal cost to transport masses from one distribution to another distribution [Villani, 2003, 2008]. While initially developed by theorists, OT has found widespread applications in statistics and machine learning (ML) (see e.g. [Ho et al., 2017, Arjovsky et al., 2017, Rabin and Papadakis, 2015, Courty et al., 2016, Solomon et al., 2014]). However, standard OT is restricted to the case where the input measures are normalized to unit masses, which facilitates the development of the Unbalanced Optimal Transport (UOT) problem between two measures of possibly different masses. The class of UOT problem relaxes OT’s marginal constraints. Specifically, it is a regularized version of Kantorovich formulation placing penalty functions on the marginal distributions based on some given divergences [Liero et al., 2017]. While there have been several divergences considered by the literature, such as squared \( \ell_2 \) norm [Blondel et al., 2018], \( \ell_1 \) norm [Caffarelli and McCann, 2010], or general \( \ell_p \) norm [Lee et al., 2019], UOT with Kullback-Leiber (KL) divergence [Chizat et al., 2017] is the most prominent and has been used in statistics and machine learning [Frogner et al., 2015], deep learning [Yang and Uhler, 2019], domain adaptation [Balaji et al., 2020, Patras et al., 2021], bioinformatics [Schiebinger et al., 2019], and OT robustness [Balaji et al., 2020, Le et al., 2021]. Throughout this paper, we refer to UOT penalized by KL divergence...
as simply UOT, unless otherwise specified. We hereby define some notations and formally present our problem of interest.

**Notations:** We let \([n]\) stand for the set \(\{1, 2, ..., n\}\) while \(\mathbb{R}_+^n\) stands for the set of all vectors in \(\mathbb{R}^n\) with nonnegative entries. For a vector \(x \in \mathbb{R}^n\) and \(p \in [1, \infty)\), we denote \(|x|_p\) as its \(\ell_p\)-norm and \(diag(x) \in \mathbb{R}^{n \times n}\) as the diagonal matrix with \(diag(x)_{ii} = x_i\). Let \(A\) and \(B\) be two matrices of size \(n \times n\), we denote their Frobenius inner product as: 
\[
\langle A, B \rangle = \sum_{i,j=1}^n A_{ij}B_{ij},
\]
\(I_n\) stands for a vector of length \(n\) with all of its components equal to 1. The KL divergence between two vectors \(x, y \in \mathbb{R}^n\) is defined as \(KL(x||y) = \sum_{i=1}^n x_i \log \left(\frac{x_i}{y_i}\right) - x_i + y_i\). The entropy of a matrix \(X \in \mathbb{R}_+^{n \times n}\) is given by \(H(X) = -\sum_{i,j=1}^n x_{ij} \log(x_{ij}) - 1\).

**Unbalanced Optimal Transport:** For a couple of finite measures with possibly different total mass \(a = (a_1, ..., a_n) \in \mathbb{R}^n_+\) and \(b = (b_1, ..., b_n) \in \mathbb{R}^n_+\), we denote \(\alpha = \sum_{i=1}^n a_i\) and \(\beta = \sum_{i=1}^n b_i\) as the total masses, and \(a_{\min} = \min_{1 \leq i \leq n} \{a_i\}\) and \(b_{\min} = \min_{1 \leq i \leq n} \{b_i\}\) as the minimum masses. The UOT problem can be written as:
\[
UOT_{KL}(a, b) = \min_{X \in \mathbb{R}_+^{n \times n}} \left\{ f(X) := \langle C, X \rangle + \tau KL(X1_n||a) + \tau KL(X^\top 1_n||b) \right\},
\]
where \(C\) is a given cost matrix, \(X\) is a transportation plan, \(\tau > 0\) is a given regularization parameter. When \(a1_n = b1_n\) and \(\tau \rightarrow \infty\), reduces to a standard OT problem.

**Definition 1.1.** For \(\varepsilon > 0\), we define \(X\) as an \(\varepsilon\)-approximate transportation plan of \(UOT_{KL}(a, b)\) if it satisfies:
\[
\langle C, X \rangle + \tau KL(X1_n||a) + \tau KL(X^\top 1_n||b) \leq \langle C, X_f \rangle + \tau KL(X_f1_n||a) + \tau KL(X^\top_f 1_n||b) + \varepsilon,
\]
where \(X_f = \arg\min_{X \in \mathbb{R}_+^{n \times n}} f(X)\) is the optimal transportation plan of the UOT problem \([1]\).

**Computational Complexity:** Recently, the existing work has approached the class of OT problems using gradient methods or variations of the Sinkhorn algorithm. Guminov et al. [2021] used an alternative minimization approach, while [Altschuler et al., 2017] proposed the Greenkhorn method, a greedy version of the Sinkhorn method, and it was further improved by [Lin et al., 2019a] and [Lin et al., 2019b]. However, the complexity theory and algorithms for UOT remain nascent despite its recent emergence. Pham et al. [2020b] shows that the Sinkhorn algorithm can solve the UOT problem in \(O\left(\frac{\tau n^2 \log(n)}{\varepsilon} \log(b_{\min}^2 \log(n))\right)\) for finding an \(\varepsilon\)-approximate solution to the discrete UOT problem between two measures with \(n\) components. Chapel et al. [2021a] proposes Majorization-Minimization (MM) algorithm as a new numerical solver for UOT achieving \(O(n^{2.5})\) empirical complexity. On the other hand, to the best of our knowledge, there is no existing analysis of gradient-based methods to solve the class of UOT problems. Yang and Uhler [2019a] is the only work that used gradient descent as a heuristic to solve the entropic regularized UOT problem.

**Approximation of OT:** Beside computational complexity, the approximability of standard OT via UOT has remained an open problem. Recently, Chapel et al. [2021] discusses the possibility to approximate OT transportation plan using that of UOT and empirically verifies it. Despite the well known fact that OT is recovered from UOT when \(\tau \rightarrow \infty\) and the masses are balanced, no work has analyzed the rate under which UOT converges to OT\([1]\) Characterization of approximation error between UOT and OT could give rise to new formal methods for OT retrieval from UOT. In this regime of large \(\tau\), Sinkhorn’s linear dependency much hinders its practicality, motivating Sejourné et al. [2022] to alleviate this issue in the specific case of 1-D UOT. However, no algorithm for general UOT problem that could achieve sublinear in \(\tau\) exists. From the practical perspective, \(\tau\) can be large in certain applications [Schiebler et al., 2019] and, for a specific case of UOT penalized by squared \(\ell_2\) norm, be even of the order of thousands [Blondel et al., 2018].

**Contributions:** In this paper, we provide a comprehensive study of UOT regarding its computational complexity and approximation of standard OT. We consider the setting where KL divergences are used to penalize the marginal constraints. Our contributions can be summarized as follows:

- We provide a novel dual formulation of the squared \(\ell_2\)-norm regularized UOT objective, which is the basis of our algorithms. This could facilitate further algorithmic development for solving UOT, since the current literature is limited to the dual formulation of entropic-regularized UOT problem [Chizat et al., 2017].
- Based on the Gradient Extrapolation Method (GEM), we propose GEM-UOT algorithm for finding an \(\varepsilon\)-approximate solution to the discrete UOT problem. We show that our algorithm is of order \(O\left(\kappa n^2 \log^2(n) \log(b_{\min}^2 \log(n))\right)\) by Sinkhorn algorithm [Pham et al., 2020a] and lifts its linear dependence on \(\tau\), a bottleneck in the regime of large \(\tau\) Sejourné et al. [2022].

The complexities of GEM-UOT and other algorithms in the literature are summarized in Table 1.

\footnote{For squared \(\ell_2\) norm penalized UOT variant, [Blondel et al., 2018] showed that the error in terms of transport distance is \(O\left(\frac{\varepsilon}{\tau}\right)\).}
Table 1: Complexity of algorithms for solving UOT problems.

| Algorithms          | Complexity                                      | Assumptions | Description                                                                 |
|---------------------|-------------------------------------------------|-------------|-----------------------------------------------------------------------------|
| MM [Chapel et al. [2021]] | $O(n^{3.27})$ - Empirical Complexity            | N.A         | Numerical solver. No theoretical guarantee.                                 |
| Sinkhorn [Pham et al. [2020b]] | $O\left((\alpha + \beta) \cdot \frac{\tau n^2 \log(n)}{\varepsilon} \cdot \log \left(\frac{\log(n)(\alpha + \beta)}{\varepsilon}\right)\right)$. | (S1), (S2), (S3), (S4) | Return a transportation plan achieving $\text{UOT}_{\text{KL}}(a, b) \pm \varepsilon$. |
| GEM-UOT (this paper)          | $O\left((\alpha + \beta) \cdot \frac{\tau n^2}{\varepsilon} \cdot \log \left(n(\alpha + \beta)\right)\right)^*$. | (A1)≡(S1), (A2)≡(S2), (A3) | Return a transportation plan achieving $\text{UOT}_{\text{KL}}(a, b) \pm \varepsilon$. |
| GEM-RUOT (this paper)          | $O\left((\alpha + \beta) \cdot \frac{\tau n^2}{\varepsilon} \cdot \log \left(n(\alpha + \beta)\right)\right)$. | (A1)≡(S1), (A2)≡(S2), (A3) | Approximate $\text{UOT}_{\text{KL}}(a, b) \pm \varepsilon$ and return a heuristic transportation plan. |

* $\kappa$ is defined in Section S.1

- To the best of our knowledge, we establish the first characterization of the approximation error between UOT and OT (in the context of KL-penalized UOT). In particular, we show that both of UOT’s transportation plan and transport distance converge to OT’s marginal constraints and transport distance with the rate $O(\frac{1}{\varepsilon})$. This result can open up directions that use UOT to approximate standard OT, where it is known that OT is more robust to outliers than OT both theoretically [Fatras et al. [2021]] and practically [Balaji et al. [2020]].

- Inspired by our results on approximation error, we present GEM-OT, which is the first algorithm with theoretical guarantee that obtains an $\varepsilon$-approximate solution to the standard OT problem retrieved from UOT solution. GEM-OT is of order $O \left(\kappa \cdot n^2 \log \left(\frac{n}{\varepsilon}\right)\right)$, which is the first complexity to achieve logarithmic dependence on $\varepsilon^{-1}$ in the literature of OT.

**Paper Organization:** The rest of the paper is organized as follows. We introduce the background of regularized UOT problems and present our novel dual formulation in Section 2. In Section 3 we analyze the complexities our proposed algorithms GEM-UOT and its practical variant GEM-RUOT. The results on approximability of OT via UOT are established in Section 4. In Section 5 we experiment on both synthetic and real datasets to compare the performances of our algorithms with the state-of-the-art Sinkhorn algorithm, and empirically verify our theories on approximation error. We conclude the paper in Section 6.

## 2 Background

In this section, we first present the entropic regularized UOT problem, used by Sinkhorn algorithm. Then we consider the squared $\ell_2$-norm regularized UOT and derive a novel dual formulation- the basis for our algorithmic development.

### 2.1 Entropic Regularized UOT

Inspired by the literature of the entropic regularized OT problem, the entropic version of the UOT problem has been considered. The problem is formulated as:

$$
\min_{X \in \mathbb{R}^{m \times n}} g(X) := \langle C, X \rangle - \eta H(X) + \tau \text{KL}(X^n_1 || a) + \tau \text{KL}(X^{\top}I_n || b),
$$

(3)

where $\eta > 0$ is a given regularization parameter. By [Chizat et al. [2017]], optimizing the Fenchel-Legendre dual of the above entropic regularized UOT is equivalent to:

$$
\min_{u, v \in \mathbb{R}^m} h(x = (u, v)) := \eta \sum_{i,j=1}^{n} \exp \left(\frac{u_i + v_j - C_{ij}}{\eta}\right) + \tau \left(e^{-u/\tau}, a\right) + \tau \left(e^{-v/\tau}, b\right).
$$

(4)

**Remark 2.1.** We note that Sinkhorn [Pham et al. [2020b]] must set small $\eta = O \left(\frac{\varepsilon}{\log(n)}\right)$, while the smoothness condition number of the above dual objective $h(x)$ is large at least at the exponential order of $\eta^{-1}$. On the other hand, the original UOT objective is non-smooth due to KL divergences. Therefore, direct application of gradient methods to solve either the primal or the dual of entropic regularized UOT would not result in competitive convergence rate, which
Moreover, we have \( \eta > 0 \) where \( \eta \) is a given regularization parameter. Let \( \eta^* = \arg\min_{X \in \mathbb{R}^{n \times n}} g_{\eta}(X) \), which is unique by the strong convexity of \( g_{\eta}(X) \). The dual formulation is given by the following Lemma 2.2 whose proof can be found in Appendix C.2.

**Lemma 2.2.** The dual problem to (5) is:

\[
\max_{(u,v,t) \in \mathcal{X}} \left\{ -\frac{1}{4\eta} \sum_{i,j=1}^{n} t_{ij}^2 - \tau \left( \mathbf{e}^\top \mathbf{e} - u^\top \mathbf{a} \right) - \tau \left( \mathbf{e}^\top \mathbf{e} - v^\top \mathbf{b} \right) + \mathbf{a}^\top \mathbf{I}_n + \mathbf{b}^\top \mathbf{I}_n \right\},
\]

where \( \mathcal{X} = \left\{ (u,v,t) \mid u, v \in \mathbb{R}^n, t \in \mathbb{R}^{n \times n} : t_{ij} \geq 0, t_{ij} \geq u_i + v_j - C_{ij} \quad \forall i, j \right\} \). Let \( x^* = (u^*, v^*, t^*) \) be an optimal solution to (6), then the primal solution to (5) is given by:

\[
X^*_{ij} = \frac{1}{2\eta} \max \left\{ 0, u^*_i + v^*_j - C_{ij} \right\}.
\]

Moreover, we have \( \forall i, j \in [n] \):

\[
\begin{align*}
-\frac{u^*_i}{\tau} + \log(a_i) &= \log\left( \sum_{k=1}^{n} X^*_{ik} \right), \\
-\frac{v^*_j}{\tau} + \log(b_j) &= \log\left( \sum_{k=1}^{n} X^*_{kj} \right), \\
t^*_{ij} &= \max \left\{ 0, u^*_i + v^*_j - C_{ij} \right\}.
\end{align*}
\]

3 Complexity Analysis of Approximating Unbalanced Optimal Transport

In this section, we provide the algorithmic development of GEM-UOT and its practical variant GEM-RUOT, and their computational complexities. We present the regularity conditions in Section 3.2 and characterize the problem as composite optimization in Section 3.3. The complexity analysis for GEM-UOT and GEM-RUOT respectively follow in Section 3.4 and 3.5. Proofs are deferred to Appendix D.

3.1 List of Quantities

Given the two masses \( a, b \in \mathbb{R}^n_+ \), we define the following notations and quantities to be used throughout the paper:

\[
\begin{align*}
a_{\text{min}} &= \min_{1 \leq i \leq n} \{ a_i \}, & b_{\text{min}} &= \min_{1 \leq j \leq n} \{ b_j \} \\
a_{\text{max}} &= \|a\|_\infty, & b_{\text{max}} &= \|b\|_\infty \\
\kappa &= \frac{1}{\min\{a_{\text{min}}, b_{\text{min}}\}}, & R &= \frac{(\alpha + \beta)^2}{4}, \\
D &= \|C\|_\infty + \eta(\alpha + \beta) + \tau \log\left( \frac{\alpha + \beta}{2} \right) - \tau \min\{\log(a_{\text{min}}), \log(b_{\text{min}})\}, \\
p &= \frac{1}{2} \min\{a_{\text{min}}, b_{\text{min}}\} e^{-\frac{D}{2}}, & q &= \alpha + \beta, \\
L_1 &= \|C\|_\infty + 2\eta q + 2\tau |\log(p)| + 2\tau |\log(q)| + \tau \max_{i} |\log(a_i)| + \tau \max_{i} |\log(b_i)|.
\end{align*}
\]
3.2 Assumptions

We hereby present the assumptions (A1-A3) required by our algorithms. For interpretation, we also restate the regularity conditions of Sinkhorn [Pham et al., 2020b] for solving UOT, where detailed discussion on how their assumptions in the original paper are equivalent to (S1-S4) is deferred to Appendix [A.1].

Regularity Conditions of this Paper

(A1) \( a_{\min} > 0, b_{\min} > 0 \)

(A2) \( |\log(a_{\min})| = O(\log(n)), |\log(b_{\min})| = O(\log(n)) \).

(A3) \( \tau = \Omega(\min\{\frac{1}{\alpha+\beta}, \|C\|_\infty\}) \).

Regularity Conditions of Sinkhorn

(S1) \( a_{\min} > 0, b_{\min} > 0 \)

(S2) \( |\log(a_{\min})| = O(\log(n)), |\log(b_{\min})| = O(\log(n)) \).

(S3) \( |\log(a_{\max})| = O(\log(n)), |\log(b_{\max})| = O(\log(n)) \).

(S4) \( \alpha, \beta, \tau \) are positive constants.

Remark 3.1. Compared to the regularity conditions of Sinkhorn, ours lift the strict assumptions (S3) and (S4) that put an upper bound on \( \tau \) and the input masses. Thus our complexity analysis supports much more flexibility for the input masses and is still suitable to applications requiring large \( \tau \) and the input masses. Thus our complexity analysis supports much more flexibility for the input masses and is still suitable to applications requiring large \( \tau \).

3.3 Characterization of the Dual Objective

Optimizing (6) is equivalent to:

\[
\min_{x \in X} h_\eta(x) = (u, v, t) := \frac{1}{4\eta} \sum_{i,j=1}^n t_{ij}^2 + \tau \left( e^{-u/\tau}, a \right) + \tau \left( e^{-v/\tau}, b \right).
\] (11)

For \( x = (u, v, t) \), we consider the functions:

\[
f_\eta(x) = \tau \left( e^{-u/\tau}, a \right) + \tau \left( e^{-v/\tau}, b \right) - \frac{\min\{a_{\min}, b_{\min}\}}{2\tau} e^{-D/\tau} (\|u\|_2^2 + \|v\|_2^2),
\] (12)

\[
w_\eta(x) = \frac{\min\{a_{\min}, b_{\min}\}}{2\tau} e^{-D/\tau} (\|u\|_2^2 + \|v\|_2^2) + \frac{1}{4\eta} \sum_{i,j=1}^n t_{ij}^2.
\] (13)

Then the problem (11) can be rewritten as:

\[
\min_{x \in X} h_\eta(x) = (u, v, t) := f_\eta(x) + w_\eta(x),
\] (14)

which can be characterized as the composite optimization over the sum of the locally convex smooth \( f_\eta(x) \) and the locally strongly convex \( w_\eta(x) \) by the following Lemma 3.2.

Lemma 3.2. Let \( V_D = \{x = (u, v, t) \in \mathbb{R}^{n^2 + 2n} : \forall i, j \in [n], \tau \log(\frac{2a_i}{\alpha+\beta}) \leq u_i \leq D, \tau \log(\frac{2b_j}{\alpha+\beta}) \leq v_j \leq D\} \). Then \( f_\eta(x) \) is convex and \( L \)-smooth in the domain \( V_D \), and \( w_\eta(x) \) is \( \mu \)-strongly convex with:

\[
L = \frac{\alpha + \beta}{2\tau} + \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D/\tau},
\]

\[
\mu = \min \left\{ \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D/\tau}, \frac{1}{2\eta} \right\}.
\]

The choice of \( D \) in the above characterization is motivated by the following Corollary 3.3.

Corollary 3.3. If \( x^* \) is an optimal solution to (14), then \( x^* \in V_D \).
We now proceed to analyze the complexity of GEM-UOT under the regularity assumptions. While minimizing the UOT dual objective, GEM-UOT hence projects its iterates onto the closed convex set \(\mathcal{V}\) through the proximal mapping \(\mathcal{M}_{\mathcal{V},\mathcal{X}}\) to enforce the smoothness and convexity of \(f_\eta(x)\) on the convex hull of all the iterates, while ensuring they are in the feasible domain. Finally, GEM-UOT outputs the approximate solution, computed from the iterates, to the original UOT problem. The steps are summarized in Algorithm 1.

### Algorithm 1 GEM-UOT

1. **Input:** \(C, a, b, \varepsilon, \tau\)
2. **Initialization:** \(x^0 = x^0 = 0, \ y^{-1} = y^0 = 0\)
3. Set \(\eta = \frac{\varepsilon}{2\tau}\)
4. Compute \(D\) based on \(\eta\) (Section 3.1)
5. Compute \(L, \mu\) based on \(D\) (Lemma 3.2)
6. \(\alpha = 1 - \frac{1}{1 + \sqrt{1 + 16\eta^2}}\)
7. \(\psi = \frac{1}{1+\alpha} - 1, \ \rho = \frac{\alpha}{1+\alpha}\mu\)
8. for \(t = 0, 1, 2, \ldots, k\) do
   9. \(y^t = y^{t-1} + \alpha(y^{t-1} - y^{t-2})\)
   10. \(x^t = \mathcal{M}_{\mathcal{V} \cap \mathcal{X}}(y^t; x^{t-1}, \rho)\)
   11. \(y^t = (1 + \psi)^{-1}(x^t + \psi x^{t-1})\)
   12. \(\theta_i = \theta_{i-1} \alpha^{-1}\)
9. end for
10. Compute: \(\tilde{x}^k = (\sum_{i=1}^k \theta_i)^{-1} \sum_{i=1}^k \theta_i x^i\)
11. \(u^k, v^k\) be the vectors such that \(\tilde{x}^k = (u^k, v^k)\)
12. for \(i,j = 1 \rightarrow n\) do
13. \(X_{ij} = \frac{1}{2\eta} \max\{0, u^k + v^k - C_{ij}\}\)
14. end for
15. Return \(X^k\)

### 3.4 Gradient Extrapolation Method for Unbalanced Optimal Transport (GEM-UOT)

#### Algorithm Description

We develop our algorithm based on Gradient Extrapolation Method (GEM), called GEM-UOT, to solve the discrete UOT problem. GEM was proposed by Lan and Zhou [2018] to address problem in the form of (14) and viewed as the dual of the Nesterov’s accelerated gradient (NAG) method.

Let us define some notations to be used throughout this section. We denote the prox-function associated with \(w\) as:

\[
P(x_0, x) := \frac{w_\eta(x) - \langle w_\eta(x_0), x - x_0 \rangle}{\mu},
\]

and the proximal mapping associated with the closed convex set \(\mathcal{Y}\) and \(w\) is defined as:

\[
\mathcal{M}_\mathcal{Y}(g, x_0, \theta) := \text{argmin}_{x \in (a, b) \cap \mathcal{X}} \{ \langle g, x \rangle + w_\eta(x) + \theta P(x_0, x) \}.
\]

One challenge for optimizing (14) is the fact that the component function \(f_\eta(x)\) is smooth and convex only on the locality \(V_D\) (Lemma 3.2). By Corollary 3.3 we can further write (14) as:

\[
\min_{x \in \mathcal{X}} h_\eta(x = (u, v, t)) = \min_{x \in V_D \cap \mathcal{X}} h_\eta(x = (u, v, t)).
\]

While minimizing the UOT dual objective, GEM-UOT hence projects its iterates onto the closed convex set \(V_D \cap \mathcal{X}\) through the proximal mapping \(\mathcal{M}_{V_D \cap \mathcal{X}}\) to enforce the smoothness and convexity of \(f_\eta(x)\) on the convex hull of all the iterates, while ensuring they are in the feasible domain. Finally, GEM-UOT outputs the approximate solution, computed from the iterates, to the original UOT problem. The steps are summarized in Algorithm 1.

The following quantities are used in our analysis:

\[
\sigma_0^2 = ||\nabla f_\eta(x^0)||_2^2,
\]

\[
\Delta_0, \sigma_0 = \mu P(x^0, x^*) + h_\eta(x^0) - h_\eta(x^*) + \sigma_0^2/\mu
\]

We now proceed to analyze the complexity of GEM-UOT under the regularity assumptions.
Complexity Analysis

We further quantify the number of iterations

$$K_0 = 4 \left(1 + \sqrt{1 + 16L^2/\mu} \right) \log \left( \frac{4n^2 \Delta_{n/0}^{1/2}}{\eta} \max \{ \frac{L}{\varepsilon}, \min \{a_{\min}, b_{\min} \}, 1 + \alpha + \beta \} \right)$$

that is required for GEM-UOT to output the $\varepsilon$-approximate solution to $\text{UOT}_{\text{KL}}(a, b)$, summarized in Theorem 3.4.

**Theorem 3.4.** If $k \geq K_0$, the output $X^k$ of Algorithm 1 is the $\varepsilon$-approximation of the UOT problem, i.e.

$$f(X^k) - f(X^*) \leq \varepsilon.$$ 

Its complexity is given by the following Corollary 3.5.

**Corollary 3.5.** Under Assumptions (A1-A3), the complexity of Algorithm 1 is:

$$O \left( (\alpha + \beta) \kappa \cdot n^2 \log \left( \frac{\tau \cdot n(\alpha + \beta)}{\varepsilon} \right) \right),$$

where $\kappa = \frac{1}{\min\{a_{\min}, b_{\min}\}}$ is defined in Section 3.1.

### 3.5 Relaxed UOT Problem and Practical Perspectives

**Motivation of Relaxed UOT Problem:** While GEM-UOT outputs a transportation plan that approximates the UOT distance with the desired accuracy $\varepsilon$, we note that in certain applications the transportation plan itself is not within interest. In particular, generative models [Genevay et al., 2018; Fatras et al., 2021] recently have adopted either OT or UOT as loss metrics for training, in which only the computation of the distance is required. To this end, we consider the relaxed UOT (RUOT) problem, where the goal is to approximate the UOT distance oblivious of the transportation plan.

**Practicality:** Gradient methods have been shown to achieve competitive complexity in the OT literature [Guminov et al., 2021; Dvurechensky et al., 2018b]. The heuristic proposed by Yang and Uhler [2019], which was based on gradient descent, achieved favorable performance for solving the UOT problem. Despite such potential, complexity analysis or development of gradient methods in the context of UOT remain nascent. Through our complexity analysis of GEM-UOT, we want to establish the preliminary understanding of the class of gradient-based optimization when applied to UOT problem. Nevertheless, the convoluted function decomposition (14) of GEM-UOT may make it hard to implement in practice. The strong convexity constant $\mu$ (Lemma 3.2) can be small on certain real datasets, thereby hindering the empirical performance. We thus develop an easy-to-implement algorithm, called GEM-RUOT, for practical purposes that avoids the dependency on $\mu$ in complexity and is specialized for the RUOT problem.

**Alternative Dual Formulation:** GEM-RUOT does not require intricate function decomposition and optimizes directly over the following simple alternative of the dual.

**Corollary 3.6.** The dual problem to (5) is:

$$\max_{x = (u, v) \in \mathbb{R}^{2n}} \left\{ f_u(x) := -\frac{1}{4\eta} \sum_{i,j=1}^{n} \max \{ 0, u_i + v_j - C_{ij} \}^2 - \tau \left( e^{-u_i/\tau}, a \right) - \tau \left( e^{-v_j/\tau}, b \right) + a^T I_n + b^T I_n \right\}. \tag{15}$$

Let $x^* = (u^*, v^*)$ be an optimal solution to (15), then the primal solution to (5) is given by:

$$X^*_{ij} = \frac{1}{2\eta} \max \{ 0, u^*_i + v^*_j - C_{ij} \}. \tag{16}$$

Moreover, we have ‘$i, j \in [n]$’:

$$-\frac{u^*_i}{\tau} + \log(a_i) = \log(\sum_{k=1}^{n} X^n_{ik}), \tag{17}$$

$$-\frac{v^*_j}{\tau} + \log(b_j) = \log(\sum_{k=1}^{n} X^n_{kj}). \tag{18}$$

Optimizing (15) is equivalent to:

$$\min_{u, v \in \mathbb{R}^{2n}} f_u(x = (u, v)) := \frac{1}{4\eta} \sum_{i,j=1}^{n} \max \{ 0, u_i + v_j - C_{ij} \}^2 + \tau \left( e^{-u_i/\tau}, a \right) + \tau \left( e^{-v_j/\tau}, b \right), \tag{19}$$

which is locally smooth by the following Lemma 3.7.
**Algorithm 2** GEM-RUOT

1. **Input:** $C, a, b, \varepsilon, \tau$
2. **Initialization:** $x^0 = y^0 = 0$, $y^{-1} = y^0 = 0$
3. Set $\eta = \frac{2}{L}$
4. Compute $D$ based on $\eta$ (Section 3.1)
5. Compute $L_a$ based on $D$ (Lemma 3.7)
6. for $t = 0, 1, 2, \ldots, k$
   7. $\alpha_t = \frac{t-1}{T}$, $\psi_t = \frac{t-1}{T}$, $\rho_t = \frac{6 \alpha_t}{\varepsilon}$
   8. $y^t = y^{t-1} + \alpha(y^{t-1} - y^{t-2})$
   9. $x' = M_{\beta} (y^t, x^{t-1}, \rho_t)$
10. $x^t = \frac{1 + \psi_t}{1} (x' + \psi_t x^{t-1})$
11. $y^t = \nabla h_a (x^t)$
12. $\theta_t = \theta_{t-1} - \alpha_t$
7. end for
8. Compute $x^k = \sum_{t=1}^{T} \theta_t \sum_{i=1}^{k} \theta_i x^t$
9. $u^k, v^k$ be the vectors such that $x^k = (u^k, v^k)$
10. for $i, j = 1 \rightarrow n$
11. $X_{ij}^k = \frac{1}{2\eta} \max \{0, u_i^k + v_j^k - C_{ij}\}$
12. end for
13. Compute $F_a(x^k)$
14. Return $F_a(x^k), X^k$

**Lemma 3.7.** Let $V_a = \{x = (u, v) \in \mathbb{R}^{2n} : \forall i, j \in [n], \tau \log \left( \frac{2a}{\alpha + \beta} \right) \leq u_i \leq D, \tau \log \left( \frac{2b_i}{\alpha + \beta} \right) \leq v_j \leq D \}$. Then $h_a(x)$ is $L_a$-smooth and convex in $V_a$ with:

$$L_a = \frac{\alpha + \beta}{\tau} + \frac{2\sqrt{n}}{\eta}.$$  

GEM-RUOT adopts the $\ell_2$ distance for its prox-function $P(x_0, x) = \frac{1}{2}||x - x_0||^2_2$, and consequently uses the standard $\ell_2$ projection operator for the proximal mapping associated with the closed convex set $Y$ as $M_Y (g, x_0, \theta) := \argmin_{x = (u, v) \in Y} \{\langle g, x \rangle + \theta P(x_0, x)\}$. Using the convex version of GEM, GEM-RUOT optimizes directly over (19) and projects its iterates onto the closed convex set $V_a$ through the proximal mapping $M_{\beta} ()$ to enforce the smoothness on the convex hull of all the iterates. Finally, it returns $F_a(x^k)$ as an $\varepsilon$-approximation of the distance $UOT_{KL}(a, b)$ and a heuristic transportation plan $X^k$. The steps are summarized in Algorithm 2. The complexity of GEM-RUOT is then given in Theorem 3.8.

**Theorem 3.8.** If $k \geq \sqrt{\frac{12 L_a n D^2}{\varepsilon}}$, then:

$$|F_a(x^k) - UOT_{KL}(a, b)| \leq \varepsilon.$$  

Under the assumptions (A1-A3), the complexity of Algorithm 2 is

$$O \left( \left( \alpha + \beta \right) \cdot \frac{\tau n^{0.75}}{\varepsilon} \cdot \log(n(\alpha + \beta)) \right).$$

## 4 Approximability of Standard OT

For balanced masses and $\tau \to \infty$, UOT problem reduces to standard OT. In this section, we establish the very first characterizations of the diminishing approximation error between UOT and OT in terms of $\tau$. To facilitate our discussion, we first formally define the standard OT problem in Section 4.1. Then in Section 4.2, we show that UOT’s transportation plan converges to the marginal constraints of standard OT in the $\ell_1$ sense with the rate $O(\frac{1}{\tau})$. Beside transportation plan, we upperbound the difference in transport plan between OT and UOT distance by $O(\frac{1}{\tau^2})$ in Section 4.3. All proofs for this section are given in Appendix E.
Algorithm 3 GEM-OT

1: **Input:** $C$, $a$, $b$, $\varepsilon$
2: $\varepsilon' = \varepsilon / 16$
3: $\eta = \frac{2\varepsilon'}{\alpha + \beta} = \varepsilon'/2$, $\gamma = \|C\|_\infty + \eta$
4: $\tau = \frac{\beta \|C\|_\infty n^2}{2\varepsilon'}$
5: $\tilde{X} = \text{GEM-UOT}(C, a, b, \varepsilon', \tau)$
6: $Y = \text{PROJ}(\tilde{X}, a, b)$
7: **Return** $Y$

4.1 Standard OT

When the masses lie in the probability simplex, i.e. $a, b \in \Delta^n := \{x \in \mathbb{R}^n : \|x\|_1 = 1\}$, the standard OT problem [Kantorovich, 1942] can be formulated as:

$$\text{OT}(a, b) = \min_{X \in \Pi(a, b)} \langle C, X \rangle$$

where $\Pi(a, b) = \{X \in \mathbb{R}^{n \times n} : X1_n = a, X^\top 1_n = b\}$.

**Definition 4.1.** For $\varepsilon > 0$, we define $X$ as an $\varepsilon$-approximation transportation plan of OT($a, b$) if it satisfies:

$$\langle C, X \rangle \leq \langle C, X^{\text{OT}} \rangle + \varepsilon,$$

where $X^{\text{OT}} = \arg\min_{X \in \Pi(a, b)} \langle C, X \rangle$ is the optimal transportation plan of the OT problem (20).

4.2 Approximation of OT Transportation Plan

**Theorem 4.2.** For $a, b \in \Delta^n$, there exists some optimal transportation plan $X_f = \arg\min_{X \in \mathbb{R}^{n \times n}} f(X)$ of the problem $\text{UOT}_{KL}(a, b)$ such that:

$$\|X_f 1_n - a\|_1 + \|X_f^\top 1_n - b\|_1 \leq \frac{2n\|C\|_\infty}{\tau}.$$  (22)

The quantity $\|X 1_n - a\|_1 + \|X^\top 1_n - b\|_1$ measures the closeness of $X$ to $\Pi(a, b)$ in $\ell_1$ distance and is also used as the stopping criteria in algorithms solving standard OT [Altschuler et al., 2018]; [Dvurechensky et al., 2018a]. Therefore, Theorem 4.2 characterizes the rate $O\left(\frac{1}{\tau}\right)$ by which the transportation plan of $\text{UOT}_{KL}(a, b)$ converges to the marginal constraints $\Pi(a, b)$ of OT($a, b$). Utilizing this observation, we present the algorithm GEM-OT (Algorithm 3), which solves the $\text{UOT}_{KL}(a, b)$ problem via GEM-UOT with fine tuned $\tau$ and finally projects the UOT solution onto $\Pi(a, b)$ (via Algorithm 4 in Appendix E), to find an $\varepsilon$-approximate solution to the standard OT problem OT($a, b$).

**Theorem 4.3.** Algorithm 3 outputs $Y \in \Pi(a, b)$ such that:

$$\langle C, Y \rangle \leq \langle C, X^{\text{OT}} \rangle + \varepsilon.$$  (23)

In other words, $Y$ is an $\varepsilon$-approximate transportation plan of OT($a, b$). Under assumptions (A1-A2), the complexity of Algorithm 3 is:

$$O\left(\kappa \cdot n^2\log\left(\frac{n\|C\|_\infty}{\varepsilon}\right)\right).$$

**Remark 4.4.** The best known complexities in the literature of OT are respectively $\tilde{O}\left(\frac{n^2}{\varepsilon}\right)$ [Dvurechensky et al., 2018a], $\tilde{O}\left(\frac{n^2}{\varepsilon^2}\right)$ [Dvurechensky et al., 2018a], and $\tilde{O}\left(\frac{n^2}{\varepsilon^{\frac{3}{2}}}\right)$ [Lin et al., 2019c] (where $r$ is a complex constant of the Bregman divergence). We highlight that the complexity achieved in Theorem 4.3 is the first complexity to achieve logarithmic dependence on $\varepsilon^{-1}$ in the literature of OT. Besides, the best known complexity $\tilde{O}\left(\frac{n^2}{\varepsilon^{\frac{3}{2}}}\right)$ in terms of $n$ and $\varepsilon$ also depends on some universal constant.

$^2\tilde{O}(\cdot)$ excludes the logarithmic terms in the complexity.
4.3 Approximation Error between UOT and OT Distances

As discussed in Section 3.5, certain applications only require the computation of either UOT or OT distances oblivious of the transportation plan, which further necessitates the understanding of approximation error between UOT and OT in the sense of purely transport distance. We establish the upper bound on the approximation error between UOT and OT distances in the following Theorem 4.5.

**Theorem 4.5.** For \( a, b \in \Delta^n \), we have:

\[
0 \leq OT(a, b) - UOT_{KL}(a, b) \leq \frac{M}{\tau},
\]

where we define:

\[
M = \log(2)\|C\|_{\infty}^2 \left( n + \frac{3}{\min\{a_{\min}, b_{\min}\}} \right)^2 + 2n\|C\|_{\infty}^2.
\]

5 Experiments

In this section, we provide experiments to demonstrate the favorable performance of our algorithms in practice, and empirically verify our theories on approximation theory. To find the ground truth value, we use the convex programming package **cvxpy** to find the exact UOT plan [Agrawal et al. 2019]. We also experiment with the CIFAR-10 dataset to illustrate the robustness of our algorithms.

5.1 Synthetic Data

We choose \( n = 25, \tau = 55, \eta = 0.01, \alpha = 4, \beta = 5 \). Then \( a_i \)'s are drawn from a uniform distribution and then rescaled to ensure \( \alpha = 4 \) while \( b_i \)'s are drawn from a normal distribution with \( \sigma = 0.1 \). The entries of \( C \) are drawn uniformly from \([10^{-5}, 10^{-4}]\). We vary \( \varepsilon \) from 1 to \( 10^{-4} \) to evaluate the empirical time complexity of GEM-UOT and Sinkhorn algorithm, reported in Figure 1. While there are some fluctuations when \( \varepsilon \approx n^{-2} \), we hypothesize that the number of iterations required for Sinkhorn to achieve an \( \varepsilon \) approximation will outgrow than that of the gradient extrapolation method. A notable behavior of the Sinkhorn algorithm is that the algorithm fluctuates as it approaches the optimal value, and hence converges slightly slower as \( \varepsilon \) gets finer. In addition, we also test the \( \tau \) dependency in Figure 2.

![Figure 1](image1.png)

**Figure 1:** Comparison in number of iterations between GEM-UOT and Sinkhorn on synthetic data for \( \varepsilon = 1 \rightarrow 10^{-3} \).

![Figure 2](image2.png)

**Figure 2:** Scalability of \( \tau \) experiment on synthetic data (left) and CIFAR-10 data (right). Note that Sinkhorn scales linearly while GEM-UOT scales logarithmically in \( \tau \).
5.2 CIFAR-10 Dataset

To validate our algorithms in real settings, we compare the GEM methods with CIFAR-10 dataset [Krizhevsky, 2009] and follow the similar settings in Pham et al. [2020b], Dvurechensky et al. [2018b]. Specifically, the marginals $a, b$ are two flattened images in a pair and the cost matrix $C$ is the matrix of $\ell_1$ distances between pixel locations. We plot the results in Figure 3 which demonstrates GEM methods’ superior performance. In addition, we also illustrate our favorable performance in $\tau$ dependency in Figure 2.

![Figure 3: Primal gap $f - f^*$ of GEM-UOT (RGEM sc), GEM-RUOT (RGEM convex) and Sinkhorn on CIFAR-10.](image)

5.3 Approximation Error between UOT and OT

We investigate the approximability of OT via UOT. In this setting, $\alpha = \beta = 1$. In Appendix F.3 we confirm the $O(\frac{1}{\tau})$ gap of Theorem 4.2. In Figure 4 we measure the gap $OT(a, b) - UOT_{KL}(a, b)$ to empirically verify Theorem 4.5 on both synthetic and real datasets. Additionally, we compare it with the gap of $UOT_{\ell_2}(a, b)$ (using squared $\ell_2$ instead of KL divergence) whose gap was also empirically tested in Blondel et al. [2018].

![Figure 4: Measuring the gap $OT(a, b) - UOT_{KL}(a, b)$ (and additionally for $UOT_{\ell_2}(a, b)$) on synthetic data (left) and CIFAR-10 data (right). Empirically, the gap scales inversely with $\tau$.](image)

6 Conclusions

We have developed a new algorithm for solving the discrete UOT problem using gradient based methods. Our method achieves the complexity of $O(\kappa n^2 \log(\frac{\tau}{\varepsilon}))$, where $\kappa$ is the condition number depending on only the two input measures. This is significantly better than the only known complexity $O(\tau n^2 \log(n) \log(\log(n)))$ of Sinkhorn algorithm [Pham et al., 2020b] to the UOT problem in terms of $\varepsilon$ and $\tau$ dependence. In addition, we are the first to theoretically analyze on the approximability of UOT problem to find an OT solution. Our numerical results show the efficiency of the new method and the tightness of our theoretical bounds. We believe that our analysis framework can be extended to study the convergence results of stochastic gradient methods, widely used in ML applications for handling large-scale datasets. Our results on approximation error can open up directions that use UOT to approximate standard OT.
APPENDIX

A Sinkhorn Algorithm

A.1 Regularity Conditions

In this section, we will show that the assumptions used by the Sinkhorn [Pham et al. 2020b] algorithm are equivalent to our stated regularity conditions (S1)-(S4).

(S4) is clearly stated in their list of regularity conditions. To claim their complexity, in [Corollary 1, Pham et al. 2020b], they further require the condition ‘\( \mathcal{R} = O(\frac{1}{n} \| C \|_\infty) \)’. Following their discussion in [Section 3.1, Pham et al. 2020b], a necessary condition for it to hold is \( \| \log(a) \|_\infty = O(\log(n)) \) (and resp. \( \| \log(b) \|_\infty = O(\log(n)) \)). To see that \( \| \log(a) \|_\infty = O(\log(n)) \) (and resp. \( \| \log(b) \|_\infty = O(\log(n)) \)) is equivalent to (S1)-(S3), we note that if \( a_{min} = 0 \) then \( \| \log(a) \|_\infty = \infty \) (contradiction!), and \( \| \log(a) \|_\infty = \max \{ |\log(a_{min})|, |\log(a_{max})| \} \).

A.2 The dependency of Sinkhorn complexity on \( \log(\alpha + \beta) \)

We note that in its pure form, Sinkhorn’s complexity has \( \log(\alpha + \beta) \), which is excluded in [Pham et al. 2020b] under their regularity condition (S4) that \( \alpha, \beta \) are constants. To see this, note that in the proof of [Corollary 1, Pham et al. 2020b], their quantity \( U \) has the complexity \( O((\alpha + \beta) \log(n)) \) and appears in Sinkhorn’s final complexity as \( \log(U) \).

B Supplementary Lemmas and Theorems

Lemma B.1. For \( x, y \in \mathbb{R} \), we have: \( |\max\{0, x\} - \max\{0, y\}| \leq |x - y| \)

Proof. We have:
\[
\max\{0, x\} = \max\{0, x - y + y\} \leq \max\{0, y\} + \max\{0, x - y\} \leq \max\{0, y\} + |x - y|
\]
\[
\max\{0, x\} - \max\{0, y\} \leq |x - y|
\]
By symmetry, \( \max\{0, y\} - \max\{0, x\} \leq |x - y| \). Therefore, \( |\max\{0, x\} - \max\{0, y\}| \leq |x - y| \).

Lemma B.2. The following identities hold:
\[
\begin{align*}
g_\eta(X^n) + 2\tau \| X^n \|_1 + \eta \| X^n \|_2^2 &= \tau(\alpha + \beta), \\
f(X_f) + 2\tau \| X_f \|_1 &= \tau(\alpha + \beta).
\end{align*}
\]

Proof. The identity (25) follows from [Pham et al. 2020a] (Lemma 4).

To prove (26), we consider the function \( g_\eta(tX^n) \), where \( t \in \mathbb{R}^+ \), we have
\[
g_\eta(tX^n) = \langle C, tX^n \rangle + \eta \| tX^n \|_2^2 + \tau \mathsf{KL}(tX^n \mathbf{1}_n || a) + \tau \mathsf{KL}((tX^n)^\top \mathbf{1}_n || b).
\]

Simple algebraic manipulation gives:
\[
\begin{align*}
\mathsf{KL}(tX^n \mathbf{1}_n || a) &= t \mathsf{KL}(X^n \mathbf{1}_n || a) + (1 - t)\alpha + \| X^n \|_1 t \log(t) \\
\mathsf{KL}((tX^n)^\top \mathbf{1}_n || b) &= t \mathsf{KL}((X^n)^\top \mathbf{1}_n || b) + (1 - t)\beta + \| X^n \|_1 t \log(t).
\end{align*}
\]
We thus obtain that:
\[
g_\eta(tX^n) = tg_\eta(X^n) + \tau(1 - t)(\alpha + \beta) + 2\tau \| X^n \|_1 t \log(t) + \eta(t^2 - t)\| X^n \|_2^2.
\]
Differentiating \( g_\eta(tX^n) \) with respect to \( t \),
\[
\frac{\partial g_\eta(tX^n)}{\partial t} = g_\eta(X^n) - \tau(\alpha + \beta) + 2\tau \| X^n \|_1 (1 + \log(t)) + \eta(2t - 1)\| X^n \|_2^2.
\]
From the above analysis, we can see that \( g_\eta(tX^n) \) is well-defined for all \( t \in \mathbb{R}^+ \) and attains its minimum at \( t = 1 \).

Setting \( \frac{\partial g_\eta(tX^n)}{\partial t} |_{t=1} = 0 \), we obtain the identity (25).
Lemma B.3. We have the following bounds:

\[ \|X^\eta\|_1 \leq \frac{\alpha + \beta}{2}, \]  \hfill (27) 
\[ \|X_f\|_1 \leq \frac{\alpha + \beta}{2}, \]  \hfill (28) 
\[ u_i^* \geq \tau \log \left( \frac{2a_i}{\alpha + \beta} \right) \quad \forall i \in [n], \]  \hfill (29) 
\[ v_j^* \geq \tau \log \left( \frac{2b_j}{\alpha + \beta} \right) \quad \forall j \in [n], \]  \hfill (30) 
\[ \min_{i,j} \left\{ \sum_{k=1}^n X_{ik}, \sum_{k=1}^n X_{kj} \right\} \geq \min\{a_{\min}, b_{\min}\} e^{-\frac{D}{2}}. \]  \hfill (32)

Proof. Noting that \( f(X) \) and \( g_\eta(X) \) are non-negative for \( X \in \mathbb{R}^{n \times n}_+ \), we directly obtain (27) and (28) from Lemma B.2.

Note that \( X_{ij}^\eta \geq 0, \forall i, j = 1, \ldots, n \), and by (27):

\[ \frac{\alpha + \beta}{2} \geq \|X^\eta\|_1 = \sum_{i,j=1}^n X_{ij}^\eta \geq \max_{j=1}^n \sum_{i=1}^n X_{ij}^\eta, \]  \hfill (33)

We obtain in view of Lemma B.2 that:

\[ \log(a_i) - \frac{u_i^*}{\tau} = \log \left( \sum_{j=1}^n X_{ij}^\eta \right) \leq \log \left( \frac{\alpha + \beta}{2} \right), \quad i = 1, \ldots, n, \]  \hfill (34) 
\[ \log(b_j) - \frac{v_j^*}{\tau} = \log \left( \sum_{i=1}^n X_{ij}^\eta \right) \leq \log \left( \frac{\alpha + \beta}{2} \right), \quad j = 1, \ldots, n. \]  \hfill (35)

which are equivalent to

\[ u_i^* \geq \tau \log \left( \frac{2a_i}{\alpha + \beta} \right) \geq \tau \left( \log(a_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right), \quad i = 1, \ldots, n, \]  \hfill (36) 
\[ v_j^* \geq \tau \log \left( \frac{2b_j}{\alpha + \beta} \right) \geq \tau \left( \log(b_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right), \quad j = 1, \ldots, n. \]  \hfill (37)

By (7), we have \( \frac{1}{2\eta}(u_i^* + v_j^* - C_{ij}) \leq X_{ij}^\eta, \forall i, j = 1, \ldots, n \). Hence,

\[ u_i^* \leq 2\eta X_{ij}^\eta + C_{ij} - v_j^* \leq 2\eta \left( \log(b_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right). \]  \hfill (38)
\[ v_j^* \leq 2\eta X_{ij}^\eta + C_{ij} - u_i^* \leq 2\eta \left( \log(a_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right). \]  \hfill (39)

Note that \( X_{ij}^\eta \leq \max_{i,j} |X_{ij}^\eta| = \|X^\eta\|_\infty \leq \|X^\eta\|_1 \leq \frac{\alpha + \beta}{2} \). We have

\[ \tau \left( \log(a_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right) \leq u_i^* \leq \eta(\alpha + \beta) + C_{ij} - \tau \left( \log(b_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right), \] 
\[ \tau \left( \log(b_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right) \leq v_j^* \leq \eta(\alpha + \beta) + C_{ij} - \tau \left( \log(a_{\min}) - \log \left( \frac{\alpha + \beta}{2} \right) \right). \]

Therefore,

\[ \max\{\|u^*\|_\infty, \|v^*\|_\infty\} \leq \|C\|_\infty + \eta(\alpha + \beta) + \tau \log \left( \frac{\alpha + \beta}{2} \right) - \tau \min\{\log(a_{\min}), \log(b_{\min})\} = D. \]
To prove (32), we note that by Lemma 2.2 \( \forall i: \)
\[
\sum_{k=1}^{n} X_{ik}^\eta = a_i \times e^{-\frac{\eta}{\tau}} \geq a_i \times e^{-\frac{\|U\|_\infty}{\tau}} \geq a_i \times e^{-\frac{\epsilon}{\tau}}.
\]
Similarly, \( \forall j: \)
\[
\sum_{k=1}^{n} X_{kj}^\eta \geq b_j \times e^{-\frac{\epsilon}{\tau}}.
\]
Therefore,
\[
\min_{i,j}\{\sum_{k=1}^{n} X_{ik}^\eta, \sum_{k=1}^{n} X_{kj}^\eta\} \geq \min\{a_{\min}, b_{\min}\} e^{-\frac{\epsilon}{\tau}}.
\]

Lemma B.4. For \( 0 < p \leq q \), let \( U_{p,q} = \{X \in \mathbb{R}^{n \times n}_{+} | \forall i, j : \sum_{k=1}^{n} X_{ik} \in [p, q], \sum_{k=1}^{n} X_{kj} \in [p, q]\} \).
Then \( g_\eta(X) \) is \( L_g \)-Lipschitz in \( U_{p,q} \), i.e. \( \forall X, Y \in U_{p,q}: \)
\[
|g_\eta(X) - g_\eta(Y)| \leq L_1 \|X - Y\|_1,
\]
where \( L_1 = \|C\|_\infty + 2\eta q + 2\tau \|g(p)\| + 2\tau \| \log(q)\| + \tau \max_i |\log(a_i)| + \tau \max_i |\log(b_i)|. \)

Proof. Take any \( X, Y \in U_{p,q} \). By Mean Value Theorem, there exists \( c \in (0, 1) \) and \( Z = (1 - c)X + cY \in U_{p,q} \) such that:
\[
|g_\eta(X) - g_\eta(Y)| = |\langle \nabla g_\eta(Z), X - Y \rangle|
\leq \| \nabla g_\eta(Z) \|_\infty \|X - Y\|_1. \quad \text{(by Holder inequality)}
\]
We also have:
\[
\frac{\partial g_\eta}{\partial Z_{ij}} = C_{ij} + 2\eta Z_{ij} + \tau \log(\sum_{k=1}^{n} Z_{ik}) + \tau \log(\sum_{k=1}^{n} Z_{kj}) - \tau \log(a_i) - \tau \log(b_j).
\]
Since \( Z \in U_{p,q} \), we obtain \( \forall i, j: \)
\[
\sum_{k=1}^{n} Z_{ik}, \sum_{k=1}^{n} Z_{kj} \in [p, q],
\]
\[
| \log(\sum_{k=1}^{n} Z_{ik})|, | \log(\sum_{k=1}^{n} Z_{kj})| \leq | \log(p)| + | \log(q)|
\]
From the above observations, we consequently obtain that:
\[
\| \nabla g_\eta(Z) \|_\infty \leq \|C\|_\infty + 2\eta q + 2\tau | \log(p)| + 2\tau | \log(q)| + \tau \max_i |\log(a_i)| + \tau \max_i |\log(b_i)| = L_g,
\]
\[
\therefore |g_\eta(X) - g_\eta(Y)| \leq L_g \|X - Y\|_1.
\]

Lemma B.5. If \( k \geq K_0 \), the output \( X^k \) of Algorithm 1 satisfies:
\[
\|X^k - X^\eta\|_1 \leq \frac{\varepsilon}{2L_1}, \quad (40)
\]
\[
g_\eta(X^k) - g_\eta(X^\eta) \leq \frac{\varepsilon}{2}. \quad (41)
\]
Proof. Because of the projection step on line 10 of Algorithm 1, we have: \[ \sum_{i,j=1}^{n} \left| t_{ij}^k - t_{ij}^* \right| \leq \frac{\sqrt{n^2 + 2n}}{2\eta} \| x^k - x^* \|_1 \]

We now proceed to prove that \( \theta^{n} = 0 \) for \( k = 0, \ldots, k_c \). By Lemma 3.2, \( f_q(x) \) is \( L \)-smooth and \( w_u(x) \) is \( \mu \)-strongly convex on \( \text{conv} \{ x^k, x^0, x^1, \ldots \} \). From the proof of Theorem 2.1 in [Lan and Zhou 2018], we have:

\[
\frac{k\alpha^k(1 - \alpha)}{1 - \alpha^k} \leq 2\alpha^{k/2},
\]

\[
\|x^k - x^*\|_2^2 \leq \frac{4\Delta_{0,\sigma_0}^2 k\alpha^k(1 - \alpha)}{\mu(1 - \alpha^k)} \leq 8\Delta_{0,\sigma_0} \alpha^{k/2}.
\]

Moreover, in view of Lemma 2.2, we have

\[
\|X^k - X^n\|_1 = \frac{1}{2\eta} \sum_{i,j=1}^{n} |t_{ij}^k - t_{ij}^*| 
\leq \frac{1}{2\eta} \|x^k - x^*\|_1 
\leq \frac{\sqrt{n^2 + 2n}}{2\eta} \|x^k - x^*\|_2 
\leq \sqrt{2n} \frac{\Delta_{0,\sigma_0}^{1/2} k^{1/4}}{\eta} 
= \frac{2n\Delta_{0,\sigma_0}^{1/2}}{\eta} \left(1 - \frac{1}{1 + \sqrt{1 + 16L/\mu}}\right)^{k/4} 
\leq \frac{2n\Delta_{0,\sigma_0}^{1/2}}{\eta} \exp\left\{-\frac{k}{4(1 + \sqrt{1 + 16L/\mu})}\right\}.
\]

The condition that \( \frac{2n\Delta_{0,\sigma_0}^{1/2}}{\eta} \exp\left\{-\frac{k}{4(1 + \sqrt{1 + 16L/\mu})}\right\} \leq \min\{\varepsilon/2, \min\{a_{\min}, b_{\min}\} e^{-D/\tau}, \alpha + \beta\} \) is equivalent to:

\[
k \geq 4(1 + \sqrt{1 + 16L/\mu}) \log \left( \frac{4\Delta_{0,\sigma_0}^{1/2}}{\eta} \max \left\{ \frac{L_1}{\varepsilon}, \frac{e^{D/\tau}}{\min\{a_{\min}, b_{\min}\}}, (\alpha + \beta)^{-1} \right\} \right) = K_0.
\]

Therefore,

\[
\|X^k - X^n\|_1 \leq \min\left\{ \frac{\varepsilon}{2L_1}, \frac{\min\{a_{\min}, b_{\min}\} e^{-D/\tau}}{2}, \frac{\alpha + \beta}{2} \right\}.
\]

For \( p = \frac{1}{2} \min\{a_{\min}, b_{\min}\} e^{-D/\tau} \) and \( q = \alpha + \beta \) given in Algorithm 1, we know that \( X^n \in U_{p,q} \) by Lemma B.3.

We now proceed to prove that \( X^k \in U_{p,q} \). Consider any \( i \in [1, n] \):

\[
\sum_{j=1}^{n} x_{ij}^k \geq \sum_{j=1}^{n} x_{ij}^n - \left| \sum_{j=1}^{n} (x_{ij}^k - x_{ij}^n) \right| \geq \sum_{j=1}^{n} x_{ij}^n - \|x^k - x^n\|_1 
\geq a_i e^{-u_i/\tau} - \|x^k - x^n\|_1 \geq a_{\min} e^{-D/\tau} - \|x^k - x^n\|_1 
\geq \frac{a_{\min} e^{-D/\tau}}{2} \geq p,
\]

\[
\sum_{j=1}^{n} x_{ij}^k \leq \sum_{j=1}^{n} x_{ij}^n + \left| \sum_{j=1}^{n} (x_{ij}^k - x_{ij}^n) \right| \leq \|x^n\|_1 + \|x^k - x^n\|_1 
\leq \alpha + \beta = \alpha + \beta.
\]

Similarly, for any \( j \in [1, n] \), we have:

\[
p \leq \sum_{i=1}^{n} x_{ij}^k \leq q,
\]
which implies $X^k \in U_{p,q}$.

Now, using Lemma B.4 with $X^k, X^\eta \in U_{p,q}$, we have:

$$g_\eta(X^k) - g_\eta(X^\eta) \leq L_1 \|X^k - X^\eta\|_1 \leq \frac{\varepsilon}{2}.$$ 

\[\square\]

C Duality

C.1 Supplementary Lemmas

**Proposition C.1** (Rockafellar [1967]). Let $(E, E^\ast)$ and $(F, F^\ast)$ be two couples of topologically paired spaces. Let $A : E \to F$ be a continuous linear operator and $A^\ast : F^\ast \to E^\ast$ its adjoint. Let $f$ and $g$ be lower semicontinuous and proper convex functions defined on $E$ and $F$ respectively. If there exists $x \in \text{dom} f$ such that $g$ is continuous at $Ax$, then

$$\sup_{x \in E} - f(-x) - g(Ax) = \inf_{y^* \in F^\ast} f^\ast(A^\ast y^*) + g^\ast(y^*)$$

Moreover, if there exists a maximizer $x \in E$ then there exists $y^* \in F^\ast$ satisfying $Ax \in \partial g^\ast(y^*)$ and $A^\ast y^* \in \partial f(-x)$.

**Lemma C.2.** For a fixed $\lambda \in \mathbb{R}_+^{n \times n}$,

$$g(\lambda, X) = g_\eta(X) - \langle \lambda, X \rangle = \langle C - \lambda, X \rangle + \eta \|X\|_2^2 + \tau \text{KL}(X I_n; a) + \tau \text{KL}(X^\top I_n; b),$$

the following duality holds:

$$\min_{X \in \mathbb{R}^{n \times n}} g(\lambda, X) = \max_{u, v \in \mathbb{R}^n} s(\lambda, u, v),$$

where:

$$s(\lambda, u, v) = -\frac{1}{4\eta} \sum_{i,j=1}^n (u_i + v_j - C_{ij} + \lambda_{ij})^2 - \tau \left( e^{-u/\tau}, a \right) - \tau \left( e^{-v/\tau}, b \right) + a^T I_n + b^T I_n.$$ 

Furthermore, for $X^\lambda = \arg\min_{X \in \mathbb{R}^{n \times n}} g(\lambda, X)$ and $(u^\lambda, v^\lambda) = \arg\max_{u, v \in \mathbb{R}^n} s(\lambda, u, v)$, we have $\forall i, j$:

$$X^\lambda_{ij} = \frac{1}{2\eta} (u^\lambda_i + v^\lambda_j - C_{ij} + \lambda_{ij}),$$

$$-\frac{u^\lambda_i}{\tau} + \log(a_i) = \log(\sum_{k=1}^n X^\lambda_{ik}),$$

$$-\frac{v^\lambda_j}{\tau} + \log(b_j) = \log(\sum_{k=1}^n X^\lambda_{kj}).$$

**Proof of Lemma C.2** Consider the functions:

$$G(X) = \frac{1}{\eta} \langle C - \lambda, X \rangle + \|X\|_2^2;$$

$$F_1(y) = \tau \text{KL}(y; a),$$

$$f_\eta(y) = \tau \text{KL}(y; b).$$

And their convex conjugates are as follows:

$$G^*(p) := \sup_{X \in \mathbb{R}^{n \times n}} \{ \langle p, X \rangle - G(X) \} = \frac{1}{4} \sum_{i,j} (p_{ij} - \frac{1}{\eta} C_{ij} + \frac{1}{\eta} \lambda_{ij})^2,$$

$$F_1^*(u) := \sup_{y \in \mathbb{R}^n} \{ \langle u, y \rangle - F_1(y) \} = \tau \left( e^{u/\tau}, a \right) - a^T I_n,$$

$$f_\eta^*(v) := \sup_{y \in \mathbb{R}^n} \{ \langle v, y \rangle - f_\eta(y) \} = \tau \left( e^{v/\tau}, a \right) - b^T I_n.$$
Consider the linear operator $A : \mathbb{R}^{2n} \to \mathbb{R}^{n \times n}$ that maps $A(u, v) = X$ with

$$X_{ij} = u_i + v_j,$$

then $A$ is continuous and its adjoint is $A^* : \mathbb{R}^{n \times n} \to \mathbb{R}^{2n}$ is $A(X) = (X_1, X^T 1_n)$.

Now note that the problem $\max_{u, v \in \mathbb{R}^n} s(\lambda, u, v)$ can be rewritten as:

$$\max_{u, v \in \mathbb{R}^n} -F^*_\eta(-u) - f^*_\eta(-v) - \eta G^*\left(\frac{A(u, v)}{\eta}\right).$$

By Proposition C.1, we obtain its Fenchel-Rockafellar dual problem as:

$$\inf_{X \in \mathbb{R}^{n \times n}} F_1(X_1, n) + f_\eta(X^T 1_n) + G(X),$$

which is the optimization problem $\min_{X \in \mathbb{R}^{n \times n}} g(\lambda, X)$. Furthermore, by Proposition C.1 we can conclude:

$$X^\lambda \in \partial G^*\left(\frac{A(u^\lambda, v^\lambda)}{\eta}\right) \implies X^\lambda = \frac{1}{2\eta}(u^\lambda_i + v^\lambda_j - C_{ij} + \lambda_{ij}),$$

$$-u^\lambda \in \partial F_1(X^\lambda 1_n) \implies \forall i : -u^\lambda_i \geq \lambda^\tau i + \log(a_i) = \log\left(\sum_{k=1}^n X^\lambda_{ik}\right),$$

$$-v^\lambda \in \partial f_\eta((X^\lambda)^T 1_n) \implies \forall j : -v^\lambda_j \geq \lambda^\tau j + \log(b_j) = \log\left(\sum_{k=1}^n X^\lambda_{kj}\right).$$

**C.2 Proof of Lemma 2.2**

For $\lambda \in \mathbb{R}^{n \times n}_+$, we consider the Lagrangian function of $g_\eta(X)$:

$$g(\lambda, X) = g_\eta(X) - \langle \lambda, X \rangle$$

(50)

The Lagrange dual problem for (5) is:

$$\min_{X \in \mathbb{R}^{n \times n}_+} g_\eta(X) = \max_{\lambda \in \mathbb{R}^{n \times n}_+} \min_{X \in \mathbb{R}^{n \times n}} g(\lambda, X),$$

$$= \max_{\lambda} \max_{u, v \in \mathbb{R}^n} s(\lambda, u, v),$$

(51)

where the second equality follows from Lemma C.2 with $s(\lambda, u, v)$ defined as:

$$s(\lambda, u, v) = -\frac{1}{4\eta} \sum_{i,j=1}^n (u_i + v_j - C_{ij} + \lambda_{ij})^2 - \tau \left(e^{-u_i/\tau}, a\right) - \tau \left(e^{-v_j/\tau}, b\right) + a^T 1_n + b^T 1_n.$$

Observe that the problem $\max_{X \in \mathbb{R}^{n \times n}_+} s(\lambda, u, v)$ can be decomposed into solving the subproblems

$$\max_{\lambda, \eta \geq 0} \left\{-\frac{1}{4\eta} (u_i + v_j - C_{ij} + \lambda_{ij})^2\right\} = -\frac{1}{4\eta} \max\{0, u_i + v_j - C_{ij}\}^2,$$

with equality at:

$$\lambda^*_{ij} = \max\{0, -(u_i + v_j - C_{ij})\}.$$  

(52)

We can then rewrite (51) as:

$$\min_{X \in \mathbb{R}^{n \times n}_+} g_\eta(X) = \max_{u, v \in \mathbb{R}^n} s(\lambda^*, u, v)$$

$$= \max_{(u, v) \in \mathcal{X}} \left\{-\frac{1}{4\eta} \sum_{i,j=1}^n t_{ij}^2 - \tau \left(e^{-u_i/\tau}, a\right) - \tau \left(e^{-v_j/\tau}, b\right) + a^T 1_n + b^T 1_n\right\}.$$  

(53)
We then obtain:

We have just rewritten our original problem (5) as the optimization problem (6) over the convex set \( \mathcal{X} \).

For the optimal solution \( x^* = (u^*, v^*, t^*) \) to (6), we must have:

Moreover, by Lemma C.2 and (52),

Here we replace \( \max \{0, u_i + v_j - C_{ij}\} \) with the dummy variable \( t_{ij} \) constrained by \( t_{ij} \geq 0 \) and \( t_{ij} \geq u_i + v_j - C_{ij} \).

Corollary 3.6 follows from the formulation in Lemma C.2.

C.3 Proof of Corollary 3.6

Now let us consider any \( x \in \mathcal{V} \), \( X = (u, v, t) \) and thereby \( \nabla f_\eta(x) \succcurlyeq 0 \) on \( \mathcal{V} \).

Now let us consider any \( x = (u, v, t), x' = (u', v', t') \) in \( \mathcal{V} \). By Mean Value Theorem, \( \exists c_i \in \text{range}(u_i, u'_i) \), such that \( e^{-u_i/\tau} - e^{-u'_i/\tau} = -\frac{1}{\tau} e^{-c_i/\tau} (u_i - u'_i) \) and \( \exists d_j \in \text{range}(v_j, v'_j) \), such that \( e^{-v_j/\tau} - e^{-v'_j/\tau} = -\frac{1}{\tau} e^{-d_j/\tau} (v_j - v'_j) \).

We then obtain:

\[
\| \nabla f_\eta(x) - \nabla f_\eta(x') \|^2 = \sum_{i=1}^{n} \left[ \frac{a_i}{\tau} e^{-c_i/\tau} - \frac{\min \{a_{\min}, b_{\min}\} e^{-D/\tau}}{\tau} (u_i - u'_i) \right]^2 + \sum_{j=1}^{n} \left[ \frac{b_j}{\tau} e^{-d_j/\tau} - \frac{\min \{a_{\min}, b_{\min}\} e^{-D/\tau}}{\tau} (v_j - v'_j) \right]^2
\]
We have:

\[ \sum_{i=1}^{n} \left( \frac{a_i}{\tau} e^{-c_i / \tau} + \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} \right) (u_i - u'_i)^2 \]

\[ + \sum_{j=1}^{n} \left( \frac{b_j}{\tau} e^{-d_j / \tau} + \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} \right) (v_j - v'_j)^2 \]

\[ \leq \sum_{i=1}^{n} \left( \frac{a_i}{\tau} \frac{\alpha + \beta}{2a_i} + \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} \right) (u_i - u'_i)^2 \]

\[ + \sum_{j=1}^{n} \left( \frac{b_j}{\tau} \frac{\alpha + \beta}{2b_j} + \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} \right) (v_j - v'_j)^2 \]

(since \( c_i \in \text{range}(u_i, u'_i), c_i \geq \tau \log\left( \frac{2a_i}{\alpha + \beta} \right) \). Similarly, \( d_j \geq \tau \log\left( \frac{2b_j}{\alpha + \beta} \right) \))

\[ \|\nabla f_\eta(x) - \nabla f_\eta(x')\| \leq \left( \frac{\alpha + \beta}{2\tau} + \frac{\min\{a_{\min}, b_{\min}\} e^{-D / \tau}}{\tau} \right) \|x - x'\|_2, \]

which implies that \( f_\eta(x) \) is \( L \)-smooth with \( L = \frac{\alpha + \beta}{2\tau} + \frac{\min\{a_{\min}, b_{\min}\} e^{-D / \tau}}{\tau} \).

Now let us consider \( w_\eta(x) \). The gradient \( w_\eta(x) \) can be computed as:

\[ \frac{\partial w}{\partial u_i} = \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} u_i, \]

\[ \frac{\partial w}{\partial v_j} = \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} v_j, \]

\[ \frac{\partial w}{\partial t_{ij}} = \frac{1}{2\eta} t_{ij}, \]

For any \( x = (u, v, t), x' = (u', v', t') \in V_D \), we have:

\[ \langle \nabla w_\eta(x) - \nabla w_\eta(x'), x - x' \rangle = \sum_{i=1}^{n} \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} (u_i - u'_i)^2 + \sum_{j=1}^{n} \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau} (v_j - v'_j)^2 \]

\[ + \sum_{i,j=1}^{n} \frac{1}{2\eta} (t_{ij} - t'_{ij})^2 \]

\[ \geq \min \left\{ \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau}, \frac{1}{2\eta} \right\} \|x - x'\|_2^2 \]

Therefore, \( w_\eta(x) \) is \( \mu \)-strongly convex on \( V_D \) with \( \mu = \min \left\{ \frac{\min\{a_{\min}, b_{\min}\}}{\tau} e^{-D / \tau}, \frac{1}{2\eta} \right\} \).

**D.2 Proof of Corollary 3.3**

This is directly implied from Lemma B.3

**D.3 Proof of Theorem 3.4**

We have:

\[ f(X^k) - f(X_f) = g_\eta(X^k) - \eta\|X^k\|_2^2 - g_\eta(X_f) + \eta\|X_f\|_2^2 \]

\[ \leq (g_\eta(X^k) - g_\eta(X^y)) + \eta(\|X_f\|_2^2 - \|X^k\|_2^2). \]

From Lemma B.3 we obtain \( \|X_f\|_2 \leq \|X_f\|_2 \leq \frac{(\alpha + \beta)^2}{4} = R \). Noting that \( \eta = \frac{\varepsilon}{2R} \), we have:

\[ \eta(\|X_f\|_2^2 - \|X^k\|_2^2) \leq \frac{\varepsilon}{2}. \]
together with Lemma [B.5] where

\[ g_\eta(X^k) - g_\eta(X^\eta) \leq \frac{\varepsilon}{2}, \]

we have

\[ f(X^k) - f(X_f) \leq \varepsilon. \]

### D.4 Proof of Corollary 3.5

Under assumptions (A1-A3), we have:

\[ D = \|C\|_\infty + \frac{\varepsilon}{\alpha + \beta} + \tau \log \left( \frac{\alpha + \beta}{2} \right) - \tau \min \{ \log(a_{min}), \log(b_{min}) \} \]

\[ \frac{D}{\tau} \leq \|C\|_\infty + \frac{\varepsilon}{\tau(\alpha + \beta)} + \log \left( \frac{\alpha + \beta}{2} \right) - \min \{ \log(a_{min}), \log(b_{min}) \} \]

\[ = O(1) + \log \left( \frac{\alpha + \beta}{2} \right) - \min \{ \log(a_{min}), \log(b_{min}) \} \]

By Lemma [3.2] \( \eta = \frac{\varepsilon}{2\tau} = \frac{2\varepsilon}{(\alpha + \beta)^2} \) (in Algorithm 1) and (A1-A2), we obtain that:

\[ L = O \left( \frac{\alpha + \beta}{2\tau} \right) \]

\[ \mu = \Omega \left( \min \left\{ \frac{(\alpha + \beta)^2}{4\varepsilon}, \frac{1}{\tau} \min \{ a_{min}, b_{min} \} e^{-\log \left( \frac{\alpha + \beta}{2} \right) + \min \{ \log(a_{min}), \log(b_{min}) \}} \right\} \right) \]

\[ = \Omega \left( \min \left\{ \frac{(\alpha + \beta)^2}{4\varepsilon}, \frac{1}{\tau} \min \{ a_{min}, b_{min} \}^2 \right\} \right) \]

\[ \geq \Omega \left( \min \left\{ \frac{n^2(2 \min \{ a_{min}, b_{min} \})^2}{4\varepsilon}, \frac{2 \min \{ a_{min}, b_{min} \}^2}{\tau(\alpha + \beta)} \right\} \right) \]

\[ = \Omega \left( \frac{\min \{ a_{min}, b_{min} \}^2}{\tau(\alpha + \beta)} \right) \text{ (since } \tau = \Omega \left( \frac{1}{\alpha + \beta} \right) \text{) } \]

\[ \sqrt{\frac{L}{\mu}} = O \left( \frac{(\alpha + \beta)}{\min \{ a_{min}, b_{min} \}} \right) = O((\alpha + \beta)\kappa). \]

For convenience, we recall that

\[ P(x_0, x) := \frac{w_\eta(x) - \left[ w_\eta(x_0) + \langle \nabla w_\eta(x_0), x - x_0 \rangle \right]}{\mu} \]

From the initialization \( x^0 = (u^0, v^0, t^0) = 0, \eta = \frac{\varepsilon}{2\tau} = \frac{2\varepsilon}{(\alpha + \beta)^2} \) (in Algorithm 1) and Lemma [B.3], we have:

\[ \frac{D}{\tau} \leq O(1) + \log \left( \frac{\alpha + \beta}{2} \right) - \min \{ \log(a_{min}), \log(b_{min}) \} \leq O(\log(n(\alpha + \beta))) \]

\[ \sigma_0^2 = \| \nabla f_\eta(x^0) \|_2^2 = \| a \|_2^2 + \| b \|_2^2 \leq (\alpha + \beta)^2 \]

\[ h_\eta(x^0) = \tau(\alpha + \beta) \geq h_\eta(x^*) \geq 0 \]

\[ w_\eta(x^0) = 0 \]

\[ 0 \leq w_\eta(x^*) \leq \frac{\min \{ a_{min}, b_{min} \}}{2\tau} \left( n\| u^* \|_2^2 + n\| v^* \|_2^2 \right) + \frac{1}{4\eta} \| X^\eta \|_2^2 \]

\[ \leq O(\min \{ a_{min}, b_{min} \} \tau n \log(n)^2 + (\alpha + \beta)^4/\varepsilon) \]

\[ \nabla w_\eta(x^0) = 0 \]

\[ \Delta_{0, \sigma_0} = \mu P(x^0, x^*) + h_\eta(x^0) - h_\eta(x^*) + \frac{\sigma_0^2}{\mu} \]
We then obtain the asymptotic complexity under (A1-A3) for the following term inside $K_0$:

$$\log \left( \frac{4n^2 \sqrt{\alpha/\eta}}{\eta} \max \left\{ \frac{L_1}{\varepsilon}, \frac{e^{\eta/\tau}}{\min \{a_{\min}, b_{\min}\}}, \frac{1}{\alpha+\beta} \right\} \right) = O \left( \log \left( \frac{\tau \cdot n(\alpha + \beta)}{\varepsilon} \right) \right)$$

Combining with the bounds of $L$ and $\mu$ above, we obtain that $K_0 = O \left( (\alpha + \beta) \kappa \log \left( \frac{\tau \cdot n(\alpha + \beta)}{\varepsilon} \right) \right)$ under the assumptions (A1-A3).

**Observation 1.** Each iteration of Algorithm 1 takes $O(n^2)$.

**Proof.** Lines 9 and 11 are $O(1)$ point-wise update steps of vectors with $n^2 + 2n$ entries, thereby incurring $O(n^2)$ complexity. The proximal mapping on line 10 is equivalent to solving a sparse separable quadratic program of size $O(n^2)$ with sparse linear constraints and thus can be solved efficiently (same structure as standard $\ell_2$ projection operator). The computation of the gradient on line 12 takes $O(n^2)$ since the partial derivative $\partial f_n$ with respect to any of its $n^2 + 2n$ variables takes $O(1)$ operations.

Since by Theorem 3.4 it takes $K_0$ iterations for Algorithm 1 to achieve the $\varepsilon$-approximation, each of which costs $O(n^2)$ per iteration by Observation 1, the total complexity of Algorithm 1 is:

$$O \left( (\alpha + \beta) \kappa \cdot n^2 \log \left( \frac{\tau \cdot n(\alpha + \beta)}{\varepsilon} \right) \right).$$

**D.5 Proof of Lemma 3.7**

The gradient of $h_a(x = (u, v))$ can be computed as:

$$\frac{\partial h_a}{\partial u_i} = -a_i e^{-u_i/\tau} + \sum_{j=1}^n \frac{\max\{0, u_i + v_j - C_{ij}\}}{2\eta},$$

$$\frac{\partial h_a}{\partial v_j} = -b_j e^{-v_j/\tau} + \sum_{i=1}^n \frac{\max\{0, u_i + v_j - C_{ij}\}}{2\eta}.$$

Let us consider any $x = (u, v), x' = (u', v') \in V_a$. By Mean Value Theorem, $\exists c_i \in \text{range}(u_i, u_i')$, such that $e^{-u_i/\tau} - e^{-u_i'/\tau} = \frac{1}{\tau} e^{-c_i/\tau}(u_i - u_i')$ and $\exists d_j \in \text{range}(v_j, v_j')$, such that $e^{-v_j/\tau} - e^{-v_j'/\tau} = \frac{1}{\tau} e^{-d_j/\tau}(v_j - v_j')$.

We then obtain:

$$\|\nabla h_a(x) - \nabla h_a(x')\|^2 \leq \sum_{i=1}^n \left[ \frac{a_i}{\tau} e^{-c_i/\tau}(u_i - u_i') \right]^2 + \sum_{j=1}^n \left[ \frac{b_j}{\tau} e^{-d_j/\tau}(v_j - v_j') \right]^2$$

$$+ \sum_{i=1}^n \frac{1}{2\eta^2} \left\{ \sum_{j=1}^n (u_i - u_i' + v_j - v_j') \right\}^2 + \sum_{j=1}^n \frac{1}{2\eta^2} \left\{ \sum_{i=1}^n (u_i - u_i' + v_j - v_j') \right\}^2$$

(By Lemma B.1 and simple algebra $\langle x + y \rangle^2 \leq 2(x^2 + y^2)$)

$$\leq \sum_{i=1}^n \left( \frac{a_i}{\tau} e^{-c_i/\tau} \right)^2 (u_i - u_i')^2 + \sum_{j=1}^n \left( \frac{b_j}{\tau} e^{-d_j/\tau} \right)^2 (v_j - v_j')^2$$

$$+ \sum_{i=1}^n \frac{1}{2\eta^2} \left\{ \sum_{j=1}^n (u_i - u_i')^2 + 2(v_j - v_j')^2 \right\} + \sum_{j=1}^n \frac{1}{2\eta^2} \left\{ \sum_{i=1}^n (u_i - u_i')^2 + 2(v_j - v_j')^2 \right\}$$

$$\leq \sum_{i=1}^n \left( \frac{a_i (\alpha + \beta)}{\tau} \right)^2 (u_i - u_i')^2 + \sum_{j=1}^n \left( \frac{b_j (\alpha + \beta)}{2\eta} \right)^2 (v_j - v_j')^2$$

$$+ \frac{2n}{\eta^2} \|x - x'\|^2.$$
The RHS of (56) is less than \( \varepsilon \). This implies that \( h_a(x) \) is \( L \)-smooth with \( L_a = \frac{\alpha + \beta}{\tau} + \frac{2\sqrt{\eta}}{\eta} \).

**D.6 Proof of theorem 3.8**

Let \( x^* \) be an optimal solution to (19) and thus (15).

Since \( \|x^*\|_\infty \leq D \) by Lemma B.3, we obtain that:

\[
\bar{P}(x^0, x^*) = \frac{1}{2} \|x^*\|^2_2 \leq \frac{1}{2} nD^2
\]

From corollary 3.6 in [Lan and Zhou 2018], we obtain that:

\[
0 \leq h_a(x^k) - h_a(x^*) = F_a(x^k) - F_a(x^*) = \text{UOT}_{KL}^\eta(a, b) - \text{UOT}_{KL}^\eta(a, b) \leq \frac{12L_a}{k(k+1)} \bar{P}(x^0, x^*) \leq \frac{6L_a nD^2}{k(k+1)} \tag{54}
\]

Using \( \|X_f\|_1 \leq \frac{\alpha + \beta}{2} \) (Lemma B.3) and \( \eta = \frac{\varepsilon}{2R} = \frac{2\varepsilon}{(\alpha + \beta)^2} \) set by the Algorithm 2, we have:

\[
\text{UOT}_{KL}^\eta(a, b) - \text{UOT}_{KL}^\eta(a, b) = g_\eta(X'^n) - f(X_f) = (f(X'^n) - f(X_f)) + \eta \|X'^n\|^2_2 \geq 0
\]

\[
\text{UOT}_{KL}^\eta(a, b) - \text{UOT}_{KL}^\eta(a, b) = g_\eta(X'^n) - g_\eta(X_f) + \eta \|X_f\|^2_2 \leq \eta \|X_f\|^2_2 \leq \eta \frac{(\alpha + \beta)^2}{4} = \frac{\varepsilon}{2}
\]

\[
\therefore 0 \leq \text{UOT}_{KL}^\eta(a, b) - \text{UOT}_{KL}^\eta(a, b) \leq \frac{\varepsilon}{2} \tag{55}
\]

Combining (54), (55), we obtain:

\[
|F_a(x^k) - \text{UOT}_{KL}^\eta(a, b)| \leq \frac{\varepsilon}{2} + \frac{6L_a nD^2}{k(k+1)} \tag{56}
\]

The RHS of (56) is less than \( \varepsilon \) if:

\[
k \geq \sqrt{\frac{12L_a nD^2}{\varepsilon}} = K_a
\]

Under the assumptions (A1-A3) and and \( \eta = \frac{\varepsilon}{2R} = \frac{2\varepsilon}{(\alpha + \beta)^2} \), we have:

\[
D = O(\tau \log(n(\alpha + \beta)))
\]

\[
L_a = O\left(\frac{(\alpha + \beta)^2 \sqrt{n}}{\varepsilon}\right)
\]

\[
\therefore K_a = O\left(\frac{(\alpha + \beta) \cdot \tau n^{0.75}}{\varepsilon} \cdot \log(n(\alpha + \beta))\right)
\]

Since it costs \( O(n^2) \) per iteration to compute the gradient \( h_a(x) \), the total complexity of Algorithm 2 is:

\[
O\left(\frac{(\alpha + \beta) \cdot \tau n^{2.75}}{\varepsilon} \cdot \log(n(\alpha + \beta))\right)
\]
E  Approximation of OT via UOT

E.1 Framework to solve OT via UOT

For any $\eta > 0$, consider the optimal solution $x^* = (u^*, v^*)$ to (15). Recall from Lemma 3.6 that $\forall i, j \in [n]$:

$$X^\eta_{ij} = \frac{1}{2\eta} \max\{0, u^*_i + v^*_j - C_{ij}\},$$  

(57)

$$-\frac{u^*_i}{\tau} + \log(a_i) = \log(\sum_{k=1}^{n} X^\eta_{ik}),$$  

(58)

$$-\frac{v^*_j}{\tau} + \log(b_j) = \log(\sum_{k=1}^{n} X^\eta_{kj}),$$  

(59)

Note that since $a, b \in \Delta^n$, we have $\alpha = \beta = 1$. By Lemma B.3, we obtain:

$$\|X^\eta\|_1 \leq 1.$$  

(60)

The following Lemmas are useful for the proofs of Theorem 4.2 and 4.3.

**Lemma E.1.** The followings hold $\forall i, j \in [n]$:

$$u^*_i \leq \|C\|_\infty + 2\eta,$$  

(61)

$$v^*_j \leq \|C\|_\infty + 2\eta.$$  

(62)

**Proof.** Firstly, we show that there must exist some $l \in [n]$ such that $v_l \geq 0$. Assume for the sake of contradiction that $\forall j \in [n] : v_j < 0$. Then from (59), we obtain $\forall j \in [n]$:

$$\log(\sum_{k=1}^{n} X^\eta_{kj}) > \log(b_j)$$

$$(X^\eta \mathbf{1}_n)_j = \sum_{k=1}^{n} X^\eta_{kj} > b_j$$

$$.\|X^\eta\|_1 = \sum_{j=1}^{n} (X^\eta \mathbf{1}_n)_j > \sum_{j=1}^{n} b_j = 1,$$

where the last line contradicts (60).

Now, from (60), (58) and the fact that $v_l \geq 0$, we have $\forall i \in [n]$:

$$1 \geq X^\eta_{il} = \frac{1}{2\eta} \max\{0, u^*_i + v^*_l - C_{il}\} \geq \frac{1}{2\eta} (u^*_i + 0 - \|C\|_\infty)$$

$$.u^*_i \leq \|C\|_\infty + 2\eta$$

Similarly, we can prove that $\forall j \in [n] : v^*_j \leq \|C\|_\infty + 2\eta.$

**Lemma E.2.** Define $\gamma = \|C\|_\infty + 2\eta$. Then the followings hold $\forall i, j \in [n]$:

$$a_i e^{-\gamma/\tau} \leq \sum_{j=1}^{n} X^\eta_{ij} = (X^\eta \mathbf{1}_n)_i \leq 1 - e^{-\gamma/\tau} (1 - a_i)$$  

(63)

$$b_j e^{-\gamma/\tau} \leq \sum_{i=1}^{n} X^\eta_{ij} = (X^\eta \mathbf{1}_n)_j \leq 1 - e^{-\gamma/\tau} (1 - b_j)$$  

(64)

**Proof.** From (58), we have $\forall i \in [n]$:

$$(X^\eta \mathbf{1}_n)_i = a_i e^{-u^*_i/\tau} \geq a_i e^{-\gamma/\tau}$$  

(65)
Algorithm 4 PROJ

1: [Algorithm 2, Altschuler et al. [2017]]
2: \textbf{Input: } \(X \in \mathbb{R}^{n \times n}, a \in \mathbb{R}^n, b \in \mathbb{R}^n\)
3: \(P \leftarrow \text{diag}(x)\) with \(x_i = \frac{a_i}{(X_1^1)_{i}}\) \& 1
4: \(X' \leftarrow PX\)
5: \(Q \leftarrow \text{diag}(y)\) with \(y_j = \frac{b_j}{(X'')_1} \& 1\)
6: \(X'' \leftarrow X'Q\)
7: \(err_r \leftarrow a - X''1_n, err_c \leftarrow b - X''^\top 1_n\)
8: \textbf{Return} \(Y \leftarrow X'' + \frac{err_rerr_c^\top}{\|err_r\|_1}\)

For the upper bound, we have \(\forall i \in [n]:\)
\[
(X''1_n)_i = \|X''\|_1 - \sum_{k \neq i} (X''1_n)_k \\
\leq 1 - \sum_{k \neq i} a_k e^{-\gamma/\tau} \\
= 1 - e^{-\gamma/\tau}(1 - a_i)
\]
We have just proved 63. Similarly, 64 is obtained.

\textbf{Lemma E.3.} Define \(\gamma = \|C\|_{\infty} + 2\eta\). Then the followings hold:
\[
\|X''1_n - a\|_1 \leq \frac{n\gamma}{\tau} \tag{66}
\]
\[
\|X''^\top 1_n - b\|_1 \leq \frac{n\gamma}{\tau} \tag{67}
\]

\textbf{Proof.} From 63, we have:
\[
-a_i (1 - e^{-\gamma/\tau}) \leq (X''1_n)_i - a_i \leq (1 - a_i)(1 - e^{-\gamma/\tau}) \\
\|X''1_n - a_i\| \leq \max\{a_i, 1 - a_i\}(1 - e^{-\gamma/\tau}) \\
\leq 1 - e^{-\gamma/\tau} \\
\therefore |(X''1_n)_i - a_i| \leq \gamma/\tau
\]
where for the last line, we use the fact that \(e^x \geq x + 1\).
We now conclude that:
\[
\|X''1_n - a\|_1 = \sum_{i=1}^{n} |(X''1_n)_i - a_i| \leq \frac{n\gamma}{\tau}
\]
Similarly,
\[
\|X''^\top 1_n - b\|_1 \leq \frac{n\gamma}{\tau}
\]

\textbf{Lemma E.4} (Lemma 7, Altschuler et al. [2017]). For the inputs \(X \in \mathbb{R}^{n \times n}, a \in \mathbb{R}^n \text{ and } b \in \mathbb{R}^n\), Algorithm 4 takes \(O(n^2)\) time to output a matrix \(Y \in \Pi(a, b)\) satisfying:
\[
\|Y - X\|_1 \leq 2\|X1_n - a\|_1 + \|X^\top 1_n - b\|_1 \tag{68}
\]
E.1.1 Proof of Theorem 4.2

From Lemma B.3 we know that \( X^n \) is bounded by \( \| X^n \|_1 \leq \frac{\alpha + \beta}{2} \). By the Bolzano–Weierstrass Theorem, there exists a convergent subsequence \( \{ X^n \}^\infty_{n=1} \) where \( \lim_{n \to \infty} \eta_n = 0 \). Define the limit of this subsequence as:

\[
X_f = \lim_{\eta_n \to 0} X^n.
\]

We first prove that \( X_f \) is a transportation plan of the \( \text{UOT}_{\text{KL}}(a, b) \) problem. Note that:

\[
\lim_{\eta_n \to 0} \text{UOT}_{KL}^n(a, b) = \text{UOT}_{KL}(a, b)
\]

By the boundedness \( \| X^n \|_1 \leq \frac{\alpha + \beta}{2} \) (Lemma B.3), we have \( \lim_{\eta_n \to 0} \eta_n \| X^n \|_2 = 0 \) and thus obtain that:

\[
\lim_{\eta_n \to 0} \text{UOT}_{KL}^n(a, b) = \lim_{\eta_n \to 0} g_{\eta_n}(X^n) = \lim_{\eta_n \to 0} [f(X^n) + \eta_n \| X^n \|_2^2] = \lim_{\eta_n \to 0} f(X^n) = f(X_f)
\]

where the last line is by the continuity of \( f \). We thus obtain that:

\[
f(X_f) = \text{UOT}_{KL}(a, b),
\]

So \( X_f \) is a transportation plan of the \( \text{UOT}_{KL}(a, b) \) problem by definition. Now by Lemma E.3 we have:

\[
\| X^n a_n - a \|_1 + \| X^n b_n - b \|_1 \leq \frac{2n(\| C \|_\infty + 2\eta_n)}{\tau}
\]

Taking \( \eta_n \to 0 \), we obtain the desired bound (22):

\[
\| X_f a_n - a \|_1 + \| X_f b_n - b \|_1 \leq \frac{2n\| C \|_\infty}{\tau}
\]

E.1.2 Proof of Theorem 4.3

We have:

\[
\langle C, Y \rangle - \langle C, X^{OT} \rangle = \langle C, X^n - X^\gamma \rangle + \langle C, X^n - X \rangle + \langle C, X - X^n - X^{OT} \rangle
\]

\[
\leq \| C \|_\infty \| X^n - X \|_1 + \| C \|_\infty \| Y - X \|_1 + \langle C, X - X^{OT} \rangle
\]

(69)

We proceed to upper-bound each of the three terms in the RHS of (69). We note that \( \bar{X} \) is the output of GEM-UOT fed with input error \( \varepsilon/16 \), and GEM-UOT (in the context of Algorithm 3) thus set \( \eta = \frac{2\varepsilon/16}{(\alpha + \beta)^2} = \frac{\varepsilon}{32} \). The first term can be directly bounded by Lemma B.3 as:

\[
\| C \|_\infty \| X^n - X \|_1 \leq \| C \|_\infty \frac{\varepsilon/16}{2L_1} \leq \frac{\varepsilon}{8}
\]

(70)

where \( L_1 \geq \| C \|_\infty \) holds directly from the definition of \( L_1 \).

For the second term, we obtain from Lemma E.4 that:

\[
\| C \|_\infty \| Y - \bar{X} \|_1 \leq 2\| C \|_\infty \| X^n a_n - a \|_1 + \| X^n b_n - b \|_1 \leq 2\| C \|_\infty \| X^n a_n - a \|_1 + \| X^n b_n - b \|_1 + 2\| C \|_\infty \| (X^n - \bar{X}) a_n \|_1 + \| (X^n - \bar{X}) b_n \|_1 \]
\]

\[
\leq 4\| C \|_\infty \eta \| X^n a_n - a \|_1 + \| X^n b_n - b \|_1 \| X^n - \bar{X} \|_1 \]
\]

\[
\leq \frac{\varepsilon}{4} + \frac{4\varepsilon}{8} = \frac{3\varepsilon}{4}
\]

(71)

where (71) follows Lemma E.3 with \( \gamma = \| C \|_\infty + 2\eta \) and (72) is by the definition of \( \tau \) and (70).
Since \(X^{OT} \in \Pi(a, b)\), we have:
\[
\text{KL}(X^{OT} \mathbf{1}_n || a) = \text{KL}(X^{OT \top} \mathbf{1}_n || b) = 0 \quad \text{and} \quad \eta \|X^{OT}\|_2^2 \leq \eta \|X^{OT}\|_1^2 = \eta.
\]
We thus obtain that:
\[
g_\eta(X^{OT}) = \langle C, X^{OT} \rangle + \eta \|X^{OT}\|_2^2 \leq \langle C, X^{OT} \rangle + \eta \tag{73}
\]
Since the KL divergence and \(\ell_2\) are non-negative, we have:
\[
g_\eta(X^n) \geq \langle C, X^n \rangle \tag{74}
\]
We recall that \(X^n = \arg\min_{X \in \mathbb{R}^{n \times n}} g_\eta(X)\), which implies:
\[
g_\eta(X^{OT}) \geq g_\eta(X^n), \tag{75}
\]
Combining (73), (74) and (75), we upper-bound the third term in the RHS of (69):
\[
\langle C, X^n - X^{OT} \rangle \leq \eta = \frac{\varepsilon}{32} \tag{76}
\]
Combining (70), (72) and (76) into (69), we have:
\[
\langle C, Y \rangle - \langle C, X^{OT} \rangle \leq \frac{\varepsilon}{8} + \frac{3\varepsilon}{4} + \frac{\varepsilon}{32} < \varepsilon.
\]

The complexity of GEM-OT is the total of \(O(n^2)\) for the PROJ() algorithm and the complexity of GEM-UOT for the specified \(\tau\) in the algorithm. By Corollary 3.5, which establishes the complexity of GEM-UOT under the assumptions (A1-A3), we conclude that the complexity of GEM-OT under (A1-A2) (where (A3) is naturally satisfied by our choice of \(\tau\)) is:
\[
O \left( \kappa \cdot n^2 \log \left( \frac{n\|C\|_\infty}{\varepsilon} \right) \right)
\]

E.2 Proof of Theorem 4.5

Proof. The lower bound is straightforward. Since \(X^{OT}\) (the transportation plan of \(\text{OT}(a, b)\)) is a feasible solution to the optimization problem \(\text{UOT}(a, b)\) and \(\text{KL}(X^{OT} \mathbf{1}_n || a) = \text{KL}(X^{OT \top} \mathbf{1}_n || b) = 0\) (as \(X^{OT} \in \Pi(a, b)\)), we obtain that:
\[
\text{UOT}(a, b) = \langle C, X^{OT} \rangle = f(X^{OT}) \geq f(X_f) = \text{UOT}(a, b)
\]
We proceed to prove the upper bound of the error approximation. We first define a variant of \(\text{UOT}_{KL}(a, b)\) problem where the KL divergence is replaced by squared \(\ell_2\) norm:
\[
\text{UOT}_{\ell_2}(a, b) = \min_{X \in \mathbb{R}^{n \times n}_+} \{ f_{\ell_2}(X) := \langle C, X \rangle + \frac{\tau}{2 \log(2)} \|X \mathbf{1}_n - a\|_2^2 \} \tag{77}
\]
By [Theorem 2, Blondel et al. (2018)], we have:
\[
0 \leq \text{UOT}(a, b) - \text{UOT}_{\ell_2}(a, b) \leq \frac{M_1}{\tau} \tag{78}
\]
where \(M_1 = \log(2)\|C\|_\infty^2 (n + 3\min\{a_{min}, b_{min}\})^2\).

Remark E.5. We highlight that deriving the approximation error for \(\text{UOT}_{KL}(a, b)\) is harder than \(\text{UOT}_{\ell_2}(a, b)\). First, KL is non-smooth while squared \(\ell_2\) norm is smooth, which is also stated in Blondel et al. (2018) as the main reason for their choice of squared \(\ell_2\) norm instead of KL to relax the marginal constraints. Second, the proof in Blondel et al. (2018) is based on the fact that the dual of \(\text{UOT}_{\ell_2}(a, b)\) can be interpreted as the sum of the dual of \(\text{OT}(a, b)\) and the additional squared \(\ell_2\)-norm regularization on the dual variables, so bounding the approximating error corresponds to simply bounding the dual variables. Third, since we consider generalized KL divergence where \(X \mathbf{1}_n\) (and resp. \(X^\top \mathbf{1}_n\)) may not be inside the probability simplex, direct application of conventional bounds on KL divergence is not possible (as standard KL divergence is defined on two probability distributions).
Note that since \( a, b \in \Delta^n \), we have \( \alpha = \beta = 1 \).

Recall that:

\[
\text{UOT}^\eta_{\text{KL}}(a, b) = \min_{X \in \mathbb{R}^n_{\times n}} \{ g_\eta(X) := \langle C, X \rangle + \eta \| X \|_2^2 + \tau \text{KL}(X_1^n || a) + \tau \text{KL}(X^\top 1_n || b) \},
\]

and \( X^n = \arg\min_{X \in \mathbb{R}^n_{\times n}} g_\eta(X) \) is the solution to \( \text{UOT}^\eta_{\text{KL}}(a, b) \).

We further define the problem \( \text{UOT}^\eta_{\text{KL}}(a, b) \) as follows:

\[
\text{UOT}^\eta_{\text{KL}}(a, b) = \min_{X \in \mathbb{R}^n_{\times n}, \| X \|_1 = 1} \{ g_\eta(X) := \langle C, X \rangle + \eta \| X \|_2^2 + \tau \text{KL}(X_1^n || a) + \tau \text{KL}(X^\top 1_n || b) \},
\]

which optimizes over the same objective cost as \( \text{UOT}^\eta_{\text{KL}}(a, b) \) yet with the additional constraint that \( X \) lies in the probability simplex, i.e. \( \| X \|_1 = 1 \). We let \( Z^n = \arg\min_{X \in \mathbb{R}^n_{\times n}, \| X \|_1 = 1} g_\eta(X) \) be the solution to \( \text{UOT}^\eta_{\text{KL}}(a, b) \).

Since \( \| Z^n \|_1 = 1 \), we have \( Z^n 1_n, Z^n 1_n \in \Delta^n \). Then Pinsker inequality gives:

\[
\text{KL}(Z^n 1_n || a) \geq \frac{1}{2 \log(2)} \| Z^n 1_n - a \|_2^2 \geq \frac{1}{2 \log(2)} \| Z^n 1_n - a \|_2^2
\]

\[
\text{KL}(Z^n^\top 1_n || b) \geq \frac{1}{2 \log(2)} \| Z^n^\top 1_n - b \|_2^2 \geq \frac{1}{2 \log(2)} \| Z^n^\top 1_n - b \|_2^2
\]

Consequently, we obtain that:

\[
\text{UOT}^\eta_{\text{KL}}(a, b) = g_\eta(Z^n) \geq f_{\ell_2}(Z^n) \geq \text{UOT}_{\ell_2}(a, b)
\]

For any \( \eta > 0 \), we consider \( Y^n = \text{PROJ}(X^n, a, b) \) as the projection of \( X^n \) onto \( \Pi(a, b) \) via Algorithm \( \text{[4]} \). Then \( \| Y^n \|_1 = a^\top 1_n = 1 \). We also have \( \| X^n \|_1 \leq 1 \) by Lemma \( \text{[B.3]} \).

By Lemma \( \text{[E.4]} \) we have:

\[
\| Y^n - X^n \|_1 \leq 2(\| X^n 1_n - a \|_1 + \| X^n^\top 1_n - b \|_1)
\]

\[
\leq 2n\| C \|_\infty (\| C \|_\infty + 2\eta) \quad \text{(by Lemma \text{[E.3]})}
\]

Since \( Y^n \) is a feasible solution to the optimization problem \( \text{UOT}^\eta_{\text{KL}}(a, b) \), we obtain that:

\[
g_\eta(Y^n) \geq \text{UOT}^\eta_{\text{KL}}(a, b),
\]

Note that \( \text{KL}(Y^n 1_n || a) = \text{KL}(Y^n^\top 1_n || b) = 0 \) as \( Y^n \in \Pi(a, b) \). We thus can write:

\[
g_\eta(Y^n) = \langle C, Y^n \rangle + \eta \| Y^n \|_2^2
\]

Now we have:

\[
g_\eta(Y^n) - \text{UOT}^\eta_{\text{KL}}(a, b) = g_\eta(Y^n) - g_\eta(X^n)
\]

\[
= \langle C, Y^n - X^n \rangle + \eta (\| Y^n \|_2^2 - \| X^n \|_2^2) - \tau \text{KL}(X^n 1_n || a) + \tau \text{KL}(X^n^\top 1_n || b)
\]

\[
\leq \| C \|_\infty \| Y^n - X^n \|_1 + \eta \| Y^n \|_2^2
\]

\[
\leq 2n\| C \|_\infty (\| C \|_\infty + 2\eta) \quad \text{(by Lemma \text{[E.3]})}
\]

where for the last line we note that \( \| Y^n \|_2 \leq \| Y^n \|_1 = 1 \). The above is equivalent to:

\[
\text{UOT}^\eta_{\text{KL}}(a, b) \geq g_\eta(Y^n) - 2n\| C \|_\infty (\| C \|_\infty + 2\eta) / \tau - \eta
\]

Now combining \( \text{[81]}, \text{[83]} \) and \( \text{[84]} \), we have \( \forall \eta > 0 \):

\[
\text{UOT}^\eta_{\text{KL}}(a, b) \geq \text{UOT}_{\ell_2}(a, b) - 2n\| C \|_\infty (\| C \|_\infty + 2\eta) / \tau - \eta
\]
Taking the limit $\eta \to 0$ on both sides and noting that $\lim_{\eta \to 0} \UOT^n_{KL}(a, b) = \UOT_{KL}(a, b)$, we obtain:

$$\UOT_{KL}(a, b) \geq \UOT_{\ell_2}(a, b) - \frac{2n\|C\|_\infty^2}{\tau}$$

$$\OT(a, b) - \UOT_{KL}(a, b) \leq \OT(a, b) - \UOT_{\ell_2}(a, b) + \frac{2n\|C\|_\infty^2}{\tau}$$

Combining with (78), we get the desired bound:

$$\OT(a, b) - \UOT_{KL}(a, b) \leq \frac{M}{\tau}$$

where $M = M_1 + 2n\|C\|_\infty^2$.

\[\Box\]

F Experiments

F.1 Sparseness of the transport plan

We empirically illustrate the sparseness of the transport plans produced by solving the entropic regularized UOT via Sinkhorn algorithm and the squared $\ell_2$ regularized UOT via GEM-UOT in Figure 5 and Figure 6.

Figure 5: [Fashion MNIST dataset] The transportation plan of the squared $\ell_2$ regularized UOT, solved by GEM-UOT, achieves 37.88% sparsity, while that of the entropic regularized UOT, solved by Sinkhorn, has 0% sparsity. Dark pixels of the figures represent positive values of the transport plan while the white pixels represent near 0 values.

In Figure 7 we compare the sparsity of the squared $\ell_2$ regularized UOT plan with that of the UOT plan (no regularization).
Figure 6: [CIFAR 10 dataset] The transportation plan of the squared $\ell_2$ regularized UOT, solved by GEM-UOT, achieves 32.6% sparsity, while that of the entropic regularized UOT, solved by Sinkhorn, has 0% sparsity. Black pixels of the figure represent positive values of the transport plan while the white pixels represent near 0 values.

Figure 7: Sparsity of the squared $\ell_2$ norm regularized UOT plan for $\eta = 10^{-3}$ and $\eta = 0$ respectively.

F.2 Fashion-MNIST dataset

We also experimented with the Fashion-MNIST dataset and report the results in Figure 8 and Figure 9. We also add a small constant to each pixel sufficiently close to 0 so that it can satisfy assumption (A2). Using similar settings to the CIFAR dataset experiment, we also observe that the GEM algorithms converge relatively faster.

Figure 8: Comparison of $f$ primal values between the Sinkhorn algorithm and GEM algorithms for 1500 iterations. We can see the Sinkhorn algorithm suffers for very unbalanced $a, b$ and converges slowly while GEM methods is mostly unaffected for unbalanced $a, b$. 
F.3 Approximability of UOT

We empirically confirm that the bound on the gap $\|X_f 1_n - a\|_1 + \|X_f^\top 1_n - b\|_1$ as shown in theorem 4.2 has the asymptotic rate of $O(\frac{1}{\tau})$. The results are plotted in Figure 10.

![Figure 10: Approximability of UOT KL: for $\alpha = \beta = 1$, entries of $C$ are drawn uniformly from $[0, 1]$, as $\tau \to \infty+$ the gap in theorem 4.2 will approach 0 with $O(\frac{1}{\tau})$.](image)

References

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems, 2019.

J. Altschuler, F. Bach, A. Rudi, and J. Weed. Approximating the quadratic transportation metric in near-linear time. *ArXiv Preprint: 1810.10046*, 2018.

Jason Altschuler, Jonathan 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.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation, 2020.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport, 2018.

Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and monge-ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010. ISSN 0003486X. URL [http://www.jstor.org/stable/20752228](http://www.jstor.org/stable/20752228).

Laetitia Chapel, Rémi Flamary, Haoran Wu, C’edric F’evotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. *ArXiv*, abs/2106.04145, 2021.
Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems, 2017.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation, 2016.

P. Dvurechensky, A. Gasnikov, and A. 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, 2018a.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn’s algorithm, 2018b.

Montacer Essid and Justin Solomon. Quadratically regularized optimal transport on graphs. SIAM Journal on Scientific Computing, 40(4):A1961–A1986, 2018. doi: 10.1137/17M1132665. URL [https://doi.org/10.1137/17M1132665](https://doi.org/10.1137/17M1132665).

Kilian Fatras, Thibault S’ejourn’e, Nicolas Courty, and Rémie Flamary. Unbalanced minibatch optimal transport; applications to domain adaptation. ArXiv, abs/2103.03606, 2021.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’15, page 2053–2061, Cambridge, MA, USA, 2015. MIT Press.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In International Conference on Artificial Intelligence and Statistics, pages 1608–1617, 2018.

Sergey Guminov, Pavel Dvurechensky, Nazarii Tupitsa, and Alexander Gasnikov. Accelerated alternating minimization, accelerated sinkhorn’s algorithm and accelerated iterative bregman projections, 2021.

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huy nh, and Dinh Phung. Multilevel clustering via wasserstein means, 2017.

L. V. Kantorovich. On the translocation of masses. In Dokl. Akad. Nauk. USSR (NS), volume 37, pages 199–201, 1942.

Alex Krizhevskiy. Learning multiple layers of features from tiny images. Technical report, 2009.

Guanghui Lan and Yi Zhou. Random gradient extrapolation for distributed and stochastic optimization. SIAM Journal on Optimization, 28(4):2753–2782, 2018.

Khang Le, Huy Nguyen, Quang Nguyen, Nhat Ho, Tung Pham, and Hung Bui. On robust optimal transport: Computational complexity, low-rank approximation, and barycenter computation, 2021.

John Lee, Nicholas P Bertrand, and Christopher J Rozell. Parallel unbalanced optimal transport regularization for large scale imaging problems. arXiv preprint arXiv:1909.00149, 2019.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. Inventiones mathematicae, 211(3):969–1117, Dec 2017. ISSN 1432-1297. doi: 10.1007/s00222-017-0759-8. URL [http://dx.doi.org/10.1007/s00222-017-0759-8](http://dx.doi.org/10.1007/s00222-017-0759-8).

T. Lin, N. Ho, and M. I. Jordan. On the acceleration of the Sinkhorn and Greenkhorn algorithms for optimal transport. ArXiv Preprint: 1906.01437, 2019a.

T. Lin, N. Ho, and M. I. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. ArXiv Preprint: 1901.06482, 2019b.

Tianyi Lin, Nhat Ho, and Michael Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In International Conference on Machine Learning, pages 3982–3991, 2019c.

K. Pham, K. Le, N. Ho, T. Pham, and H. Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In ICML, 2020a.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm, 2020b.

Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In SSVM, 2015.

R. Tyrrell Rockafellar. Duality and stability in extremum problems involving convex functions. Pacific Journal of Mathematics, 21(1):167 – 187, 1967. doi: pjm/1102992608. URL [https://doi.org/](https://doi.org/).

G. Schiebinger et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell, 176:928–943, 2019.
Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 306–314, Beijing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/solomon14.html.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe, 2022.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Cédric Villani. *Optimal transport: Old and New*. Springer, 2008.

K. D. Yang and C. Uhler. Scalable unbalanced optimal transport using generative adversarial networks. In *ICLR*, 2019.