On Unbalanced Optimal Transport: An Analysis of Sinkhorn Algorithm

Khiem Pham [* 1]   Khang Le [* 1]   Nhat Ho [2]   Tung Pham [1 3]   Hung Bui [1]

Abstract

We provide a computational complexity analysis for the Sinkhorn algorithm that solves the entropic regularized Unbalanced Optimal Transport (UOT) problem between two measures of possibly different masses with at most n components. We show that the complexity of the Sinkhorn algorithm for finding an ε-approximate solution to the UOT problem is of order $\widetilde{\mathcal{O}}(n^2/\varepsilon)$. To the best of our knowledge, this complexity is better than the best known complexity upper bound of the Sinkhorn algorithm for solving the Optimal Transport (OT) problem, which is of order $\widetilde{\mathcal{O}}(n^2/\varepsilon^2)$. Our proof technique is based on the geometric convergence rate of the Sinkhorn updates to the optimal dual solution of the entropic regularized UOT problem and scaling properties of the primal solution. It is also different from the proof technique used to establish the complexity of the Sinkhorn algorithm for approximating the OT problem since the UOT solution does not need to meet the marginal constraints of the measures.

1. Introduction

The Optimal Transport (OT) problem has a long history in mathematics and operation research, originally used to find the optimal cost to transport masses from one distribution to the other distribution (Villani, 2003). Over the last decade, OT has emerged as one of the most important tools to solve interesting practical problems in statistics and machine learning (Ho et al., 2017; Arjovsky et al., 2017; Courty et al., 2017; Srivastava et al., 2018; Peyré & Cuturi, 2019). Recently, the Unbalanced Optimal Transport (UOT) problem between two measures of possibly different masses has been used in several applications in computational biol-

---
[*]Equal contribution [1]VinAI Research [2]Department of EECS, University of California, Berkeley [3]Faculty of Mathematics, Mechanics and Informatics, Hanoi University of Science, Vietnam National University. Correspondence to: Nhat Ho <minhnhat@berkeley.edu>.

ogy (Schiebinger et al., 2019), computational imaging (Lee et al., 2019), deep learning (Yang & Uhler, 2019) and machine learning and statistics (Frogner et al., 2015; Janati et al., 2019a).

The UOT problem is a regularized version of Kantorovich formulation which places penalty functions on the marginal distributions based on some given divergences (Liero et al., 2018). When the two measures are the probability distributions, the standard OT is a limiting case of the UOT. Under the discrete setting of the OT problem where each probability distribution has at most n components, the OT problem can be recast as a linear programming problem. The benchmark methods for solving the OT problem are interior-point methods of which the most practical complexity is $\widetilde{\mathcal{O}}(n^3)$ developed by (Pele & Werman, 2009). Recently, (Lee & Sidford, 2014) used Laplacian linear system algorithms to improve the complexity of interior-point methods to $\widetilde{\mathcal{O}}(n^{5/2})$. However, the interior-point methods are not scalable when n is large.

In order to deal with the scalability of computing the OT, (Cuturi, 2013) proposed to regularize its objective function by the entropy of the transportation plan, which results in the entropic regularized OT. One of the most popular algorithms for solving the entropic regularized OT is the Sinkhorn algorithm (Sinkhorn, 1974), which was shown by (Altschuler et al., 2017) to have a complexity of $\widetilde{\mathcal{O}}(n^2/\varepsilon^3)$ when used to approximate the OT within an ε-accuracy. In the same article, (Altschuler et al., 2017) developed a greedy version of the Sinkhorn algorithm, named the Greenkhorn algorithm, that has a better practical performance than the Sinkhorn algorithm. Later, the complexity of the Greenkhorn algorithm was improved to $\widetilde{\mathcal{O}}(n^2/\varepsilon^2)$ by a deeper analysis in (Lin et al., 2019b). In order to accelerate the Sinkhorn and Greenkhorn algorithms, (Lin et al., 2019a) introduced Randkhorn and Gandkhorn algorithms that have complexity upper bounds of $\mathcal{O}(n^{7/3}/\varepsilon^{4/3})$. These complexities are better than those of Sinkhorn and Greenkhorn algorithms in terms of the desired accuracy ε. A different line of algorithms for solving the OT problem is based on primal-dual algorithms. These algorithms include accelerated primal-dual gradient descent algorithm (Dvurechensky et al., 2018), accelerated primal-dual mirror descent algorithm (Lin et al., 2019b), and accelerated primal-dual coordinate descent algorithm (Guo et al., 2019). These primal-dual

algorithms all have complexity upper bounds of $\widetilde{\mathcal{O}}(n^{2.5}/\varepsilon)$, which are better than those of Sinkhorn and Greenkhorn algorithms in terms of ε. Recently, (Jambulapati et al., 2019; Blanchet et al., 2018) developed algorithms with complexity upper bounds of $\widetilde{\mathcal{O}}(n^2/\varepsilon)$, which are believed to be optimal, based on either a dual extrapolation framework with area-convex mirror mapping or some black-box and specialized graph algorithms. However, these algorithms are quite difficult to implement. Therefore, they are less competitive than Sinkhorn and Greenkhorn algorithms in practice.

Our Contribution. While the complexity theory for OT has been rather well-understood, that for UOT is still nascent. In the paper, we establish the complexity of approximating UOT between two discrete measures with at most n components. We focus on the setting when the penalty functions are Kullback-Leiber (KL) divergences. Similar to the entropic regularized OT, in order to account for the scalability of computing UOT, we also consider an entropic version of UOT, which we refer to as *entropic regularized UOT*. The Sinkhorn algorithm is widely used to solve the entropic regularized UOT (Chizat et al., 2016); however, its complexity for approximating the UOT has not been studied. Our contribution is to prove that the Sinkhorn algorithm has a complexity of

$$\mathcal{O}\left(\frac{\tau(\alpha+\beta)n^2}{\varepsilon}\log(n)\left[\log(\|C\|_\infty)+\log(\log(n))\right.\right.$$
$$\left.\left.+\log\left(\frac{1}{\varepsilon}\right)\right]\right),$$

where C is a given cost matrix, α, β are total masses of the measures, and τ is a regularization parameter with the KL divergences in the UOT problem. This complexity is close to the probably optimal one by a factor of logarithm of n and $1/\varepsilon$.

The main difference between finding an ε-approximation solution for OT and UOT by the Sinkhorn algorithm is that the Sinkhorn algorithm for OT knows when it is close to the solution because of the constraints on the marginals, while the UOT does not have that advantage. Despite lacking that useful property, the Sinkhorn algorithm for UOT enjoys more freedom in its updates resulting in some interesting equations that relate the optimal value of the primal objective function of UOT to the masses of two measures (see Lemma 4). Those equations together with the geometric convergence rate of the dual solution of the UOT prove the almost optimal convergence of Sinkhorn algorithm to an ε-approximation solution of the UOT.

Organization. The remainder of the paper is organized as follows. In Section 2, we provide a setup for the regularized UOT with KL divergences in primal and dual forms, respectively. Based on the dual form, we show that the dual solution has a geometric convergence rate in Section 3.

We also show in Section 3 that the Sinkhorn algorithm for the UOT has a complexity of order $\tilde{\mathcal{O}}(n^2/\varepsilon)$. Section 4 presents some empirical results confirming the complexity of the Sinkhorn algorithm. Finally, we conclude in Section 5 while deferring the proofs of remaining results in the supplementary material.

Notation. We let $[n]$ stand for the set $\{1, 2, \ldots, n\}$ while $\mathbb{R}^n_+$ stands for the set of all vectors in $\mathbb{R}^n$ with nonnegative components for any $n \geq 2$. For a vector $x \in \mathbb{R}^n$ and $1 \leq p \leq \infty$, we denote $\|x\|_p$ as its ℓ_p-norm and $\mathrm{diag}(x)$ as the diagonal matrix with x on the diagonal. The natural logarithm of a vector $\mathbf{a} = (\mathbf{a}_1, ..., \mathbf{a}_n) \in \mathbb{R}^n$ is denoted $\log \mathbf{a} = (\log \mathbf{a}_1, ..., \log \mathbf{a}_n)$. $\mathbf{1}_n$ stands for a vector of length n with all of its components equal to 1. $\partial_x f$ refers to a partial gradient of f with respect to x. Lastly, given the dimension n and accuracy ε, the notation $a = \mathcal{O}(b(n, \varepsilon))$ stands for the upper bound $a \leq C \cdot b(n, \varepsilon)$ where C is independent of n and ε. Similarly, the notation $a = \widetilde{\mathcal{O}}(b(n, \varepsilon))$ indicates the previous inequality may depend on the logarithmic function of n and ε, and where $C > 0$.

2. Unbalanced Optimal Transport with Entropic Regularization

In this section, we present the primal and dual forms of the entropic regularized UOT problem and define an ε-approximation for the solution of the unregularized UOT.

For any two positive vectors $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_n) \in \mathbb{R}^n_+$ and $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_n) \in \mathbb{R}^n_+$, the UOT problem takes the form $\min_{X \in \mathbb{R}^{n \times n}_+} f(X)$ where

$$f(X) := \langle C, X \rangle + \tau \mathbf{KL}(X\mathbf{1}_n \| \mathbf{a}) \qquad (1)$$
$$+ \tau \mathbf{KL}(X^\top \mathbf{1}_n \| \mathbf{b}).$$

Here, C is a given cost matrix, X is a transportation plan, $\tau > 0$ is a given regularization parameter, and the $\mathbf{KL}$ divergence between vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathbf{KL}(\mathbf{x}\|\mathbf{y}) := \sum_{i=1}^n \mathbf{x}_i \log\left(\frac{\mathbf{x}_i}{\mathbf{y}_i}\right) - \mathbf{x}_i + \mathbf{y}_i.$$

When $\mathbf{a}^\top \mathbf{1}_n = \mathbf{b}^\top \mathbf{1}_n$ and $\tau \to \infty$, the UOT problem becomes the standard OT problem. Similar to the original OT problem, the exact computation of UOT is expensive and not scalable in terms of the number of supports n. Inspired by the recent success of the entropic regularized OT problem as an efficient approximation of the OT problem, we also consider the entropic version of the UOT problem (Frogner et al., 2015) , which we refer to as entropic regularized UOT, of finding $\min_{X \in \mathbb{R}^{n \times n}_+} g(X)$, where

$$g(X) := \langle C, X \rangle - \eta H(X) + \tau \mathbf{KL}(X\mathbf{1}_n \| \mathbf{a})$$
$$+ \tau \mathbf{KL}(X^\top \mathbf{1}_n \| \mathbf{b}). \qquad (2)$$

Here, $\eta > 0$ is a given regularization parameter and $H(X)$ is an entropic regularization given by

$$H(X) := -\sum_{i,j=1}^{n} X_{ij}(\log(X_{ij}) - 1). \tag{3}$$

For each $\eta > 0$, we can check that the entropic regularized UOT problem is strongly convex in X. Therefore, it is convenient to solve for the optimal solution of the entropic regularized UOT and uses it to approximate the original value of UOT.

Definition 1. *For any $\varepsilon > 0$, we call X an ε-approximation transportation plan if the following holds*

$$\langle C, X \rangle + \tau \mathbf{KL}(X\boldsymbol{I}_n\|\mathbf{a}) + \tau \mathbf{KL}(X^\top \boldsymbol{I}_n\|\mathbf{b})$$
$$\leq \left\langle C, \widehat{X} \right\rangle + \tau \mathbf{KL}(\widehat{X}\boldsymbol{I}_n\|\mathbf{a}) + \mathbf{KL}(\widehat{X}^\top \boldsymbol{I}_n\|\mathbf{b}) + \varepsilon,$$

where $\widehat{X}$ is an optimal transportation plan for the UOT problem (1).

We aim to develop an algorithm to obtain ε-approximation transportation plan for the UOT problem (1). In order to do that, we consider the Fenchel-Legendre dual form of the entropic regularized UOT, which is given by

$$\max_{u,v\in\mathbb{R}^n} -F^*(-u) - G^*(-v) - \eta \sum_{i,j} \exp\left(\frac{u_i + v_j - C_{ij}}{\eta}\right),$$

where the functions $F^*(.)$ and $G^*(.)$ take the following forms:

$$F^*(u) = \max_{z\in\mathbb{R}^n} z^\top u - \tau \mathbf{KL}(z\|\mathbf{a}) = \tau \left\langle e^{u/\tau}, \mathbf{a} \right\rangle - \mathbf{a}^\top \boldsymbol{1}_n,$$
$$G^*(v) = \max_{x\in\mathbb{R}^n} x^\top v - \tau \mathbf{KL}(x\|\mathbf{b}) = \tau \left\langle e^{v/\tau}, \mathbf{b} \right\rangle - \mathbf{b}^\top \boldsymbol{1}_n.$$

Since $\mathbf{a}$ and $\mathbf{b}$ are given non-negative vectors, finding the optimal solution for the above objective is equivalent to finding the optimal solution for the following objective

$$\min_{u,v\in\mathbb{R}^n} h(u,v) := \eta \sum_{i,j=1}^{n} \exp\left(\frac{u_i + v_j - C_{ij}}{\eta}\right)$$
$$+ \tau \left\langle e^{-u/\tau}, \mathbf{a} \right\rangle + \tau \left\langle e^{-v/\tau}, \mathbf{b} \right\rangle. \tag{4}$$

Problem (4) is referred to as the *dual entropic regularized UOT*.

3. Complexity Analysis of Approximating Unbalanced Optimal Transport

In this section, we provide a complexity analysis of the Sinkhorn algorithm for approximating UOT solution. We start with some notations and useful quantities followed by the lemmas and main theorems.

3.1. Notations and Assumptions

We first denote $\sum_{i=1}^{n} \mathbf{a}_i = \alpha$, $\sum_{j=1}^{n} \mathbf{b}_j = \beta$. For each $u, v \in \mathbb{R}^n$, its corresponding optimal transport in the dual form (4) is denoted by $B(u,v)$, where $B(u,v) := \mathrm{diag}(e^{u/\eta}) \, e^{-\frac{C}{\eta}} \, \mathrm{diag}(e^{v/\eta})$. The corresponding solution in (2) is denoted by $X = B(u,v)$. Let $a = B(u,v)\mathbf{1}_n$, $b = B(u,v)^\top \mathbf{1}_n$ and $\sum_{i,j=1}^{n} X_{ij} = x$.

Let (u^k, v^k) be the solution returned at the k-th iteration of the Sinkhorn algorithm and (u^*, v^*) be the optimal solution of (4). Following the above scheme, we also define X^k, a^k, b^k, x^k and X^*, a^*, b^*, x^* correspondingly. Additionally, we define $\widehat{X}$ to be the optimal solution of the unregularized objective (1) and $\sum_{i,j=1}^{n} \widehat{X}_{ij} = \widehat{x}$.

Different from the standard OT problem, the optimal solutions of the entropic regularized UOT and our complexity analysis overall also depend on the masses α, β and the $\mathbf{KL}$ regularization parameter τ. Given that, we will assume the following simple and standard regularity conditions throughout the paper.

Regularity Conditions:

(A1) α, β, τ are positive constants.

(A2) C is a matrix of non-negative entries.

Before presenting the main theorem and analysis, we define some quantities that will be used in our analysis and quantify their magnitudes under the above regularity conditions.

List of Quantities:

$$R = \max\left\{\|\log(\mathbf{a})\|_\infty, \|\log(\mathbf{b})\|_\infty\right\}$$
$$+ \max\left\{\log(n), \frac{1}{\eta}\|C\|_\infty - \log(n)\right\}, \tag{5}$$

$$\Delta_k = \max\left\{\|u^k - u^*\|_\infty, \|v^k - v^*\|_\infty\right\}, \tag{6}$$

$$\Lambda_k = \tau\left(\frac{\tau}{\tau + \eta}\right)^k R. \tag{7}$$

$$S = \frac{1}{2}(\alpha + \beta) + \frac{1}{2} + \frac{1}{4\log(n)}, \tag{8}$$

$$T = \left(\frac{\alpha + \beta}{2}\right)\left[\log\left(\frac{\alpha + \beta}{2}\right) + 2\log(n) - 1\right]$$
$$+ \log(n) + \frac{5}{2}, \tag{9}$$

$$U = \max\left\{S + T, 2\varepsilon, \frac{4\varepsilon \log(n)}{\tau}, \frac{4\varepsilon(\alpha + \beta)\log(n)}{\tau}\right\}. \tag{10}$$

As we shall see, the quantities Δ_k and Λ_k are used to establish the convergence rate of (u_k, v_k). We now consider the order of R, S and T. Since the order of the penalty function $\eta H(X)$ is $\mathcal{O}(\eta \log(n))$ and should be small for a good approximation, η is often chosen such that $\eta \log(n)$ is sufficiently small (see our choice of $\eta = \frac{\varepsilon}{U}$ in Theorem 2).

Algorithm 1 UNBALANCED_SINKHORN(C, ε)

Input: $k = 0$ and $u^0 = v^0 = 0$ and $\eta = \varepsilon/U$
while $k \le \left(\frac{\tau U}{\varepsilon} + 1\right)\left[\log(8\eta R) + \log(\tau(\tau+1)) + 3\log(\frac{U}{\varepsilon})\right]$ **do**
    $a^k = B(u^k, v^k)\mathbf{1}_n$.
    $b^k = B(u^k, v^k)^\top \mathbf{1}_n$.
    if k is even **then**
$$u^{k+1} = \left[\frac{u^k}{\eta} + \log(\mathbf{a}) - \log(a^k)\right]\frac{\eta\tau}{\eta+\tau}$$
$$v^{k+1} = v^k$$
    else
$$v^{k+1} = \left[\frac{v^k}{\eta} + \log(\mathbf{b}) - \log(b^k)\right]\frac{\eta\tau}{\eta+\tau}$$
$$u^{k+1} = u^k.$$
    end if
    $k = k + 1$.
end while
Output: $B(u^k, v^k)$.

Therefore, we can assume the dominant factor in the second term of R is $\frac{1}{\eta}\|C\|_\infty$. If $\alpha = \sum_{i=1}^n \mathbf{a}_i$ is a positive constant, then we can expect that $\mathbf{a}_i$ is as small as $\mathcal{O}(n^{-\kappa})$ for a constant $\kappa \ge 1$. In this case, $\|\log(\mathbf{a})\|_\infty = \mathcal{O}(\log(n))$. Overall, we can assume that $R = \mathcal{O}(\frac{1}{\eta}\|C\|_\infty)$ and if α, β and τ are positive constants, then $S = \mathcal{O}(1)$ and $T = \mathcal{O}(\log(n))$.

3.2. Sinkhorn Algorithm

The Sinkhorn algorithm (Chizat et al., 2016) alternatively minimizes the dual function in (4) with respect to u and v. Suppose we are at iteration $k + 1$ for $k \ge 0$ and k even, by setting the gradient to 0 we can see that given fixed v^k, the update u^{k+1} that minimizes the function in (4) satisfies

$$\exp\left(\frac{u_i^{k+1}}{\eta}\right)\sum_{j=1}^n \exp\left(\frac{v_j^k - C_{ij}}{\eta}\right) = \exp\left(-\frac{u_i^{k+1}}{\tau}\right)\mathbf{a}_i.$$

Multiplying both sides by $\exp\left(\frac{u_i^k}{\eta}\right)$, we get:

$$\exp\left(\frac{u_i^{k+1}}{\eta}\right)a_i^k = \exp\left(\frac{u_i^k}{\eta}\right)\exp\left(-\frac{u_i^{k+1}}{\tau}\right)\mathbf{a}_i.$$

Similarly with u^k fixed and k is odd:

$$\exp\left(\frac{v_j^{k+1}}{\eta}\right)b_j^k = \exp\left(\frac{v_j^k}{\eta}\right)\exp\left(-\frac{v_j^{k+1}}{\tau}\right)\mathbf{b}_j.$$

These equations translate to the pseudocode of Algorithm 1, in which we have included our stopping condition stated in Theorem 2. We now present the main theorems.

Theorem 1. *For any $k \ge 0$, the update (u^{k+1}, v^{k+1}) from Algorithm 1 satisfies the following bound*

$$\Delta_{k+1} \le \Lambda_k, \tag{11}$$

where Δ_k and Λ_k are defined in (6) and (7), respectively.

Remark 1. *Theorem 1 establishes a geometric convergence rate for the dual solution (u^k, v^k) under ℓ_∞ norm. Its geometric convergence is similar to the work of (Séjourné et al., 2019) while it is different from the work of (Chizat et al., 2016), which used the Thompson metric. The difference between the result of Theorem 1 and that in (Séjourné et al., 2019) is that we obtain a specific upper bound for the convergence rate of the Sinkhorn updates, which depends explicitly on the number of components n and all other parameters of masses and penalty functions. The convergence rate of Theorem 1 plays an important role in the complexity analysis of the Sinkhorn algorithm in the next theorem.*

Theorem 2. *Under the regularity conditions (A1-A2), with R and U defined in (5) and (10) respectively, for $\eta = \frac{\varepsilon}{U}$ and*

$$k \ge 1 + \left(\frac{\tau U}{\varepsilon} + 1\right)\left[\log(8\eta R) + \log(\tau(\tau+1))\right.$$
$$\left. + 3\log(\frac{U}{\varepsilon})\right],$$

the update X^k from Algorithm 1 is an ε-approximation of the optimal solution $\widehat{X}$ of (1).

The next corollary sums up the complexity of Algorithm 1.

Corollary 1. *Under conditions (A1-A2) and assume that $R = \mathcal{O}\left(\frac{1}{\eta}\|C\|_\infty\right), S = \mathcal{O}(1)$ and $T = \mathcal{O}(\log(n))$. Then the complexity of Algorithm 1 is*

$$\mathcal{O}\left(\frac{\tau(\alpha+\beta)n^2}{\varepsilon}\log(n)\left[\log(\|C\|_\infty) + \log(\log(n))\right.\right.$$
$$\left.\left. + \log\left(\frac{1}{\varepsilon}\right)\right]\right).$$

Proof of Corollary 1. By the assumptions on the order of R, S, T, we have

$$R = \mathcal{O}(\frac{1}{\eta}\|C\|_\infty), S = \mathcal{O}(\alpha+\beta), T = \mathcal{O}((\alpha+\beta)\log(n)).$$

Plugging the above results into the definition of U in (10), we find that

$$U = \mathcal{O}(S) + \mathcal{O}(T) + \varepsilon\mathcal{O}(\log(n)) = \mathcal{O}((\alpha+\beta)\log(n)).$$

Overall, we obtain

$$k = \mathcal{O}\left(\frac{\tau(\alpha+\beta)\log(n)}{\varepsilon}\right)\left[\log(\|C\|_\infty) + \log(\log(n))\right.$$
$$\left. + \log\left(\frac{1}{\varepsilon}\right)\right].$$

By multiplying the above bound of k with $\mathcal{O}(n^2)$ arithmetic operations per iteration, we obtain the desired final complexity of the Sinkhorn algorithm as being claimed in the conclusion of the corollary. $\square$

Remark 2. *Since α, β and τ are positive constants, we obtain the complexity $\widetilde{\mathcal{O}}\left(n^2/\epsilon\right)$ as stated in the abstract. In comparison to the best known OT's complexity of the similar order of n, i.e., (Dvurechensky et al., 2018; Lin et al., 2019b), our complexity for the UOT is better by a factor of ε. Meanwhile, among the practical algorithms for OT which have complexity of the order of $n^{7/3}$, i.e., Gankhorn and Randkhorn algorithms (Lin et al., 2019a), our bound is better by a factor of $n^{1/3}$.*

3.3. Proof of Theorem 1

The analysis of Sinkhorn algorithm for approximating unbalanced optimal transport is different from that of optimal transport, since $\mathbf{a}$ and $\mathbf{b}$ need not be probability measures. Our proof of Theorem 1 requires the geometric convergence rate of (u^k, v^k) and an upper bound on the supremum norm of the optimal dual solution (u^*, v^*). The latter result is presented in Lemma 3. Before stating that result formally, we need the following lemmas.

Lemma 1. *The optimal solution (u^*, v^*) of dual entropic regularized UOT* (4) *satisfies the following equations:*

$$\frac{u^*}{\tau} = \log(\mathbf{a}) - \log(a^*), \text{ and } \frac{v^*}{\tau} = \log(\mathbf{b}) - \log(b^*).$$

Proof. Since (u^*, v^*) is a fixed point of the updates in the Algorithm 1, we get

$$u^* = \left[\frac{u^*}{\eta} + \log(\mathbf{a}) - \log(a^*)\right]\frac{\eta\tau}{\eta + \tau}.$$

This directly leads to the stated equality for u^*, and that for v^* can be obtained similarly. $\qquad\square$

Lemma 2. *Assume that the regularity conditions (A1-A2) hold. Then, the following are true:*

(a) $\left|\log\left(\frac{a_i^*}{a_i^k}\right) - \frac{u_i^* - u_i^k}{\eta}\right| \leq \max_{1 \leq j \leq n}\frac{|v_j^* - v_j^k|}{\eta},$

(b) $\left|\log\left(\frac{b_j^*}{b_j^k}\right) - \frac{v_j^* - v_j^k}{\eta}\right| \leq \max_{1 \leq i \leq n}\frac{|u_i^* - u_i^k|}{\eta}.$

The proof is given in the supplementary material.

Lemma 3. *The sup norms of the optimal solution $\|u^*\|_\infty$ and $\|v^*\|_\infty$ are bounded by:*

$$\max\{\|u^*\|_\infty, \|v^*\|_\infty\} \leq \tau R,$$

where R is defined in equation (5).

Proof. We start with the equations for the solution u^* in Lemma 1, i.e., we have

$$\frac{u_i^*}{\tau} = \log(\mathbf{a}_i) - \log\left(\sum_{j=1}^n \exp\left(\frac{u_i^* + v_j^* - C_{ij}}{\eta}\right)\right),$$

which can be rewritten as

$$u_i^*\left(\frac{1}{\tau} + \frac{1}{\eta}\right) = \log(\mathbf{a}_i) - \log\left[\sum_{j=1}^n \exp\left(\frac{v_j^* - C_{ij}}{\eta}\right)\right].$$

The second term in the above display can be lower bounded as follows

$$\log\left[\sum_{j=1}^n \exp\left(\frac{v_j^* - C_{ij}}{\eta}\right)\right] \geq \log(n) + \min_{1 \leq j \leq n}\left\{\frac{v_j^* - C_{ij}}{\eta}\right\}$$

$$\geq \log(n) - \frac{\|v^*\|_\infty}{\eta} - \frac{\|C\|_\infty}{\eta}.$$

Additionally, we also have

$$\log\left[\sum_{j=1}^n \exp\left(\frac{v_j^* - C_{ij}}{\eta}\right)\right] \leq \log(n) + \max_{1 \leq j \leq n}\left\{\frac{v_j^* - C_{ij}}{\eta}\right\}$$

$$\leq \log(n) + \frac{\|v^*\|_\infty}{\eta}.$$

Collecting the above results, we find that

$$\left|\log\left[\sum_{j=1}^n \exp\left(\frac{v_j^* - C_{ij}}{\eta}\right)\right]\right|$$

$$\leq \frac{\|v^*\|_\infty}{\eta} + \max\left\{\log(n), \frac{\|C^*\|_\infty}{\eta} - \log(n)\right\}.$$

Hence, we obtain that

$$|u_i^*|\left(\frac{1}{\tau} + \frac{1}{\eta}\right) \leq |\log(\mathbf{a}_i)| + \frac{\|v^*\|_\infty}{\eta} +$$

$$\max\left\{\log(n), \frac{\|C^*\|_\infty}{\eta} - \log(n)\right\}.$$

By choosing an index i such that $|u_i^*| = \|u^*\|_\infty$ and combining with the fact that $|\log(\mathbf{a}_i)| \leq \max\{\|\log(\mathbf{a})\|_\infty, \|\log(\mathbf{b})\|_\infty\}$, we have

$$\|u^*\|_\infty\left(\frac{1}{\tau} + \frac{1}{\eta}\right) \leq \frac{\|v^*\|_\infty}{\eta} + R.$$

WLOG assume that $\|u^*\|_\infty \geq \|v^*\|_\infty$. Then, we obtain the stated bound in the conclusion. $\qquad\square$

Proof of Theorem 1 We first consider the case when k is even. From the update of u^{k+1} in Algorithm 1, we have:

$$u_i^{k+1} = \left(\frac{u_i^k}{\eta} + \log\mathbf{a}_i - \log a_i^k\right)\frac{\eta\tau}{\tau + \eta}$$

$$= \left\{\frac{u_i^k}{\eta} + [\log(\mathbf{a}_i) - \log(a_i^*)] + \left[\log(a_i^*) - \log(a_i^k)\right]\right\}\frac{\eta\tau}{\tau + \eta}.$$

Using Lemma 1, the above equality is equivalent to

$$u_i^{k+1} - u_i^* = \left[\eta \log\left(\frac{a_i^*}{a_i^k} \right) - (u_i^* - u_i^k) \right] \frac{\tau}{\tau + \eta}.$$

Using Lemma 2, we get

$$\left| u_i^{k+1} - u_i^* \right| \le \max_{1 \le j \le n} \left| v_j^k - v_j^* \right| \frac{\tau}{\tau + \eta}.$$

This leads to $\|u^{k+1} - u^*\|_\infty \le \frac{\tau}{\tau+\eta}\|v^k - v^*\|_\infty$. Similarly, we obtain $\|v^k - v^*\|_\infty \le \frac{\tau}{\tau+\eta}\|u^{k-1} - u^*\|_\infty$. Combining the two inequalities yields

$$\|u^{k+1} - u^*\|_\infty \le \left(\frac{\tau}{\tau + \eta} \right)^2 \|u^{k-1} - u^*\|_\infty.$$

Repeating all the above arguments alternatively, we have

$$\|u^{k+1} - u^*\|_\infty \le \left(\frac{\tau}{\tau + \eta} \right)^{k+1} \|v^0 - v^*\|_\infty$$
$$= \left(\frac{\tau}{\tau + \eta} \right)^{k+1} \|v^*\|_\infty.$$

Note that $v^{k+1} = v^k$ for k even. Therefore, we find that

$$\|v^{k+1} - v^*\|_\infty \le \frac{\tau}{\tau + \eta}\|u^{k-1} - u^*\|_\infty$$
$$\le \left(\frac{\tau}{\tau + \eta} \right)^k \|v^*\|_\infty.$$

These two results lead to $\Delta_{k+1} \le \left(\frac{\tau}{\tau+\eta} \right)^k \|v^*\|_\infty$. Similarly, for k odd we obtain $\Delta_{k+1} \le \left(\frac{\tau}{\tau+\eta} \right)^k \|v^*\|_\infty$. Thus the above inequality is true for all k. Using the fact that $\|v^*\|_\infty \le \max\{\|u^*\|_\infty, \|v^*\|_\infty\}$ and Lemma 3, we obtain the conclusion of Theorem 1.

3.4. Proof of Theorem 2

The proof is based on the upper bound for the convergence rate in Theorem 1 and an upper bound for the solutions $\widehat{x}$ and x^* of functions in equations (1) and (2), respectively, which are direct consequences of the following lemmas.

Lemma 4. *Assume that the function g attains its minimum at X^*, then*

$$g(X^*) + (2\tau + \eta)x^* = \tau(\alpha + \beta). \qquad (12)$$

Similarly, assume that f attains its minimum at $\widehat{X}$, then

$$f(\widehat{X}) + 2\tau\widehat{x} = \tau(\alpha + \beta), \qquad (13)$$

where $\sum_{i,j=1}^n X^* = x^*$ and $\sum_{i,j=1}^n \widehat{X} = \widehat{x}$.

Both equations in Lemma 4 establish the relationships between the optimal solutions of functions in equations (1) and (2) with other parameters. Those relationships are very useful for analysing the behaviour of the optimal solution of UOT, because the UOT does not have any conditions on the marginals as the OT does. Consequences of Lemma 4 include Corollary 2 which provides the upper bounds for $\widehat{x}$ and x^* as well as bounds for the entropic functions in the proof of Theorem 2. The key idea of the proof surprisingly comes from the fact that the UOT solution does not have to meet the marginal constraints. We note that equation (12) could also be proved by using the fixed point equations in Lemma 1 as in (Janati et al., 2019b). Here we offer an alternative proof that can be applied to the unregularized case as well. We now present the proof of Lemma 4.

Proof. Consider the function $g(tX^*)$, where $t \in \mathbb{R}^+$,

$$g(tX^*) = \langle C, tX^* \rangle + \tau \mathrm{KL}(tX^* \mathbf{1}_n \| \mathbf{a})$$
$$+ \tau \mathrm{KL}((tX^*)^\top \mathbf{1}_n \| \mathbf{b}) - \eta H(tX^*).$$

For the KL term of $g(tX^*)$, we have:

$$\mathrm{KL}(tX^* \mathbf{1}_n \| \mathbf{a})$$
$$= \sum_{i=1}^n (ta_i^*) \log\left(\frac{ta_i^*}{\mathbf{a}_i} \right) - \sum_{i=1}^n (ta_i^*) + \sum_{i=1}^n \mathbf{a}_i$$
$$= \sum_{i=1}^n (ta_i^*) \left(\log\left(\frac{a_i^*}{\mathbf{a}_i} \right) + \log(t) \right) - tx^* + \alpha$$
$$= t \sum_{i=1}^n \left(a_i^* \log\left(\frac{a_i^*}{\mathbf{a}_i} \right) - x^* + \mathbf{a}_i \right) + (1-t)\alpha$$
$$\qquad\qquad\qquad\qquad\qquad + x^* t \log(t)$$
$$= t\mathrm{KL}(X^* \mathbf{1}_n \| \mathbf{a}) + (1-t)\alpha + x^* t \log(t).$$

Similarly, we get

$$\mathrm{KL}(t(X^*)^\top \mathbf{1}_n \| \mathbf{b}) = t\mathrm{KL}((X^*)^\top \mathbf{1}_n \| \mathbf{b})$$
$$+ (1-t)\beta + x^* t \log(t).$$

For the entropic penalty term, we find that

$$-H(tX^*) = \sum_{i,j=1}^n tX_{ij}^* \left(\log(tX_{ij}^*) - 1 \right)$$
$$= \sum_{i,j} tX_{ij}^* \left(\log(X_{ij}^*) - 1 \right) + x^* t \log(t)$$
$$= -tH(X^*) + x^* t \log(t).$$

Putting all results together, we obtain

$$g(tX^*) = tg(X^*) + \tau(1-t)(\alpha + \beta)$$
$$+ (2\tau + \eta)x^* t \log(t).$$

Taking the derivative of $g(tX^*)$ with respect to t,

$$\frac{\partial g(tX^*)}{\partial t} = g(X^*) - \tau(\alpha + \beta) + (2\tau + \eta)x^*(1 + \log(t)).$$

The function $g(tX^*)$ is well-defined for all $t \in \mathbb{R}^+$. We know that $g(tX^*)$ attains its minimum at $t = 1$. Replace $t = 1$ into the above equation, we obtain

$$g(X^*) - \tau(\alpha + \beta) + (2\tau + \eta)x^* = 0$$
$$g(X^*) + (2\tau + \eta)x^* = \tau(\alpha + \beta).$$

The second claim is proved in the same way. □

Corollary 2. *Assume that conditions (A1-A2) hold and $\eta \log(n)$ is sufficiently small. We have the following bounds on x^* and $\widehat{x}$:*

(a) $\quad x^* \le \left(\frac{1}{2} + \frac{\eta \log(n)}{2\tau - 2\eta \log(n)} \right)(\alpha + \beta) + \frac{1}{6 \log(n)},$

(b) $\quad \widehat{x} \le \frac{\alpha + \beta}{2}.$

We defer the proof of Corollary 2 to the appendix. Next, we use the condition for k in Theorem 2 to bound some relevant quantities at the k-th iteration of the Sinkhorn algorithm.

Lemma 5. *Assume that the regularity conditions (A1-A2) hold and k satisfies the inequality in Theorem 2. The following are true*

(a) $\quad \Lambda_{k-1} \le \frac{\eta^2}{8(\tau + 1)},$

(b) $\quad |x^k - x^*| \le \frac{3}{\eta} \Delta_k \min\{x^*, x^k\},$

(c) $\quad x^k \le S,$

where S is defined in equation (8).

We are now ready to construct a proof for Theorem 2.

Proof of Theorem 2 From the definitions of f and g, we have

$$f(X^k) - f(\widehat{X})$$
$$= g(X^k) + \eta H(X^k) - g(\widehat{X}) - \eta H(\widehat{X})$$
$$= g(X^k) + \eta H(X^k) - g(\widehat{X}) - \eta H(\widehat{X}) - g(X^*) + g(X^*)$$
$$\le \left[g(X^k) - g(X^*) \right] + \eta \left[H(X^k) - H(\widehat{X}) \right], \quad (14)$$

since $g(X^*) - g(\widehat{X}) \le 0$, as X^* is the optimal solution of function (2). The above two terms can be bounded separately as follows:

Upper Bound of $H(X^k) - H(\widehat{X})$.
We first show the following inequalities

$$x - x\log(x) \le H(X) \le 2x\log(n) + x - x\log(x) \quad (15)$$

for any X that $X_{ij} \ge 0$ and $x = \sum_{ij} X_{ij}$.

Indeed, rewriting $-H(X)$ as

$$-H(X) = x \left[\sum_{i,j=1}^{n} \frac{X_{ij}}{x} \log\left(\frac{X_{ij}}{x}\right) - 1 \right] + x\log(x),$$

and using $-2\log(n) \le \sum_{i,j=1}^{n} \frac{X_{ij}}{x} \log\left(\frac{X_{ij}}{x}\right) \le 0$, we thus obtain equation (15).

Now apply the lower bound of (15) to $-H(\widehat{X})$

$$-H(\widehat{X}) \le \widehat{x}\log(\widehat{x}) - \widehat{x}$$
$$\le \max\left\{ 0, \frac{\alpha + \beta}{2} \left[\log(\frac{\alpha + \beta}{2}) - 1 \right] \right\}$$
$$\le \frac{\alpha + \beta}{2} \log(\frac{\alpha + \beta}{2}) - \frac{\alpha + \beta}{2} + 1,$$

where the second inequality is due to $x\log x - x$ being convex and $0 \le \widehat{x} \le \frac{1}{2}(\alpha + \beta)$ by Corollary 2 and the third inequality is due to $\frac{\alpha + \beta}{2} \left[\log(\frac{\alpha + \beta}{2}) - 1 \right] + 1 \ge 0$.

Similarly, apply the upper bound of equation (15) to $H(X^k)$

$$H(X^k) \le 2x^k \log(n) + x^k - x^k \log(x^k)$$
$$\le 2x^k \log(n) + 1$$
$$\le \left(\alpha + \beta + 1 + \frac{1}{2\log(n)} \right) \log(n) + 1,$$

where the last inequality is due to Lemma 5(c). By combining the two results, we have

$$H(X^k) - H(\widehat{X}) \le T, \quad (16)$$

where T is defined in equation (9).

Upper Bound of $g(X^k) - g(X^*)$.
WLOG we assume that k is odd. At step $k - 1$ of Algorithm 1, we find u^k by minimizing the dual function (4) given $\mathbf{a}$ and fixed v^{k-1}, and simply keep $v^k = v^{k-1}$. Hence, $X^k = B(u^k, v^k)$ is the optimal solution of

$$\min_{X \in \mathbb{R}_+^{n \times n}} g^k(X) := \langle C, X \rangle - \eta H(X)$$
$$+ \tau \mathbf{KL}(X\mathbf{1}_n \| \mathbf{a}) + \tau \mathbf{KL}(X^\top \mathbf{1}_n \| \mathbf{b}^k),$$

where $\mathbf{b}^k = \exp\left(\frac{v^k}{\tau}\right) \odot \left[(X^k)^T \mathbf{1}_n \right]$ with $\odot$ denoting element-wise multiplication.

Denote $\sum_{i=1}^{n} \mathbf{b}_i^k = \beta^k$. By Lemma 4,

$$g^k(X^k) = \tau(\alpha + \beta^k) - (2\tau + \eta)x^k,$$
$$g(X^*) = \tau(\alpha + \beta) - (2\tau + \eta)x^*.$$

Writing $g(X^k) - g(X^*) = [g(X^k) - g^k(X^k)] + [g^k(X^k) - g(X^*)]$, following some derivations using the above equations of $g^k(X^k)$ and $g(X^*)$ and the definitions of $g(X^k)$ and $g^k(X^k)$, we get

$$g(X^k) - g(X^*) = \left[-(2\tau + \eta)(x^k - x^*) \right] + \tau \left[\sum_{j=1}^n b_j^k \log\left(\frac{\mathbf{b}_j^k}{\mathbf{b}_j} \right) \right]. \qquad (17)$$

By part (b) of Lemma 5, the first term is bounded by $(2\tau + \eta) \frac{3}{\eta} \Delta_k x^k$.

Note that $\mathbf{b}_j^k = \exp\left(\frac{v_j^k}{\tau} \right) b_j^k$ and $\mathbf{b}_j = \exp\left(\frac{v_j^*}{\tau} \right) b_j^*$. Use part (b) of Lemma 2, we find that

$$\log\left(\frac{\mathbf{b}_j^k}{\mathbf{b}_j} \right) = \left[-\log\left(\frac{b_j^*}{b_j^k} \right) \right] + \frac{1}{\tau}(v_j^k - v_j^*),$$

$$\left| \log\left(\frac{\mathbf{b}_j^k}{\mathbf{b}_j} \right) \right| \le \left(\frac{2}{\eta} \Delta_k \right) + \left(\frac{1}{\tau} \Delta_k \right) = \left(\frac{2}{\eta} + \frac{1}{\tau} \right) \Delta_k.$$

Note that $b_j^k \ge 0$ for all j. The above inequality leads to

$$\left| \sum_{j=1}^n b_j^k \log\left(\frac{\mathbf{b}_j^k}{\mathbf{b}_j} \right) \right| \le \left(\sum_{j=1}^n b_j^k \right) \max_{1 \le j \le n} \left| \log\left(\frac{\mathbf{b}_j^k}{\mathbf{b}_j} \right) \right|$$

$$\le x^k \left(\frac{2}{\eta} + \frac{1}{\tau} \right) \Delta_k.$$

We have

$$g(X^k) - g(X^*) \le \left[(2\tau + \eta) + 3(2\tau + \eta) \right] \left(\frac{\Delta_k}{\eta} \right) x^k$$

$$\le 8(\tau + 1) \left(\frac{\Lambda_{k-1}}{\eta} \right) S,$$

where the first inequality is obtained by combining the bounds for two terms of (17) while the second inequality results from the fact that $\eta = \frac{\varepsilon}{U} \le \frac{\varepsilon}{2\varepsilon} = \frac{1}{2}$ with U defined in (10), part (c) of Lemma 5, and Theorem 1.

Using part (a) of Lemma 5, this leads to

$$g(X^k) - g(X^*) \le \eta S. \qquad (18)$$

Combining (14), (16), (18) and the fact that $\eta = \frac{\varepsilon}{U} \le \frac{\varepsilon}{S+T}$, we get

$$f(X^k) - f(\widehat{X}) \le \eta S + \eta T \le \varepsilon.$$

As a consequence, we obtain the conclusion of Theorem 2.

4. Experiments

In this section, we provide empirical evidence to illustrate our proven complexity on both synthetic data and real images. In both examples, we vary ε such that it is small relative to the minimum value of the unregularized UOT function in equation (1) which is computed in advance by using the **cvxpy** library (Agrawal et al., 2018) with the splitting conic solver option. We then report the two k values:

- The first k, denoted by k_f, follows the stopping rule in Algorithm 1.

- The second k, denoted by k_c, is defined as the minimal k_c such that for all later known iterations $k' \ge k_c$ in the experiment, Algorithm 1 returns an ε-approximation solution of the UOT problem.

4.1. Synthetic Data

For the simulated example, we choose $n = 100$ and $\tau = 5$. The elements of the cost matrix C are drawn uniformly from the closed interval $[1, 50]$ while those of the marginal vectors $\mathbf{a}$ and $\mathbf{b}$ are drawn uniformly from $[0.1, 1]$ and then normalized to have masses 2 and 4, respectively. By varying ε from 1.0 to 10^{-4} (here we uniformly vary $\log(\frac{1}{\varepsilon})$ in the corresponding range for visualization purpose), we follow the scheme presented in the beginning of the section, and report values of k_f and k_c in Figure 1.
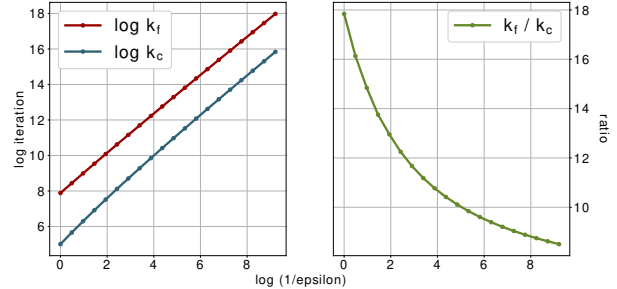

Figure 1. Comparison between k_c and k_f on the synthetic data when varying ε from 1 to 10^{-4} (and η from $3.7(10)^{-2}$ to $3.7(10)^{-6}$ as computed from Theorem 2). The optimal value here is $f(\widehat{X}) = 17.28$, yielding a relative error from $5.8(10)^{-2}$ to $5.8(10)^{-6}$. Both x-axes represent $\log(\frac{1}{\varepsilon})$; the y-axis of the left plot is the natural logarithm of the numbers of iteration (i.e. k_f, k_c) and the y-axis of the right plot is the ratio between them.

Figure 1 shows the log values of k_f, k_c stated above when varying ε. When ε becomes smaller, the left plot indicates that the gap between k_f and k_c becomes narrower, while the right plot shows that the ratio $\frac{k_f}{k_c}$ decreases ($\frac{k_f}{k_c} \approx 18$ at $\varepsilon = 1$, going down to about 8 at $\varepsilon = 10^{-4}$). We hypothesize from this trend that our bound becomes more and more accurate as ε approaches 0.

4.2. MNIST Data

For the MNIST dataset[1], we follow similar settings in (Dvurechensky et al., 2018; Altschuler et al., 2017). In particular, the marginals $\mathbf{a}, \mathbf{b}$ are two flattened images in a pair and the cost matrix C is the matrix of ℓ_1 distances between pixel locations. We also add a small constant 10^{-6} to each pixel with intensity 0, except we do not normalize the marginals. We average the results over 10 randomly chosen image pairs and plot the results in Figure 2. The results on MNIST dataset confirm our theoretical results on the bound of k in Theorem 2. It also shows that the smaller ε in the approximation, the closer the empirical result to the theoretical result.
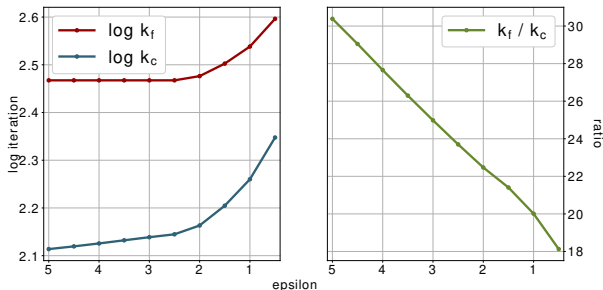

Figure 2. Comparison between $\log(k_f)$ and $\log(k_c)$ on MNIST when varying ε from 5 to 0.5. We used higher ε to keep the relative error similar to the first experiment, due to a higher optimal value (among 10 chosen pairs, the minimum was 117.524 and the maximum was 459.297.)

4.3. A Further Analysis for Synthetic Data

In order to investigate how challenging it is to improve the theoretical bound for the number of required iterations, we carry out a deeper analysis on the synthetic example. In particular, we set $\eta = 0.5$, $\tau = 5$ and compute the ratios $\frac{\|v^k - v^*\|_\infty}{\|u^{k+1} - u^*\|_\infty}$ and $\frac{\|u^{k-1} - u^*\|_\infty}{\|v^k - v^*\|_\infty}$ for even k in range $[0, 100]$ and plot them in Figure 3. As has been proved in Theorem 1, these ratios are no less than $\frac{\tau+\eta}{\tau}$. The main reason for this choice is that these differences are used to construct bounds for many key quantities in lemmas and theorems. These ratios, which are extremely close to 1.1 for most of the iterations, are consistent with the ratio $\frac{\tau+\eta}{\tau} = 1.1$. Consequently, it is difficult to improve our inequalities in Theorem 1.

5. Discussion

In this paper, we prove the near-optimal upper bound of order $\widetilde{\mathcal{O}}(n^2/\varepsilon)$ of the complexity of the Sinkhorn algorithm for approximating the unbalanced optimal transport problem. That complexity is better than the complexity of the
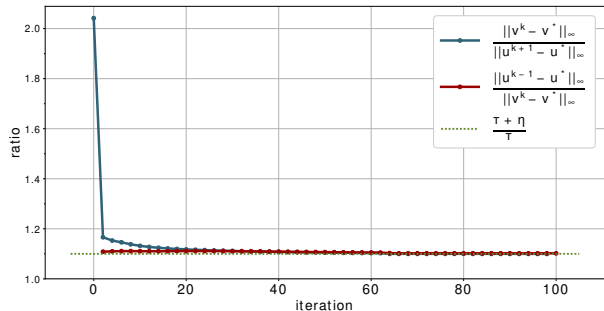
---

[1] http://yann.lecun.com/exdb/mnist/

Figure 3. The ratios as observed empirically remain close to our geometric factor for most of the iterations.

Sinkhorn algorithm for approximating the OT problem. In our analysis, some inequalities might not be tight, since we prefer to keep them in simple forms for easier presentation. These sub-optimalities perhaps lead to the inclusion of the logarithmic terms of ε and n in our complexity upper bound of the Sinkhorn algorithm. We now discuss a few future directions that can serve as natural follow-ups of our work. First, our analysis could be used in the multi-marginal case of UOT by applying Algorithm 1 repeatedly to every pair of marginals. Second, since the UOT barycenter problem has found several applications in recent years (Janati et al., 2019a; Schiebinger et al., 2019), it is desirable to establish the complexity analysis of algorithms for approximating it. Finally, similar to the OT problem, the Sinkhorn algorithm for solving UOT also suffers from the curse of dimensionality, namely, when the supports of the measures lie in high dimensional spaces. An important direction is to study efficient dimension reduction scheme with the UOT problem and optimal algorithms for solving it.

References

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223, 2017.

Blanchet, J., Jambulapati, A., Kent, C., and Sidford, A. Towards optimal running times for optimal transport. *ArXiv Preprint: 1810.07717*, 2018.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F. Scaling algorithms for unbalanced transport problems. *ArXiv Preprint: 1607.05816*, 2016.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International conference on machine learning*, pp. 1367–1376, 2018.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Guo, W., Ho, N., and Jordan, M. I. Accelerated primal-dual coordinate descent for computational optimal transport. *ArXiv Preprint: 1905.09952*, 2019.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In *Proceedings of the International Conference on Machine Learning*, 2017.

Jambulapati, A., Sidford, A., and Tian, K. A direct $\widetilde{O}(1/\varepsilon)$ iteration parallel algorithm for optimal transport. *ArXiv Preprint: 1906.00618*, 2019.

Janati, H., Cuturi, M., and Gramfort, A. Wasserstein regularization for sparse multi-task regression. In *AISTATS*, 2019a.

Janati, H., Cuturi, M., and Gramfort, A. Spatio-temporal alignments: Optimal transport through space and time. *arXiv preprint arXiv:1910.03860*, 2019b.

Lee, J., Bertrand, N. P., and Rozell, C. J. Parallel unbalanced optimal transport regularization for large scale imaging problems. *arXiv preprint arXiv:1909.00149*, 2019.

Lee, Y. T. and Sidford, A. Path finding methods for linear programming: Solving linear programs in $\widetilde{O}(\sqrt{rank})$ iterations and faster algorithms for maximum flow. In *FOCS*, pp. 424–433. IEEE, 2014.

Liero, M., Mielke, A., and Savaré, M. I. Optimal entropy-transport problemsand a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211:969–1117, 2018.

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

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

Pele, O. and Werman, M. Fast and robust earth mover's distance. In *ICCV*. IEEE, 2009.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176:928–943, 2019.

Séjourné, T., Feydy, J., Vialard, F.-X., Trouvé, A., and Peyré, G. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Sinkhorn, R. Diagonal equivalence to matrices with prescribed row and column sums. *Proceedings of the American Mathematical Society*, 45(2):195–198, 1974.

Srivastava, S., Li, C., and Dunson, D. Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018.

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

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