On the Acceleration of the Sinkhorn and Greenkhorn Algorithms for Optimal Transport

Tianyi Lin[‡] Nhat Ho[◇] Michael I. Jordan[◇,†]

Department of Electrical Engineering and Computer Sciences[◇]
Department of Industrial Engineering and Operations Research[‡]
Department of Statistics[†]
University of California, Berkeley

August 4, 2019

Abstract

We propose and analyze a novel approach to accelerate the Sinkhorn and Greenkhorn algorithms for solving the entropic regularized optimal transport (OT) problems. Focusing on the discrete setting where the probability distributions have at most n atoms, and letting $\varepsilon \in (0, 1)$ denote the tolerance, we introduce accelerated algorithms that have complexity bounds of $\widetilde{\mathcal{O}}\left(n^{7/3}\varepsilon^{-1}\right)$. This improves on the best known complexity bound of $\widetilde{\mathcal{O}}\left(n^2\varepsilon^{-2}\right)$ for the Sinkhorn and Greenkhorn algorithms in terms of ε and that of $\widetilde{\mathcal{O}}\left(n^{5/2}\varepsilon^{-1}\right)$ for the practical accelerated first-order primal-dual algorithms in terms of n. We provide an extensive experimental comparison on both synthetic and real datasets to explore the relative advantages of the new algorithms.

1 Introduction

From its origins in work by Monge and Kantorovich in the eighteenth and twentieth centuries, respectively, and through to the present day, the optimal transport (OT) problem has played a determinative role in the theory of optimization [38]. It also has found a wide range of applications in problem domains beyond the original setting in logistics. In the current era, the strong and increasing linkage between optimization and machine learning has brought new applications of OT to the fore; see for example, [4, 8, 30, 31, 37, 12]. In these applications, the focus is on the probability distributions underlying the OT formulation. These distributions are generally either empirical distributions, obtained by placing unit masses at data points, or are probability models of a putative underlying data-generating process. The OT problem accordingly often has a direct inferential meaning—as the definition of an estimator, the definition of a likelihood, or as a robustification of an estimator. The key challenge is computational. Indeed, in machine learning applications the underlying distributions generally involve high-dimensional data sets and complex probability models.

We study the OT problem in a discrete setting, where we assume that the target and source probability distributions each have at most n atoms. In this setting, the benchmark methods for solving OT problems are interior-point methods, reflecting the linear-programming formulation of the OT problem. A specialized interior-point method [32] delivers a complexity bound of $\widetilde{\mathcal{O}}\left(n^3\right)$. Lee and Sidford [24] have improved this to $\widetilde{\mathcal{O}}\left(n^{5/2}\right)$ via an appeal to Laplacian linear system algorithms. Neither method, however, provides an effective practical solution to large-scale machine learning problems; the former because of scalability issues and the latter because efficient practical implementations of Laplacian approach are yet unknown.

Cuturi [9] initiated a productive line of research in which he used an entropic regularizer to replace the nonnegative constraints in the transportation plan. This OT problem is referred to as *entropic regularized optimal transport* or *regularized OT*. The key advantage of regularized OT is that its dual representation has structure that can be exploited computationally. In particular, [9] showed that a dual coordinate ascent algorithm for solving regularized OT is equivalent to the celebrated *Sinkhorn algorithm* [35, 22, 20, 7]. Further progress in this vein was presented by [3], who proposed and analyzed a greedy alternative to the Sinkhorn algorithm that they referred to as the *Greenkhorn algorithm*. The best known complexity bounds shown by [13, 26] are $\widetilde{\mathcal{O}}\left(n^2\varepsilon^{-2}\right)$ for both Sinkhorn and Greenkhorn algorithms, which remain the current baseline solution methods in practice [15].

Further progress has been made by considering other algorithmic procedure for the OT problem [13, 26, 17, 19, 11, 6, 16, 1, 2, 23]. While the primal-dual schemes along with gradient descent [13], mirror descent [26] and coordinate descent [17] all lead to the complexity bound $\widetilde{\mathcal{O}}\left(n^{5/2}\varepsilon^{-1}\right)$, Jambulapati et.al. [19] has designed an algorithm with the complexity bound $\widetilde{\mathcal{O}}\left(n^2\varepsilon^{-1}\right)$ by incorporating the dual extrapolation framework with area-convex mirror mapping [34]. This complexity bound is believed to be optimal in [5] and also achieved by some black-box algorithms [5, 33] and a specialized graph algorithm [23]. Despite the better theoretical complexity bound, the lack of the simplicity and ease-of-implementation makes these algorithms less competitive with Sinkhorn and Greenkhorn algorithms in practice.

Another line of related work builds on Nesterov [29], who developed a randomized coordinate-descent algorithm with an overall iteration complexity of $\mathcal{O}(\varepsilon^{-1/2})$ in terms of the convex objective gap. Subsequently, several researchers extended Nesterov's technique and analysis to a variety of other problem settings [28, 14, 25]. Very recently, Lu et al. [27] has shown that a novel variant of accelerated greedy coordinate descent algorithm also achieves the improved complexity bound of $\mathcal{O}(\varepsilon^{-1/2})$.

In this paper, we show that the Sinkhorn and Greenkhorn algorithms can be accelerated directly by firstly considering the monotone constant step scheme of accelerated randomized coordinate descent. The resulting complexity bound is commensurate with the more specialized acceleration techniques. More specifically, we develop an accelerated randomized scheme for the Sinkhorn algorithm, which we refer to as the *Randkhorn algorithm*. The Randkhorn algorithm involves exact minimization for the main iterates accompanied by an auxiliary sequence of iterates that are based on a randomized coordinate gradient update and a monotone constant step scheme. We establish the complexity bound of $\widetilde{\mathcal{O}}\left(n^{7/3}\varepsilon^{-1}\right)$ for the Randkhorn algorithm, which is better than the complexity bound of $\widetilde{\mathcal{O}}\left(n^2\varepsilon^{-2}\right)$ achieved by the Sinkhorn algorithm in terms of ε and that of $\widetilde{\mathcal{O}}\left(n^{5/2}\varepsilon^{-1}\right)$ achieved by the accelerated first-order primal-dual algorithms in terms of the number of atoms n. We also accelerate the Greenkhorn algorithm, yielding an algorithm that we refer to as the *Gandkhorn algorithm*, and obtain the same improved complexity bound.

Organization. The remainder of the paper is organized as follows. In Section 2 we present the formulation of entropic regularized OT and its dual form. We discuss the properties of optimal solutions of these objective functions. In Section 3, we derive the Randkhorn algorithm and establish its complexity bound. We turn to the Gandkhorn algorithm and its theoretical guarantee in Section 4. Extensive simulation studies of these algorithms with both synthetic and real data are presented in Section 5. We conclude in Section 6 and defer the proof of remaining key results in the paper to Appendix A.

Notation. For any $n \geq 2$, we start with Δ^n a probability simplex in $n-1$ dimensions, namely $\Delta^n := \{v = (v_1, \ldots, v_n) \in \mathbb{R}^n : \sum_{i=1}^n v_i = 1,\ v \geq 0\}$. For $x \in \mathbb{R}^n$ and $1 \leq p \leq \infty$, the notation $\|x\|_p$ stands for ℓ_p-norm while $\|x\|$ indicates an ℓ_2-norm. Furthermore, we define $[n] := \{1, 2, \ldots, n\}$. For any $n \geq 1$, $\mathbb{R}^n_+$ is the set of all vectors in the space $\mathbb{R}^n$ with nonnegative coordinates. The notation $\text{diag}(x)$ is standard diagonal matrix whose has the vector x on its diagonal. The notation $\mathbf{1}$ is a vector with all components take value 1. The notation $\nabla_x f$ denotes a partial derivative of f in terms of x. Finally, for any dimension n and desired accuracy ε, two notation $a = \mathcal{O}\left(b(n, \varepsilon)\right)$ and $a = \Omega\left(b(n, \varepsilon)\right)$ respectively indicate the upper and lower bounds $a \leq C_1 \cdot b(n, \varepsilon)$ and $a \geq C_2 \cdot b(n, \varepsilon)$, where C_1 and C_2 are independent of n and ε. Given these notation, $a = \Theta\left(b(n, \varepsilon)\right)$ if and only if $a = \mathcal{O}\left(b(n, \varepsilon)\right)$ and $a = \Omega\left(b(n, \varepsilon)\right)$. Similarly, we denote $a = \widetilde{\mathcal{O}}(b(n, \varepsilon))$ to indicate that the inequality with $\mathcal{O}\left(b(n, \varepsilon)\right)$ may depend on some logarithmic function of both n and ε.

2 Problem Setup

In this section, we provide some background on the problem of computing the OT distance between two discrete probability measures with at most n atoms. In particular, we discuss the *entropic regularized OT* problem and its dual formulation.

2.1 Entropic regularized OT

According to [21], the problem of approximating the optimal transportation distance is equivalent to solving the following linear programming problem:

$$\min_{X \in \mathbb{R}^{n \times n}} \langle C, X \rangle \qquad \text{s.t.} \;\; X\mathbf{1} = r, \;\; X^\top \mathbf{1} = l, \;\; X \geq 0, \tag{1}$$

where X refers to the *transportation plan*, $C = (C_{ij}) \in \mathbb{R}^{n \times n}_+$ stands for a cost matrix with nonnegative components, and r and l refer to two known probability distributions in the simplex Δ^n. The goal of the paper is to find a transportation plan $\hat{X} \in \mathbb{R}^{n \times n}_+$ satisfying marginal distribution constraints $\hat{X}\mathbf{1} = r$ and $\hat{X}^\top \mathbf{1} = l$ and the following bound

$$\langle C, \hat{X} \rangle \;\leq\; \langle C, X^* \rangle + \varepsilon. \tag{2}$$

Here, X^* is defined as an optimal transportation plan for the OT problem (1). For the sake of presentation, we respectively denote $\langle C, \hat{X} \rangle$ an ε-*approximation* and $\hat{X}$ an ε-*approximate transportation plan* for the original optimal transportation distance.

Since problem (1) is a linear programming problem, we can solve it by means of the interior-point method; however, this method performs poorly on large-scale problems due to its high per-iteration computational cost. Seeking a formulation for OT distance that is more amenable to computationally efficient algorithms, Cuturi [9] proposed to solve an entropic regularized version of the OT problem (1), which is given by

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times n}} \quad & \langle C, X \rangle - \eta H(X) \\ \text{s.t.} \quad & X\mathbf{1} = r, \;\; X^\top \mathbf{1} = l. \end{aligned} \tag{3}$$

Here, $\eta > 0$ in the above display stands for the *regularization parameter* while $H(X)$ refers to an entropic regularization admitting the following formulation

$$H(X) \;=\; -\sum_{i,j=1}^{n} X_{ij} \log(X_{ij}). \tag{4}$$

Altschuler et al. [3] have shown that an ε-approximate transportation plan can be obtained by solving (3) with $\eta = \frac{\varepsilon}{4\log(n)}$.

It is clear that the entropic regularized OT problem (3) is a convex optimization problem with affine constraints. We demonstrate below that its dual problem is in fact an unconstrained optimization problem. Such a nice structure of this dual problem is favorable to not only the algorithmic development but also the theoretical complexity analysis of algorithms. Simple algebra indicate that the Lagrangian function admits the following formulation

$$\mathcal{L}(X, \alpha, \beta) \;=\; \langle \alpha, r\rangle + \langle \beta, l\rangle + \langle C, X\rangle - \eta H(X) - \langle \alpha, X\mathbf{1}\rangle - \langle \beta, X^\top \mathbf{1}\rangle.$$

To obtain a dual form of entropic regularized OT, we need to solve $\min_{X\in\mathbb{R}^{n\times n}} \mathcal{L}(X, \alpha, \beta)$. Since the Lagrangian function $\mathcal{L}(\cdot, \alpha, \beta)$ is both strictly convex and differentiable, we can solve the previous optimization problem by setting $\partial_X \mathcal{L}(X, \alpha, \beta) = 0$. It is equivalent to the following equations

$$C_{ij} + \eta\left(1 + \log(X_{ij})\right) - \alpha_i - \beta_j = 0, \quad \forall i, j \in [n].$$

The above equations lead to the following value of transportation plan X:

$$X_{ij} \;=\; e^{\frac{-C_{ij}+\alpha_i+\beta_j}{\eta}-1}, \quad \forall i, j \in [n].$$

We substitute this solution into the Lagrangian function and define

$$\varphi(\alpha, \beta) \;:=\; \min_{X\in\mathbb{R}^{n\times n}} \mathcal{L}(X, \alpha, \beta) \;=\; -\eta \sum_{i,j=1}^{n} e^{-\frac{C_{ij}-\alpha_i-\beta_j}{\eta}-1} + \langle \alpha, r\rangle + \langle \beta, l\rangle\,. \tag{5}$$

To further simplify the notation, we set $u_i = \frac{\alpha_i}{\eta} - \frac{1}{2}\mathbf{1}$ and $v_j = \frac{\beta_j}{\eta} - \frac{1}{2}\mathbf{1}$, which yields a new form of function φ as follows:

$$\varphi(u, v) \;=\; \eta\left(-\sum_{i,j=1}^{n} e^{-\frac{C_{ij}}{\eta}+u_i+v_j} + \langle u, r\rangle + \langle v, l\rangle + 1\right).$$

Letting $B(u, v) := \mathrm{diag}(e^u)\; e^{-\frac{C}{\eta}}\; \mathrm{diag}(e^v)$, the dual problem $\max_{u,v\in\mathbb{R}^n} \varphi(u, v)$ reduces to

$$\min_{u,v\in\mathbb{R}^n}\; f(u, v) \;:=\; \mathbf{1}^\top B(u, v)\mathbf{1} - \langle u, r\rangle - \langle v, l\rangle. \tag{6}$$

The problem (6) is called the *dual (entropic) regularized OT* problem. We denote (u^*, v^*) the optimal solution of this problem.

2.2 Some key properties

We notice that problem (3) is a special case of the following problem:

$$\min_{x\in\mathbb{R}^n} f(x), \quad \text{s.t. } Ax = b, \tag{7}$$

where $\|A\|_1 = 2$ and f is strongly convex with respect to the ℓ_1-norm:

$$f(x_2) - f(x_1) - \langle \nabla f(x_1), x_2 - x_1\rangle \geq \frac{\eta}{2}\left\|x_2 - x_1\right\|_1^2.$$

By [26, Lemma 4.1], the dual objective φ satisfies the following inequality with $\lambda_i = (\alpha_i, \beta_i)$,

$$\varphi(\alpha_1, \beta_1) - \varphi(\alpha_2, \beta_2) - \left\langle \nabla\varphi(\alpha_2, \beta_2), \begin{pmatrix}\alpha_1 \\ \beta_1\end{pmatrix} - \begin{pmatrix}\alpha_2 \\ \beta_2\end{pmatrix} \right\rangle \leq \frac{\|A\|_1^2}{2\eta} \left\| \begin{pmatrix}\alpha_1 \\ \beta_1\end{pmatrix} - \begin{pmatrix}\alpha_2 \\ \beta_2\end{pmatrix} \right\|_\infty^2.$$

Setting $u_i = \frac{\alpha_i}{\eta} - \frac{1}{2}\mathbf{1}$ and $v_j = \frac{\beta_j}{\eta} - \frac{1}{2}\mathbf{1}$, we have

$$\begin{aligned} f(u_1, v_1) - f(u_2, v_2) - \left\langle \nabla f(u_2, v_2), \begin{pmatrix}u_1 \\ v_1\end{pmatrix} - \begin{pmatrix}u_2 \\ v_2\end{pmatrix} \right\rangle &\leq \frac{\eta \|A\|_1^2}{2} \left\| \begin{pmatrix}u_1 \\ v_1\end{pmatrix} - \begin{pmatrix}u_2 \\ v_2\end{pmatrix} \right\|_\infty^2 \\ &\leq 2\eta \left\| \begin{pmatrix}u_1 \\ v_1\end{pmatrix} - \begin{pmatrix}u_2 \\ v_2\end{pmatrix} \right\|^2. \end{aligned} \tag{8}$$

Therefore, the objective in the dual entropic regularized OT (6) is smooth with respect to the ℓ_2-norm and the Lipschitz constant is 2η. This further implies the following relationship between the norm of the gradient and the objective gap.

Lemma 2.1. *For any optimal solution (u^*, v^*) of the dual regularized OT problem in (6), we have*

$$\|\nabla f(u, v)\|^2 \leq 8\eta \left(f(u, v) - f(u^*, v^*)\right).$$

Proof. Let $(u_1, v_1) = (u, v) - \frac{1}{4\eta}\nabla f(u, v)$ and $\tau_2 = (u, v)$ in (8) , we have

$$\begin{aligned} f\left((u, v) - \frac{1}{4\eta}\nabla f(u, v)\right) &\leq f(u, v) - \left\langle \nabla f(u, v), \frac{1}{4\eta}\nabla f(u, v) \right\rangle + 2\eta \left\| \frac{1}{4\eta}\nabla f(u, v) \right\|^2 \\ &= f(u, v) - \frac{1}{8\eta} \|\nabla f(u, v)\|^2. \end{aligned}$$

Since τ^* is an optimal solution, we must have $f\left((u, v) - \frac{1}{4\eta}\nabla f(u, v)\right) \geq f(u^*, v^*)$. Putting these pieces together yields the desired inequality. □

Finally, we present an upper bound for an optimal solution to the dual regularized OT problem (6).

Lemma 2.2. *For the dual regularized OT problem in (6), there exists an optimal solution (u^*, v^*) such that*

$$\|u^*\| \leq \sqrt{n}R, \qquad \|v^*\| \leq \sqrt{n}R, \tag{9}$$

where $R > 0$ is defined as

$$R := \frac{\|C\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1 \leq i,j \leq n} \{r_i, l_j\}\right).$$

Proof. By [26, Lemma 3.2], it holds that $\|u^*\|_\infty \leq R$ and $\|v^*\|_\infty \leq R$. Therefore, the desired results follow from the definition of ℓ_2-norm and ℓ_∞-norm. □

3 Randkhorn: Accelerated Randomized Sinkhorn Algorithm

In this section, we present the Randkhorn algorithm and its complexity analysis. The key idea behind the algorithm is to interpret the Sinkhorn algorithm as a block coordinate descent algorithm for the dual regularized OT problem (6). Then, we improve the algorithm by incorporating an estimated sequence. The complexity analysis for the Randkhorn algorithm yields a complexity bound of $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$, which improves on the best known complexity bound $\mathcal{O}\left(\frac{n^2\|C\|_\infty^2\log(n)}{\varepsilon^2}\right)$ for the Sinkhorn algorithm [13] in terms of desired accuracy ε and the complexity bound of $\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty\sqrt{\log(n)}}{\varepsilon}\right)$ for the accelerated first-order primal-dual algorithms [26] in terms of the number of atoms n. To ease the ensuing discussion, we present the pseudocode of Randkhorn algorithm in Algorithm 1 and its application to regularized OT in Algorithm 2.

Similar to the Sinkhorn algorithm, the Randkhorn algorithm can be viewed as an accelerated randomized coordinate descent algorithm for the dual regularized OT problem (6). More specifically, the update for the main iterates, (u, v), is an exact minimization (cf. Step 2 in the algorithm) while that for the estimated iterates, $(\tilde{u}, \tilde{v})$, is based on randomized coordinate gradient (cf. Step 3 in the algorithm). This is in contrast to existing accelerated randomized algorithms, which are based purely on the coordinate gradient updates [29, 28, 14, 25, 27]. Quantifying the per-iteration progress of the Randkhorn algorithm accordingly turns out to be more challenging than that of other accelerated randomized coordinate descent algorithms, and we needed to improve the current proof techniques by further exploiting the problem structure of dual regularized OT.

The presentation of the Randkhorn algorithm in Algorithm 1 makes use of a function $\rho : \mathbb{R}_+^n \times \mathbb{R}_+^n \to [0, +\infty]$ given by:

$$\rho(a, b) \ := \ \mathbf{1}^\top (b - a) + \sum_{i=1}^n a_i \log\left(\frac{a_i}{b_i}\right).$$

The function ρ measures the progress in the dual regularized OT objective (6) between two consecutive iterates of the Randkhorn algorithm. It is easy to check that:

$$\rho(a, b) \ \geq \ 0, \quad \text{for all } a, b \in \mathbb{R}_+^n,$$

with equality holding true if and only if $a = b$.

The optimality condition for the dual regularized OT problem (6) is given by:

$$B(u, v)\mathbf{1} - r \ = \ 0, \qquad B(u, v)^\top \mathbf{1} - l \ = \ 0.$$

This suggests that a natural quantity to measure the error of the k-th iterate of the Randkhorn algorithm as follows:

$$E^k \ := \ \mathbb{E}\left[\|B(u^k, v^k)\mathbf{1} - r\|_1 + \|B(u^k, v^k)^\top \mathbf{1} - l\|_1\right], \tag{10}$$

where the expectation is taken with respect to the Bernoulli distributions in Step 4 of Randkhorn algorithm. Finally, we show how to apply the Randkhorn algorithm to regularized OT in Algorithm 2, where we have made use of standard parameter settings from [3]. More specifically, we set $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$ and

$$\left(\tilde{r}, \tilde{l}\right) = \left(1 - \frac{\varepsilon'}{8}\right)(r, l) + \frac{\varepsilon'}{8n}(\mathbf{1}, \mathbf{1}),$$

which are supported by the complexity analysis in the sequel.

Algorithm 1: RANDKHORN$(C, \eta, r, l, \varepsilon')$

Input: $k = 0$, $\theta_{-1} \in (0, 2]$ and $y_u^0 = y_v^0 = \tilde{u}_0 = \tilde{v}_0 = \mathbf{0}$.

while $E^k > \varepsilon'$ **do**

Step 1: $\theta_k = \frac{\theta_{k-1}}{2}\left(\sqrt{\theta_{k-1}^2 + 4} - \theta_{k-1}\right)$.

Step 2: $\begin{pmatrix} \bar{u}^k \\ \bar{v}^k \end{pmatrix} = (1 - \theta_k)\begin{pmatrix} y_u^k \\ y_v^k \end{pmatrix} + \theta_k \begin{pmatrix} \tilde{u}^k \\ \tilde{v}^k \end{pmatrix}$.

Step 3: $r(\bar{u}^k, \bar{v}^k) = B(\bar{u}^k, \bar{v}^k)\mathbf{1}$ and $l(\bar{u}^k, \bar{v}^k) = B(\bar{u}^k, \bar{v}^k)^\top \mathbf{1}$.

if $\rho\left(r, r(\bar{u}^k, \bar{v}^k)\right) \geq \rho\left(l, l(\bar{u}^k, \bar{v}^k)\right)$ **then**

$\hat{u}^{k+1} = \bar{u}^k + \log(r) - \log\left(r(\bar{u}^k, \bar{v}^k)\right)$ and $\hat{v}^{k+1} = \bar{v}^k$.

else

$\hat{v}^{k+1} = \bar{v}^k + \log(l) - \log\left(l(\bar{u}^k, \bar{v}^k)\right)$ and $\hat{u}^{k+1} = \bar{u}^k$.

end if

Step 4: Randomly sample $\xi^k \sim \text{Bernoulli}(\frac{1}{2})$, a Bernoulli random variable with parameter $\frac{1}{2}$.

if $\xi^k = 0$ **then**

$\tilde{u}^{k+1} = \tilde{u}^k - \dfrac{r(\bar{u}^k, \bar{v}^k) - r}{8\theta_k \eta}$ and $\tilde{v}^{k+1} = \tilde{v}^k$.

else

$\tilde{v}^{k+1} = \tilde{v}^k - \dfrac{l(\bar{u}^k, \bar{v}^k) - l}{8\theta_k \eta}$ and $\tilde{u}^{k+1} = \tilde{u}^k$.

end if

Step 5: $\begin{pmatrix} u^k \\ v^k \end{pmatrix} = \text{argmin}\left\{ f(u, v) \mid \begin{pmatrix} u \\ v \end{pmatrix} \in \left\{ \begin{pmatrix} y_u^k \\ y_v^k \end{pmatrix}, \begin{pmatrix} \hat{u}^{k+1} \\ \hat{v}^{k+1} \end{pmatrix} \right\} \right\}$.

Step 6: $r(u^k, v^k) = B(u^k, v^k)\mathbf{1}$ and $l(u^k, v^k) = B(u^k, v^k)^\top \mathbf{1}$.

if $\rho\left(r, r(u^k, v^k)\right) \geq \rho\left(l, l(u^k, v^k)\right)$ **then**

$y_u^{k+1} = u^k + \log(r) - \log\left(r(u^k, v^k)\right)$ and $y_v^{k+1} = v^k$.

else

$y_v^{k+1} = v^k + \log(l) - \log\left(l(u^k, v^k)\right)$ and $y_u^{k+1} = u^k$.

end if

Step 7: $k = k + 1$.

end while

Output: $B(u^k, v^k)$.

3.1 Technical lemmas

In this section, we provide several technical lemmas for bounding the dual objective gap δ_k in the Randkhorn algorithm: $\delta_k := \mathbb{E}[f(y_u^k, y_v^k) - f(u^*, v^*)]$ where (y_u^k, y_v^k) are defined in Algorithm 1. Our analysis hinges upon two key sequences of iterates. The first sequence is obtained by performing a full gradient descent step with a step size $1/(8\eta)$ and a starting point $(\bar{u}^k, \bar{v}^k)^\top$:

$$\begin{pmatrix} s_u^{k+1} \\ s_v^{k+1} \end{pmatrix} := \begin{pmatrix} \bar{u}^k \\ \bar{v}^k \end{pmatrix} - \frac{1}{8\eta}\nabla f(\bar{u}^k, \bar{v}^k). \tag{11}$$

The second sequence is obtained, on the other hand, by taking a full gradient descent step with a different step size $1/(8\eta\theta_k)$, where θ_k is given in Step 1 of Algorithm 1, and making use

Algorithm 2: Approximating OT by RANDKHORN

Input: Regularization parameter $\eta = \frac{\varepsilon}{4\log(n)}$ and error $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.

Step 1: Choose $\tilde{r} \in \Delta_n$ and $\tilde{l} \in \Delta_n$ as

$$\left(\tilde{r}, \tilde{l}\right) = \left(1 - \frac{\varepsilon'}{8}\right)(r, l) + \frac{\varepsilon'}{8n}(\mathbf{1}, \mathbf{1}).$$

Step 2: Compute $\tilde{X} = \text{RANDKHORN}\left(C, \eta, \tilde{r}, \tilde{l}, \varepsilon'/2\right)$

Step 3: Given Algorithm 2 in [3], we round $\tilde{X}$ to $\hat{X}$ to satisfy $\hat{X}\mathbf{1} = r$ and $\hat{X}^\top \mathbf{1} = l$.

Output: $\hat{X}$.

of a different starting point $(\tilde{u}^k, \tilde{v}^k)^\top$:

$$\begin{pmatrix} \tilde{s}_u^{k+1} \\ \tilde{s}_v^{k+1} \end{pmatrix} \ := \ \begin{pmatrix} \tilde{u}^k \\ \tilde{v}^k \end{pmatrix} - \frac{1}{8\eta\theta_k}\nabla f(\bar{u}^k, \bar{v}^k). \tag{12}$$

Given the definition of these two sequences, we first establish a key descent inequality regarding the values of dual regularized OT at Randkhorn updates.

Lemma 3.1. *For each iteration $k > 0$ of the Randkhorn algorithm, we have*

$$\begin{aligned} f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ &\leq \ \theta_k f(u^*, v^*) + (1-\theta_k) f(y_u^k, y_v^k) \\ &+ \ 4\eta\theta_k^2 \left(\left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix} \right\|^2 \right). \end{aligned} \tag{13}$$

Proof. We claim that the following inequalities hold:

$$f(\bar{u}^k, \bar{v}^k) - f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ \geq \ \frac{1}{2}\left(\rho\left(r, r(\bar{u}^k, \bar{v}^k)\right) + \rho\left(l, l(\bar{u}^k, \bar{v}^k)\right)\right), \tag{14}$$

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ \leq \ f(\bar{u}^k, \bar{v}^k) - \frac{1}{16\eta}\|\nabla f(\bar{u}^k, \bar{v}^k)\|^2. \tag{15}$$

Assuming that these inequalities hold for the moment, we invoke the definition of s_u^{k+1} and s_v^{k+1} in (11) and obtain the following equations:

$$\begin{aligned} \|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 \ &= \ 2\|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 - \|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 \\ &\overset{(11)}{=} \ 16\eta \begin{pmatrix} \bar{u}^k - s_u^{k+1} \\ \bar{v}^k - s_v^{k+1} \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) - 16\eta \left\| \begin{pmatrix} \bar{u}^k - s_u^{k+1} \\ \bar{v}^k - s_v^{k+1} \end{pmatrix} \right\|^2. \end{aligned}$$

Plugging this equality into (15) and rearranging yields the following inequality:

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ \leq \ f(\bar{u}^k, \bar{v}^k) + \begin{pmatrix} s_u^{k+1} - \bar{u}^k \\ s_v^{k+1} - \bar{v}^k \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) + 4\eta \left\| \begin{pmatrix} s_u^{k+1} - \bar{u}^k \\ s_v^{k+1} - \bar{v}^k \end{pmatrix} \right\|^2. \tag{16}$$

Furthermore, based on the definition of $\tilde{s}_u^{k+1}$ and $\tilde{s}_v^{k+1}$ in (12), we have

$$\begin{pmatrix} s_u^{k+1} - \bar{u}^k \\ s_v^{k+1} - \bar{v}^k \end{pmatrix} \ = \ \theta_k \begin{pmatrix} \tilde{s}_u^{k+1} - \tilde{u}^k \\ \tilde{s}_v^{k+1} - \tilde{v}^k \end{pmatrix}.$$

Plugging into (16) yields:

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) \;\leq\; f(\bar{u}^k, \bar{v}^k) + \theta_k \left[\begin{pmatrix} \tilde{s}_u^{k+1} - \tilde{u}^k \\ \tilde{s}_v^{k+1} - \tilde{v}^k \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) + 4\eta\theta_k \left\| \begin{pmatrix} \tilde{s}_u^{k+1} - \tilde{u}^k \\ \tilde{s}_v^{k+1} - \tilde{v}^k \end{pmatrix} \right\|^2 \right]. \quad (17)$$

Invoking the definition of $\tilde{s}_u^{k+1}$ and $\tilde{s}_v^{k+1}$ again, we find that

$$\begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix}^\top \left[\begin{pmatrix} \tilde{s}_u^{k+1} - \tilde{u}^k \\ \tilde{s}_v^{k+1} - \tilde{v}^k \end{pmatrix} + \frac{1}{8\eta\theta_k} \nabla f(\bar{u}^k, \bar{v}^k) \right] \;=\; 0.$$

Rearranging the terms yields:

$$\frac{1}{4\eta\theta_k} \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) \;=\; \left\| \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \tilde{s}_u^{k+1} - \tilde{u}^k \\ \tilde{s}_v^{k+1} - \tilde{v}^k \end{pmatrix} \right\|^2. \quad (18)$$

Finally, by plugging the result from (18) into (17), we arrive at the following:

$$\begin{aligned} & f(\hat{u}^{k+1}, \hat{v}^{k+1}) && (19) \\ \leq \;& f(\bar{u}^k, \bar{v}^k) + \theta_k \left[\begin{pmatrix} \tilde{s}_u^{k+1} - \tilde{u}^k \\ \tilde{s}_v^{k+1} - \tilde{v}^k \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) + \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) \right] \\ + \;& 4\eta\theta_k^2 \left(\left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix} \right\|^2 \right) \\ = \;& f(\bar{u}^k, \bar{v}^k) + \theta_k \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) + 4\eta\theta_k^2 \left(\left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix} \right\|^2 \right). \end{aligned}$$

By simple algebra, we can check that

$$\theta_k \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \;=\; \theta_k \begin{pmatrix} u^* - \bar{u}^k \\ v^* - \bar{v}^k \end{pmatrix} + (1 - \theta_k) \begin{pmatrix} y_u^k - \bar{u}^k \\ y_v^k - \bar{v}^k \end{pmatrix}.$$

Thus, we obtain that

$$f(\bar{u}^k, \bar{v}^k) + \theta_k \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) \;\leq\; \theta_k f(u^*, v^*) + (1 - \theta_k) f(y_u^k, y_v^k). \quad (20)$$

By plugging the result from (20) into (19), we obtain the desired inequality (13), which proves the lemma.

Proof of claim (14): Without loss of generality, we assume that $\rho\left(r, r(\bar{u}^k, \bar{v}^k)\right) \geq \rho\left(l, l(\bar{u}^k, \bar{v}^k)\right)$ as the proof argument for the other case is similar. Given that assumption, we have $\hat{u}^{k+1} = \bar{u}^k + \log(r) - \log\left(r(\bar{u}^k, \bar{v}^k)\right)$ and $\hat{v}^{k+1} = \bar{v}^k$. This leads to the following equation

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) \;=\; 1 - \langle \bar{u}^k, r \rangle - \langle \bar{v}^k, l \rangle - \sum_{i=1}^n r_i \log\left(\frac{r_i}{r_i(\bar{u}^k, \bar{v}^k)} \right).$$

Furthermore, we have

$$f(\bar{u}^k, \bar{v}^k) \;=\; \mathbf{1}^\top r(\bar{u}^k, \bar{v}^k) - \langle \bar{u}^k, r \rangle - \langle \bar{v}^k, l \rangle.$$

Therefore, we conclude that

$$\begin{aligned} f(\bar{u}^k, \bar{v}^k) - f(\hat{u}^{k+1}, \hat{v}^{k+1}) &= \mathbf{1}^\top \left(r(\bar{u}^k, \bar{v}^k) - r\right) + \sum_{i=1}^n r_i \log\left(\frac{r_i}{r_i(\bar{u}^k, \bar{v}^k)}\right) \\ &= \rho\left(r, r(\bar{u}^k, \bar{v}^k)\right). \end{aligned}$$

Combining with the assumption that $\rho\left(r, r(\bar{u}^k, \bar{v}^k)\right) \geq \rho\left(l, l(\bar{u}^k, \bar{v}^k)\right)$, we obtain the desired inequality (14).

Proof of claim (15): By the definition of $\rho\left(r, r(\bar{u}^k, \bar{v}^k)\right)$ and $\rho\left(l, l(\bar{u}^k, \bar{v}^k)\right)$, we have

$$\begin{aligned} \rho\left(r, r(\bar{u}^k, \bar{v}^k)\right) &= f(\bar{u}^k, \bar{v}^k) - \operatorname*{argmin}_{u \in \mathbb{R}^n} f(u, \bar{v}^k) \;\geq\; f(\bar{u}^k, \bar{v}^k) - f\left(\bar{u}^k - \frac{1}{4\eta}\nabla_u f(\bar{u}^k, \bar{v}^k), \bar{v}^k\right) \\ \rho\left(l, l(\bar{u}^k, \bar{v}^k)\right) &= f(\bar{u}^k, \bar{v}^k) - \operatorname*{argmin}_{v \in \mathbb{R}^n} f(\bar{u}^k, v) \;\geq\; f(\bar{u}^k, \bar{v}^k) - f\left(\bar{u}^k, \bar{v}^k - \frac{1}{4\eta}\nabla_u f(\bar{u}^k, \bar{v}^k)\right). \end{aligned}$$

Applying (8) yields that

$$\begin{aligned} f(\bar{u}^k, \bar{v}^k) - f\left(\bar{u}^k - \frac{1}{4\eta}\nabla_u f(\bar{u}^k, \bar{v}^k), \bar{v}^k\right) &\geq \frac{1}{8\eta}\|\nabla_u f(\bar{u}^k, \bar{v}^k)\|^2, \\ f(\bar{u}^k, \bar{v}^k) - f\left(\bar{u}^k, \bar{v}^k - \frac{1}{4\eta}\nabla_u f(\bar{u}^k, \bar{v}^k)\right) &\geq \frac{1}{8\eta}\|\nabla_v f(\bar{u}^k, \bar{v}^k)\|^2. \end{aligned}$$

Combining with (14), we achieve the conclusion of claim (15). □

We are now ready to bound the dual objective gap δ^k.

Lemma 3.2. *For the iterates $\{(u^k, v^k)\}_{k\geq 0}$ returned by the Randkhorn algorithm, we have*

$$\delta_k \;\leq\; \frac{\left(32 + 8/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(k+1)^2}. \tag{21}$$

Proof. We claim that we can replace $(\tilde{s}_u^{k+1}, \tilde{s}_v^{k+1})^\top$ by $(\tilde{u}^{k+1}, \tilde{v}^{k+1})^\top$ on the right-hand side of inequality (13) as follows:

$$\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \left\|\begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix}\right\|^2 \;=\; 2\left(\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \mathbb{E}_{\xi^k}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right). \tag{22}$$

Assume that this claim is true for the moment. Putting together Lemma 3.1 and equality (22) leads to the following inequality

$$\begin{aligned} f(\hat{u}^{k+1}, \hat{v}^{k+1}) \;\leq\; & \theta_k f(u^*, v^*) + (1-\theta_k) f(y_u^k, y_v^k) \\ & + 8\eta\theta_k^2\left(\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \mathbb{E}_{\xi^k}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right). \end{aligned}$$

Subtracting $f(u^*, v^*)$ from both sides and taking an expectation with respect to $\{\xi^j\}_{j=1}^{k-1}$ yields

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) - f(u^*, v^*) \;\leq\; (1-\theta_k)\delta^k + 8\eta\theta_k^2\left(\mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2\right] - \mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right).$$

Furthermore, by the definition of $\begin{pmatrix} u^k \\ v^k \end{pmatrix}$ and $\begin{pmatrix} y_u^{k+1} \\ y_v^{k+1} \end{pmatrix}$, we have

$$f(y_u^{k+1}, y_v^{k+1}) \;\leq\; f(u^k, v^k) \;\leq\; f(\hat{u}^{k+1}, \hat{v}^{k+1}).$$

The above result implies that

$$\delta^{k+1} \;\leq\; (1-\theta_k)\delta^k + 8\eta\theta_k^2 \left(\mathbb{E}\left[\left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 \right] - \mathbb{E}\left[\left\| \begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix} \right\|^2 \right] \right).$$

Since θ_k satisfies $\theta_{k+1} = \frac{\theta_k}{2}\left(\sqrt{\theta_k^2+4} - \theta_k\right)$, we obtain that $\frac{1}{\theta_{k-1}^2} \;=\; \frac{1-\theta_k}{\theta_k^2}$ and

$$\frac{\delta_{k+1}}{\theta_k^2} \;\leq\; \frac{\delta_k}{\theta_{k-1}^2} + 8\eta \left(\mathbb{E}\left[\left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 \right] - \mathbb{E}\left[\left\| \begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix} \right\|^2 \right] \right).$$

Changing the count k to i and summing the inequality over $i = 0, 1, \ldots, k-1$, we obtain:

$$\delta_k \;\leq\; \theta_{k-1}^2 \left(\frac{\delta_0}{\theta_{-1}^2} + 8\eta \left(\|u^*\|^2 + \|v^*\|^2 \right) \right).$$

Furthermore, since f is smooth with respect to ℓ_2-norm (cf. inequality (8)), we have

$$\delta_0 \;\leq\; 2\eta \left\| \begin{pmatrix} u^0 \\ v^0 \end{pmatrix} - \begin{pmatrix} u^* \\ v^* \end{pmatrix} \right\|^2 \;=\; 2\eta \left(\|u^*\|^2 + \|v^*\|^2 \right).$$

We now use an induction argument to demonstrate that $\theta_k \leq \frac{2}{k+2}$ for all $k \geq -1$. Indeed, the hypothesis holds when $k = -1$ as we have $\theta_{-1} \in (0, 2]$. Assume that the hypothesis holds for $k \geq -1$; i.e., $\theta_k \leq \frac{2}{k+2}$. We obtain:

$$\theta_{k+1} \;=\; \frac{2}{1 + \sqrt{1 + \frac{4}{\theta_k^2}}} \;\leq\; \frac{2}{1 + \sqrt{1 + (k+2)^2}} \;\leq\; \frac{2}{k+3}.$$

Therefore, the hypothesis holds for $k+1$. Putting all these pieces together yields the following inequality

$$\delta_k \;\leq\; \frac{\left(32 + 8/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(k+1)^2},$$

which establishes the lemma.

Proof of claim (22): By the definition of $\tilde{u}^{k+1}$ and $\tilde{v}^{k+1}$, we have

$$\begin{aligned} \mathbb{E}_{\xi^k}\left[\left\| \begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix} \right\|^2 \right] \;&=\; \frac{1}{2}\left(\left\| \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix} \right\|^2 \right) \\ &=\; \frac{1}{2}\left(\left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix} \right\|^2 \right). \end{aligned}$$

This directly implies the desired equality (22). $\square$

3.2 Main results

In this section, we provide an upper bound for the complexity of Randkhorn algorithm. First, we derive the iteration complexity of Randkhorn algorithm based on the results of Lemma 2.2 and Lemma 3.2.

Theorem 3.3. *The Randkhorn algorithm returns a matrix $B(u^K, v^K)$ that satisfies the condition $E^K \leq \varepsilon'$ with the number of iterations K satisfying the following upper bound*

$$K \;\leq\; 1 + 4\left(224 + 56/\theta_{-1}^2\right)^{1/3}\left(\frac{\sqrt{\eta n}R}{\varepsilon'}\right)^{2/3} \tag{23}$$

where R is defined in Lemma 2.2 to control optimal solutions of dual regularized OT problem (6).

Proof. By the similar argument for proving (14), we have

$$f(u^k, v^k) - f(y_u^{k+1}, y_v^{k+1}) \;\geq\; \frac{1}{2}\left(\rho\left(r, r(u^k, v^k)\right) + \rho\left(l, l(u^k, v^k)\right)\right). \tag{24}$$

Furthermore, by the definition of $\begin{pmatrix} u^k \\ v^k \end{pmatrix}$, we obtain that $f(u^k, v^k) \leq f(y_u^k, y_v^k)$. This together with (24) yields that

$$f(y_u^k, y_v^k) - f(y_u^{k+1}, y_v^{k+1}) \;\geq\; \frac{1}{2}\left(\rho\left(r, r(u^k, v^k)\right) + \rho\left(l, l(u^k, v^k)\right)\right).$$

Given any $1 \leq j \leq K$, by summing the above inequality over $k = j, j+1, \ldots, K$, we find that

$$f(y_u^j, y_v^j) - f(y_u^{K+1}, y_v^{K+1}) \;\geq\; \frac{1}{2}\sum_{k=j}^{K}\left[\left(\rho\left(r, r(u^k, v^k)\right) + \rho\left(l, l(u^k, v^k)\right)\right)\right].$$

Since $f(u^*, v^*) \leq f(y_u^{K+1}, y_v^{K+1})$, the above inequality leads to

$$\begin{aligned} f(y_u^j, y_v^j) - f(u^*, v^*) \;&\geq\; \frac{1}{2}\sum_{k=j}^{K}\left[\rho\left(r, B(u^k, v^k)\mathbf{1}\right) + \rho\left(l, B(u^k, v^k)^\top\mathbf{1}\right)\right] \\ &\geq\; \frac{1}{14}\sum_{k=j}^{K}\left(\|r - B(u^k, v^k)\mathbf{1}\|_1^2 + \|l - B(u^k, v^k)^\top\mathbf{1}\|_1^2\right), \end{aligned}$$

where the second inequality comes from [3, Lemma 6]. Taking an expectation on both sides with respect to the Bernoulli random variables $\{\xi^j\}_{j=1}^k$ yields that

$$\delta_j \;\geq\; \frac{1}{14}\sum_{k=j}^{K}\left(\mathbb{E}\left[\|r - B(u^k, v^k)\mathbf{1}\|_1^2 + \|l - B(u^k, v^k)^\top\mathbf{1}\|_1^2\right]\right) \overset{(10)}{\geq} \frac{1}{28}\sum_{k=j}^{K} E_k^2.$$

Combining the above result with that from Lemma 3.2 leads to the following inequality

$$\sum_{k=j}^{K} E_k^2 \;\leq\; \frac{\left(896 + 224/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(j+1)^2}.$$

On the other hand, according to Lemma 2.2, we have $\|u^*\| \leq \sqrt{n}R$ and $\|v^*\| \leq \sqrt{n}R$. Furthermore, $E_k \geq \varepsilon'$ holds true as soon as the stopping criterion is not fulfilled. Therefore, the following inequality holds:

$$(\varepsilon')^2 \;\leq\; \frac{\left(896 + 224/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(j+1)^2(K-j+1)} \;\leq\; \frac{\left(1792 + 448/\theta_{-1}^2\right)\eta n R^2}{(j+1)^2(K-j+1)}.$$

Since the above inequality holds true for any $1 \leq j \leq K$, we assume without loss of generality that K is even and let $j = K/2$. Then, we obtain that

$$K \;\leq\; 1 + 4\left(224 + 56/\theta_{-1}^2\right)^{1/3}\left(\frac{\sqrt{\eta n}R}{\varepsilon'}\right)^{2/3}$$

As a consequence, we conclude that the number of iterations k satisfies (23). $\square$

Equipped with the result of Theorem 3.3 and the scheme of Algorithm 2 for approximating OT by Randkhorn algorithm, we obtain the following result regarding the complexity of the Randkhorn algorithm.

Theorem 3.4. *The Randkhorn algorithm for approximating the optimal transport problem (Algorithm 2) returns a transportation plan $\hat{X} \in \mathbb{R}^{n\times n}$ satisfying the constraints $\hat{X}\mathbf{1} = r$, $\hat{X}^\top\mathbf{1} = l$ and criterion (2) in a total of*

$$\mathcal{O}\left(\frac{n^{7/3}\,\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$$

arithmetic operations.

The proof of Theorem 3.4 is provided in Appendix A.1. The complexity of the Randkhorn algorithm improves upon the best known complexity bound $\mathcal{O}\left(\frac{n^2\|C\|_\infty^2\log(n)}{\varepsilon^2}\right)$ for the Sinkhorn algorithm [13] when ε is sufficiently small and outperforms the complexity bound $\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty\sqrt{\log(n)}}{\varepsilon}\right)$ for the accelerated first-order primal-dual algorithms [13, 26]. This is supported empirically by the comparative performance of the Randkhorn algorithm with both synthetic and real data in Section 5.

4 Gandkhorn: Greedy Randkhorn Algorithm

We now turn to the Gandkhorn algorithm, a *greedy Randkhorn algorithm*. The motivation for this algorithm stems from the fact that the Greenkhorn algorithm [3, 26], a greedy coordinate version of Sinkhorn algorithm, has been shown to have favorable practical performance and a comparable theoretical guarantee with respect to the Sinkhorn algorithm. We present the pseudocode for the Gandkhorn algorithm in Algorithm 3 and its application to approximate regularized OT in Algorithm 4.

The algorithmic design of the Gandkhorn algorithm is similar to that of the Randkhorn algorithm; both are based on coordinate descent for the dual regularized OT problem (6) and the estimated sequences. We wish to remark that the Gandkhorn algorithm differs from existing accelerated randomized algorithms based on coordinate gradient in that the update for the main iterates (u_I, v_J) of the Gandkhorn algorithm in Algorithm 3 is an exact minimization (cf. Step 2 in Algorithm 3).

Algorithm 3: GANDKHORN$(C, \eta, r, l, \varepsilon')$

Input: $k = 0$, $\theta_{-1} \in (0, 2]$ and $y_u^0 = y_v^0 = \tilde{u}_0 = \tilde{v}_0 = \mathbf{0}$.

while $E^k > \varepsilon'$ **do**

Step 1: $\theta_k = \frac{\theta_{k-1}}{2}\left(\sqrt{\theta_{k-1}^2 + 4} - \theta_{k-1}\right)$.

Step 2: $\begin{pmatrix} \bar{u}^k \\ \bar{v}^k \end{pmatrix} = (1 - \theta_k)\begin{pmatrix} y_u^k \\ y_v^k \end{pmatrix} + \theta_k \begin{pmatrix} \tilde{u}^k \\ \tilde{v}^k \end{pmatrix}$.

Step 3: Compute

$$\begin{aligned} I &= \operatorname{argmax}_{1 \le i \le n} \ |r_i - r_i(\bar{u}^k, \bar{v}^k)|, \quad r(\bar{u}^k, \bar{v}^k) = B(\bar{u}^k, \bar{v}^k)\mathbf{1}, \\ J &= \operatorname{argmax}_{1 \le j \le n} \ |l_j - l_j(\bar{u}^k, \bar{v}^k)|, \quad l(\bar{u}^k, \bar{v}^k) = B(\bar{u}^k, \bar{v}^k)^\top \mathbf{1}. \end{aligned}$$

if $\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) \ge \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)$ **then**

$\hat{u}_I^{k+1} = \bar{u}_I^k + \log(r_I) - \log\left(r_I(\bar{u}^k, \bar{v}^k)\right)$, $\hat{u}_i^{k+1} = \bar{u}_i^k$ for $i \ne I$ and $\hat{v}^{k+1} = \bar{v}^k$.

else

$\hat{v}_J^{k+1} = \bar{v}_J^k + \log(l_J) - \log\left(l_J(\bar{u}^k, \bar{v}^k)\right)$, $\hat{v}_j^{k+1} = \bar{v}_j^k$ for $j \ne J$ and $\hat{u}^{k+1} = \bar{u}^k$.

end if

Step 4: Randomly sample $\xi^k \sim \text{Bernoulli}(\frac{1}{2})$, a Bernoulli random variable with parameter $\frac{1}{2}$, and $\pi \in \{1, 2, \ldots, n\}$ at uniform.

if $\xi^k = 0$ **then**

$\tilde{u}_\pi^{k+1} = \tilde{u}_\pi^k - \dfrac{r_\pi(\bar{u}^k, \bar{v}^k) - r_\pi}{8n\eta\theta_k}$, $\tilde{u}_i^{k+1} = \tilde{u}_i^k$ for $i \ne \pi$ and $\tilde{v}^{k+1} = \tilde{v}^k$.

else

$\tilde{v}_\pi^{k+1} = \tilde{v}_\pi^k - \dfrac{l_\pi(\bar{u}^k, \bar{v}^k) - l_\pi}{8n\eta\theta_k}$, $\tilde{v}_j^{k+1} = \tilde{v}_j^k$ for $j \ne \pi$ and $\tilde{u}^{k+1} = \tilde{u}^k$.

end if

Step 5: $\begin{pmatrix} u^k \\ v^k \end{pmatrix} = \operatorname{argmin}\left\{ f(u, v) \mid \begin{pmatrix} u \\ v \end{pmatrix} \in \left\{ \begin{pmatrix} y_u^k \\ y_v^k \end{pmatrix}, \begin{pmatrix} \hat{u}^{k+1} \\ \hat{v}^{k+1} \end{pmatrix} \right\} \right\}$.

Step 6: Compute

$$\begin{aligned} I &= \operatorname{argmax}_{1 \le i \le n} \ |r_i - r_i(u^k, v^k)|, \quad r(u^k, v^k) = B(u^k, v^k)\mathbf{1}, \\ J &= \operatorname{argmax}_{1 \le j \le n} \ |l_j - l_j(u^k, v^k)|, \quad l(u^k, v^k) = B(u^k, v^k)^\top \mathbf{1}. \end{aligned}$$

if $\hat{\rho}\left(r_I, r_I(u^k, v^k)\right) \ge \hat{\rho}\left(l_J, l_J(u^k, v^k)\right)$ **then**

$(y_u)_I^{k+1} = u_I^k + \log(r_I) - \log\left(r_I(u^k, v^k)\right)$, $(y_u)_i^{k+1} = u_i^k$ for $i \ne I$ and $y_v^{k+1} = v^k$.

else

$(y_v)_J^{k+1} = v_I^k + \log(l_J) - \log\left(l_J(u^k, v^k)\right)$, $(y_v)_j^{k+1} = v_j^k$ for $j \ne J$ and $y_u^{k+1} = u^k$.

end if

Step 7: $k = k + 1$.

end while

Output: $B(u^k, v^k)$.

We present a complexity analysis for the Gandkhorn algorithm that yields a complexity bound of $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$, which is better than the best existing complexity bound $\mathcal{O}\left(\frac{n^2\|C\|_\infty^2\log(n)}{\varepsilon^2}\right)$ for the Greenkhorn algorithm [26] in terms of the desired accuracy ε.

The presentation of the Gandkhorn algorithm in Algorithm 3 makes use of a function

---

Algorithm 4: Approximating OT by GANDKHORN

Input: Regularization parameter $\eta = \frac{\varepsilon}{4\log(n)}$ and error $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.

Step 1: Choose $\tilde{r} \in \Delta_n$ and $\tilde{l} \in \Delta_n$ as

$$\left(\tilde{r}, \tilde{l}\right) = \left(1 - \frac{\varepsilon'}{8}\right)(r, l) + \frac{\varepsilon'}{8n}(\mathbf{1}, \mathbf{1}).$$

Step 2: Compute $\tilde{X} = \text{GANDKHORN}\left(C, \eta, \tilde{r}, \tilde{l}, \varepsilon'/2\right)$.

Step 3: Given Algorithm 2 in [3], we round $\tilde{X}$ to $\hat{X}$ to satisfy $\hat{X}\mathbf{1} = r$ and $\hat{X}^\top \mathbf{1} = l$.

Output: $\hat{X}$.

$\hat{\rho} : \mathbb{R}_+ \times \mathbb{R}_+ \to [0, +\infty]$ given by

$$\hat{\rho}(a, b) \ := \ b - a + a\log\left(\frac{a}{b}\right),$$

which measures the progress in the dual objective between two consecutive iterates of the Gandkhorn algorithm. Note that $\rho(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n \hat{\rho}(a_i, b_i)$ for any $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, where ρ is defined in Section 3. Step 3 and Step 6 of the Gandkhorn algorithm differ from Step 3 and Step 6 of the Randkhorn algorithm in that they are designed specifically to choose the most promising coordinates based on $\hat{\rho}$ distance. In the theoretical analysis of the Gandkhorn algorithm, we also use the quantity E^k defined in (10) to measure the error of the k-th iterate for the Gandkhorn algorithm. Finally, we describe the application of the Gandkhorn algorithm to approximate OT in Algorithm 4 by introducing a standard scheme from [3].

4.1 Technical lemmas

In this section, we provide several technical lemmas for bounding the following dual objective gap δ_k in the Gandkhorn algorithm: $\delta_k = \mathbb{E}[f(y_u^k, y_v^k) - f(u^*, v^*)]$ where (y_u^k, y_v^k) are defined in Algorithm 3. We modify the two sequences of iterates defined in (11) and (12) as follows:

$$\begin{pmatrix} q_u^{k+1} \\ q_v^{k+1} \end{pmatrix} \ := \ \begin{pmatrix} \bar{u}^k \\ \bar{v}^k \end{pmatrix} - \frac{1}{8n\eta}\nabla f(\bar{u}^k, \bar{v}^k), \tag{25}$$

and

$$\begin{pmatrix} \tilde{q}_u^{k+1} \\ \tilde{q}_v^{k+1} \end{pmatrix} \ := \ \begin{pmatrix} \tilde{u}^k \\ \tilde{v}^k \end{pmatrix} - \frac{1}{8n\eta\theta_k}\nabla f(\bar{u}^k, \bar{v}^k). \tag{26}$$

Given the formulations of these sequences, we present a key descent inequality for the values of dual regularized OT at the Gandkhorn updates.

Lemma 4.1. *For each iteration $k > 0$ of the Gandkhorn algorithm, we have*

$$\begin{aligned} f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ \leq \ & \theta_k f(u^*, v^*) + (1-\theta_k) f(y_u^k, y_v^k) \\ & + 4n\eta\theta_k^2 \left(\left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1} \end{pmatrix} \right\|^2 \right). \end{aligned} \tag{27}$$

Proof. The proof technique of the lemma is similar to that of Lemma 3.1. We nonetheless provide the details for completeness. Assume that the following inequalities hold:

$$f(\bar{u}^k, \bar{v}^k) - f(u^{k+1}, v^{k+1}) \ \geq \ \frac{1}{2}\left(\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) + \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)\right), \tag{28}$$

$$f(u^{k+1}, v^{k+1}) \ \leq \ f(\bar{u}^k, \bar{v}^k) - \frac{1}{16n\eta}\|\nabla f(\bar{u}^k, \bar{v}^k)\|^2. \tag{29}$$

Using the definitions of q_u^{k+1} and q_v^{k+1} in (25), we find that

$$\begin{aligned}\|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 \quad &= \quad 2\|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 - \|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 \\ &\overset{(25)}{=} \quad 16n\eta \begin{pmatrix} \bar{u}^k - q_u^{k+1} \\ \bar{v}^k - q_v^{k+1}\end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) - 64n^2\eta^2 \left\|\begin{pmatrix} \bar{u}^k - q_u^{k+1} \\ \bar{v}^k - q_v^{k+1}\end{pmatrix}\right\|^2.\end{aligned}$$

Plugging the above equality into (29) and rearranging yields the following inequality

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ \leq \ f(\bar{u}^k, \bar{v}^k) + \begin{pmatrix} q_u^{k+1} - \bar{u}^k \\ q_v^{k+1} - \bar{v}^k\end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) + 4n\eta \left\|\begin{pmatrix} q_u^{k+1} - \bar{u}^k \\ q_v^{k+1} - \bar{v}^k\end{pmatrix}\right\|^2. \tag{30}$$

Furthermore, invoking the definitions of $\tilde{q}_u^{k+1}$ and $\tilde{q}_v^{k+1}$ in (26), we obtain the following equality:

$$\begin{pmatrix} q_u^{k+1} - \bar{u}^k \\ q_v^{k+1} - \bar{v}^k\end{pmatrix} \ = \ \theta_k \begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k\end{pmatrix},$$

and combining this equality with (30) leads to:

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ \leq \ f(\bar{u}^k, \bar{v}^k) + \theta_k \left[\begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k\end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) + 4n\eta\theta_k \left\|\begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k\end{pmatrix}\right\|^2\right]. \tag{31}$$

Based on the definitions of $\tilde{s}_u^{k+1}$ and $\tilde{s}_v^{k+1}$ in (26), we can check that

$$\begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1}\end{pmatrix}^\top \left[\begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k\end{pmatrix} + \frac{1}{8n\eta\theta_k}\nabla f(\bar{u}^k, \bar{v}^k)\right] \ = \ 0.$$

Rearranging the terms yields that

$$\frac{1}{4n\eta\theta_k}\begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1}\end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) \ = \ \left\|\begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1}\end{pmatrix}\right\|^2 - \left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k\end{pmatrix}\right\|^2 + \left\|\begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k\end{pmatrix}\right\|^2. \tag{32}$$

By plugging (32) into (31), we arrive at the following result:

$$\begin{aligned} f(\hat{u}^{k+1}, \hat{v}^{k+1}) \quad &\leq \quad f(\bar{u}^k, \bar{v}^k) + \theta_k \left[\begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k\end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) + \begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1}\end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k)\right] \\ &+ \quad 4n\eta\theta_k^2 \left(\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k\end{pmatrix}\right\|^2 - \left\|\begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1}\end{pmatrix}\right\|^2\right) \\ &= \quad f(\bar{u}^k, \bar{v}^k) + \theta_k \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k\end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) \\ &+ \quad 4n\eta\theta_k^2 \left(\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k\end{pmatrix}\right\|^2 - \left\|\begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1}\end{pmatrix}\right\|^2\right). \end{aligned} \tag{33}$$

Simple algebra indicates that

$$\theta_k \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \ = \ \theta_k \begin{pmatrix} u^* - \bar{u}^k \\ v^* - \bar{v}^k \end{pmatrix} + (1-\theta_k) \begin{pmatrix} y_u^k - \bar{u}^k \\ y_v^k - \bar{v}^k \end{pmatrix}.$$

Collecting the previous results, we arrive at the following inequality:

$$f(\bar{u}^k, \bar{v}^k) + \theta_k \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) \ \leq \ \theta_k f(u^*, v^*) + (1-\theta_k) f(y_u^k, y_v^k). \tag{34}$$

Therefore, we conclude the desired inequality (27) by plugging (34) into (33).

Proof of claim (28)**:** First, we assume that $\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) \geq \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)$. We then have $u_I^{k+1} = \bar{u}_I^k + \log(r_I) - \log\left(r_I(\bar{u}^k, \bar{v}^k)\right)$. This implies that

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ = \ 1 - \langle \bar{u}^k, r\rangle - \langle \bar{v}^k, l\rangle - r_I \log\left(\frac{r_I}{r_I(\bar{u}^k, \bar{v}^k)}\right).$$

Furthermore, we also have $f(\bar{u}^k, \bar{v}^k) = \mathbf{1}^\top r(\bar{u}^k, \bar{v}^k) - \langle \bar{u}^k, r\rangle - \langle \bar{v}^k, l\rangle$. which implies:

$$\begin{aligned} f(\bar{u}^k, \bar{v}^k) - f(\hat{u}^{k+1}, \hat{v}^{k+1}) \ &= \ r_I(\bar{u}^k, \bar{v}^k) - r_I + r_I \log\left(\frac{r_I}{r_I(\bar{u}^k, \bar{v}^k)}\right) \\ &= \ \hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right). \end{aligned}$$

Using the assumption $\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) \geq \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)$ yields the desired inequality (28). A similar argument holds true for the case $\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) < \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)$. As a consequence, we obtain the conclusion of claim (28).

Proof of claim (29)**:** By the definition of $\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right)$ and $\hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)$, we have

$$\begin{aligned} \hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) \ &= \ f(\bar{u}^k, \bar{v}^k) - \operatorname*{argmin}_{u_I \in \mathbb{R}} f(u_I, \bar{u}^k_{i\neq I}, \bar{v}^k) \\ &\geq \ f(\bar{u}^k, \bar{v}^k) - f\left(\{\bar{u}_I^k - \frac{(\nabla_u f(\bar{u}^k, \bar{v}^k))_I}{4\eta}, \bar{u}^k_{i\neq I}\}, \bar{v}^k\right), \\ \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right) \ &= \ f(\bar{u}^k, \bar{v}^k) - \operatorname*{argmin}_{v_J \in \mathbb{R}} f(\bar{u}^k, v_J, \bar{v}^k_{j\neq J}) \\ &\geq \ f(\bar{u}^k, \bar{v}^k) - f\left(\bar{u}^k, \{\bar{v}_J^k - \frac{(\nabla_u f(\bar{u}^k, \bar{v}^k))_J}{4\eta}, \bar{v}^k_{j\neq J}\}\right). \end{aligned}$$

Applying (8) leads to the following inequalities

$$\begin{aligned} f(\bar{u}^k, \bar{v}^k) - f\left(\{\bar{u}_I^k - \frac{(\nabla_u f(\bar{u}^k, \bar{v}^k))_I}{4\eta}, \bar{u}^k_{i\neq I}\}, \bar{v}^k\right) \ &\geq \ \frac{1}{8\eta}\|(\nabla_u f(\bar{u}^k, \bar{v}^k))_I\|^2, \\ f(\bar{u}^k, \bar{v}^k) - f\left(\bar{u}^k, \{\bar{v}_J^k - \frac{(\nabla_u f(\bar{u}^k, \bar{v}^k))_J}{4\eta}, \bar{v}^k_{j\neq J}\}\right) \ &\geq \ \frac{1}{8\eta}\|(\nabla_v f(\bar{u}^k, \bar{v}^k))_J\|^2. \end{aligned}$$

By the definition of f, we have

$$\nabla_u f(\bar{u}^k, \bar{v}^k) \ = \ B(\bar{u}^k, \bar{v}^k)\mathbf{1} - r, \quad \nabla_v f(\bar{u}^k, \bar{v}^k) \ = \ B(\bar{u}^k, \bar{v}^k)^\top \mathbf{1} - l.$$

Thus, the definition of I and J in Step 3 of the Gandkhorn algorithm implies that

$$\begin{aligned}
\|(\nabla_u f(\bar{u}^k, \bar{v}^k))_I\|^2 &\geq \frac{1}{n}\|\nabla_u f(\bar{u}^k, \bar{v}^k)\|^2, \\
\|(\nabla_v f(\bar{u}^k, \bar{v}^k))_J\|^2 &\geq \frac{1}{n}\|\nabla_v f(\bar{u}^k, \bar{v}^k)\|^2.
\end{aligned}$$

Putting these pieces together with (28) yields the result of claim (29). □

Given the result of Lemma 4.1, we are now ready to bound the dual objective gap δ^k.

Lemma 4.2. *For the iterates $\{(u^k, v^k)\}_{k\geq 0}$ returned by the Gandkhorn algorithm, we have*

$$\delta_k \leq \frac{\left(32n^2 + 8/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(k+1)^2}. \tag{35}$$

Proof. First, we estimate the third term of the right-hand side of (27) using a similar approach to the proof of claim (22). In particular, we have the following equality:

$$\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \left\|\begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix}\right\|^2 = 2n\left(\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \mathbb{E}_{\xi^k}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right). \tag{36}$$

which follows directly from the definition of the Gandkhorn algorithm. We now put the result of Lemma 4.1 and equality (36) together, which leads to the following inequality:

$$\begin{aligned}
f(\hat{u}^{k+1}, \hat{v}^{k+1}) \leq\ & \theta_k f(u^*, v^*) + (1-\theta_k) f(y_u^k, y_v^k) \\
& + 8\eta n^2\theta_k^2\left(\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \mathbb{E}_{\xi^k}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right).
\end{aligned}$$

Subtracting $f(u^*, v^*)$ from both sides and taking an expectation with respect to $\{\xi^j\}_{j=1}^{k-1}$ yields:

$$f(\hat{u}^{k+1}, \hat{v}^{k+1}) - f(u^*, v^*) \leq (1-\theta_k)\delta^k + 8\eta n^2\theta_k^2\left(\mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2\right] - \mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right).$$

Furthermore, by the definition of $\begin{pmatrix} u^k \\ v^k \end{pmatrix}$ and $\begin{pmatrix} y_u^{k+1} \\ y_v^{k+1} \end{pmatrix}$, we have

$$f(y_u^{k+1}, y_v^{k+1}) \leq f(u^k, v^k) \leq f(\hat{u}^{k+1}, \hat{v}^{k+1}),$$

which implies that

$$\delta^{k+1} \leq (1-\theta_k)\delta^k + 8\eta n^2\theta_k^2\left(\mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2\right] - \mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right).$$

Dividing both sides of the inequality by θ_k^2, we find that

$$\frac{\delta_{k+1}}{\theta_k^2} \leq \frac{\delta_k}{\theta_{k-1}^2} + 8\eta n^2\left(\mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2\right] - \mathbb{E}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right).$$

Changing the count k to i and summing the inequality over $i = 0, 1, \ldots, k-1$ gives the following inequality:

$$\delta_k \ \leq \ \theta_{k-1}^2 \left(\frac{\delta_0}{\theta_{-1}^2} + 8\eta n^2 \left(\|u^*\|^2 + \|v^*\|^2 \right) \right).$$

Recall from the proof of Lemma 3.2 that, we have

$$\delta_0 \ \leq \ 2\eta \left(\|u^*\|^2 + \|v^*\|^2 \right), \quad \text{and} \quad \theta_k \leq \frac{2}{k+2} \text{ for } k \geq -1.$$

Putting the pieces together leads to

$$\delta_k \ \leq \ \frac{\left(32n^2 + 8/\theta_{-1}^2\right) \eta \left(\|u^*\|^2 + \|v^*\|^2 \right)}{(k+1)^2},$$

which proves the lemma. $\square$

4.2 Main results

In this section, we first provide an upper bound for the number of iterations K to achieve a desired tolerance ε' for the iterates of the Gandkhorn algorithm.

Theorem 4.3. *The Gandkhorn algorithm returns a matrix $B(u^K, v^K)$ that satisfies the condition $E^K \leq \varepsilon'$ with the number of iterations K satisfying the following upper bound*

$$K \ \leq \ 1 + 4\left(224n^3 + 56n/\theta_{-1}^2\right)^{1/3} \left(\frac{\sqrt{\eta n} R}{\varepsilon'} \right)^{2/3}, \tag{37}$$

where R is defined in Lemma 2.2 to control optimal solutions of dual regularized OT problem (6).

Proof. Arguing similarly as in Theorem 3.3, we have

$$\delta_j \ \geq \ \frac{1}{14n} \sum_{k=j}^{K} \left(\mathbb{E}\left[\|r - B(u^k, v^k)\mathbf{1}\|_1^2 + \|l - B(u^k, v^k)^\top \mathbf{1}\|_1^2 \right] \right) \ \overset{(10)}{\geq} \ \frac{1}{28n} \sum_{k=j}^{K} E_k^2.$$

Combining with the results of Lemma 4.2 yields that

$$\sum_{k=j}^{K} E_k^2 \ \leq \ \frac{\left(896n^3 + 224n/\theta_{-1}^2\right) \eta \left(\|u^*\|^2 + \|v^*\|^2 \right)}{(j+1)^2}.$$

On the other hand, we have $\|u^*\| \leq \sqrt{n}R$ and $\|v^*\| \leq \sqrt{n}R$. Also, $E_k \geq \varepsilon'$ holds true as soon as the stopping criterion is not fulfilled. Therefore, the following inequality holds:

$$(\varepsilon')^2 \ \leq \ \frac{\left(896n^3 + 224n/\theta_{-1}^2\right) \eta \left(\|u^*\|^2 + \|v^*\|^2 \right)}{(j+1)^2(K-j+1)} \ \leq \ \frac{\left(1792n^3 + 448n/\theta_{-1}^2\right) \eta n R^2}{(j+1)^2(K-j+1)}.$$

Since the above inequality holds true for any $1 \leq j \leq K$, we assume without loss of generality that K is even and let $j = K/2$. Then, we obtain that

$$K \ \leq \ 1 + 4\left(224n^3 + 56n/\theta_{-1}^2\right)^{1/3} \left(\frac{\sqrt{\eta n} R}{\varepsilon'} \right)^{2/3}$$

Therefore, we conclude that the number of iterations K satisfies (23). $\square$

Equipped with the result of Theorem 4.3 and the scheme of Algorithm 4, we are able to establish the following result for the complexity of the Gandkhorn algorithm.

Theorem 4.4. *The Gandkhorn algorithm for approximating the optimal transport problem (Algorithm 4) returns a transportation plan $\hat{X} \in \mathbb{R}^{n\times n}$ satisfying the constraints $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$ and criterion (2) in a total of*

$$\mathcal{O}\left(\frac{n^{7/3}\,\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$$

arithmetic operations.

The proof of Theorem 4.4 is provided in Appendix A.2. The complexity bound of the Gandkhorn algorithm in Theorem 4.4 is comparable to that of Randkhorn algorithm and improves on the best known complexity bound $\mathcal{O}\left(\frac{n^2\|C\|_\infty^2\log(n)}{\epsilon^2}\right)$ for the Greenkhorn algorithm [26] when ε is sufficiently small.

5 Experiments

In this section, we conduct careful comparative experiments with the Randkhorn and Gandkhorn algorithms on synthetic images and real images from the MNIST Digits dataset[1]. For comparison purposes, we use the Sinkhorn and Greenkhorn algorithms as baselines [9, 3]. We also compare the Gandkhorn algorithm with a practical implementation of Greenkhorn algorithm with normalization, which we refer to as the normalized Greenkhorn algorithm. This algorithm has been widely used in real application [15]. To obtain the optimal value of the original optimal transport problem without entropic regularization, we employ the default linear programming solver in MATLAB.

We will demonstrate that, while the Randkhorn algorithm consistently outperforms the Sinkhorn algorithm, the comparison between the Gandkhorn algorithm and two different implementations of Greenkhorn algorithms need to be discussed case by case.

5.1 Experiments on synthetic images

To generate the synthetic images we adopt the process from [3]. We evaluate the performance of different algorithms on these synthetic images following the procedures in [26]. Note that the transportation distance is defined between two synthetic images while the cost matrix is defined based on the ℓ_1 distances among locations of pixel in the images.

The synthetic images are of size 20 by 20 pixels and are generated by means of randomly placing a foreground square in a black background. Furthermore, a uniform distribution on $[0, 1]$ is used for the intensities of the pixels in the background while a uniform distribution on $[0, 50]$ is employed for the pixels in the foreground. Here, we fix the proportion of the size of the foreground square as 10% of the whole images and implement all of the aforementioned algorithms on these synthetic images.

We use standard metrics to assess the performance of different algorithms. The first metric is the ℓ_1 distance (cf. [3] for an argument of choosing ℓ_1 distance) between the output of some algorithm X and the corresponding transportation polytope, which is given by

$$d(X) \;:=\; \|r(X) - r\|_1 + \|l(X) - l\|_1.$$

Here, $r(X)$ and $l(X)$ in the above display are the row and column obtained from the output of the algorithm X while r and l are the given row and column vectors of the OT problem. The

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

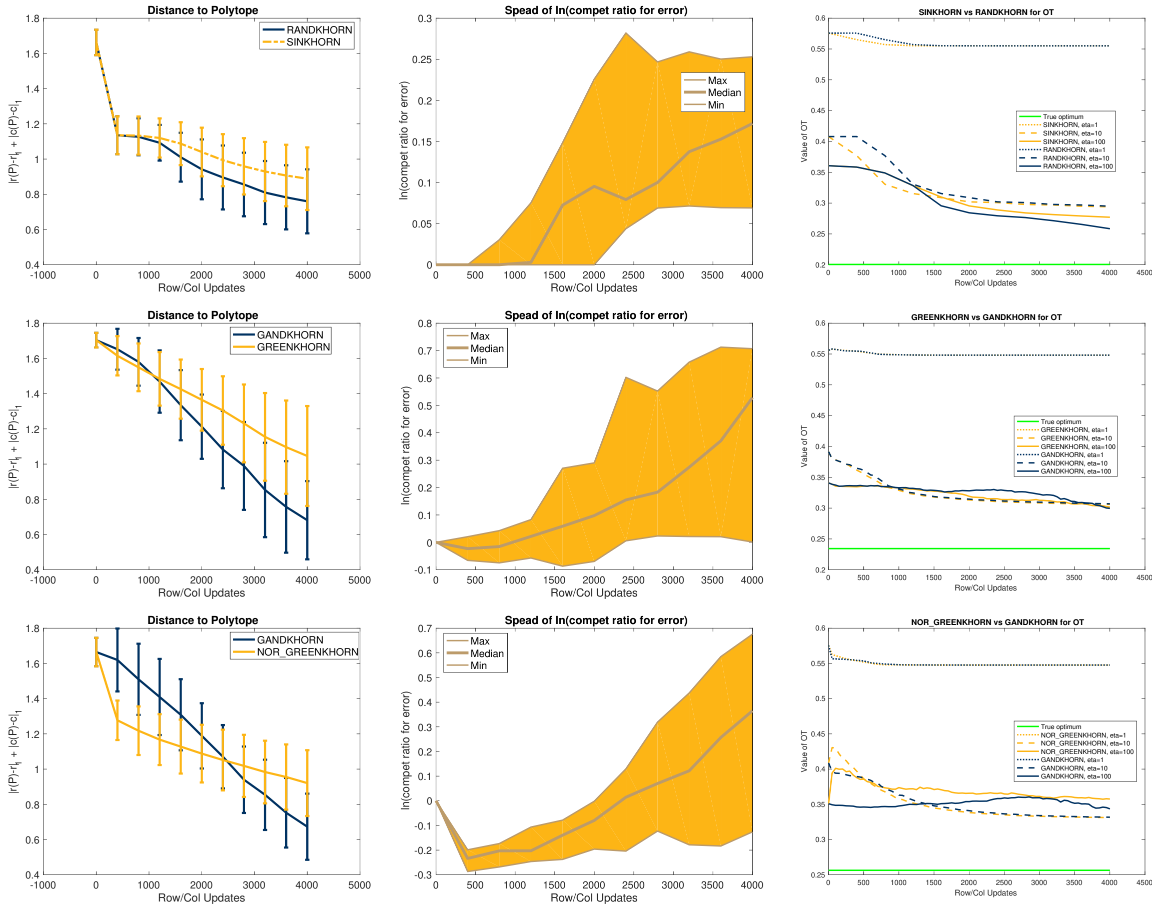

Figure 1: Comparative performance of the Sinkhorn versus Randkhorn algorithms, the Greenkhorn versus Gandkhorn algorithms and the normalized Greenkhorn versus Gandkhorn algorithms on the synthetic images. For the images in the first row, we compare the performance of the Sinkhorn and Randkhorn algorithms based on the number of iteration counts. In the leftmost image of that row, the comparison is based on using distance to transportation polytope $d(X)$ where X are Sinkhorn and Randkhorn algorithms. In the middle image of that row, the maximum, median and minimum values of the competitive ratios on ten pairs of images are utilized for the comparison between Sinkhorn and Randkhorn algorithms. In the rightmost image of that row, we vary the regularization parameter $\eta \in \{1, 10, 100\}$ with these algorithms and using the value of the optimal transport problem (without the entropic regularization term) as the baseline. Similarly, the second rows of images present comparative results for the Greenkhorn versus Gandkhorn algorithms and the normalized Greenkhorn versus Gandkhorn algorithms.

second metric is defined as $\log(d(X_1)/d(X_2))$, which is termed *the competitive ratio*, where $d(X_1)$ and $d(X_2)$ are respectively the distances between the outputs of two algorithms X_1 and X_2 and the corresponding transportation polytope.

We perform three pairwise comparative experiments: Sinkhorn versus Randkhorn, Greenkhorn versus Gandkhorn, normalized Greenkhorn versus Gandkhorn, on ten randomly selected pairs of synthetic images. To have further evaluations with these algorithms, we also compare their performance with different choices of regularization parameter η in the specific set $\{1, 10, 100\}$

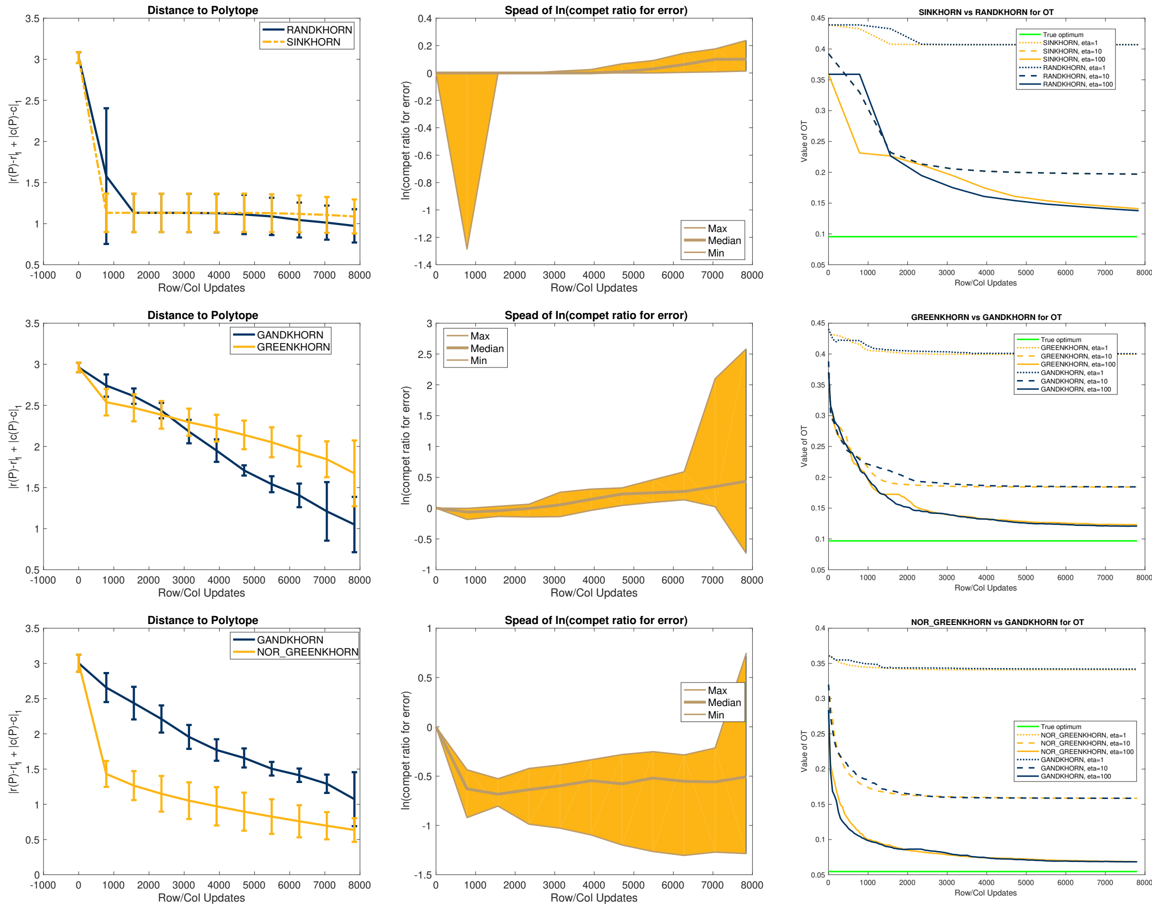

Figure 2: Comparative performance of the Sinkhorn versus Randkhorn algorithms, the Greenkhorn versus Gandkhorn algorithms and the normalized Greenkhorn versus Gandkhorn algorithms on the MNIST real images. See the caption of Figure 1 for more detail.

while using the value of the optimal transport problem (without entropic regularization term) as the baseline. For all the algorithms, the total number of iterations is set as $T = 10$.

We present experimental results in Figure 1 for different choices of regularization parameters. We observe that the Randkhorn algorithm consistently outperforms the Sinkhorn algorithm in terms of iteration count. This demonstrates the improvement achieved by the proposed algorithms for solving the dual regularized OT problem, and provides support for our theoretical assertion that the proposed algorithms achieves a better complexity bound than the Sinkhorn algorithm. Similarly, the Gandkhorn algorithm behaves the best, followed by the normalized Greenkhorn algorithm, outperforming the Greenkhorn algorithm in terms of iteration count. This supports the better theoretical complexity of the proposed Gandkhorn algorithm over the Greenkhorn algorithm.

5.2 Experiments on MNIST images

In this section, we use the same evaluation metrics as in Section 5.1 to compare the performance of different algorithms on real images from MNIST dataset. Note that the MNIST dataset contains 60,000 images of handwritten digits with the given size of 28 by 28 pixels. To understand more precisely the dependence on the dimension n of our algorithms, following the procedure in [3, 26], we add a very small noise term of the order 10^{-6} to all the zero elements and then perform a normalization step to guarantee that their sum becomes one.

We present experimental results of our algorithms in Figure 2 with several choices of regularization parameter η. In Figure 2, the Randkhorn algorithm also outperforms the Sinkhorn algorithm. Besides that, all the comparative results on real images are consistent with those on the synthetic images. Therefore, we conclude that the Randkhorn algorithm has favorable practical performance relative to that of the Sinkhorn algorithm. However, while the Gandkhorn algorithm still outperforms the Greenkhorn algorithm, it behaves worse than the normalized Greenkhorn algorithm. It is possibly because the normalization technique can alleviate the ill-conditioning which often occurs on real datasets. Unfortunately, the direct application of normalization technique to the Gandkhorn algorithm seems invalid and makes the algorithm divergent. To this end, it remains open whether the normalized Gandkhorn is possible. Therefore, we conclude that the normalized Greenkhorn algorithm is still favorable in practice despite the lack of theoretical guarantee.

6 Conclusion

In the paper, we proposed several novel accelerated versions of the Sinkhorn and Greenkhorn algorithms for solving optimal transport problems. In particular, we introduced an accelerated, monotone randomized version of the Sinkhorn algorithm, which we named the Randkhorn algorithm. The algorithm was shown to have a complexity bound of $\widetilde{\mathcal{O}}\left(\frac{n^{7/3}}{\varepsilon}\right)$. This is more favorable than that of the Sinkhorn algorithm in terms of desired accuracy ε and that of the accelerated first-order primal-dual algorithms in terms of the number of atoms n. Similarly, a greedy version of Randkhorn algorithm, which we referred to as the Gandkhorn algorithm, was proposed to accelerate Greenkhorn algorithm. This algorithm was demonstrated to have a complexity bound of $\widetilde{\mathcal{O}}\left(\frac{n^{7/3}}{\varepsilon}\right)$, which is comparable to that of Randkhorn algorithm and faster than that of Greenkhorn algorithm in terms of ε.

This work lays the foundations for several research directions. First, the proposed algorithms are specific for optimal transport distance between two discrete probability distributions with dimension at most n. However, in several practical applications, one of these measures can have infinite dimension; i.e., it may have uncountable or even continuous support. Subsampling methods have been widely employed to approximate optimal transport problem between such measures [36] by the optimal transport distance between their corresponding empirical measures, which are discrete. Since the number of atoms of these empirical measures needs to be sufficiently large to give a good approximation of the original OT, it is of practical interest to investigate whether the accelerated algorithms proposed in the paper can realize computational advantages over the Sinkhorn or Greenkhorn algorithms when being used to compute the OT between these measures.

The Wasserstein barycenter problem is closely related to the optimal transport problem, and it has also been shown to useful in various applications of machine learning and statistics [37, 18]. While a variety of algorithms have been proposed to solve the Wasserstein barycenter problem [10, 12], accelerated versions of Sinkhorn and Greenkhorn algorithms for this problem have not yet been developed. Given the favorable practical performance of the proposed accelerated algorithms for solving the optimal transport problem, it is of significant interest to extend these algorithms to the Wasserstein barycenter problem.

Acknowledgements

This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

A Technical Proofs

In this appendix, we provide proofs of the remaining results in the paper.

A.1 Proof of Theorem 3.4

We follow the proof argument of Theorem 1 in [3]. Here, we provide the detail proof of Theorem 3.4 for the completeness. In particular, simple algebra lead to the following bound

$$\left\langle C, \hat{X} \right\rangle - \langle C, X^* \rangle \quad \leq \quad \frac{\varepsilon}{2} + 4\left(\left\|\tilde{X}\mathbf{1} - r\right\|_1 + \left\|\tilde{X}^\top \mathbf{1} - l\right\|_1\right)\|C\|_\infty,$$

where transportation plan $\hat{X}$ is an output of Algorithm 2 and X^* is the optimal transportation plan of the OT problem (1). Furthermore, $\tilde{X}$ is a transportation plan returned by the Randkhorn algorithm (Algorithm 1) with the choice of $\tilde{r}$, $\tilde{l}$ and $\varepsilon'/2$ are given in Step 3 of Algorithm 2. Now, by means of triangle inequality with ℓ_1 distance, we derive that

$$\left\|\tilde{X}\mathbf{1} - r\right\|_1 + \left\|\tilde{X}^\top \mathbf{1} - l\right\|_1 \leq \left\|\tilde{X}\mathbf{1} - \tilde{r}\right\|_1 + \left\|\tilde{X}^\top \mathbf{1} - \tilde{l}\right\|_1 + \|r - \tilde{r}\|_1 + \left\|l - \tilde{l}\right\|_1 \leq \varepsilon'.$$

Putting the above results together, we obtain that $\langle C, \hat{X}\rangle - \langle C, X^*\rangle \leq \varepsilon$ where $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$. Given the previous bound, our remaining task is to analyze the complexity bound of Randkhorn algorithm in terms of the number iterations K to reach the condition $E_K \leq \varepsilon'$ with the given values of $\tilde{r}$, $\tilde{l}$. Based on the result of Theorem 3.3, we find that

$$\begin{aligned}
K \quad & \leq \quad 1 + 4\left(224 + 56/\theta_{-1}^2\right)^{1/3}\left(\frac{\sqrt{\eta n}R}{\varepsilon'}\right)^{2/3} \\
& \leq \quad 1 + 4\left(224 + 56/\theta_{-1}^2\right)^{1/3}\left(\frac{8\sqrt{\eta n}\|C\|_\infty}{\varepsilon}\left(\frac{\|C\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\left\{\tilde{r}_i, \tilde{l}_j\right\}\right)\right)\right)^{2/3} \\
& \leq \quad 1 + 16\left(224 + 56/\theta_{-1}^2\right)^{1/3}\left(\sqrt{\frac{n\|C\|_\infty^2}{4\log(n)\varepsilon}}\left(\frac{4\|C\|_\infty\log(n)}{\varepsilon} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\left\{\tilde{r}_i, \tilde{l}_j\right\}\right)\right)\right)^{2/3} \\
& = \quad \mathcal{O}\left(\frac{n^{1/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right).
\end{aligned}$$

Based on the above upper bound with K, the total iteration number of the Randkhorn algorithm is bounded by $\mathcal{O}\left(\frac{n^{1/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$. Each iteration of the Randkhorn algorithm only requires $\mathcal{O}(n^2)$ arithmetic operations. Combining these two results, a total number of arithmetic operations is of order $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$. Furthermore, the vectors $\tilde{r}$ and $\tilde{l}$ in Step 2 of Algorithm 2 can be approximated within $\mathcal{O}(n)$ arithmetic operations [3, Algorithm 2]. Therefore, the required number of arithmetic operations is of order $O(n^2)$. Putting all the results together, we conclude the desired complexity bound of the Randkhorn algorithm.

A.2 Proof of Theorem 4.4

The proof is nearly the same as that of Theorem 3.4. The only difference is to analyze the complexity bound of Gandkhorn algorithm in terms of the number iterations K to reach the condition $E_K \leq \varepsilon'$ with the given values of $\tilde{r}$, $\tilde{l}$. According to the result of Theorem 4.3, we obtain that

$$
\begin{aligned}
K &\leq 1 + 4\left(224n^3 + 56n/\theta_{-1}^2\right)^{1/3}\left(\frac{\sqrt{\eta n}R}{\varepsilon'}\right)^{2/3} \\
&\leq 1 + 4\left(224n^3 + 56n/\theta_{-1}^2\right)^{1/3}\left(\frac{8\sqrt{\eta n}\|C\|_\infty}{\varepsilon}\left(\frac{\|C\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\left\{\tilde{r}_i, \tilde{l}_j\right\}\right)\right)\right)^{2/3} \\
&\leq 1 + 16\left(224n^3 + 56n/\theta_{-1}^2\right)^{1/3}\left(\sqrt{\frac{n\|C\|_\infty^2}{4\log(n)\varepsilon}}\left(\frac{4\|C\|_\infty\log(n)}{\varepsilon} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\left\{\tilde{r}_i, \tilde{l}_j\right\}\right)\right)\right)^{2/3} \\
&= \mathcal{O}\left(\frac{n^{4/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right).
\end{aligned}
$$

Based on the above upper bound with K, the total iteration number of the Gandkhorn algorithm is bounded by $\mathcal{O}\left(\frac{n^{4/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$. Each iteration of the Gandkhorn algorithm only requires $\mathcal{O}(n)$ arithmetic operations. Combining these two results, a total number of arithmetic operations is of order $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{4/3}\log^{1/3}(n)}{\varepsilon}\right)$. By the similar argument as in Theorem 3.4, we conclude the desired complexity bound of the Gandkhorn algorithm.

References

[1] J. Altschuler, F. Bach, A. Rudi, and J. Weed. Approximating the quadratic transportation metric in near-linear time. *ArXiv Preprint: 1810.10046*, 2018. (Cited on page 2.)

[2] J. Altschuler, F. Bach, A. Rudi, and J. Weed. Massively scalable Sinkhorn distances via the Nyström method. *ArXiv Preprint: 1812.05189*, 2018. (Cited on page 2.)

[3] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *NIPS*, pages 1964–1974, 2017. (Cited on pages 2, 4, 6, 8, 12, 13, 15, 20, 23, and 24.)

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. (Cited on page 1.)

[5] J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. Towards optimal running times for optimal transport. *ArXiv Preprint: 1810.07717*, 2018. (Cited on page 2.)

[6] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In *AISTATS*, pages 880–889, 2018. (Cited on page 2.)

[7] D. Chakrabarty and S. Khanna. Better and simpler error analysis of the Sinkhorn-Knopp algorithm for matrix scaling. *ArXiv Preprint: 1801.02790*, 2018. (Cited on page 2.)

[8] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. (Cited on page 1.)

[9] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013. (Cited on pages 2, 3, and 20.)

[10] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *ICML*, pages 685–693, 2014. (Cited on page 23.)

[11] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320343, 2016. (Cited on page 2.)

[12] P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *NIPS*, pages 10783–10793, 2018. (Cited on pages 1 and 23.)

[13] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorns algorithm. In *ICML*, pages 1367–1376, 2018. (Cited on pages 2, 6, and 13.)

[14] O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015. (Cited on pages 2 and 6.)

[15] R. Flamary and N. Courty. POT: Python optimal transport library, 2017. (Cited on pages 2 and 20.)

[16] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *NIPS*, pages 3440–3448, 2016. (Cited on page 2.)

[17] W. Guo, N. Ho, and M. I. Jordan. Accelerated primal-dual coordinate descent for computational optimal transport. *ArXiv Preprint: 1905.09952*, 2019. (Cited on page 2.)

[18] N. Ho, V. Huynh, D. Phung, and M. I. Jordan. Probabilistic multilevel clustering via composite transportation distance. *AISTATS*, 2019. (Cited on page 23.)

[19] A. Jambulapati, A. Sidford, and K. Tian. A direct $\tilde{o}$ $(1/\varepsilon)$ iteration parallel algorithm for optimal transport. *ArXiv Preprint: 1906.00618*, 2019. (Cited on page 2.)

[20] B. Kalantari, I. Lari, F. Ricca, and B. Simeone. On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. *Mathematical Programming*, 112(2):371–401, 2008. (Cited on page 2.)

[21] L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942. (Cited on page 3.)

[22] P. A. Knight. The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. (Cited on page 2.)

[23] N. Lahn, D. Mulchandani, and S. Raghvendra. A graph theoretic additive approximation of optimal transport. *ArXiv Preprint: 1905.11830*, 2019. (Cited on page 2.)

[24] Y. T. Lee and A. Sidford. Path finding methods for linear programming: Solving linear programs in $\widetilde{O}(\text{sqrt}(\text{rank}))$ iterations and faster algorithms for maximum flow. In *FOCS*, pages 424–433. IEEE, 2014. (Cited on page 1.)

[25] Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015. (Cited on pages 2 and 6.)

[26] 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*, 2019. (Cited on pages 2, 5, 6, 13, 14, 20, and 23.)

[27] H. Lu, R. Freund, and V. Mirrokni. Accelerating greedy coordinate descent methods. In *ICML*, pages 3263–3272, 2018. (Cited on pages 2 and 6.)

[28] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015. (Cited on pages 2 and 6.)

[29] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. (Cited on pages 2 and 6.)

[30] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013. (Cited on page 1.)

[31] X. Nguyen. Borrowing strength in hierarchical Bayes: posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016. (Cited on page 1.)

[32] O. Pele and M. Werman. Fast and robust earth movers distance. In *ICCV*. IEEE, 2009. (Cited on page 1.)

[33] K. Quanrud. Approximating optimal transport with linear programs. In *SOSA*, pages 61–69, 2019. (Cited on page 2.)

[34] J. Sherman. Area-convexity, ℓ_∞ regularization, and undirected multicommodity flow. In *STOC*, pages 452–460. ACM, 2017. (Cited on page 2.)

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

[36] M. Sommerfeld, J. Schrieber, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation withex. *ArXiv Preprint: 1802.05570*, 2018. (Cited on page 23.)

[37] S. Srivastava, C. Li, and D. Dunson. Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018. (Cited on pages 1 and 23.)

[38] C. Villani. *Optimal transport: Old and New*. Springer, 2008. (Cited on page 1.)