# 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

June 21, 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 $\widetilde{\mathcal{O}}\left(n^{5/2}/\varepsilon^{3/2}\right)$. This improves on the known complexity bound of $\widetilde{\mathcal{O}}\left(n^2/\varepsilon^2\right)$ for the Sinkhorn and Greenkhorn algorithms. We also present two hybrid algorithms that use the new accelerated algorithms to initialize the Sinkhorn and Greenkhorn algorithms, and we establish complexity bounds of $\widetilde{\mathcal{O}}\left(n^{7/3}/\varepsilon\right)$ for these hybrid algorithms. 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 $\widetilde{\mathcal{O}}\left(n^{5/2}/\varepsilon\right)$, Jambulapati et.al. [19] has designed an algorithm with the complexity bound $\widetilde{\mathcal{O}}\left(n^2/\varepsilon\right)$ 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 the paper, we bring these threads of research together and show that the Sinkhorn and Greenkhorn algorithms can be accelerated directly, and the resulting complexity bound is commensurate with the more specialized acceleration techniques. Our specific contributions can be summarized as follows:

1. We develop an accelerated randomized scheme for the Sinkhorn algorithm, which we refer to as the *Randkhorn algorithm*, which involves exact minimization for the main iterates accompanied by an auxiliary sequence of iterates that are based on a randomized coordinate gradient update. We establish the complexity bound of $\widetilde{\mathcal{O}}\left(n^{5/2}/\varepsilon^{3/2}\right)$ 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 algorithms in terms of $\varepsilon$. We also accelerate the Greenkhorn algorithm, yielding an algorithm that we refer to as the *Gandkhorn algorithm*, and obtain the same improved complexity bound.

2. We present two hybrid algorithms–*hybrid Sinkhorn* and *hybrid Greenkhorn*–in which the Randkhorn and Gandkhorn algorithms are employed to initialize the Sinkhorn and Greenkhorn algorithms. The resulting algorithms are faster than the *Randkhorn* and *Gandkhorn* algorithms in theory, possessing a complexity bound of $\widetilde{\mathcal{O}}\left(n^{7/3}/\varepsilon\right)$.

**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. Based on these algorithms, we introduce the hybrid Sinkhorn and

Greenkhorn algorithms and study their complexity bounds in Section 5. Extensive simulation studies of these algorithms with both synthetic and real data are presented in Section 6. We conclude in Section 7 and defer the proof of remaining key results in the paper to Appendix A.

**Notation.** For any $n \geq 2$, we start with $\Delta^n$ 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 $\ell_p$-norm while $\|x\|$ indicates an $\ell_2$-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 $\varepsilon$, 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 $\varepsilon$. 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 $\varepsilon$.

# 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 $\Delta^n$. 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 *$\varepsilon$-approximation* and $\hat{X}$ an *$\varepsilon$-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 $\varepsilon$-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 $\varphi$ 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 $\ell_1$-norm:

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

By [26, Lemma 4.1], the dual objective $\varphi$ 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 $\ell_2$-norm and the Lipschitz constant is $2\eta$. 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 $\tau^*$ 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 $\ell_2$-norm and $\ell_\infty$-norm. □

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

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

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

  **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**

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

  **else**

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

  **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**

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

  **else**

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

  **end if**

  **Step 5:** $k = k + 1$.

**end while**

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

## 3 Randkhorn: A 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 $\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$, 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 $\varepsilon$. 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.

**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:** $\hat{X}$.

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 $\rho$ 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.

### 3.1 Technical lemmas

In this section, we provide several technical lemmas for bounding the dual objective gap $\delta_k$ in the Randkhorn algorithm: $\delta_k := \mathbb{E}[f(u^k, v^k) - f(u^*, v^*)]$ where $(u^k, v^k)$ 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 $\theta_k$ is given in Step 1 of Algorithm 1, and making use 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*

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

*Proof.* We will show that the following inequalities hold:

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

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

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:

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

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:

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

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:

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

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

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

By simple algebra, we can check that

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

Thus we obtain

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

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 $u^{k+1} = \bar{u}^k + \log(r) - \log\left(r(\bar{u}^k, \bar{v}^k)\right)$ and $v^{k+1} = \bar{v}^k$. This leads to the following equation

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

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

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

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 $\delta^k$.

**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

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

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

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

Additionally, since $\theta_k$ 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 $\ell_2$-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). □

### 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*

$$k \;\leq\; 1 + \frac{4R\sqrt{\eta n \left(28 + 7/\theta_{-1}^2\right)}}{\varepsilon'}, \tag{23}$$

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

*Proof.* Given the iterate $(u^k, v^k)$, we define

$$\hat{u}^k := \operatorname*{argmin}_{u \in \mathbb{R}^n} \; f(u, v^k), \quad \text{and} \quad \hat{v}^k := \operatorname*{argmin}_{v \in \mathbb{R}^n} \; f(u^k, v).$$

We find that

$$\begin{aligned} f(u^k, v^k) - f(u^*, v^*) &\geq f(u^k, v^k) - f(\hat{u}^k, v^k) = \rho\left(r, B(u^k, v^k)\mathbf{1}\right), \\ f(u^k, v^k) - f(u^*, v^*) &\geq f(u^k, v^k) - f(u^k, \hat{v}^k) = \rho\left(l, B(u^k, v^k)^\top \mathbf{1}\right). \end{aligned}$$

Therefore, we conclude that

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

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

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

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

$$E_k^2 \leq \frac{\left(448 + 112/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(k+1)^2}.$$

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,

$$\begin{aligned}
k &\leq 1 + \frac{\sqrt{\left(448 + 112/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}}{\varepsilon'} \\
&\leq 1 + \frac{4R\sqrt{\eta n\left(28 + 7/\theta_{-1}^2\right)}}{\varepsilon'}.
\end{aligned}$$

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

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*

$$\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$$

*arithmetic operations.*

The proof of Theorem 3.4 is provided in Section A.1. The complexity of the Randkhorn algorithm improves upon the best known complexity bound $\mathcal{O}\left(\frac{n^2\|C\|_\infty^2\log(n)}{\epsilon^2}\right)$ for the Sinkhorn algorithm [13] when $\varepsilon$ is sufficiently small. This is supported empirically by the comparative performance of the Randkhorn algorithm with both synthetic and real data in Section 6.

# 4 Gandkhorn: Randomized Greenkhorn Algorithm

We now turn to the Gandkhorn algorithm, a *randomized Greenkhorn algorithm.* The motivation for this algorithm stems from the fact that the Greenkhorn algorithm [3, 26], a greedy coordinate

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

**Input:** $k = 0$, $\theta_{-1} \in (0, 2]$ and $u^0 = v^0 = \tilde{u}_0 = \tilde{v}_0 = \mathbf{0}$.
**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:** $\begin{pmatrix} \bar{u}^k \\ \bar{v}^k \end{pmatrix} = (1 - \theta_k)\begin{pmatrix} u^k \\ v^k \end{pmatrix} + \theta_k \begin{pmatrix} \tilde{u}^k \\ \tilde{v}^k \end{pmatrix}$.
**Step 3:** Compute

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

**if** $\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) \geq \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)$ **then**
$u_I^{k+1} = \bar{u}_I^k + \log(r_I) - \log\left(r_I(\bar{u}^k, \bar{v}^k)\right)$.
**else**
$v_J^{k+1} = \bar{v}_J^k + \log(l_J) - \log\left(l_J(\bar{u}^k, \bar{v}^k)\right)$.
**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**
$\tilde{u}_\pi^{k+1} = \tilde{u}_\pi^k - \dfrac{r_\pi(\bar{u}^k, \bar{v}^k) - r_\pi}{8n\eta\theta_k}$.
**else**
$\tilde{v}_\pi^{k+1} = \tilde{v}_\pi^k - \dfrac{l_\pi(\bar{u}^k, \bar{v}^k) - l_\pi}{8n\eta\theta_k}$.
**end if**
**Step 5:** $k = k + 1$.
**end while**
**Output:** $B(u^k, v^k)$.

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).

We present a complexity analysis for the Gandkhorn algorithm that yields a complexity bound of $\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$, 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 $\varepsilon$.

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 $\rho$ is defined in Section 3. Step 3 of the Gandkhorn algorithm differs from Step 3 of the Randkhorn algorithm in that it is 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 $\delta_k$ in the Gandkhorn algorithm: $\delta_k = \mathbb{E}[f(u^k, v^k) - f(u^*, v^*)]$ where $(u^k, v^k)$ are defined in Algorithm 3. We modify the two sequences of iterates defined in (11) and (12) as follows:

$$\begin{pmatrix} q_u^{k+1} \\ q_v^{k+1} \end{pmatrix} \ := \ \begin{pmatrix} \bar{u}^k \\ \bar{v}^k \end{pmatrix} - \frac{1}{8n\eta}\nabla f(\bar{u}^k, \bar{v}^k), \tag{24}$$

and

$$\begin{pmatrix} \tilde{q}_u^{k+1} \\ \tilde{q}_v^{k+1} \end{pmatrix} \ := \ \begin{pmatrix} \tilde{u}^k \\ \tilde{v}^k \end{pmatrix} - \frac{1}{8n\eta\theta_k}\nabla f(\bar{u}^k, \bar{v}^k). \tag{25}$$

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*

$$\begin{aligned} f(u^{k+1}, v^{k+1}) \ &\leq \ \theta_k f(u^*, v^*) + (1 - \theta_k) f(u^k, v^k) \\ &\quad + 4n\eta\theta_k^2 \left( \left\| \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^* - \tilde{q}_u^{k+1} \\ v^* - \tilde{q}_v^{k+1} \end{pmatrix} \right\|^2 \right). \end{aligned} \tag{26}$$

*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:

$$f(\bar{u}^k, \bar{v}^k) - f(u^{k+1}, v^{k+1}) \;\geq\; \frac{1}{2}\left(\hat{\rho}\left(r_I, r_I(\bar{u}^k, \bar{v}^k)\right) + \hat{\rho}\left(l_J, l_J(\bar{u}^k, \bar{v}^k)\right)\right), \tag{27}$$

$$f(u^{k+1}, v^{k+1}) \;\leq\; f(\bar{u}^k, \bar{v}^k) - \frac{1}{16n\eta}\|\nabla f(\bar{u}^k, \bar{v}^k)\|^2. \tag{28}$$

Using the definitions of $q_u^{k+1}$ and $q_v^{k+1}$ in (24), we find that

$$\begin{aligned} \|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 \;&=\; 2\|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 - \|\nabla f(\bar{u}^k, \bar{v}^k)\|^2 \\ &\overset{(24)}{=}\; 16n\eta \begin{pmatrix} \bar{u}^k - q_u^{k+1} \\ \bar{v}^k - q_v^{k+1} \end{pmatrix}^{\top} \nabla f(\bar{u}^k, \bar{v}^k) - 64n^2\eta^2 \left\| \begin{pmatrix} \bar{u}^k - q_u^{k+1} \\ \bar{v}^k - q_v^{k+1} \end{pmatrix} \right\|^2. \end{aligned}$$

Plugging the above equality into (28) and rearranging yields the following inequality

$$f(u^{k+1}, v^{k+1}) \;\leq\; f(\bar{u}^k, \bar{v}^k) + \begin{pmatrix} q_u^{k+1} - \bar{u}^k \\ q_v^{k+1} - \bar{v}^k \end{pmatrix}^{\top} \nabla f(\bar{u}^k, \bar{v}^k) + 4n\eta \left\| \begin{pmatrix} q_u^{k+1} - \bar{u}^k \\ q_v^{k+1} - \bar{v}^k \end{pmatrix} \right\|^2. \tag{29}$$

Furthermore, invoking the definitions of $\tilde{q}_u^{k+1}$ and $\tilde{q}_v^{k+1}$ in (25), 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 (29) leads to:

$$f(u^{k+1}, v^{k+1}) \;\leq\; f(\bar{u}^k, \bar{v}^k) + \theta_k \left[ \begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k \end{pmatrix}^{\top} \nabla f(\bar{u}^k, \bar{v}^k) + 4n\eta\theta_k \left\| \begin{pmatrix} \tilde{q}_u^{k+1} - \tilde{u}^k \\ \tilde{q}_v^{k+1} - \tilde{v}^k \end{pmatrix} \right\|^2 \right]. \tag{30}$$

Based on the definitions of $\tilde{s}_u^{k+1}$ and $\tilde{s}_v^{k+1}$ in (25), 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

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

By plugging (31) into (30), we arrive at the following result:

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

Simple algebra indicates that

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

Collecting the results, we arrive at the following inequality:

$$f(\bar{u}^k, \bar{v}^k) + \theta_k \begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}^\top \nabla f(\bar{u}^k, \bar{v}^k) \;\leq\; \theta_k f(u^*, v^*) + (1-\theta_k) f(u^k, v^k). \tag{32}$$

Therefore, we conclude the desired inequality (26) by plugging (32) into (32).

**Proof of claim** (27)**:** 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

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

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:

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

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 (27). 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 (27).

**Proof of claim** (28)**:** 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

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

Applying (8) leads to the following inequalities

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

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 (27) yields the result of claim (28). $\square$

We are now ready to bound the dual objective gap $\delta^k$.

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

$$\delta_k \;\leq\; \frac{\left(32n^2 + 8/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(k+1)^2}. \tag{33}$$

*Proof.* First, we estimate the third term of the right-hand side of (26) using a similar approach to the proof of claim (22). In particular, we have the following equality:

$$\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \left\|\begin{pmatrix} u^* - \tilde{s}_u^{k+1} \\ v^* - \tilde{s}_v^{k+1} \end{pmatrix}\right\|^2 \;=\; 2n\left(\left\|\begin{pmatrix} u^* - \tilde{u}^k \\ v^* - \tilde{v}^k \end{pmatrix}\right\|^2 - \mathbb{E}_{\xi^k}\left[\left\|\begin{pmatrix} u^* - \tilde{u}^{k+1} \\ v^* - \tilde{v}^{k+1} \end{pmatrix}\right\|^2\right]\right). \tag{34}$$

which follows directly from the definition of the Gandkhorn algorithm. We now put the result of Lemma 4.1 and equality (34) together, which leads to the following inequality:

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

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

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

Dividing both sides of the inequality by $\theta_k^2$, 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 $\varepsilon'$ 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*

$$k \;\leq\; 1 + \frac{4R\sqrt{\eta n\left(28n^2 + 7/\theta_{-1}^2\right)}}{\varepsilon'}, \tag{35}$$

*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

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

Combining with the results of Lemma 4.2 yields that

$$E_k^2 \;\leq\; \frac{\left(448n^2 + 112/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{(k+1)^2}.$$

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,

$$\begin{aligned} k \;&\leq\; 1 + \frac{\sqrt{\left(448n^2 + 112/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}}{\varepsilon'} \\ &\leq\; 1 + \frac{4R\sqrt{\eta n\left(28n^2 + 7/\theta_{-1}^2\right)}}{\varepsilon'}. \end{aligned}$$

Therefore, we conclude that the number of iterations $k$ satisfies (23). □

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*

$$\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$$

*arithmetic operations.*

The proof of Theorem 4.4 is provided in Section 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 $\varepsilon$ is sufficiently small. Later, in Section 6, we demonstrate empirically that the Gandkhorn algorithm has better performance than the Randkhorn and Greenkhorn algorithms with both synthetic and real data.

**Algorithm 5:** Approximating OT by Hybrid SINKHORN

**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 $\bar{X} = \text{RANDKHORN}\left(C, \eta, \tilde{r}, \tilde{l}, \cdot\right)$ with $\bar{\delta} = n^{1/3}\varepsilon'$.

**Step 3:** Compute $\tilde{X} = \text{SINKHORN}\left(C, \eta, \tilde{r}, \tilde{l}, \varepsilon'/2\right)$ with $\bar{X}$ as an initialization.

**Step 4:** Given Algorithm 2 in [3], we round $\tilde{X}$ to $\hat{X}$ to satisfy $\hat{X}\mathbf{1} = r$ and $\hat{X}^\top\mathbf{1} = l$.

**Algorithm 6:** Approximating OT by Hybrid GREENKHORN

**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 $\bar{X} = \text{GANDKHORN}\left(C, \eta, \tilde{r}, \tilde{l}, \cdot\right)$ with $\bar{\delta} = n^{1/3}\varepsilon'$.

**Step 3:** Compute $\tilde{X} = \text{GREENKHORN}\left(C, \eta, \tilde{r}, \tilde{l}, \varepsilon'/2\right)$ with $\bar{X}$ as an initialization.

**Step 4:** 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:** $\hat{X}$.

## 5 Hybrid Sinkhorn and Greenkhorn Algorithms

In this section, we propose two faster algorithms for the dual regularized OT problem (6) that we refer to as *hybrid Sinkhorn* and *hybrid Greenkhorn*. These algorithms utilize the outputs of Randkhorn and Gandkhorn algorithms as initializations for Sinkhorn and Greenkhorn algorithms, respectively. The pseudocode for these algorithms is presented in Algorithm 5 and Algorithm 6. We demonstrate that both the hybrid Sinkhorn and Greenkhorn algorithms achieve a complexity bound of $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right)$, which improves on the accelerated gradient algorithms in [13, 26, 17] in terms of the dimension $n$. Additionally, this complexity bound is better than that of Randkhorn and Gandkhorn algorithms in terms of both $n$ and $\varepsilon$.

The hybrid algorithms rely on the original Sinkhorn and Greenkhorn algorithms, incorporating the Randkhorn and Gandkhorn algorithms as subroutines in order to achieve acceleration. More precisely, we first run the Randkhorn and Gandkhorn algorithms until the following stopping criterion is satisfied: $\delta_k \leq \bar{\delta} = n^{1/3}\varepsilon'$ where $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$. We then run the Sinkhorn or Greenkhorn algorithms by using the current outputs of Randkhorn and Gandkhorn algorithms as initializations until the criterion $E_k \leq \varepsilon$ is satisfied.

## 5.1 Complexity analysis for the hybrid Sinkhorn algorithm

We first provide an upper bound for the number of iterations $k$ to achieve a desired tolerance $\varepsilon'$ for the iterates of the hybrid Sinkhorn algorithm.

**Theorem 5.1.** *The hybrid Sinkhorn algorithm with $\bar{\delta} = n^{1/3}\varepsilon'$ returns a matrix $B(u^k, v^k)$ that satisfies $E_k \leq \varepsilon'$ with the number of iterations $k$ satisfying the following upper bound*

$$k \;\leq\; 2 + \frac{4n^{1/3}R\sqrt{\eta\left(4 + 1/\theta_{-1}^2\right)}}{\sqrt{\varepsilon'}} + \frac{n^{1/3}}{\varepsilon'}, \tag{36}$$

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

*Proof.* We first estimate the number of iterations $k_1$ required by the Randkhorn algorithm in Step 2 of hybrid Sinkhorn algorithm to reach the criterion $\delta_{k_1} \leq n^{1/3}\varepsilon'$. Invoking the results from Lemma 2.2 and Lemma 3.2, this number of steps satisfies:

$$\begin{aligned} k_1 \;&\leq\; 1 + \sqrt{\frac{\left(32 + 8/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{n^{1/3}\varepsilon'}} \\ &\leq\; 1 + \frac{4R\sqrt{\eta n\left(4 + 1/\theta_{-1}^2\right)}}{n^{1/6}\sqrt{\varepsilon'}} \\ &=\; 1 + \frac{4n^{1/3}R\sqrt{\eta\left(4 + 1/\theta_{-1}^2\right)}}{\sqrt{\varepsilon'}}. \end{aligned}$$

Next, we need to estimate the number of iterations $k_2$ required by the Sinkhorn algorithm in Step 3 to reach the stopping condition $E_{k_2} \leq \varepsilon'$. A direct application of [3, Lemma 2 and Lemma 4] leads to:

$$k_2 \;\leq\; 1 + \frac{\delta_{k_1}}{(\varepsilon')^2} \;\leq\; \frac{n^{1/3}\varepsilon}{(\varepsilon')^2} \;=\; \frac{n^{1/3}}{\varepsilon'}.$$

Since the per-iteration cost of the Sinkhorn and Randkhorn algorithms are the same, we obtain that the number of iterations $k$ such that the hybrid Sinkhorn algorithm returns a matrix $B(u^k, v^k)$ satisfying $E_k \leq \varepsilon'$, where with $\bar{\delta} = n^{1/3}\varepsilon'$, can be upper bounded as follows:

$$k \;=\; k_1 + k_2 \;\leq\; 2 + \frac{4n^{1/3}R\sqrt{\eta\left(4 + 1/\theta_{-1}^2\right)}}{\sqrt{\varepsilon'}} + \frac{n^{1/3}}{\varepsilon'}.$$

As a consequence, we achieve the conclusion of the theorem. □

Equipped with the result of Theorem 5.1 and the scheme of hybrid Sinkhorn algorithm in Algorithm 5, we are able to establish the following result for the complexity of the hybrid Sinkhorn algorithm:

**Theorem 5.2.** *The hybrid Sinkhorn algorithm for approximating the optimal transport problem (Algorithm 5) 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 condition (2) in a total of*

$$\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right).$$

*arithmetic operations.*

The proof of Theorem 5.2 is provided in Section A.3. The result of Theorem 5.2 indicates that the complexity bound for the hybrid Sinkhorn algorithm is better than that of the Sinkhorn algorithm, which is $\widetilde{\mathcal{O}}\left(\frac{n^2}{\varepsilon^2}\right)$, in terms of $\varepsilon$.

### 5.2 Complexity analysis for the hybrid Greenkhorn algorithm

We provide an upper bound for the complexity of the hybrid Greenkhorn. First, we bound the number of iterations $k$ to achieve a desired tolerance $\varepsilon'$.

**Theorem 5.3.** *The hybrid Greenkhorn algorithm with $\bar{\delta} = n^{1/3}\varepsilon'$ returns a matrix $B(u^k, v^k)$ that satisfies $E_k \leq \varepsilon'$ with the number of iterations $k$ satisfying the following upper bound*

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

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

*Proof.* The proof of Theorem 5.3 is similar to that of Theorem 5.1. Based on the results of Lemma 2.2 and Lemma 4.2, the number of iterations $k_1$ required by the Gandkhorn algorithm in Step 2 of hybrid Greenkhorn algorithm to reach the criterion $\delta_{k_1} \leq n^{1/3}\varepsilon'$ satisfies

$$\begin{aligned} k_1 \;&\leq\; 1 + \sqrt{\frac{\left(32n^2 + 8/\theta_{-1}^2\right)\eta\left(\|u^*\|^2 + \|v^*\|^2\right)}{n^{1/3}\varepsilon'}} \\ &\leq\; 1 + \frac{4R\sqrt{\eta n\left(4n^2 + 1/\theta_{-1}^2\right)}}{n^{1/6}\sqrt{\varepsilon'}} \\ &=\; 1 + \frac{4n^{1/3}R\sqrt{\eta\left(4n^2 + 1/\theta_{-1}^2\right)}}{\sqrt{\varepsilon'}}. \end{aligned}$$

Invoking [3, Lemma 5 and Lemma 6], the number of iterations $k_2$ required by the Greenkhorn algorithm in Step 3 of hybrid Greenkhorn algorithm to reach $E_{k_2} \leq \varepsilon'$ can be upper bounded as

$$k_2 \;\leq\; 1 + \frac{n\delta_{k_1}}{(\varepsilon')^2} \;\leq\; \frac{n^{1/3}\varepsilon}{(\varepsilon')^2} \;=\; \frac{n^{4/3}}{\varepsilon'}.$$

Since the per-iteration cost of the Greenkhorn and Gandkhorn algorithms are the same, we obtain that the hybrid Greenkhorn algorithm with $\bar{\delta} = n^{1/3}\varepsilon'$ returns a matrix $B(u^k, v^k)$ that satisfies $E_k \leq \varepsilon'$ in the number of iterations $k$ satisfying

$$k \;=\; k_1 + k_2 \;\leq\; 2 + \frac{4n^{1/3}R\sqrt{\eta\left(4n^2 + 1/\theta_{-1}^2\right)}}{\sqrt{\varepsilon'}} + \frac{n^{4/3}}{\varepsilon'}.$$

As a consequence, we achieve the conclusion of the theorem. □

Putting together the result of Theorem 5.3 and the scheme of Algorithm 6, the hybrid Greenkhorn algorithm has the following complexity upper bound:

**Theorem 5.4.** *The hybrid Greenkhorn algorithm for approximating the optimal transport problem (Algorithm 6) returns a transportation plan* $\hat{X} \in \mathbb{R}^{n\times n}$ *satisfying* $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$ *and condition (2) in a total of*

$$\mathcal{O}\left(\frac{n^{7/3}\,\|C\|_\infty^{3/2}\,\sqrt{\log(n)}}{\varepsilon}\right).$$

*arithmetic operations.*

The proof of Theorem 5.4 is provided in Section A.4. The complexity upper bound of hybrid Greenkhorn algorithm is similar to that of hybrid Sinkhorn algorithm while it is better than that of Greenkhorn algorithm, which is $\widetilde{\mathcal{O}}\left(\frac{n^2}{\varepsilon^2}\right)$, in terms of $\varepsilon$.

# 6 Experiments

In this section, careful comparative experiments with the Randkhorn, Gandkhorn, and hybrid Sinkhorn and Greenkhorn algorithms are conducted 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]. 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 see that, while the Randkhorn algorithm and the hybrid Sinkhorn algorithm consistently outperform the Sinkhorn algorithm, the comparison between the Greenkhorn, Gandkhorn and the hybrid Greenkhorn algorithms need to be discussed case by case.

## 6.1 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 $\ell_1$ 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 $\ell_1$ distance (cf. [3] for an argument of choosing $\ell_1$ 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 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.

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

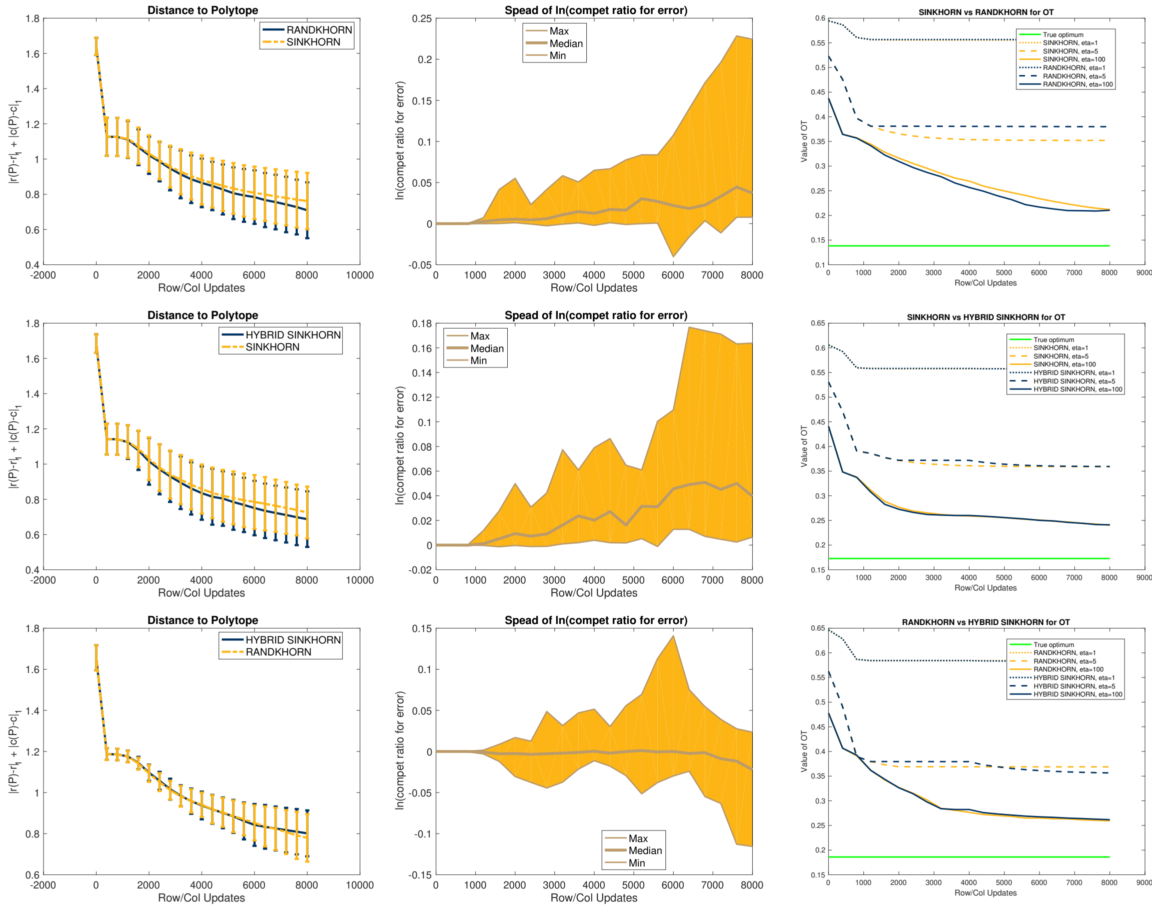


Figure 1: Comparative performance of the Sinkhorn, Randkhorn and hybrid Sinkhorn algorithms on the synthetic images. For the images in the first row, we compare the performance of the Sinkhorn and Randkhorn algorithms based on the number of iteration counts. In the leftmost image of that row, the comparison 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 $\eta \in \{1, 5, 100\}$ with these algorithms and using the value of the optimal transport problem (without the entropic regularization term) as the baseline. Similarly, the second and third rows of images present comparative results for Sinkhorn versus hybrid Sinkhorn and Randkhorn versus hybrid Sinkhorn algorithms.

We perform six pairwise comparative experiments: Sinkhorn versus Randkhorn, Sinkhorn versus hybrid Sinkhorn, Randkhorn versus hybrid Sinkhorn, Greenkhorn versus Gandkhorn, Greenkhorn versus hybrid Greenkhorn, and Gandkhorn versus hybrid Greenkhorn, 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 $\eta$ in the specific set $\{1, 5, 100\}$ 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 $T = 20$. For the hybrid Sinkhorn algorithm, we set $\theta_0 = 0.01$ and the number of iterations for each of the Randkhorn and Sinkhorn subroutines as $3T/4 = 15$. For the hybrid Greenkhorn algorithm, we set $\theta_0 = 0.01$ and the number of iterations for each of

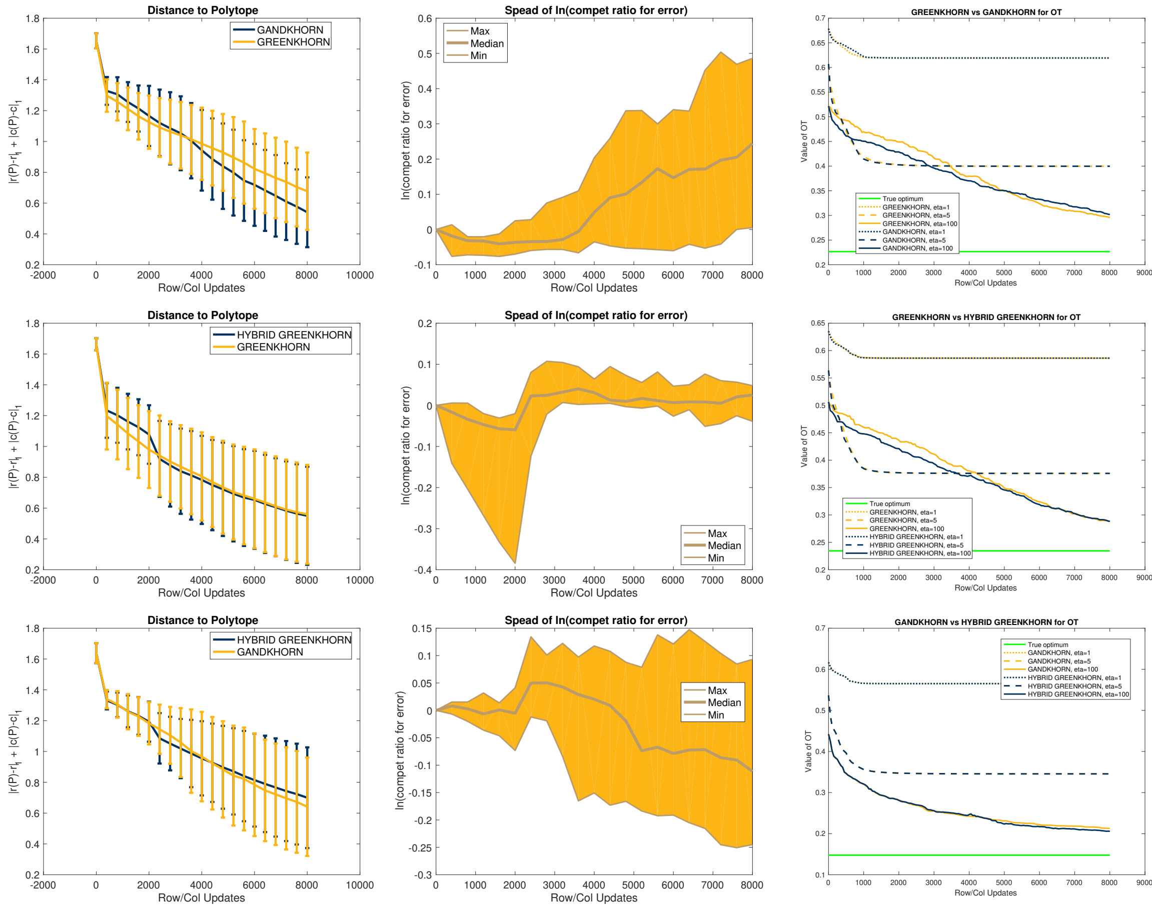


Figure 2: Comparative performance of the Greenkhorn, Gandkhorn and hybrid Greenkhorn algorithms on the synthetic images. The images in the first row, second row, and third row respectively serve for comparing the performance of Greenkhorn and Gandkhorn, Greenkhorn and hybrid Greenkhorn, and Gandkhorn and hybrid Greenkhorn algorithms in terms of iteration counts.

the Gandkhorn and Greenkhorn subroutines as $T/4 = 5$.

We present experimental results in Figure 1 and Figure 2 for different choices of regularization parameters. As can be seen in Figure 1, the Randkhorn and hybrid Sinkhorn algorithms outperform 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. Furthermore, the Randkhorn algorithm slightly outperforms the hybrid Sinkhorn algorithm, suggesting that the complexity bound of the Randkhorn algorithm can be further improved.

Figure 2 shows that the Gandkhorn and hybrid Greenkhorn algorithms outperform the Greenkhorn algorithm in terms of iteration count, supporting the better theoretical complexity of the proposed algorithms over the Greenkhorn algorithm. One interesting phenomenon is that the Gandkhorn and hybrid Greenkhorn algorithms do not behave well at the initial stage. This is possibly because that the Gandkhorn algorithm is not a descent algorithm and is more sensitive to the initial point than the Greenkhorn algorithm. Finally, the behavior of the hybrid Greenkhorn algorithm is similar to that of the Greenkhorn algorithm but slightly worse

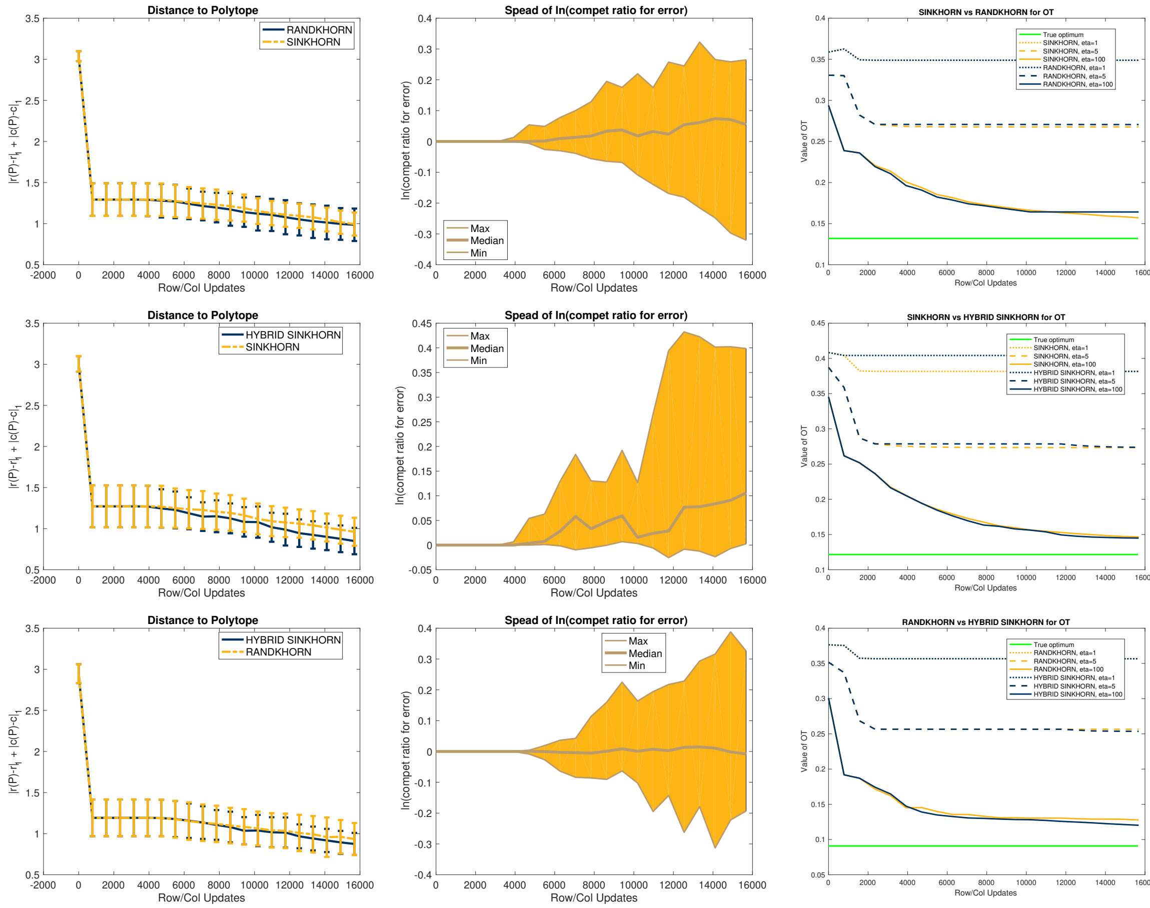


Figure 3: Comparative performance of the Sinkhorn, Randkhorn and hybrid Sinkhorn algorithms on the MNIST real images. See the caption of Figure 1 for more detail.

than the Gankdhorn algorithm.

### 6.2 Experiments on MNIST images

In this section, we use the same evaluation metrics as in Section 6.1 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 3 and Figure 4 with several choices of regularization parameter $\eta$. In Figure 3, the hybrid Sinkhorn algorithm is followed by the Randkhorn algorithm, both outperforming the Sinkhorn algorithm. All the comparative results on real images are consistent with those on the synthetic images. We conclude that both Randkhorn and hybrid Sinkhorn algorithms have favorable practical performance relative to that of the Sinkhorn algorithm.

Figure 4 shows that the performance of the Greenkhorn, Gandkhorn and hybrid Greenkhorn algorithms on MNIST real images is different from that on the synthetic images. The Greenkhorn algorithm performs significantly better than the Gandkhorn algorithm in terms

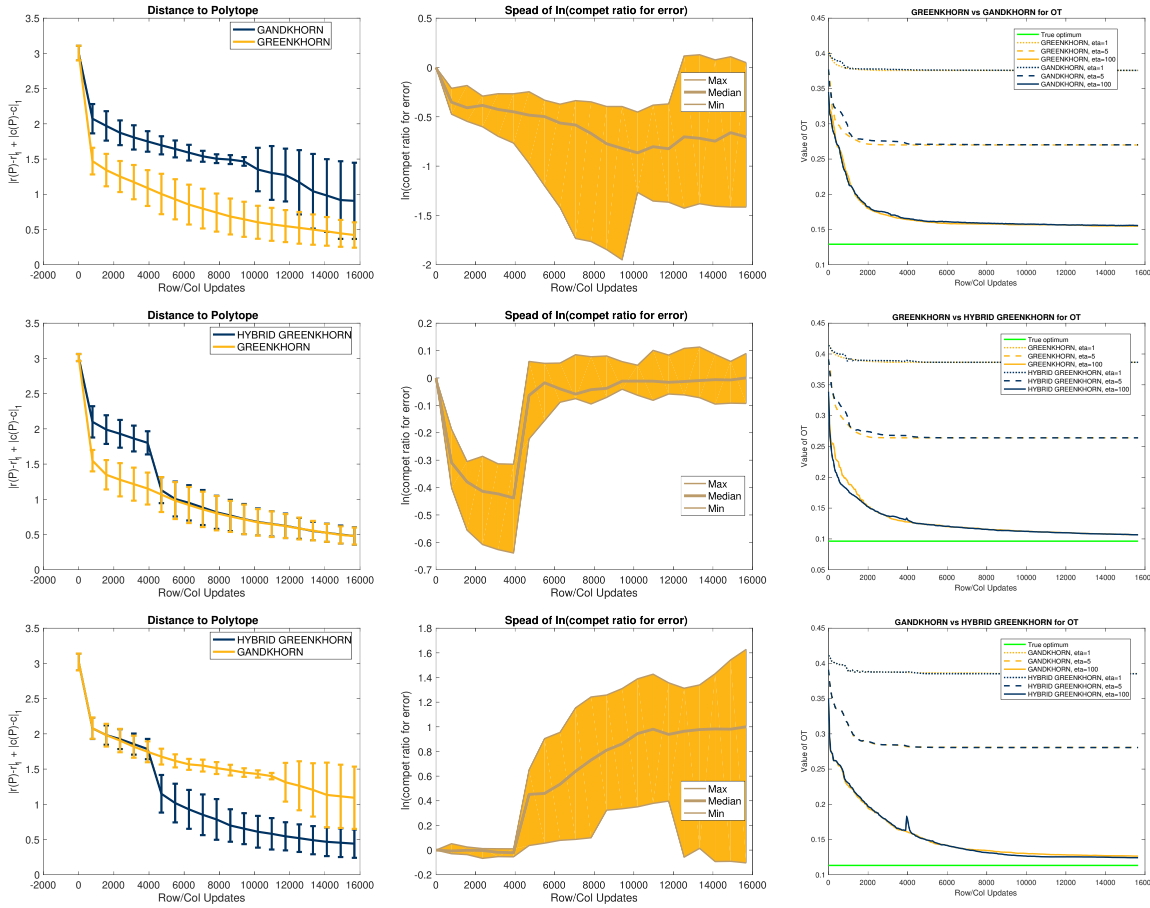


Figure 4: Comparative performance of the Greenkhorn, Gandkhorn and hybrid Greenkhorn algorithms on the MNIST real images. See the caption of Figure 1 for more detail.

of iteration count, while the behavior of the hybrid Greenkhorn algorithm is similar to that of the Greenkhorn algorithm after a few iterations. This phenomenon holds because the problem dimension $n$ of the MNIST real images is larger than that of synthetic image and the dependence of $n$ in the complexity bound of the Gandkhorn algorithm is worse than that of the Greenkhorn and hybrid Greenkhorn algorithms. The hybrid Greenkhorn algorithm achieves a trade-off between $n$ and the tolerance $\varepsilon$ but still suffers from the worse dependence of $n$ in the complexity bound than the Greenkhorn algorithm. Finally, the superior performance of the Greenkhorn algorithm on MNIST real images also tells that the best existing complexity bound, from [26], may not be tight.

# 7 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, randomized version of the Sinkhorn algorithm, which we named the Randkhorn algorithm, and which was shown to have a complexity bound of $\widetilde{\mathcal{O}}\left(\frac{n^{5/2}}{\varepsilon^{3/2}}\right)$. This is more favorable than that of the Sinkhorn algorithm in terms of desired accuracy $\varepsilon$. 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 $\widetilde{\mathcal{O}}\left(\frac{n^{5/2}}{\varepsilon^{3/2}}\right)$, which is comparable to that of Randkhorn algorithm and faster than that of Greenkhorn algorithm in terms of $\varepsilon$. By viewing these new algorithms as subroutines in Sinkhorn and Greenkhorn algorithms, we obtained hybrid Sinkhorn and Greenkhorn algorithms. These hybrid algorithms were demonstrated to achieve a complexity bound of $\widetilde{\mathcal{O}}\left(\frac{n^{7/3}}{\varepsilon}\right)$, which is better than that of Sinkhorn and Greenkhorn algorithms in terms of $\varepsilon$ and that of Randkhorn and Gandkhorn algorithms both in the number of atoms $n$ and the accuracy $\varepsilon$.

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 $\ell_1$ 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

$$
\begin{aligned}
k &\leq 1 + \frac{4R\sqrt{\eta n \left(28 + 7/\theta_{-1}^2\right)}}{\varepsilon'} \\
&\leq 1 + \frac{32\sqrt{n}\,\|C\|_\infty}{\varepsilon}\left(\frac{\|C\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\left\{\tilde{r}_i, \tilde{l}_j\right\}\right)\right)\sqrt{\eta\left(28 + 7/\theta_{-1}^2\right)} \\
&\leq 1 + \frac{32\sqrt{n}\,\|C\|_\infty}{\varepsilon}\left(\frac{4\,\|C\|_\infty \log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\,\|C\|_\infty}\right)\right)\sqrt{\frac{\varepsilon\left(28 + 7/\theta_{-1}^2\right)}{4\log(n)}} \\
&= \mathcal{O}\left(\frac{\sqrt{n}\,\|C\|_\infty^2 \sqrt{\log(n)}}{\varepsilon^{3/2}}\right).
\end{aligned}
$$

Given the above upper bound with $k$, the Randkhorn algorithm has its total iteration complexity bounded by $\mathcal{O}\left(\frac{\sqrt{n}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$. In addition, 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 $\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$. 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

$$
\begin{aligned}
k &\leq 1 + \frac{4R\sqrt{\eta n \left(28n^2 + 7/\theta_{-1}^2\right)}}{\varepsilon'} \\
&\leq 1 + \frac{32\sqrt{n}\,\|C\|_\infty}{\varepsilon}\left(\frac{\|C\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\left\{\tilde{r}_i, \tilde{l}_j\right\}\right)\right)\sqrt{\eta\left(28n^2 + 7/\theta_{-1}^2\right)} \\
&\leq 1 + \frac{32\sqrt{n}\,\|C\|_\infty}{\varepsilon}\left(\frac{4\,\|C\|_\infty \log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\,\|C\|_\infty}\right)\right)\sqrt{\frac{\varepsilon\left(28n^2 + 7/\theta_{-1}^2\right)}{4\log(n)}} \\
&= \mathcal{O}\left(\frac{n^{3/2}\,\|C\|_\infty^2 \sqrt{\log(n)}}{\varepsilon^{3/2}}\right).
\end{aligned}
$$

Based on the above upper bound with $k$, the Gandkhorn algorithm has its total iteration complexity bounded by $\mathcal{O}\left(\frac{n^{3/2}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$. In addition, 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 $\mathcal{O}\left(\frac{n^{5/2}\|C\|_\infty^2\sqrt{\log(n)}}{\varepsilon^{3/2}}\right)$. By the similar argument as in Theorem 3.4, we conclude the desired complexity bound of the Gandkhorn algorithm.

## A.3 Proof of Theorem 5.2

The proof argument is nearly the same as that of Theorem 3.4 in Section A.1. Here, we provide the proof of Theorem 5.2 for completeness. It follows from Theorem 5.1 that the hybrid Sinkhorn algorithm with $\delta_k \leq n^{1/3}\varepsilon'$ returns a matrix $B(u^k, v^k)$ that satisfies $E_k \leq \varepsilon'$ in a number of iterations $k$ satisfying

$$\begin{aligned}
k &\leq 2 + \frac{4n^{1/3}R\sqrt{\eta\left(4 + 1/\theta_{-1}^2\right)}}{\sqrt{\varepsilon'}} + \frac{n^{1/3}}{\varepsilon'} \\
&\leq 2 + \frac{8n^{1/3}\sqrt{2\left\|C\right\|_\infty}}{\sqrt{\varepsilon}}\left(\frac{\left\|C\right\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\left\{\tilde{r}_i, \tilde{l}_j\right\}\right)\right)\sqrt{\eta\left(4 + 1/\theta_{-1}^2\right)} \\
&\qquad + \frac{8n^{1/3}\left\|C\right\|_\infty}{\varepsilon} \\
&\leq 2 + \frac{8n^{1/3}\sqrt{2\left\|C\right\|_\infty}}{\sqrt{\varepsilon}}\left(\frac{4\left\|C\right\|_\infty\log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\left\|C\right\|_\infty}\right)\right)\sqrt{\frac{\varepsilon\left(4 + 1/\theta_{-1}^2\right)}{4\log(n)}} \\
&\qquad + \frac{8n^{1/3}\left\|C\right\|_\infty}{\varepsilon} \\
&= \mathcal{O}\left(\frac{n^{1/3}\left\|C\right\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right).
\end{aligned}$$

In light of the above upper bound with $k$, the hybrid Sinkhorn algorithm has its total iteration complexity bounded by $\mathcal{O}\left(\frac{n^{1/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right)$. Additionally, each iteration of the hybrid Sinkhorn algorithm only requires $\mathcal{O}(n^2)$ arithmetic operations. Combining these two results, a total number of arithmetic operations is of order $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right)$. By a similar argument as that of Theorem 3.4 in Section A.1, the total number of arithmetic operations required for the hybrid Sinkhorn algorithm is of the order $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right)$.

## A.4 Proof of Theorem 5.4

Based on the result of Theorem 5.3, the hybrid Greenkhorn algorithm with $\delta_k \leq n^{1/3}\varepsilon'$ returns a matrix $B(u^k, v^k)$ that satisfies $E_k \leq \varepsilon'$ in a number of iterations $k$ satisfying

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

$$
\begin{aligned}
&\leq \quad 2 + \frac{8n^{1/3}\sqrt{2\left\|C\right\|_\infty}}{\sqrt{\varepsilon}}\left(\frac{4\left\|C\right\|_\infty \log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\left\|C\right\|_\infty}\right)\right)\sqrt{\frac{\varepsilon\left(4n^2 + 1/\theta_{-1}^2\right)}{4\log(n)}} \\
&\qquad + \frac{8n^{4/3}\left\|C\right\|_\infty}{\varepsilon} \\
&= \quad \mathcal{O}\left(\frac{n^{4/3}\left\|C\right\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right).
\end{aligned}
$$

Given the above upper bound with $k$, the hybrid Greenkhorn algorithm has its total iteration complexity bounded by $\mathcal{O}\left(\frac{n^{4/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right)$. Additionally, each iteration of the hybrid Greenkhorn algorithm only requires $\mathcal{O}(n)$ arithmetic operations. Combining these two results, a total number of arithmetic operations is of order $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right)$. Invoking a similar proof argument as that of Theorem 3.4, the total number of arithmetic operations required for the hybrid Greenkhorn algorithm is of the order $\mathcal{O}\left(\frac{n^{7/3}\|C\|_\infty^{3/2}\sqrt{\log(n)}}{\varepsilon}\right)$.

## 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, 7, 12, 14, 19, 20, 21, 22, 25, 27, and 28.)

[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 22.)

[10] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *ICML*, pages 685–693, 2014. (Cited on page 27.)

[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 27.)

[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, 12, and 19.)

[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 page 2.)

[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 pages 2 and 19.)

[18] N. Ho, V. Huynh, D. Phung, and M. I. Jordan. Probabilistic multilevel clustering via composite transportation distance. *AISTATS*, 2019. (Cited on page 27.)

[19] A. Jambulapati, A. Sidford, and K. Tian. A direct õ $(1/\varepsilon)$ 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, 12, 13, 18, 19, 22, 25, and 26.)

[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, $\ell_\infty$ 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 27.)

[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 27.)

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