Journal of Machine Learning Research 23 (2022) 1-42 Submitted 3/20; Revised 2/22; Published 5/22

On the Efficiency of Entropic Regularized Algorithms for Optimal Transport

Tianyi Lin darren_lin@berkeley.edu
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Nhat Ho minhnhat@utexas.edu
Department of Statistics and Data Sciences
University of Texas, Austin
Texas, 78712-1823, USA

Michael I. Jordan jordan@cs.berkeley.edu
Department of Electrical Engineering and Computer Science and Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Editor: Suvrit Sra

Abstract

We present several new complexity results for the entropic regularized algorithms that approximately solve the optimal transport (OT) problem between two discrete probability measures with at most n atoms. First, we improve the complexity bound of a greedy variant of Sinkhorn, known as *Greenkhorn*, from $\widetilde{O}(n^2\varepsilon^{-3})$ to $\widetilde{O}(n^2\varepsilon^{-2})$. Notably, our result can match the best known complexity bound of Sinkhorn and help clarify why Greenkhorn significantly outperforms Sinkhorn in practice in terms of row/column updates as observed by Altschuler et al. (2017). Second, we propose a new algorithm, which we refer to as *APDAMD* and which generalizes an adaptive primal-dual accelerated gradient descent (APDAGD) algorithm (Dvurechensky et al., 2018) with a prespecified mirror mapping ϕ. We prove that APDAMD achieves the complexity bound of $\widetilde{O}(n^2\sqrt{\delta}\varepsilon^{-1})$ in which $\delta > 0$ stands for the regularity of ϕ. In addition, we show by a counterexample that the complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ proved for APDAGD before is invalid and give a refined complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$. Further, we develop a *deterministic* accelerated variant of Sinkhorn via appeal to estimated sequence and prove the complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$. As such, we see that accelerated variant of Sinkhorn outperforms Sinkhorn and Greenkhorn in terms of $1/\varepsilon$ and APDAGD and accelerated alternating minimization (AAM) (Guminov et al., 2021) in terms of n. Finally, we conduct the experiments on synthetic and real data and the numerical results show the efficiency of Greenkhorn, APDAMD and accelerated Sinkhorn in practice.

Keywords: Optimal transport, computational complexity, Sinkhorn, Greenkhorn, accelerated primal-dual algorithm, accelerated Sinkhorn.

1. Introduction

From its origins in the seminal works by Monge (1781) and Kantorovich (1942) respectively in the eighteenth and twentieth centuries, and through to present day, the optimal transport (OT) problem has played a *determinative* role in the theory of mathematics (Villani, 2009). 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, e.g., Nguyen, 2013; Cuturi and Doucet, 2014; Srivastava et al., 2015; Rolet et al., 2016; Peyré et al., 2016; Nguyen, 2016; Carrière et al., 2017; Arjovsky et al., 2017; Gulrajani et al., 2017; Courty et al., 2017; Srivastava et al., 2018; Dvurechenskii et al., 2018; Tolstikhin et al., 2018; Sommerfeld et al., 2019; Lin et al., 2019b; Ho et al., 2019). In these data-driven applications, the focus is on the probability distributions underlying the OT formulation; indeed, these distributions are either empirical distributions which are 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 (Dudley, 1969; Fournier and Guillin, 2015; Weed and Bach, 2019; Lei, 2020), the definition of a likelihood (Sommerfeld and Munk, 2018; Bernton et al., 2019; Blanchet and Murthy, 2019), or as the robust variant of an estimator (Blanchet et al., 2019; Paty and Cuturi, 2019; Balaji et al., 2020). The key challenge is computational (Peyré and Cuturi, 2019). Indeed, the underlying distributions generally involve high-dimensional data and complex models in machine learning (ML) applications.

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 OT problem can be solved exactly using linear programming (LP) solver based on specialized interior-point methods (Pele and Werman, 2009; Lee and Sidford, 2014; van den Brand et al., 2021), reflecting the LP formulation of the OT problem. In this context, van den Brand et al. (2021) have provided a bunch of randomized interior-point algorithms with improved runtimes for solving linear programs with two-sided constraints, leading to a new OT algorithm based on the Laplacian system solvers that achieved the best known complexity bounds of $\tilde{O}(n^2)$. However, it does not provide an effective solution to large-scale machine learning problems in practice since efficient implementations of Laplacian approach are yet unknown. Furthermore, many combinatorial techniques give exact algorithms for the OT problem. Indeed, the Hungarian algorithm (Kuhn, 1955, 1956; Munkres, 1957) solves the assignment problem in $O(n^3)$ time while there are several combinatorial algorithms that can solve the OT problem exactly in $\tilde{O}(n^{2.5})$ time (Gabow and Tarjan, 1991; Orlin and Ahuja, 1992). Combined with the scaling technique, the network simplex algorithms (Orlin et al., 1993; Orlin, 1997) can be used to solve the OT problem exactly in $\tilde{O}(n^3)$ time and Lahn et al. (2019) have recently developed a faster approximation algorithm for the OT problem via appeal to the modification of the algorithm developed in Gabow and Tarjan (1991). However, computing the OT problem exactly results in an output that is *not* differentiable with respect to measures' locations or weights (Bertsimas and Tsitsiklis, 1997). Moreover, OT suffers from the curse of dimensionality (Dudley, 1969; Fournier and Guillin, 2015) and is thus likely to be meaningless when used on samples from high-dimensional densities.

An alternative to solve the OT problem is a class of approximation algorithms based on the entropy regularization which has been investigated in optimization and transportation science long before (Sinkhorn, 1974; Schneider and Zenios, 1990; Kalantari and Khachiyan, 1996; Knight, 2008; Kalantari et al., 2008; Chakrabarty and Khanna, 2018). It was Cuturi (2013) that popularized the use of entropy regularization for OT in the machine learning community and then initiated a productive line of research where an entropic regularization was imposed to approximate the non-negative constraints in the original OT problem. The resulting problem is referred to as *entropic regularized OT* and the corresponding class of approximation algorithms are called *entropic regularized algorithms*. It is worth mentioning that the entropic regularized OT has many favorable properties that the OT does not enjoy, motivating us to study the computational efficiency of entropic regularized algorithms in this paper. More specifically, from a statistical point of view, the entropic regularized OT enjoys significantly better sample complexity that is polynomial in the dimension (Genevay et al., 2019; Mena and Niles-Weed, 2019; Chizat et al., 2020), demonstrating that adding an entropy regularization will reduce the curse of dimensionality. Even from a computational point of view, such regularization in OT leads to *Sinkhorn* which attains a first near-linear time guarantee for the OT problem (Cuturi, 2013; Altschuler et al., 2017; Dvurechensky et al., 2018), and also makes the problem differentiable with regards to distributions (Feydy et al., 2019); hence, the entropic regularized algorithms are more easily applicable to deep learning applications (Courty et al., 2017; Cuturi et al., 2019; Balaji et al., 2020) as opposed to combinatorial algorithms. This point was highlighted in Dong et al. (2020) and further necessitated the development of faster entropic regularized algorithms. In this regard, the greedy variant of Sinkhorn – Greenkhorn – was proposed and shown to outperform Sinkhorn empirically (Altschuler et al., 2017). However, a sizable gap exists here since the best known complexity bound of $\widetilde{O}(n^2\varepsilon^{-3})$ for Greenkhorn (Altschuler et al., 2017) is worse than that of $\widetilde{O}(n^2\varepsilon^{-2})$ for Sinkhorn (Dvurechensky et al., 2018).

Further progress has been made by adapting first-order optimization algorithms for the OT problem (Cuturi and Peyré, 2016; Genevay et al., 2016; Blondel et al., 2018; Dvurechensky et al., 2018; Altschuler et al., 2019; Guo et al., 2020; Guminov et al., 2021). Among these approaches, two of representatives are an adaptive primal-dual accelerated gradient descent (APDAGD) algorithm (Dvurechensky et al., 2018) with the claimed complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ and an accelerated alternating minimization (AAM) algorithm (Guminov et al., 2021) with the complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$. Moreover, there are several second-order optimization algorithms (Allen-Zhu et al., 2017; Cohen et al., 2017) which are adapted for the OT problem (Blanchet et al., 2018; Quanrud, 2019) and guaranteed to achieve the improved complexity bound of $\widetilde{O}(n^2\varepsilon^{-1})$. However, the aforementioned second-order algorithms do not provide effective solutions to large-scale machine learning problems due to the lack of efficient implementations in practice.

Contributions. Given the advantages of entropic regularization in OT, we focus in his paper the computational efficiency of a class of entropic regularized algorithms for the OT problem and our theoretical analysis lead to several improvements over the state-of-the-art algorithms in the literature. We summarize the contributions as follows:

1. We improve the complexity bound of Greenkhorn from $\widetilde{O}(n^2\varepsilon^{-3})$ to $\widetilde{O}(n^2\varepsilon^{-2})$, which matches the best existing bound of Sinkhorn. The proof techniques are new and differ-

ent from that used in Dvurechensky et al. (2018) for analyzing Sinkhorn. In particular, Greenkhorn only updates a single row or column at each iteration and quantifying the per-iteration progress is more difficult than the measurement in Sinkhorn.

2. We propose an adaptive primal-dual accelerated mirror descent (APDAMD) algorithm which generalizes APDAGD with a prespecified mirror mapping ϕ and prove that APDAMD achieves the complexity bound of $\widetilde{O}(n^2\sqrt{\delta}\varepsilon^{-1})$ where $\delta > 0$ refers to the regularity of ϕ w.r.t. ℓ_∞ norm. We show by a counterexample that the complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ proved for APDAGD (Dvurechensky et al., 2018) is invalid and give a refined complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$ which is worse than the claimed bound in terms of n.

3. We propose a deterministic accelerated variant of Sinkhorn via appeal to an estimated sequence and prove the complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$. In particular, accelerated Sinkhorn consists in an exact minimization for main iterates accompanied by another sequence of iterates based on coordinate gradient updates and monotone search. Our results show that accelerated Sinkhorn outperforms Sinkhorn and Greenkhorn in terms of $1/\varepsilon$ and APDAGD and AAM in terms of n.

We note that a preliminary version with only the analysis for Greenkhorn and APDAMD has been accepted by ICML (Lin et al., 2019a). After our conference paper was published, some new algorithms were developed for solving the OT problem (Jambulapati et al., 2019; Lahn et al., 2019). In particular, Jambulapati et al. (2019) developed a dual extrapolation algorithm with the complexity bound $\widetilde{O}\left(n^2\varepsilon^{-1}\right)$ using an area-convex mapping (Sherman, 2017). Despite the theoretically sound complexity bound, the lack of simplicity and ease-of-implementation make this algorithm less competitive with Sinkhorn and Greenkhorn which remain the baseline solution methods in practice (Flamary and Courty, 2017).

Different from the algorithm in Jambulapati et al. (2019), the combinatorial algorithm in Lahn et al. (2019) is a practical solution method for the OT problem. It is worth mentioning that the algorithm in Lahn et al. (2019) and other combinatorial algorithms, e.g., the Hungarian algorithm, outperform our algorithms in practice. This is in consistence with the observation in Dong et al. (2020) who pointed out that combinatorial algorithms can outperform entropic regularized algorithms in speed even the latter ones are asymptotically faster for OT (i.e., the case of large n). However, we believe our results are still valuable due to the importance of entropic regularized algorithms as mentioned before.

Organization. The remainder is organized as follows. In Section 2, we present the basic setup for the primal and dual form of the entropic regularized OT problem. In Section 3, we provide the complexity analysis for Greenkhorn. In Section 4, we propose APDAMD for solving entropic regularized OT and provide several results on the complexity bound of APDAGD and APDAMD. In Section 5, we propose and analyze an accelerated variant of Sinkhorn. In Section 6, we conduct the experiments on synthetic and real data and the numerical results show the efficiency of our algorithms. We conclude this paper in Section 7.

Notation. For $n \geq 2$, we let $[n]$ be the set $\{1, 2, \ldots, n\}$ and $\mathbb{R}^n_+$ be the set of all vectors in $\mathbb{R}^n$ with non-negative coordinates. The notation $\Delta^n = \{v \in \mathbb{R}^n_+ : \sum_{i=1}^n v_i = 1\}$ stands for a probability simplex in $n-1$ dimensions. For a vector $x \in \mathbb{R}^n$ and let $1 \leq p < +\infty$, the notation $\|x\|_p$ stands for the ℓ_p-norm and $\|x\|$ indicates an ℓ_2-norm. $\text{diag}(x)$ is a diagonal matrix which has the vector x on its diagonal. $\mathbf{1}_n$ and $\mathbf{0}_n$ are n-dimensional vector with all components being 1 and 0. For a matrix $A \in \mathbb{R}^{n\times n}$, we denote $\text{vec}(A)$ as the vector in $\mathbb{R}^{n^2}$ obtained from concatenating the rows and columns of A. The notation $\|A\|_{1\to 1}$ stands for $\sup_{\|x\|_1=1} \|Ax\|_1$ and the notations $r(A) = A\mathbf{1}_n$ and $c(A) = A^\top \mathbf{1}_n$ stand for the row and column sums respectively. For a function f, the notation $\nabla_x f$ denotes a partial derivative with respect to x. For the dimension n and tolerance $\varepsilon > 0$, the notations $a = O(b(n,\varepsilon))$ and $a = \Omega(b(n,\varepsilon))$ indicate that $a \leq C_1 \cdot b(n,\varepsilon)$ and $a \geq C_2 \cdot b(n,\varepsilon)$ respectively where C_1 and C_2 are independent of n and ε. We also denote $a = \Theta(b(n,\varepsilon))$ iff $a = O(b(n,\varepsilon)) = \Omega(b(n,\varepsilon))$. Similarly, we denote $a = \widetilde{O}(b(n,\varepsilon))$ to indicate the previous inequality where C_1 depends on some logarithmic function of n and ε.

2. Problem Setup

In this section, we first present the linear programming (LP) representation of the optimal transport (OT) problem as well as a specification of an approximate transportation plan. We also present an entropic regularized variant of the OT problem and derive the dual form where the objective function is in the form of the logarithm of sum of exponents. Finally, we establish several properties of that dual form which are useful for the subsequent analysis.

2.1 Linear programming representation

According to Kantorovich (1942), the problem of approximating the OT distance is equivalent to solving the following linear programming (LP) problem:

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

In the above formulation, X refers to the *transportation plan*, $C = (C_{ij}) \in \mathbb{R}_+^{n\times n}$ stands for a cost matrix with non-negative components, and $r \in \mathbb{R}^n$ and $c \in \mathbb{R}^n$ are two probability distributions in the simplex Δ^n.

We see from Eq. (1), that the OT problem is a LP with $2n$ equality constraints and n^2 variables and can be solved by the interior-point method; however, this method performs poorly on large-scale problems due to its high per-iteration computational cost. In general, the solution that the algorithms aim at achieving is an ε-approximate transportation plan $\widehat{X} \in \mathbb{R}_+^{n\times n}$ satisfying the marginal distribution constraints $\widehat{X}\mathbf{1}_n = r$ and $\widehat{X}^\top\mathbf{1}_n = c$ and the inequality given by

$$\langle C, \widehat{X}\rangle \leq \langle C, X^\star\rangle + \varepsilon.$$

Here $X^\star$ is defined as an optimal transportation plan for the OT problem. For simplicity, we respectively denote $\langle C, \hat{X}\rangle$ an *ε-approximate transportation cost* and $\hat{X}$ an *ε-approximate transportation plan* for the original problem. Formally, we have the following definition of ε-approximate transportation plan.

Definition 1. *The matrix $\hat{X} \in \mathbb{R}_+^{n\times n}$ is called an ε-approximate transportation plan if $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top\mathbf{1}_n = c$ and the following inequality holds true,*

$$\langle C, \widehat{X}\rangle \leq \langle C, X^\star\rangle + \varepsilon.$$

where $X^\star$ is defined as an optimal transportation plan for the OT problem.

With this definition in mind, the goal of this paper is to study the OT problem from a computational point of view. Indeed, we hope to derive an improved complexity bound of the current state-of-the-art algorithms and seek new practical algorithms whose running time required to obtain an ε-approximate transportation plan has better dependence on $1/\varepsilon$ than the benchmark algorithms in the literature. The aforementioned new algorithms are favorable in the machine learning applications where high precision (ε is small) is necessary.

2.2 Entropic regularized OT and its dual form

Seeking another formulation for OT distance that is more amenable to computationally efficient algorithms, Cuturi (2013) proposed to solve an entropic regularized version of the OT problem in Eq. (1), which is given by

$$\min_{X \in \mathbb{R}^{n \times n}} \langle C, X \rangle - \eta H(X), \quad \text{s.t. } X\mathbf{1}_n = r, X^\top \mathbf{1}_n = c, \tag{2}$$

where $\eta > 0$ denotes the regularization parameter and $H(X)$ denotes the entropic regularization term, which is given by:

$$H(X) := -\langle X, \log(X) - \mathbf{1}_{n \times n} \rangle.$$

Note that, the optimal solution of the entropic regularized OT problem exists since the objective function $\langle C, X \rangle - \eta H(X)$ is continuous and the feasible region $\{X \in \mathbb{R}^{n \times n} : X \geq 0, X\mathbf{1}_n = r, X^\top \mathbf{1}_n = c\}$ is compact. Furthermore, that optimal solution is also unique since the objective function $\langle C, X \rangle - \eta H(X)$ is strongly convex over the feasible region with respect to ℓ_1-norm. However, the optimal value of the entropic regularized OT problem (cf. Eq (2)) yields a poor approximation to the unregularized OT problem if η is large. An additional issue of entropic regularization is that the sparsity of the solution is lost. Even though an ε-approximate transportation plan can be found efficiently, it is not clear how different the sparsity pattern of this solution is with respect to the solution of the actual OT problem. In contrast, the actual OT distance suffers from the curse of dimensionality (Dudley, 1969; Fournier and Guillin, 2015; Weed and Bach, 2019) and is significantly worse than its entropic regularized version in terms of the sample complexity (Genevay et al., 2019; Mena and Niles-Weed, 2019; Chizat et al., 2020).

While there is an ongoing debate in the literature on the merits of solving the OT problem *v.s.* its entropic regularized version, we adopt here the viewpoint that reaching an additive approximation of the actual OT cost matters and therefore propose to scale η as a function of the desired accuracy of the approximation. Then, we proceed to derive the dual form of the entropic regularized OT problem in Eq. (2) and show that it remains an unconstrained smooth optimization problem. By introducing the dual variables $\alpha, \beta \in \mathbb{R}^n$, we define the Lagrangian function of the entropic regularized OT problem as follows:

$$\mathcal{L}(X, \lambda_1, \ldots, \lambda_m) = \langle C, X \rangle - \eta H(X) - \alpha^\top (X\mathbf{1}_n - r) - \beta^\top (X^\top \mathbf{1}_n - c). \tag{3}$$

In order to derive the smooth dual objective function, we consider the following minimization problem:

$$\min_{X : \|X\|_1 = 1} \langle C, X \rangle - \eta H(X) - \alpha^\top (X\mathbf{1}_n - r) - \beta^\top (X^\top \mathbf{1}_n - c).$$

The above objective function is strongly convex over the domain $\{X \in \mathbb{R}_+^{n\times n} \mid \|X\|_1 = 1\}$. Thus, the optimal solution is unique. After the simple calculations, the optimal solution $\bar{X} = X(\alpha, \beta)$ has the following form:

$$\bar{X}_{ij} = \frac{e^{\eta^{-1}(\alpha_i+\beta_j-C_{ij})}}{\sum_{1\le i,j\le n} e^{\eta^{-1}(\alpha_i+\beta_j-C_{ij})}}. \tag{4}$$

Plugging Eq. (4) into Eq. (3) yields that the dual form is:

$$\max_{\alpha,\beta} \left\{ -\eta \log\left(\sum_{1\le i,j\le n} e^{\eta^{-1}(\alpha_i+\beta_j-C_{ij})} \right) + \alpha^\top r + \beta^\top c \right\}.$$

In order to streamline our presentation, we perform a change of variables, $u = \eta^{-1}\alpha$ and $v = \eta^{-1}\beta$, and reformulate the above problem as

$$\min_{\alpha,\beta} \ \varphi(\alpha, \beta) := \log\left(\sum_{1\le i,j\le n} e^{u_i+v_j-\frac{C_{ij}}{\eta}} \right) - u^\top r - v^\top c.$$

To further simplify the notation, we define $B(u, v) := (B_{ij})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$ by

$$B_{ij} = e^{u_i+v_j-\frac{C_{ij}}{\eta}}.$$

To this end, we obtain the *dual entropic regularized OT problem* defined by

$$\min_{u,v} \ \varphi(u, v) := \log(\|B(u, v)\|_1) - u^\top r - v^\top c. \tag{5}$$

Remark 2. *The first part of the objective function φ is in the form of the logarithm of sum of exponents while the second part is a linear function. This is different from the objective function used in previous dual entropic regularized OT problem (Cuturi, 2013; Altschuler et al., 2017; Dvurechensky et al., 2018; Lin et al., 2019a). Notably, Eq.* (5) *is a special instance of a softmax minimization problem, and the objective function φ is known to be smooth (Nesterov, 2005). Finally, we point out that the same formulation has been derived in Guminov et al. (2021) for analyzing AAM.*

In the remainder of the paper, we also denote $(u^\star, v^\star) \in \mathbb{R}^{2n}$ as an optimal solution of the dual entropic regularized OT problem in Eq. (5).

2.3 Properties of dual entropic regularized OT

We present several useful properties of the dual entropic regularized OT in Eq. (5). In particular, we show that there exists an optimal solution $(u^\star, v^\star) \in \mathbb{R}^{2n}$ such that it has an upper bound in terms of the ℓ_∞-norm.

Lemma 3. *For the dual entropic regularized OT problem in Eq.* (5)*, there exists an optimal solution $(u^\star, v^\star)$ such that*

$$\|u^\star\|_\infty \le R, \qquad \|v^\star\|_\infty \le R,$$

where $R := \eta^{-1}\|C\|_\infty + \log(n) - \log(\min_{1\le i,j\le n}\{r_i, c_j\})$ depends on C, r and c.

Proof First, we claim that there exists an optimal solution $(u^\star, v^\star)$ such that

$$\max_{1 \leq i \leq n} u_i^\star \geq 0 \geq \min_{1 \leq i \leq n} u_i^\star, \qquad \max_{1 \leq i \leq n} v_i^\star \geq 0 \geq \min_{1 \leq i \leq n} v_i^\star. \tag{6}$$

Indeed, letting $(\widehat{u}^\star, \widehat{v}^\star)$ be an optimal solution to Eq. (5), the claim holds true if $(\widehat{u}^\star, \widehat{v}^\star)$ satisfies Eq. (6). Otherwise, we define the shift term given by

$$\begin{aligned} \widehat{\Delta}_u &= \frac{\max_{1 \leq i \leq n} \widehat{u}_i^\star + \min_{1 \leq i \leq n} \widehat{u}_i^\star}{2}, \\ \widehat{\Delta}_v &= \frac{\max_{1 \leq i \leq n} \widehat{v}_i^\star + \min_{1 \leq i \leq n} \widehat{v}_i^\star}{2}, \end{aligned}$$

and define $(u^\star, v^\star)$ by

$$u^\star = \widehat{u}^\star - \widehat{\Delta}_u \mathbf{1}_n, \qquad v^\star = \widehat{v}^\star - \widehat{\Delta}_v \mathbf{1}_n.$$

By definition, we have $(u^\star, v^\star)$ satisfies Eq. (6). Since $\mathbf{1}_n^\top r = \mathbf{1}_n^\top c = 1$, we have $(u^\star)^\top r = (\widehat{u}^\star)^\top r - \widehat{\Delta}_u$ and $(v^\star)^\top c = (\widehat{v}^\star)^\top c - \widehat{\Delta}_v$. In addition, $\log(\|B(u^\star, v^\star)\|_1) = \log(\|B(\widehat{u}^\star, \widehat{v}^\star)\|_1) + \widehat{\Delta}_u + \widehat{\Delta}_v$. Putting these pieces together yields $\varphi(u^\star, v^\star) = \varphi(\widehat{u}^\star, \widehat{v}^\star)$. Therefore, $(u^\star, v^\star)$ is an optimal solution of the dual entropic regularized OT that satisfies Eq. (6).

Then, we show that

$$\max_{1 \leq i \leq n} u_i^\star - \min_{1 \leq i \leq n} u_i^\star \leq \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1 \leq i, j \leq n} \{r_i, c_j\}\right), \tag{7}$$

$$\max_{1 \leq i \leq n} v_i^\star - \min_{1 \leq i \leq n} v_i^\star \leq \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1 \leq i, j \leq n} \{r_i, c_j\}\right). \tag{8}$$

Indeed, for any $1 \leq i \leq n$, we derive from the optimality condition of $(u^\star, v^\star)$ that

$$\frac{e^{u_i^\star}(\sum_{j=1}^n e^{v_j^\star - \eta^{-1} C_{ij}})}{\|B(u^\star, v^\star)\|_1} = r_i, \quad \text{for all } i \in [n].$$

Since $C_{ij} \geq 0$ for all $1 \leq i, j \leq n$ and $r_i \geq \min_{1 \leq i, j \leq n}\{r_i, c_j\}$ for all $1 \leq i \leq n$, we have

$$u_i^\star \geq \log\left(\min_{1 \leq i, j \leq n} \{r_i, c_j\}\right) - \log\left(\sum_{j=1}^n e^{v_j^\star}\right) + \log(\|B(u^\star, v^\star)\|_1), \quad \text{for all } i \in [n].$$

Since $0 < r_i \leq 1$ and $C_{ij} \leq \|C\|_\infty$, we have

$$u_i^\star \leq \frac{\|C\|_\infty}{\eta} - \log\left(\sum_{j=1}^n e^{v_j^\star}\right) + \log(\|B(u^\star, v^\star)\|_1), \quad \text{for all } i \in [n].$$

Putting these pieces together yields Eq. (7). By the similar argument, we can prove Eq. (8).

Finally, we prove our main results. Indeed, Eq. (6) and Eq. (7) imply that

$$-\frac{\|C\|_\infty}{\eta} + \log\left(\min_{1 \leq i, j \leq n} \{r_i, c_j\}\right) \leq \min_{1 \leq i \leq n} u_i^\star \leq 0,$$

and

$$0 \leq \max_{1 \leq i \leq n} u_i^\star \leq \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1 \leq i,j \leq n} \{r_i, c_j\}\right).$$

Combining the above two inequalities with the definition of R implies that $\|u^\star\|_\infty \leq R$. By the similar argument, we can prove that $\|v^\star\|_\infty \leq R$. As a consequence, we obtain the conclusion of the lemma. ■

The upper bound for the ℓ_∞-norm of an optimal solution of dual entropic regularized OT in Lemma 3 directly leads to the following direct bound for the ℓ_2-norm.

Corollary 4. *For the dual entropic regularized OT problem in Eq.* (5)*, there exists an optimal solution* $(u^\star, v^\star)$ *such that*

$$\|u^\star\| \leq \sqrt{n}R, \qquad \|v^\star\| \leq \sqrt{n}R,$$

where $R > 0$ *is defined in Lemma 3.*

Since the function $-H(X)$ is strongly convex with respect to the ℓ_1-norm on the probability simplex $Q \subseteq \mathbb{R}^{n \times n}$, the entropic regularized OT problem in Eq. (2) is a special case of the following linearly constrained convex optimization problem:

$$\min_{x \in Q} \ f(x), \quad \text{s.t. } Ax = b,$$

where f is strongly convex with respect to the ℓ_1-norm on the set Q:

$$f(x') - f(x) - (x' - x)^\top \nabla f(x) \geq \frac{\eta}{2}\|x' - x\|_1^2 \text{ for any } x', x \in Q.$$

By Nesterov (2005, Theorem 1) with the ℓ_2-norm for the dual space of the Lagrange multipliers, the dual objective function $\tilde{\varphi}$ satisfies the following inequality:

$$\widetilde{\varphi}(\alpha', \beta') - \widetilde{\varphi}(\alpha, \beta) - \begin{pmatrix} \alpha' - \alpha \\ \beta' - \beta \end{pmatrix}^\top \nabla\widetilde{\varphi}(\alpha, \beta) \leq \frac{\|A\|_{1 \to 2}^2}{2\eta} \left\| \begin{pmatrix} \alpha' - \alpha \\ \beta' - \beta \end{pmatrix} \right\|^2 \text{ for any } (\alpha', \beta'), (\alpha, \beta) \in \mathbb{R}^{2n}.$$

Recall that the function $\tilde{\varphi}$ is given by

$$\widetilde{\varphi}(\alpha, \beta) = -\eta \log\left(\sum_{1 \leq i,j \leq n} e^{\eta^{-1}(\alpha_i + \beta_j - C_{ij})}\right) + \alpha^\top r + \beta^\top c. \tag{9}$$

We notice that the function φ in Eq. (5) satisfies that $\varphi(u, v) = -\eta^{-1}\widetilde{\varphi}(\eta u, \eta v)$. After some simple calculations, we have

$$\varphi(u', v') - \varphi(u, v) - \begin{pmatrix} u' - u \\ v' - v \end{pmatrix}^\top \nabla\varphi(u, v) \leq \left(\frac{\|A\|_{1 \to 2}^2}{2}\right) \left\| \begin{pmatrix} u' - u \\ v' - v \end{pmatrix} \right\|^2. \tag{10}$$

In the entropic regularized OT problem, each column of the matrix A contains no more than two nonzero elements which are equal to one. Since $\|A\|_{1 \to 2}$ is equal to maximum ℓ_2-norm of the column of this matrix, we have $\|A\|_{1 \to 2} = \sqrt{2}$. Thus, the dual objective function φ is 2-gradient Lipschitz with respect to the ℓ_2-norm.

Algorithm 1 GREENKHORN$(C, \eta, r, c, \varepsilon')$

Input: $t = 0$ and $u^0 = v^0 = \mathbf{0}_n$.
while $E_t > \varepsilon'$ **do**
  Compute $I = \text{argmax}_{1 \leq i \leq n} \rho(r_i, r_i(B(u^t, v^t)))$ where $\rho(a, b) = b - a + a\log(a/b)$ and $(B(u^t, v^t))_{i'j'} = e^{u^t_{i'} + v^t_{j'} - \frac{C_{i'j'}}{\eta}}$ for all (i', j').
  Compute $J = \text{argmax}_{1 \leq j \leq n} \rho(c_j, c_j(B(u^t, v^t)))$.
  if $\rho(r_i, r_i(B(u^t, v^t))) > \rho(c_j, c_j(B(u^t, v^t)))$ **then**
    $u^{t+1}_I = u^t_I + \log(r_I) - \log(r_I(B(u^t, v^t)))$.
  else
    $v^{t+1}_J = v^t_J + \log(c_J) - \log(c_J(B(u^t, v^t)))$.
  end if
  Increment by $t = t + 1$.
end while
Output: $B(u^t, v^t)$.

3. Greenkhorn

In this section, we present a complexity analysis for Greenkhorn. In particular, we improve the existing best known complexity bound $O(n^2\|C\|_\infty^3 \log(n)\varepsilon^{-3})$ (Altschuler et al., 2017) to $O(n^2\|C\|_\infty^2 \log(n)\varepsilon^{-2})$, which matches the current state-of-the-art complexity bound for Sinkhorn (Dvurechensky et al., 2018).

To facilitate the subsequent discussion, we present the pseudocode of Greenkhorn in Algorithm 1 and its application to regularized OT in Algorithm 2. The function for quantifying the progress in the dual objective value between two consecutive iterates is given by $\rho(a, b) = b - a + a\log(a/b)$ and we recall that (u, v) is an optimal solution of the dual entropic regularized OT problem in Eq. (5) if $r(B(u, v)) - r = \mathbf{0}_n$ and $c(B(u, v)) - c = \mathbf{0}_n$. This leads to the quantity which measures the error of the t-th iterate in Algorithm 1:

$$E_t := \|r(B(u^t, v^t)) - r\|_1 + \|c(B(u^t, v^t)) - c\|_1.$$

Both Sinkhorn and Greenkhorn can be interpreted as coordinate descent for minimizing the following convex function (Cuturi, 2013; Altschuler et al., 2017; Dvurechensky et al., 2018; Lin et al., 2019a):

$$f(u, v) := \|B(u, v)\|_1 - u^\top r - v^\top c. \tag{11}$$

Comparing to the scheme of Sinkhorn that consists in the updates of *all* rows and columns, Algorithm 1 updates only *one* row or column at each step. As such, Algorithm 1 updates only $O(n)$ entries per iteration rather than $O(n^2)$ in Sinkhorn. It is also worth mentioning that Algorithm 1 can be implemented such that each iteration runs in only $O(n)$ arithmetic operations (Altschuler et al., 2017).

Despite cheap per-iteration computational cost, it is difficult to quantify the per-iteration progress of Algorithm 1 and the proof techniques for Sinkhorn in Dvurechensky et al. (2018) are not applicable here. This motivates us to investigate another proof strategy which will be elaborated in the sequel.

Algorithm 2 Approximating OT by Algorithm 1

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.
Step 2: Compute $\tilde{X} = \textsc{Greenkhorn}(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$.
Step 3: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. (2017, Algorithm 2) such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$.
Output: $\hat{X}$.

3.1 Complexity analysis—bounding dual objective values

Given the definition of E_t, we first prove the following lemma which yields an upper bound for the objective values of the iterates.

Lemma 5. *Letting $\{(u^t, v^t)\}_{t \geq 0}$ be the iterates generated by Algorithm 1, we have*

$$f(u^t, v^t) - f(u^\star, v^\star) \leq 2E_t(\|u^\star\|_\infty + \|v^\star\|_\infty),$$

where (u^*, v^*) is a point that minimizes $f(u, v) = \|B(u, v)\|_1 - u^\top r - v^\top c$.

Proof By the definition, we have

$$f(u, v) = \sum_{1 \leq i,j \leq n} e^{u_i + v_j - \frac{C_{ij}}{\eta}} - \sum_{i=1}^n u_i r_i - \sum_{j=1}^n v_j c_j.$$

By definition, we have $\nabla_u f(u^t, v^t) = B(u^t, v^t)\mathbf{1}_n - r$ and $\nabla_v f(u^t, v^t) = B(u^t, v^t)^\top \mathbf{1}_n - c$. Thus, we have $E_t = \|\nabla_u f(u^t, v^t)\|_1 + \|\nabla_v f(u^t, v^t)\|_1$. Since f is convex and minimized at $(u^\star, v^\star)$, we have

$$f(u^t, v^t) - f(u^\star, v^\star) \leq (u^t - u^\star)^\top \nabla_u f(u^t, v^t) + (v^t - v^\star)^\top \nabla_v f(u^t, v^t).$$

Combining Hölder's inequality and the definition of E_t yields

$$f(u^t, v^t) - f(u^\star, v^\star) \leq E_t(\|u^t - u^\star\|_\infty + \|v^t - v^\star\|_\infty). \tag{12}$$

Thus, it suffices to show that

$$\|u^t - u^\star\|_\infty + \|v^t - v^\star\|_\infty \leq 2\|u^\star\|_\infty + 2\|v^\star\|_\infty.$$

The next result is the key observation that makes our analysis work for Greenkhorn. We use an induction argument to establish the following bound:

$$\max\{\|u^t - u^\star\|_\infty, \|v^t - v^\star\|_\infty\} \leq \max\{\|u^0 - u^\star\|_\infty, \|v^0 - v^\star\|_\infty\}. \tag{13}$$

It is clear that Eq. (13) holds true when $t = 0$. Suppose that the inequality holds true for $t \leq k_0$, we show that it also holds true for $t = k_0 + 1$. Without loss of generality, let I be the index chosen at the $(k_0 + 1)$-th iteration. Then

$$\|u^{k_0+1} - u^\star\|_\infty \leq \max\{\|u^{k_0} - u^\star\|_\infty, |u_I^{k_0+1} - u_I^\star|\}, \tag{14}$$

$$\|v^{k_0+1} - v^\star\|_\infty = \|v^{k_0} - v^\star\|_\infty. \tag{15}$$

By the updating formula for $u_I^{k_0+1}$ and the optimality condition for $u_I^\star$, we have

$$e^{u_I^{k_0+1}} = \frac{r_I}{\sum_{j=1}^n e^{-\frac{C_{ij}}{\eta}+v_j^{k_0}}}, \qquad e^{u_I^\star} = \frac{r_I}{\sum_{j=1}^n e^{-\frac{C_{ij}}{\eta}+v_j^\star}}.$$

Putting these pieces together with the inequality that $\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \leq \max_{1\leq j\leq n} \frac{a_i}{b_i}$ for all $a_i, b_i > 0$ yields

$$|u_I^{k_0+1} - u_I^\star| = \left|\log\left(\frac{\sum_{j=1}^n e^{-\eta^{-1}C_{Ij}+v_j^{k_0}}}{\sum_{j=1}^n e^{-\eta^{-1}C_{Ij}+v_j^\star}}\right)\right| \leq \|v^{k_0} - v^\star\|_\infty. \tag{16}$$

Combining Eq. (14) and Eq. (16) yields

$$\|u^{k_0+1} - u^\star\|_\infty \leq \max\{\|u^{k_0} - u^\star\|_\infty, \|v^{k_0} - v^\star\|_\infty\}. \tag{17}$$

Combining Eq. (15) and Eq. (17) further implies Eq. (13). This together with $u^0 = v^0 = \mathbf{0}_n$ implies

$$\|u^t - u^\star\|_\infty + \|v^t - v^\star\|_\infty \leq 2(\|u^0 - u^\star\|_\infty + \|v^0 - v^\star\|_\infty) = 2\|u^\star\|_\infty + 2\|v^\star\|_\infty. \tag{18}$$

Putting Eq. (12) and Eq. (18) together yields the desired result. ■

Our second lemma shows that at least one optimal solution $(u^\star, v^\star)$ of f has an upper bound of $\eta^{-1}\|C\|_\infty + \log(n) - 2\log(\min_{1\leq i,j\leq n}\{r_i, c_j\})$ in ℓ_∞-norm. This result is stronger than Dvurechensky et al. (2018, Lemma 1) and generalizes Blanchet et al. (2018, Lemma 10).

Lemma 6. *There exists an optimal solution $(u^\star, v^\star)$ of the function f defined in Eq. (11) such that the following inequality holds true,*

$$\|u^\star\|_\infty \leq R, \qquad \|v^\star\|_\infty \leq R,$$

where $R := \eta^{-1}\|C\|_\infty + \log(n) - 2\log(\min_{1\leq i,j\leq n}\{r_i, c_j\})$ depends on C, r and c.

Proof By using the similar argument as in Lemma 3, we can first show that there exists an optimal solution pair $(u^\star, v^\star)$ such that (but not for $v^\star$ simultaneously)

$$\max_{1\leq i\leq n} u_i^\star \geq 0 \geq \min_{1\leq i\leq n} u_i^\star. \tag{19}$$

Then, we proceed to establish the bounds that are analogous to Eq. (7) and (8):

$$\max_{1\leq i\leq n} u_i^\star - \min_{1\leq i\leq n} u_i^\star \leq \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right), \tag{20}$$

$$\max_{1\leq i\leq n} v_i^\star - \min_{1\leq i\leq n} v_i^\star \leq \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right). \tag{21}$$

Indeed, for each $1 \leq i \leq n$, we have

$$e^{-\eta^{-1}\|C\|_\infty + u_i^\star}\left(\sum_{j=1}^n e^{v_j^\star}\right) \leq \sum_{j=1}^n e^{-\eta^{-1}C_{ij}+u_i^\star+v_j^\star} = [B(u^\star, v^\star)\mathbf{1}_n]_i = r_i \leq 1,$$

which implies $u_i^\star \leq \eta^{-1}\|C\|_\infty - \log(\sum_{j=1}^n e^{v_j^\star})$. Furthermore, we have

$$e^{u_i^\star}\left(\sum_{j=1}^n e^{v_j^\star}\right) \geq \sum_{j=1}^n e^{-\eta^{-1}C_{ij}+u_i^\star+v_j^\star} = [B(u^\star, v^\star)\mathbf{1}_n]_i = r_i \geq \min_{1\leq i,j\leq n}\{r_i, c_j\},$$

which implies $u_i^\star \geq \log(\min_{1\leq i,j\leq n}\{r_i, c_j\}) - \log(\sum_{j=1}^n e^{v_j^\star})$. Putting these pieces together yields Eq. (20). Using the similar argument, we can prove Eq. (21) holds true.

Finally, we prove our main results. Since $\max_{1\leq i\leq n} u_i^\star \geq 0 \geq \min_{1\leq i\leq n} u_i^\star$, we derive from Eq. (20) that

$$-\frac{\|C\|_\infty}{\eta} + \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right) \leq \min_{1\leq i\leq n} u_i^\star \leq \max_{1\leq i\leq n} u_i^\star \leq \frac{\|C\|_\infty}{\eta} - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right).$$

This implies that $\|u^\star\|_\infty \leq R$. Then, we bound $\|v^\star\|_\infty$ by considering two different cases.

For the former case, we assume that $\max_{1\leq i\leq n} v_i^\star \geq 0$. Note that the optimality condition is $\sum_{i,j=1}^n e^{-\eta^{-1}C_{ij}+u_i^\star+v_j^\star} = 1$ and further implies that

$$\max_{1\leq i\leq n} u_i^\star + \max_{1\leq i\leq n} v_i^\star \leq \log\left(\max_{1\leq i,j\leq n} e^{\eta^{-1}C_{ij}}\right) = \frac{\|C\|_\infty}{\eta}.$$

Since $\max_{1\leq i\leq n} u_i^* \geq 0$ and $\max_{1\leq i\leq n} v_i^* \geq 0$, we have $0 \leq \max_{1\leq i\leq n} v_i^\star \leq \frac{\|C\|_\infty}{\eta}$. Combining $\max_{1\leq i\leq n} v_i^* \geq 0$ with Eq. (21) yields that

$$\min_{1\leq i\leq n} v_i^\star \geq -\frac{\|C\|_\infty}{\eta} + \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right).$$

which implies that $\|v^\star\|_\infty \leq R$.

For the latter case, we assume that $\max_{1\leq i\leq n} v_i^\star \leq 0$. Then, we have

$$\min_{1\leq i\leq n} v_i^\star \geq \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right) - \log\left(\sum_{i=1}^n e^{u_i^\star}\right).$$

This together with $\|u^\star\|_\infty \leq \frac{\|C\|_\infty}{\eta} - \log(\min_{1\leq i,j\leq n}\{r_i, c_j\})$ yields that $\|v^\star\|_\infty \leq R$. ■

Putting Lemma 5 and 6 together, we have the following straightforward consequence:

Corollary 7. *Letting $\{(u^t, v^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 1, we have*

$$f(u^t, v^t) - f(u^\star, v^\star) \leq 4RE_t.$$

Remark 8. *The notation R is also used in Dvurechensky et al. (2018) and has the same order as ours since R in our paper only involves an term* $\log(n) - \log(\min_{1\leq i,j\leq n}\{r_i, c_j\})$.

Remark 9. *We further comment on the proof techniques in this paper and Dvurechensky et al. (2018). Indeed, the proof for Dvurechensky et al. (2018, Lemma 2) depends on taking full advantage of the shift property of Sinkhorn; namely, either $B(\overline{u}^t, \overline{v}^t)\mathbf{1}_n = r$ or $B(\overline{u}^t, \overline{v}^t)^\top \mathbf{1}_n = c$ where $(\overline{u}^t, \overline{v}^t)$ stands for the iterate generated by Sinkhorn. Unfortunately, Greenkhorn does not enjoy such a shift property. We have thus proposed a different approach for bounding $f(u^t, v^t) - f(u^\star, v^\star)$ via appeal to the ℓ_∞-norm of the solution $(u^\star, v^\star)$.*

3.2 Complexity analysis—bounding the number of iterations

We proceed to provide an upper bound for the iteration number to achieve a desired tolerance ε' in Algorithm 1. First, we start with a lower bound for the difference of function values between two consecutive iterates of Algorithm 1:

Lemma 10. *Letting $\{(u^t, v^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 1, we have*

$$f(u^t, v^t) - f(u^{t+1}, v^{t+1}) \geq \frac{(E_t)^2}{28n}.$$

Proof Combining Altschuler et al. (2017, Lemma 5) and the fact that the row or column update is chosen in a greedy manner, we have

$$f(u^t, v^t) - f(u^{t+1}, v^{t+1}) \geq \frac{1}{2n}\left(\rho(r, r(B(u^t, v^t)) + \rho(c, c(B(u^t, v^t)))\right).$$

Furthermore, Altschuler et al. (2017, Lemma 6) implies that

$$\rho(r, r(B(u^t, v^t)) + \rho(c, c(B(u^t, v^t)) \geq \frac{1}{7}\left(\|r - r(B(u^t, v^t))\|_1^2 + \|c - c(B(u^t, v^t))\|_1^2\right).$$

Putting these pieces together yields that

$$f(u^t, v^t) - f(u^{t+1}, v^{t+1}) \geq \frac{1}{14n}\left(\|r - r(B(u^t, v^t))\|_1^2 + \|c - c(B(u^t, v^t))\|_1^2\right).$$

Combining the above inequality with the definition of E_t implies the desired result. ∎

We are now able to derive the iteration complexity of Algorithm 1.

Theorem 11. *Letting $\{(u^t, v^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 1, the number of iterations required to satisfy $E_t \leq \varepsilon'$ is upper bounded by $t \leq 2 + \frac{112nR}{\varepsilon'}$ where $R > 0$ is defined in Lemma 6.*

Proof Letting $\delta_t = f(u^t, v^t) - f(u^\star, v^\star)$, we derive from Corollary 7 and Lemma 10 that

$$\delta_t - \delta_{t+1} \geq \max\left\{\frac{\delta_t^2}{448nR^2}, \frac{(\varepsilon')^2}{28n}\right\},$$

where $E_t \geq \varepsilon'$ as soon as the stopping criterion is not fulfilled. In the following step we apply a switching strategy introduced by Dvurechensky et al. (2018). Given any $t \geq 1$, we have two estimates:

(i) Considering the process from the first iteration and the t-th iteration, we have

$$\frac{\delta_{t+1}}{448nR^2} \leq \frac{1}{t + 448nR^2\delta_1^{-2}} \implies t \leq 1 + \frac{448nR^2}{\delta_t} - \frac{448nR^2}{\delta_1}.$$

(ii) Considering the process from the $(t+1)$-th iteration to the $(t+m)$-th iteration for any $m \geq 1$, we have

$$\delta_{t+m} \leq \delta_t - \frac{(\varepsilon')^2 m}{28n} \implies m \leq \frac{28n(\delta_t - \delta_{t+m})}{(\varepsilon')^2}.$$

We then minimize the sum of two estimates by an optimal choice of $s \in (0, \delta_1]$:

$$t \leq \min_{0<s\leq\delta_1} \left(2 + \frac{448nR^2}{s} - \frac{448nR^2}{\delta_1} + \frac{28ns}{(\varepsilon')^2}\right) = \begin{cases} 2 + \frac{224nR}{\varepsilon'} - \frac{448nR^2}{\delta_1}, & \delta_1 \geq 4R\varepsilon', \\ 2 + \frac{28n\delta_1}{(\varepsilon')^2}, & \delta_1 \leq 4R\varepsilon'. \end{cases}$$

This implies that $t \leq 2 + \frac{112nR}{\varepsilon'}$ in both cases and completes the proof. ■

Equipped with the result of Theorem 11 and the scheme of Algorithm 2, we are able to establish the following result for the complexity of Algorithm 2:

Theorem 12. *The Greenkhorn scheme for approximating optimal transport (Algorithm 2) returns an ε-approximate transportation plan (cf. Definition 1) in*

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

arithmetic operations.

Proof We follow the proof steps in (Altschuler et al., 2017, Theorem 1) and obtain that the transportation plan $\hat{X}$ returned by Algorithm 2 satisfies that

$$\begin{aligned} \langle C, \hat{X}\rangle - \langle C, X^\star\rangle &\leq 2\eta\log(n) + 4(\|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1)\|C\|_\infty \\ &\leq \frac{\varepsilon}{2} + 4(\|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1)\|C\|_\infty, \end{aligned}$$

where $X^\star$ is an optimal solution to the OT problem and $\tilde{X} = \textsc{Greenkhorn}(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$. The last inequality in the above display holds true since $\eta = \frac{\varepsilon}{4\log(n)}$. Furthermore,

$$\begin{aligned} \|\tilde{X}\mathbf{1}_n - r\|_1 + \|\tilde{X}^\top\mathbf{1}_n - c\|_1 &\leq \|\tilde{X}\mathbf{1}_n - \tilde{r}\|_1 + \|\tilde{X}^\top\mathbf{1}_n - \tilde{c}\|_1 + \|r - \tilde{r}\|_1 + \|c - \tilde{c}\|_1 \\ &\leq \frac{\varepsilon'}{2} + \frac{\varepsilon'}{4} + \frac{\varepsilon'}{4} = \varepsilon'. \end{aligned}$$

Putting these pieces together with $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$ yields that $\langle C, \hat{X}\rangle - \langle C, X^\star\rangle \leq \varepsilon$.

The remaining step is to analyze the complexity bound. It follows from Theorem 11 and the definition of $\tilde{r}$ and $\tilde{c}$ in Algorithm 2 that

$$\begin{aligned} t \ \leq \ 2 + \frac{112nR}{\varepsilon'} &\leq 2 + \frac{96n\|C\|_\infty}{\varepsilon}\left(\frac{\|C\|_\infty}{\eta} + \log(n) - 2\log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right)\right) \\ &\leq 2 + \frac{96n\|C\|_\infty}{\varepsilon}\left(\frac{4\|C\|_\infty\log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\|C\|_\infty}\right)\right) \\ &= O\left(\frac{n\|C\|_\infty^2\log(n)}{\varepsilon^2}\right). \end{aligned}$$

The total iteration complexity in Step 2 of Algorithm 2 is bounded by $O(n\|C\|_\infty^2\log(n)\varepsilon^{-2})$. Each iteration of Algorithm 1 requires $O(n)$ arithmetic operations. Thus, the total number of arithmetic operations is $O(n^2\|C\|_\infty^2\log(n)\varepsilon^{-2})$. Moreover, $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 2 can be found in $O(n)$ arithmetic operations and Altschuler et al. (2017, Algorithm 2) requires $O(n^2)$ arithmetic operations. Therefore, we conclude that the total number of arithmetic

Algorithm 3 APDAMD$(\varphi, A, b, \varepsilon')$

Input: $t = 0$.
Initialization: $\bar{\alpha}^0 = \alpha^0 = 0$, $z^0 = \mu^0 = \lambda^0 = \mathbf{0}_{2n}$ and $L^0 = 1$.
repeat
  Set $M^t = \frac{L^t}{2}$.
  repeat
    Set $M^t = 2M^t$.
    Compute the stepsize: $\alpha^{t+1} = \frac{1+\sqrt{1+4\delta M^t \bar{\alpha}^t}}{2\delta M^t}$.
    Compute the average coefficient: $\bar{\alpha}^{t+1} = \bar{\alpha}^t + \alpha^{t+1}$.
    Compute the first average step: $\mu^{t+1} = \frac{\alpha^{t+1} z^t + \bar{\alpha}^t \lambda^t}{\bar{\alpha}^{t+1}}$.
    Compute the mirror descent: $z^{t+1} = \operatorname{argmin}_{z \in \mathbb{R}^n} \{(z - \mu^{t+1})^\top \nabla\varphi(\mu^{t+1}) + \frac{B_\phi(z, z^t)}{\alpha^{t+1}}\}$.
    Compute the second average step: $\lambda^{t+1} = \frac{\alpha^{t+1} z^{t+1} + \bar{\alpha}^t \lambda^t}{\bar{\alpha}^{t+1}}$.
  until $\varphi(\lambda^{t+1}) - \varphi(\mu^{t+1}) - (\lambda^{t+1} - \mu^{t+1})^\top \nabla\varphi(\mu^{t+1}) \leq \frac{M^t}{2}\|\lambda^{t+1} - \mu^{t+1}\|_\infty^2$.
  Compute the main average step: $x^{t+1} = \frac{\alpha^{t+1} x(\mu^{t+1}) + \bar{\alpha}^t x^t}{\bar{\alpha}^{t+1}}$.
  Set $L^{t+1} = \frac{M^t}{2}$.
  Set $t = t + 1$.
until $\|Ax^t - b\|_1 \leq \varepsilon'$.
Output: X^t where $x^t = \text{vec}(X^t)$.

operations is $O(n^2\|C\|_\infty^2 \log(n)\varepsilon^{-2})$. ■

The complexity results presented in Theorem 12 improve the best known complexity bound $\widetilde{O}(n^2\epsilon^{-3})$ of Greenkhorn (Altschuler et al., 2017; Abid and Gower, 2018), Notably, it matches the best known complexity bound of Sinkhorn (Dvurechensky et al., 2018). The key feature of our analysis is that the per-iteration progress of Greenkhorn can be lower bounded by a new quantity (cf. Lemmas 5 and 6). It allows us to apply the switching strategy in Theorem 11 to improve the complexity upper bound of Greenkhorn.

In practice, Greenkhorn has been reported to outperform Sinkhorn (Altschuler et al., 2017) in terms of row/column updates and our improved complexity bound can provide the theoretical justification for this phenomenon.

4. Adaptive Primal-Dual Accelerated Mirror Descent

In this section, we propose an adaptive primal-dual accelerated mirror descent (APDAMD) for solving the entropic regularized OT problem in Eq. (2). APDAMD and its application to the OT problem are presented in Algorithm 3 and 4. We prove the complexity bound of $O(n^2\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$ where $\delta > 0$ stands for the regularity of the mirror mapping ϕ.

4.1 General setup

We follow the setup in Section 2 and consider the following generalization of the entropic regularized OT problem in Eq. (2):

$$\min_{x \in Q} \; f(x), \quad \text{s.t. } Ax = b, \tag{22}$$

where f is strongly convex with respect to the ℓ_1-norm on the set Q:

$$f(x') - f(x) - (x' - x)^\top \nabla f(x) \geq \frac{\eta}{2}\|x' - x\|_1^2 \text{ for any } x', x \in Q.$$

Note that, in the specific setting of the entropic regularized OT problem, the function $f(x) = \sum_{i,j} C_{ij} x_{j+n(i-1)} + \eta \cdot x_{j+n(i-1)} \log(x_{j+n(i-1)})$ where $x_{j+n(i-1)} = X_{ij}$ for any i, j where X is the transportation plan in equation (2), and the vector $b \in \mathbb{R}^{2n \times 1}$ is defined as: $b_i = r_i$ as $1 \leq i \leq n$ and $b_i = c_{i-n}$ when $n+1 \leq i \leq 2n$. Furthermore, the matrix $A = (A_{ij}) \in \mathbb{R}^{2n \times n^2}$ is defined as: When $1 \leq i \leq n$, we denote $A_{ij} = 1$ if $1+n(i-1) \leq j \leq n \cdot i$ and 0 otherwise; When $n+1 \leq i \leq 2n$, we define $A_{ij} = 1$ if $j \in \{i-n+n(l-1) : 1 \leq l \leq n\}$ and 0 otherwise. To be consistent with the notations in Algorithms 4 and 5, we specifically denote A_{ot} as the matrix A corresponding to the entropic regularized OT problem.

After some calculations with the general problem (22), we obtain that the dual problem is as follows:

$$\min_{\lambda \in \mathbb{R}^{2n}} \widetilde{\varphi}(\lambda) := \{\langle \lambda, b \rangle + \max_{x \in \mathbb{R}^{n^2}} \{-f(x) - \langle A^\top \lambda, x \rangle\}\}, \tag{23}$$

and $\nabla \widetilde{\varphi}(\lambda) = b - Ax(\lambda)$ where $x(\lambda) = \text{argmax}_{x \in \mathbb{R}^n} \{-f(x) - \langle A^\top \lambda, x \rangle\}$; see the explicit form in Eq. (9) with $\lambda = (\alpha, \beta)$. By Nesterov (2005, Theorem 1) with ℓ_1-norm for the dual space of the Lagrange multipliers, the dual objective function $\tilde{\varphi}$ satisfies the following inequality:

$$\widetilde{\varphi}(\lambda') - \widetilde{\varphi}(\lambda) - (\lambda' - \lambda)^\top \nabla \widetilde{\varphi}(\lambda) \leq \frac{\|A\|_{1 \to 1}^2}{2\eta} \|\lambda' - \lambda\|_\infty^2. \tag{24}$$

In the entropic regularized OT problem, each column of the matrix A_{ot} contains no more than two nonzero elements which are equal to one. Since $\|A_{\text{ot}}\|_{1 \to 1}$ is equal to maximum ℓ_1-norm of the column of this matrix, we have $\|A_{\text{ot}}\|_{1 \to 1} = 2$. Thus, the dual objective function $\widetilde{\varphi}$ is $\frac{4}{\eta}$-gradient Lipschitz with respect to the ℓ_∞-norm.

In addition, we define the Bregman divergence $B_\phi : \mathbb{R}^{2n} \times \mathbb{R}^{2n} \mapsto [0, +\infty)$ by

$$B_\phi(\lambda', \lambda) := \phi(\lambda') - \phi(\lambda) - (\lambda' - \lambda)^\top \nabla \phi(\lambda),$$

where the mirror mapping ϕ is $\frac{1}{\delta}$-strongly convex and 1-smooth on $\mathbb{R}^{2n}$ with respect to ℓ_∞-norm; that is,

$$\frac{1}{2\delta}\|\lambda' - \lambda\|_\infty^2 \leq \phi(\lambda') - \phi(\lambda) - (\lambda' - \lambda)^\top \nabla \phi(\lambda) \leq \frac{1}{2}\|\lambda' - \lambda\|_\infty^2.$$

For example, we can choose $\phi(\lambda) = \frac{1}{2n}\|\lambda\|^2$ and $B_\phi(\lambda', \lambda) = \frac{1}{2n}\|\lambda' - \lambda\|^2$ in APDAMD where $\delta = n$. As such, $\delta > 0$ is a function of n in general and it will appear in the complexity bound of APDAMD for approximating the OT problem (cf. Theorem 17). It is worth noting that our algorithm uses a regularizer that acts only in the dual and our complexity bound is the best existing one among this group of algorithms (Dvurechensky et al., 2018; Guo et al., 2020; Guminov et al., 2021). A very recent work of Jambulapati et al. (2019) showed that the complexity bound can be improved to $\widetilde{O}(n^2 \varepsilon^{-1})$ using a more advanced area-convex mirror mapping (Sherman, 2017).

Algorithm 4 Approximating OT by Algorithm 3

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.
Step 2: Let $A_{\text{ot}} \in \mathbb{R}^{2n\times n^2}$ and $b \in \mathbb{R}^{2n}$ be defined by $A_{\text{ot}}\text{vec}(X) = \begin{pmatrix} X\mathbf{1}_n \\ X^\top\mathbf{1}_n \end{pmatrix}$ and $b = \begin{pmatrix} \tilde{r} \\ \tilde{c} \end{pmatrix}$.
Step 3: Compute $\tilde{X} = \textsc{Apdamd}(\widetilde{\varphi}, A_{\text{ot}}, b, \frac{\varepsilon'}{2})$ where $\widetilde{\varphi}$ is defined by Eq. (23).
Step 4: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. (2017, Algorithm 2) such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top\mathbf{1}_n = c$.
Output: $\hat{X}$.

4.2 Properties of APDAMD

We present several important properties of Algorithm 3 that can be used later for entropic regularized OT problems. First, we prove the following result regarding the number of line search iterations in Algorithm 3 for solving the entropic regularized OT problem:

Lemma 13. *The number of line search iterations in Algorithm 3 for solving the entropic OT problem is finite. Furthermore, the total number of gradient oracle calls after the t-th iteration is bounded as*

$$N_t \leq 4t + 4 + \frac{2\log(\frac{8}{\eta}) - 2\log(L^0)}{\log 2}.$$

Proof First, we observe that multiplying M^t by two will not stop until the line search stopping criterion is satisfied. Then, Eq. (24) implies that the number of line search iterations in the line search strategy is finite and $M^t \leq \frac{2\|A_{\text{ot}}\|_{1\to 1}^2}{\eta}$ holds true for all $t \geq 0$. Otherwise, the line search stopping criterion is satisfied with $\frac{M^t}{2}$ since $\frac{M^t}{2} \geq \frac{\|A_{\text{ot}}\|_{1\to 1}^2}{\eta}$.

Letting i_j denote the total number of multiplication at the j-th iteration, we have

$$i_0 \leq 1 + \frac{\log(\frac{M^0}{L^0})}{\log 2}, \qquad i_j \leq 2 + \frac{\log(\frac{M^j}{M^{j-1}})}{\log 2}.$$

Then, the total number of line search iterations is bounded by

$$\sum_{j=0}^t i_j \leq 1 + \frac{\log(\frac{M^0}{L^0})}{\log 2} + \sum_{j=1}^t \left(2 + \frac{\log(\frac{M^j}{M^{j-1}})}{\log 2}\right) \leq 2t + 1 + \frac{\log(\frac{2\|A_{\text{ot}}\|_{1\to 1}^2}{\eta}) - \log(L^0)}{\log 2}.$$

Since each line search contains two gradient oracle calls and $\|A_{\text{ot}}\|_{1\to 1} = 2$, we conclude the desired upper bound for the total number of gradient oracle calls after the t-th iteration. ■

The next lemma presents a property of the function $\widetilde{\varphi}$ in Algorithm 3.

Lemma 14. *For each iteration t of Algorithm 3 and any $z \in \mathbb{R}^{2n}$, we have*

$$\bar{\alpha}^t\widetilde{\varphi}(\lambda^t) \leq \sum_{j=0}^t (\alpha^j(\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla\widetilde{\varphi}(\mu^j))) + \|z\|_\infty^2.$$

Proof First, we claim that it holds for any $z \in \mathbb{R}^n$:

$$\alpha^{t+1}(z^t - z)^\top \nabla \widetilde{\varphi}(\mu^{t+1}) \leq \bar{\alpha}^{t+1}(\widetilde{\varphi}(\mu^{t+1}) - \widetilde{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}). \tag{25}$$

Indeed, the optimality condition in mirror descent implies that, for any $z \in \mathbb{R}^{2n}$, we have

$$(z - z^{t+1})^\top \left(\nabla \widetilde{\varphi}(\mu^{t+1}) + \tfrac{\nabla \phi(z^{t+1}) - \nabla \phi(z^t)}{\alpha^{t+1}}\right) \geq 0.$$

By definition, we have $B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - B_\phi(z^{t+1}, z^t) = (z - z^{t+1})^\top(\nabla\phi(z^{t+1}) - \nabla\phi(z^t))$ and $B_\phi(z^{t+1}, z^t) \geq \frac{1}{2\delta}\|z^{t+1} - z^t\|_\infty^2$. Putting these pieces together yields that

$$\begin{aligned}
\alpha^{t+1}(z^t - z)^\top \nabla \widetilde{\varphi}(\mu^{t+1}) &= \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^{t+1} - z)^\top \nabla \widetilde{\varphi}(\mu^{t+1}) \\
&\leq \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}) + (z - z^{t+1})^\top(\nabla\phi(z^{t+1}) - \nabla\phi(z^t)) \\
&= \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - B_\phi(z^{t+1}, z^t) \\
&\leq \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - \tfrac{\|z^{t+1} - z^t\|_\infty^2}{2\delta}.
\end{aligned} \tag{26}$$

The update formulas of μ^{t+1}, λ^{t+1}, α^{t+1} and $\bar{\alpha}^{t+1}$ imply that

$$\lambda^{t+1} - \mu^{t+1} = \frac{\alpha^{t+1}}{\bar{\alpha}^{t+1}}(z^{t+1} - z^t), \qquad \delta M^t(\alpha^{t+1})^2 = \bar{\alpha}^{t+1}.$$

Therefore, we have

$$\alpha^{t+1}(z^t - z^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}) = \bar{\alpha}^{t+1}(\mu^{t+1} - \lambda^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}),$$

and

$$\|z^{t+1} - z^t\|_\infty^2 = (\tfrac{\bar{\alpha}^{t+1}}{\alpha^{t+1}})^2\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2 = \delta M^t \bar{\alpha}^{t+1}\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2.$$

Putting these pieces together with Eq. (26) yields that

$$\begin{aligned}
&\alpha^{t+1}(z^t - z)^\top \nabla \widetilde{\varphi}(\mu^{t+1}) \\
&\leq \bar{\alpha}^{t+1}(\mu^{t+1} - \lambda^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - \tfrac{\bar{\alpha}^{t+1} M^t}{2}\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2 \\
&= \bar{\alpha}^{t+1}\left((\mu^{t+1} - \lambda^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1}) - \tfrac{M^t}{2}\|\mu^{t+1} - \lambda^{t+1}\|_\infty^2\right) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) \\
&\leq \bar{\alpha}^{t+1}(\widetilde{\varphi}(\mu^{t+1}) - \widetilde{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}),
\end{aligned}$$

where the last inequality comes from the stopping criterion in the line search. This implies that Eq. (25) holds true.

The next step is to bound the iterative objective gap given by

$$\begin{aligned}
&\bar{\alpha}^{t+1}\widetilde{\varphi}(\lambda^{t+1}) - \bar{\alpha}^t \widetilde{\varphi}(\lambda^t) \\
&\leq \alpha^{t+1}(\widetilde{\varphi}(\mu^{t+1}) + (z - \mu^{t+1})^\top \nabla \widetilde{\varphi}(\mu^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}).
\end{aligned} \tag{27}$$

Indeed, by combining $\bar{\alpha}^{t+1} = \bar{\alpha}^t + \alpha^{t+1}$ and the update formula of μ^{t+1}, we have

$$\alpha^{t+1}(\mu^{t+1} - z^t) = (\bar{\alpha}^{t+1} - \bar{\alpha}^t)\mu^{t+1} - \alpha^{t+1}z^t = \alpha^{t+1}z^t + \bar{\alpha}^t\lambda^t - \bar{\alpha}^t\mu^{t+1} - \alpha^{t+1}z^t = \bar{\alpha}^t(\lambda^t - \mu^{t+1}).$$

This together with the convexity of $\widetilde{\varphi}$ implies that

$$\begin{aligned}
& \alpha^{t+1}(\mu^{t+1}-z)^\top \nabla\widetilde{\varphi}(\mu^{t+1}) \\
= \ & \alpha^{t+1}(\mu^{t+1}-z^t)^\top \nabla\widetilde{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^t-z)^\top \nabla\widetilde{\varphi}(\mu^{t+1}) \\
= \ & \bar{\alpha}^t(\lambda^t-\mu^{t+1})^\top \nabla\widetilde{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^t-z)^\top \nabla\widetilde{\varphi}(\mu^{t+1}) \\
\leq \ & \bar{\alpha}^t(\widetilde{\varphi}(\lambda^t)-\widetilde{\varphi}(\mu^{t+1})) + \alpha^{t+1}(z^t-z)^\top \nabla\widetilde{\varphi}(\mu^{t+1}).
\end{aligned}$$

Furthermore, we derive from Eq. (25) and $\bar{\alpha}^{t+1}=\bar{\alpha}^t+\alpha^{t+1}$ that

$$\begin{aligned}
& \bar{\alpha}^t(\widetilde{\varphi}(\lambda^t)-\widetilde{\varphi}(\mu^{t+1})) + \alpha^{t+1}(z^t-z)^\top \nabla\widetilde{\varphi}(\mu^{t+1}) \\
\leq \ & \bar{\alpha}^t(\widetilde{\varphi}(\lambda^t)-\widetilde{\varphi}(\mu^{t+1})) + \bar{\alpha}^{t+1}(\widetilde{\varphi}(\mu^{t+1})-\widetilde{\varphi}(\lambda^{t+1})) + B_\phi(z,z^t) - B_\phi(z,z^{t+1}) \\
= \ & \bar{\alpha}^t\widetilde{\varphi}(\lambda^t) - \bar{\alpha}^{t+1}\widetilde{\varphi}(\lambda^{t+1}) + \alpha^{t+1}\widetilde{\varphi}(\mu^{t+1}) + B_\phi(z,z^t) - B_\phi(z,z^{t+1}).
\end{aligned}$$

Putting these pieces together yields that Eq. (27) holds true.

Finally, we prove our main results. By changing the index t to j in Eq. (27) and summing up the resulting inequality over $j=0,1,\ldots,t-1$, we have

$$\bar{\alpha}^t\widetilde{\varphi}(\lambda^t) - \bar{\alpha}^0\widetilde{\varphi}(\lambda^0) \leq \sum_{j=0}^{t-1}(\alpha^{j+1}(\widetilde{\varphi}(\mu^{j+1}) + (z-\mu^{j+1})^\top \nabla\widetilde{\varphi}(\mu^{j+1}))) + B_\phi(z,z^0) - B_\phi(z,z^t).$$

Since $\alpha^0=\bar{\alpha}^0=0$, $B_\phi(z,z^t)\geq 0$ and ϕ is 1-smooth with respect to ℓ_∞-norm, we have

$$\begin{aligned}
\bar{\alpha}^t\widetilde{\varphi}(\lambda^t) \ &\leq \ \sum_{j=0}^{t}(\alpha^j(\widetilde{\varphi}(\mu^j) + (z-\mu^j)^\top \nabla\widetilde{\varphi}(\mu^j))) + B_\phi(z,z^0) \\
&\leq \ \sum_{j=0}^{t}(\alpha^j(\widetilde{\varphi}(\mu^j) + (z-\mu^j)^\top \nabla\widetilde{\varphi}(\mu^j))) + \|z-z^0\|_\infty^2 \\
&\overset{z^0=0}{=} \ \sum_{j=0}^{t}(\alpha^j(\widetilde{\varphi}(\mu^j) + (z-\mu^j)^\top \nabla\widetilde{\varphi}(\mu^j))) + \|z\|_\infty^2.
\end{aligned}$$

This completes the proof. ■

The final lemma provides us with a key lower bound for the accumulating parameter.

Lemma 15. *For each iteration t of Algorithm 3, we have $\bar{\alpha}^t \geq \frac{\eta(t+1)^2}{32\delta}$.*

Proof For $t=1$, we have $\bar{\alpha}^1=\alpha^1=\frac{1}{\delta M^1}\geq\frac{\eta}{8\delta}$ since $M^1\leq\frac{8}{\eta}$ was proven in Lemma 13. Thus, the desired result holds true when $t=1$. Then we proceed to prove that it holds true for $t\geq 1$ using the induction. Indeed, we have

$$\begin{aligned}
\bar{\alpha}^{t+1} \ &= \ \bar{\alpha}^t + \alpha^{t+1} \ = \ \bar{\alpha}^t + \frac{1+\sqrt{1+4\delta M^t\bar{\alpha}^t}}{2\delta M^t} \\
&= \ \bar{\alpha}^t + \frac{1}{2\delta M^t} + \sqrt{\frac{1}{4(\delta M^t)^2} + \frac{\bar{\alpha}^t}{\delta M^t}} \\
&\geq \ \bar{\alpha}^t + \frac{1}{2\delta M^t} + \sqrt{\frac{\bar{\alpha}^t}{\delta M^t}} \\
&\geq \ \bar{\alpha}^t + \frac{\eta}{16\delta} + \sqrt{\frac{\eta\bar{\alpha}^t}{8\delta}},
\end{aligned}$$

where the last inequality comes from $M^t \leq \frac{8}{\eta}$ as shown in Lemma 13. Suppose that the desired result holds true for $t = k_0$, we have

$$\bar{\alpha}^{k_0+1} \geq \frac{\eta(k_0+1)^2}{32\delta} + \frac{\eta}{16\delta} + \sqrt{\frac{\eta^2(k_0+1)^2}{256\delta^2}} = \frac{\eta((k_0+1)^2+2+2(k_0+1))}{32\delta} \geq \frac{\eta(k_0+2)^2}{32\delta}.$$

This completes the proof. ■

4.3 Complexity analysis for APDAMD

We are now ready to establish the complexity bound of APDAMD for solving the entropic regularized OT problem. Indeed, we recall that $\widetilde{\varphi}(\lambda)$ is defined with $\lambda = (\alpha, \beta)$ by

$$\widetilde{\varphi}(\alpha, \beta) = -\eta \log\left(\sum_{1 \leq i,j \leq n} e^{\eta^{-1}(\alpha_i + \beta_j - C_{ij})}\right) + \alpha^\top r + \beta^\top c.$$

Since (α, β) can be obtain by $\alpha_i = \eta u_i$ and $\beta_j = \eta v_j$, we derive from Lemma 3 that

$$\|\alpha^\star\|_\infty \leq \eta R, \quad \|\beta^\star\|_\infty \leq \eta R.$$

where R is defined accordingly. Then, we proceed to the following key result determining an upper bound for the number of iterations for Algorithm 3 to reach a desired accuracy ε':

Theorem 16. *Letting $\{X^t\}_{t\geq 0}$ be the iterates generated by Algorithm 3, the number of iterations required to satisfy $\|A_{ot}\text{vec}(X^t) - b\|_1 \leq \varepsilon'$ is upper bounded by*

$$t \leq 1 + \sqrt{\frac{128\delta R}{\varepsilon'}},$$

where $R > 0$ is defined in Lemma 3.

Proof From Lemma 14, we have

$$\bar{\alpha}^t \widetilde{\varphi}(\lambda^t) \leq \min_{z \in B_\infty(2\eta R)} \left\{\sum_{j=0}^t (\alpha^j(\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla\widetilde{\varphi}(\mu^j))) + \|z\|_\infty^2\right\},$$

where $B_\infty(r) := \{\lambda \in \mathbb{R}^n \mid \|\lambda\|_\infty \leq r\}$. This implies that

$$\bar{\alpha}^t \widetilde{\varphi}(\lambda^t) \leq \min_{z \in B_\infty(2\eta R)} \left\{\sum_{j=0}^t (\alpha^j(\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla\widetilde{\varphi}(\mu^j)))\right\} + 4\eta^2 R^2.$$

Since $\widetilde{\varphi}$ is the objective function of dual entropic regularized OT problem, we have

$$\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla\widetilde{\varphi}(\mu^j) = -f(x(\mu^j)) + z^\top(b - A_{\text{ot}} x(\mu^j)).$$

Therefore, we conclude that

$$\begin{aligned}
\bar{\alpha}^t \widetilde{\varphi}(\lambda^t) &\leq \min_{z \in B_\infty(2\eta R)} \left\{ \sum_{j=0}^{t} (\alpha^j (\widetilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla \widetilde{\varphi}(\mu^j))) \right\} + 4\eta^2 R^2 \\
&\leq 4\eta^2 R^2 - \bar{\alpha}^t f(x^t) + \min_{z \in B_\infty(2\eta R)} \{ \bar{\alpha}^t z^\top (b - A_{\text{ot}} x^t) \} \\
&= 4\eta^2 R^2 - \bar{\alpha}^t f(x^t) - 2\bar{\alpha}^t \eta R \| A_{\text{ot}} x^t - b \|_1,
\end{aligned}$$

where the second inequality comes from the convexity of f and the last equality comes from the fact that ℓ_1-norm is the dual norm of ℓ_∞-norm. That is to say,

$$f(x^t) + \widetilde{\varphi}(\lambda^t) + 2\eta R \| A_{\text{ot}} x^t - b \|_1 \leq \frac{4\eta^2 R^2}{\bar{\alpha}^t}.$$

Suppose that $\lambda^\star$ is an optimal solution to dual entropic regularized OT problem satisfying $\|\lambda^\star\|_\infty \leq \eta R$, we have

$$\begin{aligned}
f(x^t) + \widetilde{\varphi}(\lambda^t) &\geq f(x^t) + \widetilde{\varphi}(\lambda^\star) \;=\; f(x^t) + b^\top \lambda^\star + \max_{x \in \mathbb{R}^{n^2}} \left\{ -f(x) - (\lambda^\star)^\top A_{\text{ot}} x \right\} \\
&\geq f(x^t) + b^\top \lambda^\star - f(x^t) - (\lambda^\star)^\top A_{\text{ot}} x^t \;=\; (b - A_{\text{ot}} x^t) \lambda^\star \\
&\geq -\eta R \| A_{\text{ot}} x^t - b \|_1,
\end{aligned}$$

Therefore, we conclude that

$$\| A_{\text{ot}} x^t - b \|_1 \leq \frac{4\eta R}{\bar{\alpha}^t} \leq \frac{128 \delta R}{(t+1)^2}.$$

This completes the proof. ∎

Now, we are ready to present the complexity bound of Algorithm 4 for approximating the OT problem.

Theorem 17. *The APDAMD scheme for approximating optimal transport (Algorithm 4) returns an ε-approximate transportation plan (cf. Definition 1) in*

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

arithmetic operations.

Proof Using the same argument as in Theorem 12, we have

$$\langle C, \hat{X} \rangle - \langle C, X^\star \rangle \leq \frac{\varepsilon}{2} + 4(\|\tilde{X} \mathbf{1}_n - r\|_1 + \|\tilde{X}^\top \mathbf{1}_n - c\|_1) \|C\|_\infty,$$

where $\hat{X}$ is returned by Algorithm 4, X^* is a solution to the OT problem and $\tilde{X} = \textsc{Apdamd}(\widetilde{\varphi}, A_{\text{ot}}, b, \frac{\varepsilon'}{2})$. Since $\|\tilde{X} \mathbf{1}_n - r\|_1 + \|\tilde{X}^\top \mathbf{1}_n - c\|_1 \leq \varepsilon'$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$, we have $\langle C, \hat{X} \rangle - \langle C, X^\star \rangle \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$.

The remaining step is to analyze the complexity bound. If follows from Lemma 13 and Theorem 16 that

$$\begin{aligned} N_t &\leq 4t + 4 + \frac{2\log(\frac{8}{\eta}) - 2\log(L^0)}{\log 2} \\ &\leq 8 + \sqrt{\frac{2048\delta R}{\varepsilon'}} + \frac{2\log(\frac{8}{\eta}) - 2\log(L^0)}{\log 2} \\ &= 8 + 256\sqrt{\frac{\delta R\|C\|_\infty \log(n)}{\varepsilon}} + \frac{2\log(\frac{32\log(n)}{\varepsilon}) - 2\log(L^0)}{\log 2}. \end{aligned}$$

Combining the definition of R in Lemma 3 with the definition of η, $\tilde{r}$ and $\tilde{c}$ in Algorithm 4, we have

$$R \leq \frac{4\|C\|_\infty \log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\|C\|_\infty}\right).$$

Therefore, we conclude that

$$\begin{aligned} N_t \;\leq\; & 256\sqrt{\frac{\delta\|C\|_\infty \log(n)}{\varepsilon}}\sqrt{\frac{4\|C\|_\infty \log(n)}{\varepsilon} + \log(n) - 2\log\left(\frac{\varepsilon}{64n\|C\|_\infty}\right)} \\ & + \frac{2\log(\frac{32\log(n)}{\varepsilon}) - 2\log(L^0)}{\log 2} + 8 \;=\; O\left(\frac{\sqrt{\delta}\,\|C\|_\infty \log(n)}{\varepsilon}\right). \end{aligned}$$

The total iteration complexity in Step 3 of Algorithm 4 is bounded by $O(\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$. Each iteration of Algorithm 3 requires $O(n^2)$ arithmetic operations. Therefore, the total number of arithmetic operations is $O(n^2\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$. Moreover, $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 4 can be found in $O(n)$ arithmetic operations and Altschuler et al. (2017, Algorithm 2) requires $O(n^2)$ arithmetic operations. Therefore, we conclude that the total number of arithmetic operations is $O(n^2\sqrt{\delta}\|C\|_\infty \log(n)\varepsilon^{-1})$. ■

The complexity results in Theorem 17 suggests an interesting feature of the (regularized) OT problem. Indeed, the dependence of that bound on δ manifests the necessity of ℓ_∞-norm in the understanding of the complexity of the entropic regularized OT problem. This view is also in harmony with the proof technique of running time for Greenkhorn in Section 3, where we rely on ℓ_∞-norm of optimal solutions of the dual entropic regularized OT problem to measure the progress in the objective value among the successive iterates.

4.4 Revisiting APDAGD

We revisit APDAGD (Dvurechensky et al., 2018) for the entropic regularized OT problem. First, we point out that the current complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ is not valid by a simple counterexample. Then, we establish a new complexity bound of APDAGD using our techniques in Section 4.3. Despite the issue with entropic regularized OT, we wish to emphasize that APDAGD is still an interesting and efficient accelerated algorithm for general linearly constrained convex optimization problem with solid theoretical guarantee. More precisely, Dvurechensky et al. (2018, Theorem 3) is not applicable to entropic regularized OT since no dual solution exists with a constant bound in ℓ_2-norm. However, it can be used for analyzing other problems with bounded optimal dual solution.

Algorithm 5 Approximating OT by Dvurechensky et al. (2018, Algorithm 3)

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

Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.

Step 2: Let $A_{\text{ot}} \in \mathbb{R}^{2n \times n^2}$ and $b \in \mathbb{R}^{2n}$ be defined by $A_{\text{ot}}\text{vec}(X) = \begin{pmatrix} X\mathbf{1}_n \\ X^\top \mathbf{1}_n \end{pmatrix}$ and $b = \begin{pmatrix} \tilde{r} \\ \tilde{c} \end{pmatrix}$.

Step 3: Compute $\tilde{X} = \textsc{Apdagd}(\widetilde{\varphi}, A_{\text{ot}}, b, \frac{\varepsilon'}{2})$ where $\widetilde{\varphi}$ is defined by Eq. (23).

Step 4: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. (2017, Algorithm 2) such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$.

To facilitate the ensuing discussion, we first present the complexity bound for entropic regularized OT in Dvurechensky et al. (2018) using our notation. Indeed, we recall that APDAGD is developed for solving the optimization problem with the objective function $\widehat{\varphi}$ defined as follows,

$$\min_{\alpha, \beta \in \mathbb{R}^n} \ \widehat{\varphi}(\alpha, \beta) := \eta \left(\sum_{i,j=1}^{n} e^{-\frac{C_{ij} - \alpha_i - \beta_j}{\eta} - 1} \right) - \alpha^\top r - \beta^\top c. \tag{28}$$

Theorem 18 (Theorem 4 in Dvurechensky et al. (2018))**.** *Let $\overline{R} > 0$ be defined such that there exists an optimal solution to the dual entropic regularized OT problem in Eq.* (23)*, denoted by $(u^\star, v^\star)$, satisfying $\|(u^\star, v^\star)\| \leq \overline{R}$, the APDAGD scheme for approximating optimal transport (cf. Algorithm 5) returns an ε-approximate transportation plan (cf. Definition 1) in*

$$O\left(\min\left\{\frac{n^{9/4}\sqrt{\overline{R}\|C\|_\infty \log(n)}}{\varepsilon}, \frac{n^2\overline{R}\|C\|_\infty \log(n)}{\varepsilon^2}\right\}\right),$$

arithmetic operations.

From the above theorem, Dvurechensky et al. (2018) claims that the complexity bound for APDAGD is $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$. However, there are two issues:

1. The upper bound $\overline{R}$ is assumed to be independent of n, which is not correct; see our counterexample in Proposition 19.

2. The known upper bound $\overline{R}$ for the optimal solution depends on $\min_{1 \leq i,j \leq n}\{r_i, c_j\}$ (cf. Dvurechensky et al. (2018, Lemma 1) or Lemma 3 in our paper). This implies that the valid algorithm needs to take the rounding error with r and c into account.

Corrected upper bound $\overline{R}$. Corollary 4 and Lemma 6 imply that a straightforward upper bound for $\overline{R}$ is $\widetilde{O}(\sqrt{n})$. Given a tolerance $\varepsilon \in (0, 1)$, we further show that $\overline{R}$ is indeed $\Omega(\sqrt{n})$ by using a specific entropic regularized OT problem as follows.

Proposition 19. *Suppose that $C = \mathbf{1}_n\mathbf{1}_n^\top$ and $r = c = \frac{1}{n}\mathbf{1}_n$. Given a tolerance $\varepsilon \in (0, 1)$ and the regularization term $\eta = \frac{\varepsilon}{4\log(n)}$, all the optimal solutions of the dual entropic regularized OT problem in Eq.* (28) *satisfy that $\|(\alpha^*, \beta^*)\| \gtrsim \sqrt{n}$.*

Proof By the definition r, c and η, we rewrite the dual function $\widehat{\varphi}(\alpha, \beta)$ as follows:

$$\widehat{\varphi}(\alpha, \beta) = \frac{\varepsilon}{4e\log(n)} \sum_{1 \leq i,j \leq n} e^{-\frac{4\log(n)(1-\alpha_i-\beta_j)}{\varepsilon}} - \frac{\alpha^\top \mathbf{1}_n}{n} - \frac{\beta^\top \mathbf{1}_n}{n}.$$

Since $(\alpha^\star, \beta^\star)$ is an optimal solution of dual entropic regularized OT problem, we have

$$e^{\frac{4\log(n)\alpha_i^\star}{\varepsilon}} \sum_{j=1}^n e^{-\frac{4\log(n)(1-\beta_j^\star)}{\varepsilon}} = e^{\frac{4\log(n)\beta_i^\star}{\varepsilon}} \sum_{j=1}^n e^{-\frac{4\log(n)(1-\alpha_j^\star)}{\varepsilon}} = \frac{e}{n} \quad \text{for all } i \in [n]. \tag{29}$$

This implies $\alpha_i^\star = \alpha_j^\star$ and $\beta_i^\star = \beta_j^\star$ for all $i, j \in [n]$, and $\alpha_i^\star + \beta_i^\star$ are the same for all $i \in [n]$. Without loss of generality, we can let $\alpha_i^\star = 0$ in Eq. (29) and obtain that

$$\beta_i^\star = 1 + \frac{\varepsilon}{4\log(n)}\left(1 - 2\log(n)\right) = 1 + \frac{\varepsilon}{4\log(n)} - \frac{\varepsilon}{2}.$$

which implies that $\alpha_i^\star + \beta_i^\star = 1 + \frac{\varepsilon}{4\log(n)} - \frac{\varepsilon}{2} \geq \frac{1}{2}$ for all $i \in [n]$. Thus, we have

$$\|(\alpha^\star, \beta^\star)\| \geq \sqrt{\frac{\sum_{i=1}^n (\alpha_i^\star + \beta_i^\star)^2}{2}} = \frac{1}{2}\sqrt{\frac{n}{2}} \gtrsim \sqrt{n}.$$

As a consequence, we achieve the conclusion of the proposition. ∎

Approximation algorithm for OT by APDAGD. It is worth noting that the rounding procedure is missing in Dvurechensky et al. (2018, Algorithm 4) and we improve it to Algorithm 5. In particular, Dvurechensky et al. (2018, Algorithm 3) is used in Step 3 of Algorithm 5 for another function $\widetilde{\varphi}$ defined in Eq. (9). Given the corrected upper bound $\overline{R}$ and Algorithm 5 for approximating OT, we provide a new complexity bound of Algorithm 5 in the following proposition.

Proposition 20. *The APDAGD scheme for approximating optimal transport (Algorithm 5) returns an ε-approximate transportation plan (cf. Definition 1) in*

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

arithmetic operations.

Proof The proof is a simple modification of the proof for Dvurechensky et al. (2018, Theorem 4) and we only give a proof sketch here. In particular, we can obtain that the number of iterations for Algorithm 5 required to reach the tolerance ε is

$$t \leq O\left(\max\left\{\min\left\{\frac{n^{1/4}\sqrt{\overline{R}}\|C\|_\infty \log(n)}{\varepsilon}, \frac{\overline{R}\|C\|_\infty \log(n)}{\varepsilon^2}\right\}, \frac{\overline{R}\sqrt{\log n}}{\varepsilon}\right\}\right). \tag{30}$$

Moreover, we have $\overline{R} \leq \sqrt{n}\eta R$ where $R = \eta^{-1}\|C\|_\infty + \log(n) - 2\log(\min_{1 \leq i,j \leq n}\{r_i, c_j\})$. Therefore, the total iteration complexity in Step 3 of Algorithm 5 is $O(\sqrt{n\log(n)}\|C\|_\infty \varepsilon^{-1})$.

Each iteration of APDAGD requires $O(n^2)$ arithmetic operations. Therefore, the total number of arithmetic operations is $O(n^{5/2}\|C\|_\infty\sqrt{\log(n)}\varepsilon^{-1})$. Note that $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 5 can be found in $O(n)$ arithmetic operations and Altschuler et al. (2017, Algorithm 2) requires $O(n^2)$ arithmetic operations. Therefore, we conclude that the total number of arithmetic operations is $O(n^{5/2}\|C\|_\infty\sqrt{\log(n)}\varepsilon^{-1})$. ■

Remark 21. *As indicated in Proposition 20, the corrected complexity bound of APDAGD for the entropic regularized OT is similar to that of APDAMD when we choose* $\phi(\cdot) = \frac{1}{2n}\|\cdot\|^2$ *and have* $\delta = n$*. From this perspective, our algorithm can be viewed as a generalization of APDAGD. Since our algorithm utilizes* ℓ_∞*-norm in the line search criterion, it is more robust than APDAGD in practice; see Section 6 for the details.*

5. Accelerating Sinkhorn

In this section, we present an accelerated variant of Sinkhorn for solving the entropic regularized OT problem in Eq. (2). Combined with a rounding scheme, our algorithm can be used for solving the OT problem in Eq. (1) and achieves a complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$, which improves that of Sinkhorn in terms of $1/\varepsilon$ and APDAGD and AAM (Guminov et al., 2021) in terms of n. The idea comes from a novel combination of Nesterov's estimated sequence and the techniques for analyzing Sinkhorn.

5.1 Algorithmic procedure

We present the pseudocode of accelerated Sinkhorn in Algorithm 6. This algorithm achieves the acceleration by using Nesterov's estimate sequences (Nesterov, 2018). While our algorithm can be interpreted as an accelerated block coordinate descent algorithm, it is worth mentioning that our algorithm is purely *deterministic* and thus differs from other accelerated randomized algorithms (Lin et al., 2015; Fercoq and Richtárik, 2015; Lu et al., 2018) in the optimization literature.

Algorithm 6 is a novel combination of Nesterov's estimate sequences, a monotone search step, the choice of greedy coordinate and two coordinate updates. It is applied to solve the dual entropic regularized OT problem in Eq. (5):

$$\min_{u,v} \ \varphi(u,v) := \log(\|B(u,v)\|_1) - u^\top r - v^\top c.$$

More specifically, Nesterov's estimate sequences are responsible for optimizing a dual objective function φ in a fast rate. The coordinate update guarantees that $\varphi(\widehat{u}^t, \widehat{v}^t) \leq \varphi(\grave{u}^t, \grave{v}^t)$ and $\|B(\widehat{u}^t, \widehat{v}^t)\|_1 = 1$. The monotone search step guarantees that $\varphi(u^t, v^t) \leq \varphi(\widehat{u}^t, \widehat{v}^t)$. The greedy coordinate update guarantees that $\varphi(\check{u}^{t+1}, \check{v}^{t+1}) \leq \varphi(u^t, v^t)$ with sufficient progress.

Furthermore, we also use the same quantity as that in Greekhorn to measure the per-iteration residue of Algorithm 6:

$$E_t = \|r(B(u^t, v^t)) - r\|_1 + \|c(B(u^t, v^t)) - c\|_1. \tag{31}$$

The computationally expensive step is to compute $\frac{r(B(\bar{u}^t, \bar{v}^t))}{\|B(\bar{u}^t, \bar{v}^t)\|_1}$ and $\frac{c(B(\bar{u}^t, \bar{v}^t))}{\|B(\bar{u}^t, \bar{v}^t)\|_1}$. Since $B(\bar{u}^t, \bar{v}^t)$ does not have any special property, it is difficult to design some implementation trick to

Algorithm 6 Accelerated Sinkhorn$(C, \eta, r, c, \varepsilon')$

Input: $t = 0$, $\theta_0 = 1$ and $\check{u}^0 = \tilde{u}^0 = \check{v}^0 = \tilde{v}^0 = \mathbf{0}_n$.
while $E_t > \varepsilon'$ **do**
Compute $\begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} = (1-\theta_t) \begin{pmatrix} \check{u}^t \\ \check{v}^t \end{pmatrix} + \theta_t \begin{pmatrix} \tilde{u}^t \\ \tilde{v}^t \end{pmatrix}$.
Compute $\tilde{u}^{t+1}$ and $\tilde{v}^{t+1}$ by

$$\tilde{u}^{t+1} = \tilde{u}^t - \frac{1}{2\theta_t}\left(\frac{r(B(\bar{u}^t, \bar{v}^t))}{\|B(\bar{u}^t, \bar{v}^t)\|_1} - r\right), \qquad \tilde{v}^{t+1} = \tilde{v}^t - \frac{1}{2\theta_t}\left(\frac{c(B(\bar{u}^t, \bar{v}^t))}{\|B(\bar{u}^t, \bar{v}^t)\|_1} - c\right).$$

Compute $\begin{pmatrix} \grave{u}^t \\ \grave{v}^t \end{pmatrix} = \begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} + \theta_t \left(\begin{pmatrix} \tilde{u}^{t+1} \\ \tilde{v}^{t+1} \end{pmatrix} - \begin{pmatrix} \tilde{u}^t \\ \tilde{v}^t \end{pmatrix}\right)$.
if t is even **then**
$\widehat{u}^t = \grave{u}^t + \log(r) - \log(r(B(\grave{u}^t, \grave{v}^t)))$ and $\widehat{v}^t = \grave{v}^t$.
else
$\widehat{u}^t = \grave{u}^t$ and $\widehat{v}^t = \grave{v}^t + \log(c) - \log(c(B(\grave{u}^t, \grave{v}^t)))$.
end if
Compute $\begin{pmatrix} u^t \\ v^t \end{pmatrix} = \operatorname{argmin}\left\{\varphi(u, v) \middle| \begin{pmatrix} u \\ v \end{pmatrix} \in \left\{\begin{pmatrix} \check{u}^t \\ \check{v}^t \end{pmatrix}, \begin{pmatrix} \widehat{u}^t \\ \widehat{v}^t \end{pmatrix}\right\}\right\}$.
if t is even **then**
$\check{u}^{t+1} = u^t + \log(r) - \log(r(B(u^t, v^t)))$ and $\check{v}^{t+1} = v^t$.
else
$\check{u}^{t+1} = u^t$ and $\check{v}^{t+1} = v^t + \log(c) - \log(c(B(u^t, v^t)))$.
end if
Compute $\theta_{t+1} = \frac{\theta_t(\sqrt{\theta_t^2+4} - \theta_t)}{2}$.
Set $t = t + 1$.
end while
Output: $B(u^t, v^t)$.

Algorithm 7 Approximating OT by Algorithm 6

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

Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{c} \in \Delta_n$ be defined by $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon'}{8})(r, c) + \frac{\varepsilon'}{8n}(\mathbf{1}_n, \mathbf{1}_n)$.
Step 2: Compute $\tilde{X} =$ Accelerated Sinkhorn$(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$.
Step 3: Round $\tilde{X}$ to $\hat{X}$ using Altschuler et al. (2017, Algorithm 2) such that $\hat{X}\mathbf{1}_n = r$ and $\hat{X}^\top \mathbf{1}_n = c$.
Output: $\widehat{X}$.

reduce the order of n. As such, the arithmetic operations for each iteration is $O(n^2)$ and is exactly the same as Sinkhorn (Cuturi, 2013), APDAGD (Dvurechensky et al., 2018) and AAM (Guminov et al., 2021). Combining Algorithm 6 and Altschuler et al. (2017, Algorithm 2), we are ready to present the pseudocode of our main algorithm in Algorithm 7. The regularization parameter η is set as before, and Step 1 is necessary since accelerated Sinkhorn is not well behaved if the marginal distributions have sparse support.

5.2 Technical lemmas

We first present two technical lemmas which are essential in the analysis of Algorithm 6. The first lemma provides an inductive relationship on the quantity

$$\delta_t = \varphi(\check{u}^t, \check{v}^t) - \varphi(u^\star, v^\star), \tag{32}$$

where $(u^\star, v^\star)$ is an optimal solution of the dual entropic regularized OT problem in Eq. (5) that satisfies Lemma 4. To facilitate the discussion, we recall Eq. (10) with $\|A\|_{1\to 2} = \sqrt{2}$ as follows,

$$\varphi(u', v') - \varphi(u, v) - \begin{pmatrix} u' - u \\ v' - v \end{pmatrix}^\top \nabla\varphi(u, v) \leq \left\| \begin{pmatrix} u' - u \\ v' - v \end{pmatrix} \right\|^2, \tag{33}$$

which will be used in the proof of the first lemma.

Lemma 22. *Let $\{(\check{u}^t, \check{v}^t)\}_{t\geq 0}$ be the iterates generated by Algorithm 6 and $(u^\star, v^\star)$ be an optimal solution of the dual entropic regularized OT problem. Then, we have*

$$\delta_{t+1} \leq (1-\theta_t)\delta_t + \theta_t^2 \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right).$$

Proof Using Eq. (33) with $(u', v') = (\grave{u}^t, \grave{v}^t)$ and $(u, v) = (\bar{u}^t, \bar{v}^t)$, we have

$$\varphi(\grave{u}^t, \grave{v}^t) \leq \varphi(\bar{u}^t, \bar{v}^t) + \theta_t \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) + \theta_t^2 \left\| \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right\|^2.$$

After simple calculations, we find that

$$\begin{aligned} \varphi(\bar{u}^t, \bar{v}^t) &= (1-\theta_t)\varphi(\bar{u}^t, \bar{v}^t) + \theta_t\varphi(\bar{u}^t, \bar{v}^t), \\ \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) &= -\begin{pmatrix} \tilde{u}^t - \bar{u}^t \\ \tilde{v}^t - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) + \begin{pmatrix} \tilde{u}^{t+1} - \bar{u}^t \\ \tilde{v}^{t+1} - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t). \end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned} \varphi(\grave{u}^t, \grave{v}^t) \leq \ & \theta_t \left(\underbrace{\varphi(\bar{u}^t, \bar{v}^t) + \begin{pmatrix} \tilde{u}^{t+1} - \bar{u}^t \\ \tilde{v}^{t+1} - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) + \theta_t \left\| \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right\|^2}_{\mathbf{I}} \right) \\ & + \underbrace{(1-\theta_t)\varphi(\bar{u}^t, \bar{v}^t) - \theta_t \begin{pmatrix} \tilde{u}^t - \bar{u}^t \\ \tilde{v}^t - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t)}_{\mathbf{II}}. \end{aligned} \tag{34}$$

We first bound the term $\mathbf{I}$. Indeed, by the update formula for $(\tilde{u}^{t+1}, \tilde{v}^{t+1})$ and the definition of $\nabla\varphi$, we have

$$\begin{pmatrix} u - \tilde{u}^{t+1} \\ v - \tilde{v}^{t+1} \end{pmatrix}^\top \left(\nabla\varphi(\bar{u}^t, \bar{v}^t) + 2\theta_t \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right) = 0 \text{ for all } (u, v) \in \mathbb{R}^{2n}.$$

Letting $(u, v) = (u^\star, v^\star)$ and rearranging the resulting equation yields that

$$\begin{aligned} \begin{pmatrix} \tilde{u}^{t+1} - \bar{u}^t \\ \tilde{v}^{t+1} - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) &= \begin{pmatrix} u^\star - \bar{u}^t \\ v^\star - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) \\ &+ \theta_t \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} \tilde{u}^{t+1} - \tilde{u}^t \\ \tilde{v}^{t+1} - \tilde{v}^t \end{pmatrix} \right\|^2 \right). \end{aligned}$$

Using the convexity of φ, we have

$$\begin{pmatrix} u^\star - \bar{u}^t \\ v^\star - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) \leq \varphi(u^\star, v^\star) - \varphi(\bar{u}^t, \bar{v}^t).$$

Putting these pieces together yields that

$$\mathbf{I} \leq \varphi(u^\star, v^\star) + \theta_t \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right). \tag{35}$$

We then bound the term $\mathbf{II}$. Indeed, we see from the definition of $(\bar{u}^t, \bar{v}^t)$ that

$$-\theta_t \begin{pmatrix} \tilde{u}^t - \bar{u}^t \\ \tilde{v}^t - \bar{v}^t \end{pmatrix} = \theta_t \begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} + (1 - \theta_t) \begin{pmatrix} \check{u}^t \\ \check{v}^t \end{pmatrix} - \begin{pmatrix} \bar{u}^t \\ \bar{v}^t \end{pmatrix} = (1 - \theta_t) \begin{pmatrix} \check{u}^t - \bar{u}^t \\ \check{v}^t - \bar{v}^t \end{pmatrix}.$$

Combining the above equation with the convexity of φ, we have

$$\mathbf{II} = (1 - \theta_t) \left(\varphi(\bar{u}^t, \bar{v}^t) + \begin{pmatrix} \check{u}^t - \bar{u}^t \\ \check{v}^t - \bar{v}^t \end{pmatrix}^\top \nabla\varphi(\bar{u}^t, \bar{v}^t) \right) \leq (1 - \theta_t)\varphi(\check{u}^t, \check{v}^t). \tag{36}$$

Plugging Eq. (35) and Eq. (36) into Eq. (34) yields that

$$\varphi(\grave{u}^t, \grave{v}^t) \leq (1 - \theta_t)\varphi(\check{u}^t, \check{v}^t) + \theta_t \varphi(u^\star, v^\star) + \theta_t^2 \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right).$$

Since $(\check{u}^{t+1}, \check{v}^{t+1})$ is obtained by a coordinate update from (u^t, v^t), we have $\varphi(u^t, v^t) \geq \varphi(\check{u}^{t+1}, \check{v}^{t+1})$. By the definition of (u^t, v^t), we have $\varphi(\widehat{u}^t, \widehat{v}^t) \geq \varphi(u^t, v^t)$. Since $(\widehat{u}^t, \widehat{v}^t)$ is obtained by a coordinate update from $(\grave{u}^t, \grave{v}^t)$, we have $\varphi(\grave{u}^t, \grave{v}^t) \geq \varphi(\widehat{u}^t, \widehat{v}^t)$. Collecting all of these results leads to

$$\begin{aligned} \varphi(\check{u}^{t+1}, \check{v}^{t+1}) - \varphi(u^\star, v^\star) &\leq (1 - \theta_t)(\varphi(\check{u}^t, \check{v}^t) - \varphi(u^\star, v^\star)) \\ &\quad + \theta_t^2 \left(\left\| \begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix} \right\|^2 \right). \end{aligned}$$

This completes the proof. ■

The second lemma provides an upper bound for δ_t defined by Eq. (32) where $\{(\check{u}^t, \check{v}^t)\}_{t \geq 0}$ are generated by Algorithm 6 and $(u^\star, v^\star)$ is an optimal solution defined by Corollary 4.

Lemma 23. *Let $\{(\check{u}^t, \check{v}^t)\}_{t \geq 0}$ be the iterates generated by Algorithm 6 and $(u^\star, v^\star)$ be an optimal solution of the dual entropic regularized OT problem satisfying that $\|(u^\star, v^\star)\| \leq \sqrt{2n}R$ where R is defined in Corollary 4. Then, we have*

$$\delta_t \leq \frac{8nR^2}{(t+1)^2}.$$

Proof By simple calculations, we derive from the definition of θ_t that $\frac{\theta_{t+1}}{\theta_t} = \sqrt{1-\theta_{t+1}}$. Therefore, we conclude from Lemma 22 that

$$\left(\frac{1-\theta_{t+1}}{\theta_{t+1}^2}\right)\delta_{t+1} - \left(\frac{1-\theta_t}{\theta_t^2}\right)\delta_t \leq \left\|\begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix}\right\|^2 - \left\|\begin{pmatrix} u^\star - \tilde{u}^{t+1} \\ v^\star - \tilde{v}^{t+1} \end{pmatrix}\right\|^2.$$

Equivalently, we have

$$\left(\frac{1-\theta_t}{\theta_t^2}\right)\delta_t + \left\|\begin{pmatrix} u^\star - \tilde{u}^t \\ v^\star - \tilde{v}^t \end{pmatrix}\right\|^2 \leq \left(\frac{1-\theta_0}{\theta_0^2}\right)\delta_0 + \left\|\begin{pmatrix} u^\star - \tilde{u}^0 \\ v^\star - \tilde{v}^0 \end{pmatrix}\right\|^2.$$

Since $\theta_0 = 1$ and $\tilde{u}^0 = \tilde{v}^0 = \mathbf{0}_n$, we have $\delta_t \leq \theta_{t-1}^2\|(u^\star, v^\star)\|^2 \leq 2nR^2\theta_{t-1}^2$.

The remaining step is to show that $0 < \theta_t \leq \frac{2}{t+2}$. Indeed, the claim holds when $t = 0$ as we have $\theta_0 = 1$. Assume that the claim holds for $t \leq t_0$, i.e., $\theta_{t_0} \leq \frac{2}{t_0+2}$, we have

$$\theta_{t_0+1} = \frac{2}{1 + \sqrt{1 + \frac{4}{\theta_{t_0}^2}}} \leq \frac{2}{t_0+3}.$$

Putting these pieces together yields the desired inequality for δ_t. ■

5.3 Main results

We present an upper bound for the number of iterations required by Algorithm 6. Note that the per-iteration progress of Algorithm 6 is measured by the function $\rho : \mathbb{R}_+^n \times \mathbb{R}_+^n \rightarrow \mathbb{R}_+$ given by: $\rho(a, b) := \mathbf{1}_n^\top(b-a) + \sum_{i=1}^n a_i \log(\frac{a_i}{b_i})$.

Theorem 24. *Let $\{(u^t, v^t)\}_{t \geq 0}$ be the iterates generated by Algorithm 6. The number of iterations required to reach the stopping criterion $E_t \leq \varepsilon'$ satisfies*

$$t \leq 1 + \left(\frac{16\sqrt{n}R}{\varepsilon'}\right)^{2/3},$$

where $R > 0$ is defined in Lemma 3.

Proof We first claim that

$$\varphi(u^t, v^t) - \varphi(\check{u}^{t+1}, \check{v}^{t+1}) \geq \frac{1}{2}\left(\|r(B(u^t, v^t)) - r\|_1^2 + \|c(B(u^t, v^t)) - c\|_1^2\right). \tag{37}$$

By the definition of φ, we have

$$\begin{aligned} \varphi(u^t, v^t) - \varphi(\check{u}^{t+1}, \check{v}^{t+1}) \ &= \ \log(\|B(u^t, v^t)\|_1) \\ &- \log(\|B(\check{u}^{t+1}, \check{v}^{t+1})\|_1) - (u^t - \check{u}^{t+1})^\top r - (v^t - \check{v}^{t+1})^\top c. \end{aligned} \tag{38}$$

From the update formula for $(\widehat{u}^t, \widehat{v}^t)$ and $(\check{u}^{t+1}, \check{v}^{t+1})$, it is clear that $\|B(\widehat{u}^t, \widehat{v}^t)\|_1 = 1$ and $\|B(\check{u}^{t+1}, \check{v}^{t+1})\|_1 = 1$ for all $t \geq 0$. Then, we derive from the update formula for (u^t, v^t) that $\|B(u^t, v^t)\|_1 = 1$ for all $t \geq 1$. Therefore, we have

$$\begin{aligned}\varphi(u^t, v^t) - \varphi(\check{u}^{t+1}, \check{v}^{t+1}) &= -(u^t - \check{u}^{t+1})^\top r - (v^t - \check{v}^{t+1})^\top c \\ &= (\log(r) - \log(r(B(u^t, v^t))))^\top r + (\log(c) - \log(c(B(u^t, v^t))))^\top c.\end{aligned}$$

Since $\mathbf{1}_n^\top r = \mathbf{1}_n^\top r(B(u^t, v^t)) = \mathbf{1}_n^\top c = \mathbf{1}_n^\top c(B(u^t, v^t)) = 1$, we have

$$\varphi(u^t, v^t) - \varphi(\check{u}^{t+1}, \check{v}^{t+1}) = \rho(r, r(B(u^t, v^t))) + \rho(c, c(B(u^t, v^t))).$$

Using Altschuler et al. (2017, Lemma 4), we derive Eq. (37) as desired.

By the definition of (u^t, v^t), we have $\varphi(\check{u}^t, \check{v}^t) \geq \varphi(u^t, v^t)$. Plugging this inequality into Eq. (37) together with the Cauchy-Schwarz inequality yields

$$\varphi(\check{u}^t, \check{v}^t) - \varphi(\check{u}^{t+1}, \check{v}^{t+1}) \geq \frac{1}{4} E_t^2.$$

Therefore, we conclude that

$$\varphi(\check{u}^t, \check{v}^t) - \varphi(\check{u}^{t+1}, \check{v}^{t+1}) \geq \frac{1}{4}\left(\sum_{i=j}^t E_i^2\right) \text{ for any } j \in \{1, 2, \ldots, t\}.$$

Since $\varphi(\check{u}^{t+1}, \check{v}^{t+1}) \geq \varphi(u^\star, v^\star)$ for all $t \geq 1$, we have $\varphi(\check{u}^j, \check{v}^j) - \varphi(\check{u}^{t+1}, \check{v}^{t+1}) \leq \delta_j$. Then, it follows from Lemma 23 that

$$\sum_{i=j}^t E_i^2 \leq \frac{32nR^2}{(j+1)^2}.$$

Putting these pieces together with the fact that $E_t \geq \varepsilon'$ as soon as the stopping criterion is not fulfilled yields

$$\frac{32nR^2}{(j+1)^2(t-j+1)} \geq (\varepsilon')^2.$$

Since this inequality holds true for all $j \in \{1, 2, \ldots, t\}$, we assume without loss of generality that t is even and let $j = t/2$. Then, we obtain that

$$t \leq 1 + \left(\frac{16\sqrt{n}R}{\varepsilon'}\right)^{2/3}.$$

This completes the proof of the theorem. ■

We are ready to present the complexity bound of Algorithm 7 for solving the OT problem in Eq. (1). Note that $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$ is defined using the desired accuracy $\varepsilon > 0$.

Theorem 25. *The accelerated Sinkhorn scheme for approximating optimal transport (Algorithm 7) returns an ε-approximate transportation plan (cf. Definition 1) in*

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

arithmetic operations.

Proof Applying the same argument which is used in Theorem 12, we obtain that $\langle C, \widehat{X}\rangle - \langle C, X^\star\rangle \leq \varepsilon$ where $\tilde{X} = \text{Accelerated Sinkhorn}(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2})$ in Step 2 of Algorithm 7.

It remains to bound the number of iterations required by Algorithm 6 to reach the stopping criterion $E_t \leq \frac{\varepsilon'}{2}$. Using Theorem 24, we have

$$t \leq 1 + \left(\frac{32\sqrt{n}R}{\varepsilon'}\right)^{2/3}.$$

By the definition of R (cf. Lemma 3), $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$, we have

$$\begin{aligned}
t &\leq 1 + \left(\frac{32\sqrt{n}R}{\varepsilon'}\right)^{2/3} \\
&\leq 1 + \left(\frac{256\sqrt{n}\|C\|_\infty}{\varepsilon}\left(\frac{\|C\|_\infty}{\eta} + \log(n) - \log\left(\min_{1\leq i,j\leq n}\{r_i, c_j\}\right)\right)\right)^{2/3} \\
&\leq 1 + \left(\frac{256\sqrt{n}\|C\|_\infty}{\varepsilon}\left(\frac{4\log(n)\|C\|_\infty}{\varepsilon} + \log(n) - \log\left(\frac{\varepsilon}{64n\|C\|_\infty}\right)\right)\right)^{2/3} \\
&= O\left(\frac{n^{1/3}\|C\|_\infty^{4/3}(\log(n))^{1/3}}{\varepsilon^{4/3}}\right).
\end{aligned}$$

Since each iteration of Algorithm 6 requires $O(n^2)$ arithmetic operations, the total number of arithmetic operations required by Step 2 of Algorithm 7 is $O(n^{7/3}\|C\|_\infty^{4/3}(\log(n))^{1/3}\varepsilon^{-4/3})$. Computing two vectors $\tilde{r}$ and $\tilde{c}$ in Step 1 of Algorithm 7 requires $O(n)$ arithmetic operations and Altschuler et al. (2017, Algorithm 2) requires $O(n^2)$ arithmetic operations. Therefore, the complexity bound of Algorithm 7 is $O(n^{7/3}\|C\|_\infty^{4/3}(\log(n))^{1/3}\varepsilon^{-4/3})$. ■

Remark 26. *Theorem 25 shows that the complexity bound of accelerated Sinkhorn is better than that of Sinkhorn and Greenkhorn in terms of $1/\varepsilon$ but appears not to be near-linear in n^2. Thus, our algorithm is recommended when $n \ll 1/\varepsilon$. This occurs if the desired solution accuracy is relatively small, saying 10^{-4}, and the examples include the application problems from economics and operations research. In contrast, Sinkhorn and Greenkhorn are recommended when $n \gg 1/\varepsilon$. This occurs if the desired solution accuracy is relatively large, saying 10^{-2}, and the examples include the application problems from image processing.*

6. Experiments

In this section, we conduct the experiments to evaluate Greenkhorn, accelerated Sinkhorn and APDAMD on synthetic data and real images from the MNIST Digits dataset[1]. We use Sinkhorn (Cuturi, 2013), APDAGD (Dvurechensky et al., 2018) and GCPB[2] (Genevay et al., 2016) as the baseline approaches. Since the focus of this paper is the entropic regularized

1. http://yann.lecun.com/exdb/mnist/
2. GCPB is simply an application of stochastic averaged gradient (Schmidt et al., 2017) for solving the dual entropic regularized OT problem.

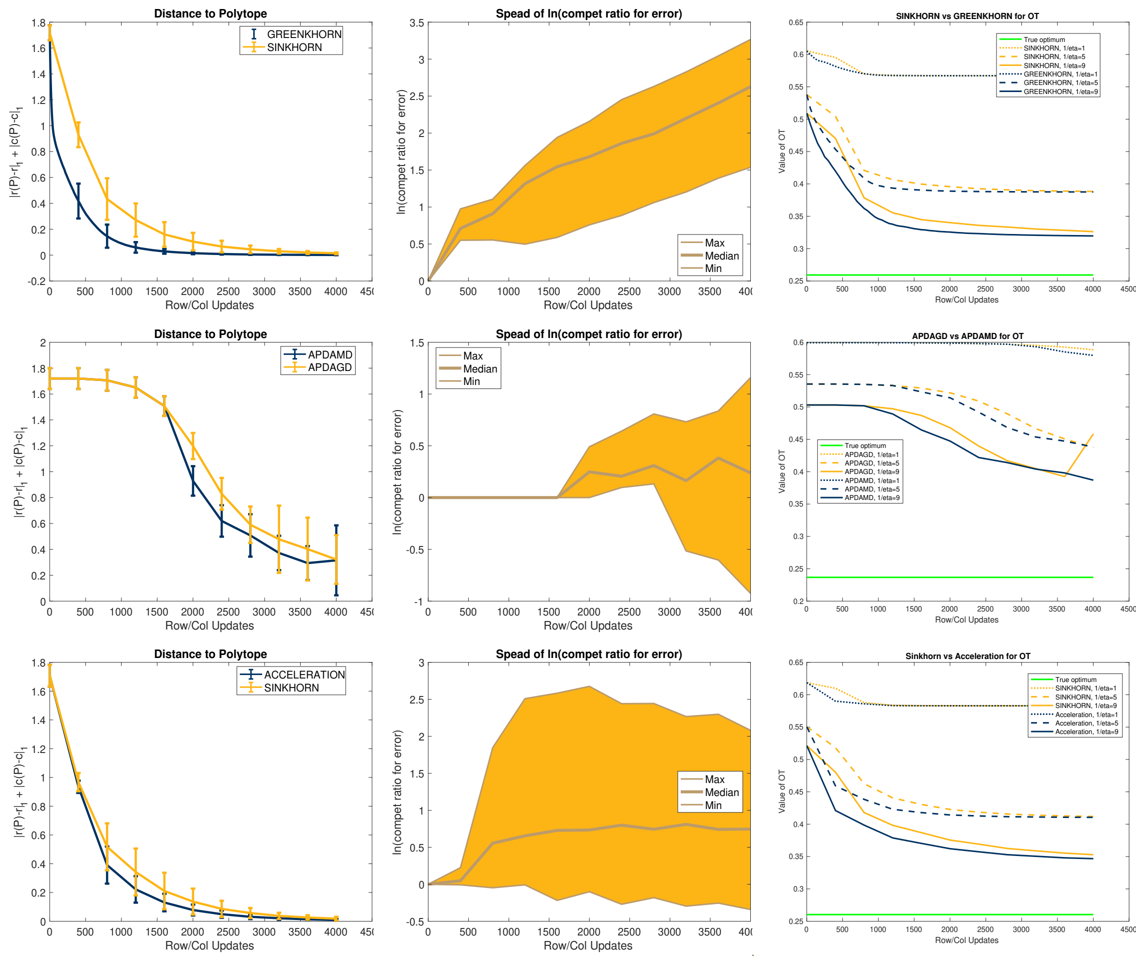

Figure 1: Comparative performance of Sinkhorn v.s. Greenkhorn, APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn on synthetic images.

algorithms, we exclude the combinatorial algorithms from our experiment and refer to Dong et al. (2020) for an excellent comparative study.

In the literature, Greenkhorn and APDAGD were shown to outperform the Sinkhorn algorithm in terms of row/column updates (Altschuler et al., 2017; Dvurechensky et al., 2018) and we repeat the comparisons for the sake of completeness. For parameter tuning in the implementation of Greenkhorn, accelerated Sinkhorn and APDAMD, we follow most of the setups as shown in Algorithm 1, 3 and 6 but employ more aggressive choice of stepsize for the coordinate gradient updates in Algorithm 6. To obtain an optimal value of the OT problem, we employ the default LP solver in MATLAB.

6.1 Synthetic images

To generate the synthetic images, we adopt the process from Altschuler et al. (2017) and evaluate the performance of different algorithms on these synthetic images. The transportation distance is defined between two synthetic images while the cost matrix is defined based on the ℓ_1 distances among locations of pixel in the images. Each image is of size 20 by 20 pixels and 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. We fix the proportion of the size of the foreground square as 10% of the whole images and implement all candidate algorithms.

We use the standard metrics to assess the performance of all the candidate algorithms. The first metric $d(.)$ is an ℓ_1 distance between the row, column outputs of some algorithm $\mathcal{A}$ and the corresponding transportation polytope of the probability measures, which is given by:

$$d(\mathcal{A}) := \|r(\mathcal{A}) - r\|_1 + \|c(\mathcal{A}) - c\|_1$$

where $r(\mathcal{A})$ and $c(\mathcal{A})$ are the row and column obtained from the output of the algorithm $\mathcal{A}$ and r and c are row and column vectors of the original probability measures. The second metric is defined as competitive ratio $\log(d(\mathcal{A}_1)/d(\mathcal{A}_2))$ where $d(\mathcal{A}_1)$ and $d(\mathcal{A}_2)$ are the distances between the row, column outputs of algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ and the transportation polytope. We perform three pairwise comparative experiments on 10 randomly generated data: Sinkhorn v.s. Greekhorn, APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn. To further evaluate these algorithms, we compare their performance with respect to different choices of regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$ while using the value of the OT problem as the baseline approach. The maximum number of iterations is $T = 5$. Figure 1 summarizes the experimental results. The images in the first row show the comparative performance of Sinkhorn and Greenkhorn in terms of the row/column updates. In the leftmost image, the comparison uses distance to transportation polytope $d(\mathcal{A})$ where $\mathcal{A}$ is either Sinkhorn or Greenkhorn. In the middle image, the maximum, median and minimum values of the competitive ratios $\log(d(\mathcal{A}_1)/d(\mathcal{A}_2))$ on 10 images are utilized for the comparison where $\mathcal{A}_1$ is Sinkhorn and $\mathcal{A}_2$ is Greenkhorn. In the rightmost image, we vary the regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$ with these algorithms and using the value of the unregularized OT problem as the baseline. The other rows of images present comparative results for APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn. We find that (i) Greenkhorn outperforms Sinkhorn in terms of row/column updates, illustrating the improvement from *greedy coordinate descent*; (ii) APDAMD with $\delta = n$ and $\phi = (1/2n)\| \cdot \|^2$ is more robust than APDAGD, illustrating the advantage of using *mirror descent* and line search with $\| \cdot \|_\infty$; (iii) accelerated Sinkhorn outperforms Sinkhorn in terms of row/column updates, illustrating the improvement from *estimated sequence* and *monotone search*.

6.2 MNIST images

We proceed to the comparison between different algorithms on real images, using essentially the same evaluation metrics as in the synthetic images. The MNIST dataset consists of 60,000 images of handwritten digits of size 28 by 28 pixels. To ensure that the masses of probability measures are dense, which leads to a tight dependence on n for our algorithms, we add a very small noise term (10^{-6}) to all zero elements in the measures and then normalize them so that their sum is 1. The maximum number of iterations is $T = 5$.

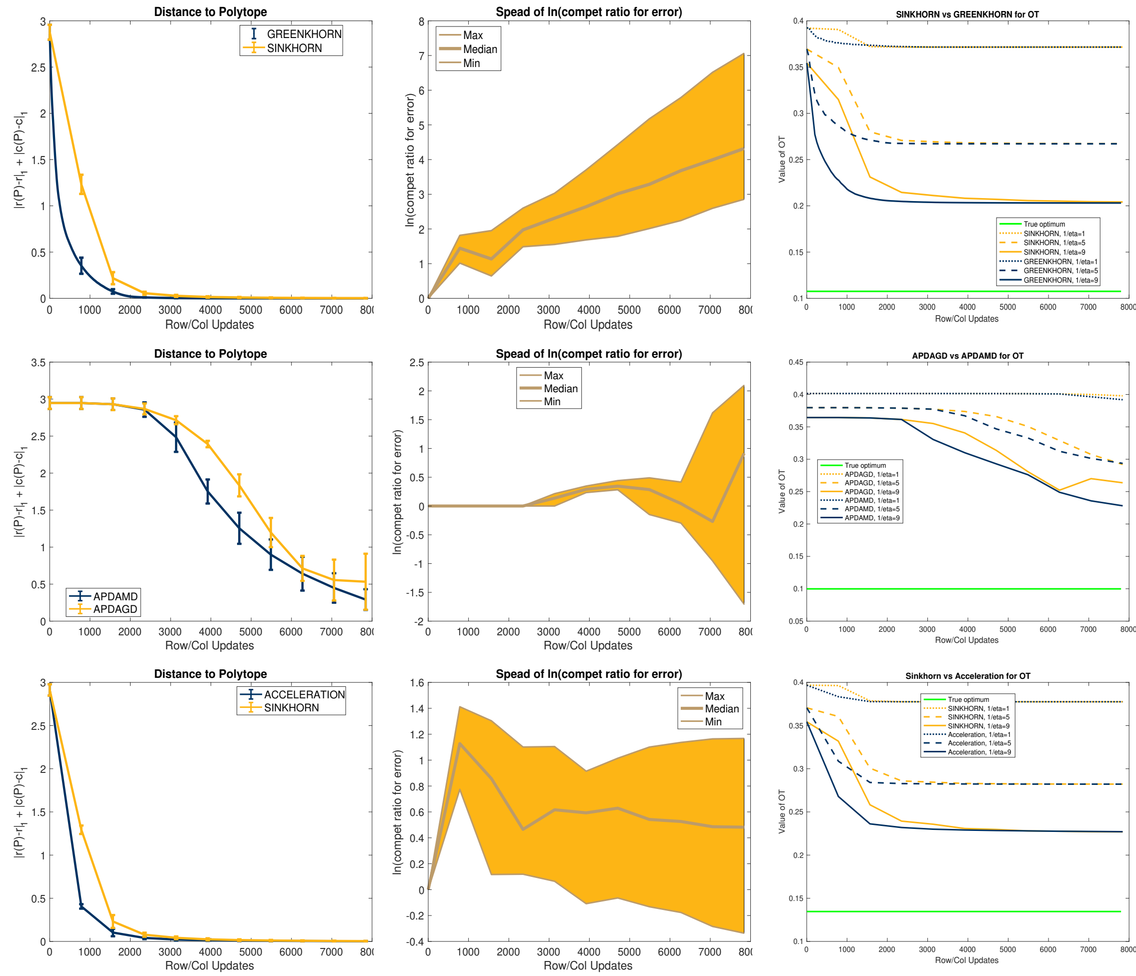

Figure 2: Comparative performance of Sinkhorn v.s. Greenkhorn, APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn on the MNIST real images.

Figures 2 and 3 summarize the experimental results on MNIST. In the first row of Figure 2, we compare Sinkhorn and Greenkhorn in terms of row/column updates. The leftmost image specifies the distances $d(\mathcal{A})$ to the transportation polytope for the algorithm $\mathcal{A}$, which is either Sinkhorn or Greenkhorn; the middle image specifies the maximum, median and minimum of competitive ratios $\log(d(\mathcal{A}_1)/d(\mathcal{A}_2))$ on ten random pairs of MNIST images, where $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively correspond to Sinkhorn and Greenkhorn; the rightmost image specifies the values of the entropic regularized OT problem with varying regularization parameters $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$. The remaining rows present comparative results for APDAGD v.s.APDAMD and Sinkhorn v.s.accelerated Sinkhorn. We observe that (i) the comparative performances of Sinkhorn v.s.Greenkhorn and APDAGD v.s.APDAMD are consistent with those on synthetic images; (ii) accelerated Sinkhorn deteriorates but remains better than Sinkhorn; (iii) APDAMD is more robust than APDAGD and GCPB.

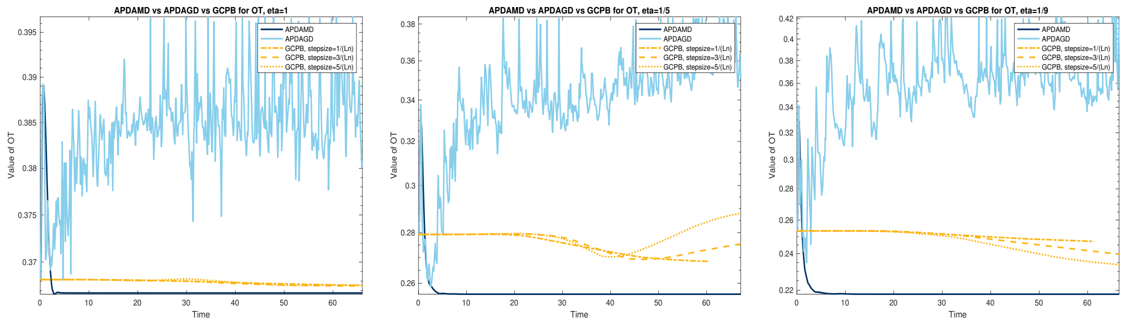

Figure 3: Performance of GCPB, APDAGD and APDAMD in term of time on the MNIST real images. These images specify the values of entropic regularized OT with varying regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$, demonstrating the robustness of APDAMD.

7. Conclusion

We first show that the complexity bound of Greenkhorn can be improved to $\widetilde{O}(n^2\varepsilon^{-2})$, which matches the best known bound of Sinkhorn. Then, we propose APDAMD by generalizing APDAGD with a prespecified mirror mapping ϕ and show that it achieves the complexity bound of $\widetilde{O}(n^2\sqrt{\delta}\varepsilon^{-1})$ where $\delta > 0$ refers to the regularity of ϕ. We prove that the complexity bound of $\widetilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ proved for APDAGD is invalid and prove a refined complexity bound of $\widetilde{O}(n^{5/2}\varepsilon^{-1})$. Moreover, we propose a *deterministic* accelerated variant of Sinkhorn via appeal to estimate sequence techniques and prove the complexity bound of $\widetilde{O}(n^{7/3}\varepsilon^{-4/3})$. As such, we see that accelerated Sinkhorn outperforms Sinkhorn and Greenkhorn in terms of $1/\varepsilon$ and APDAGD and AAM in terms of n. Experiments on synthetic data and real images demonstrate the efficiency of our algorithms.

There are a few promising future directions arising from this work. First, it is important to develop fast algorithms to compute dimension-reduced versions of OT. Indeed, the OT suffers from the curse of dimensionality (Dudley, 1969; Fournier and Guillin, 2015), which means that a large amount of samples from two continuous measures is necessary to approximate the true OT between them. This can be mitigated when data lie on low-dimensional manifolds (Weed and Bach, 2019; Paty and Cuturi, 2019) but the sample complexity still remain pessimistic even in that case. This motivates recent works on efficient dimension-reduced OT, e.g., the sliced OT (Bonneel et al., 2015), generalized sliced OT (Kolouri et al., 2019), distributional sliced OT (Nguyen et al., 2021), further inspiring us to explore the application of our algorithms to these settings and eventually automatic differentiation schemes. Second, there have been several application problems arising from the interplay between OT and adversarial ML; see Bhagoji et al. (2019) and Pydi and Jog (2020) for example. However, it is known that OT has robustness issues when there are outliers in the supports of probability measures. Robust OT had been introduced to deal with these robustness issues (Balaji et al., 2020) where the idea is to relax the marginal constraints via certain probability divergences, such as KL divergence. It is to limit the amount of masses that the transportation plan will assign for the outliers in the supports of measures. Similar to OT, a key practical question with robust OT is computational. As such, we manage to develop efficient algorithms for the robust OT problem in the future work.

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 to MJ, and by the NSF IFML 2019844 award and research gifts by UT Austin ML grant to NH.

References

B. K. Abid and R. M. Gower. Greedy stochastic algorithms for entropy-regularized optimal transport problems. In *AISTATS*, 2018.

Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson. Much faster algorithms for matrix scaling. In *FOCS*, pages 890–901. IEEE, 2017.

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *NeurIPS*, pages 1964–1974, 2017.

J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *NeurIPS*, pages 4429–4439, 2019.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In *NeurIPS*, 2020.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. In *NeurIPS*, pages 7498–7510, 2019.

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

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

J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In *AISTATS*, pages 880–889, 2018.

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

M. Carrière, M. Cuturi, and S. Oudot. Sliced Wasserstein kernel for persistence diagrams. In *ICML*, pages 1–10, 2017.

D. Chakrabarty and S. Khanna. Better and simpler error analysis of the Sinkhorn-Knopp algorithm for matrix scaling. In *1st Symposium on Simplicity in Algorithms*, 2018.

L. Chizat, P. Roussillon, F. Léger, F-X. Vialard, and G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In *NeurIPS*, pages 2257–2269, 2020.

M. B. Cohen, A. Madry, D. Tsipras, and A. Vladu. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *FOCS*, pages 902–913. IEEE, 2017.

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.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *ICML*, pages 685–693, 2014.

M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

M. Cuturi, O. Teboul, and J-P. Vert. Differentiable ranks and sorting using optimal transport. In *NeurIPS*, pages 6861–6871, 2019.

Y. Dong, Y. Gao, R. Peng, I. Razenshteyn, and S. Sawlani. A study of performance of optimal transport. *ArXiv Preprint: 2005.01182*, 2020.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *NeurIPS*, pages 10783–10793, 2018.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *ICML*, pages 1367–1376, 2018.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

J. Feydy, T. Séjourné, F-X. Vialard, S-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *AISTATS*, pages 2681–2690. PMLR, 2019.

R. Flamary and N. Courty. POT: Python optimal transport library, 2017. URL `https://github.com/rflamary/POT`.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

H. N. Gabow and R. E. Tarjan. Faster scaling algorithms for general graph matching problems. *Journal of the ACM (JACM)*, 38(4):815–853, 1991.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *NeurIPS*, pages 3440–3448, 2016.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *AISTATS*, pages 1574–1583. PMLR, 2019.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *NeurIPS*, pages 5767–5777, 2017.

S. Guminov, P. Dvurechensky, N. Tupitsa, and A. Gasnikov. On a combination of alternating minimization and Nesterov's momentum. In *ICML*, pages 3886–3898. PMLR, 2021.

W. Guo, N. Ho, and M. Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *AISTATS*, pages 2088–2097. PMLR, 2020.

N. Ho, V. Huynh, D. Phung, and M. I. Jordan. Probabilistic multilevel clustering via composite transportation distance. *AISTATS*, pages 3149–3157, 2019.

A. Jambulapati, A. Sidford, and K. Tian. A direct tilde {O}(1/epsilon) iteration parallel algorithm for optimal transport. In *NeurIPS*, pages 11355–11366, 2019.

B. Kalantari and L. Khachiyan. On the complexity of nonnegative matrix scaling. *Linear Algebra and its Applications*, 240:87–103, 1996.

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.

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

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

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *NeurIPS*, pages 261–272, 2019.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

H. W. Kuhn. Variants of the Hungarian method for assignment problems. *Naval Research Logistics Quarterly*, 3(4):253–258, 1956.

N. Lahn, D. Mulchandani, and S. Raghvendra. A graph theoretic additive approximation of optimal transport. In *NeurIPS*, pages 13813–13823, 2019.

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.

J. Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

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.

T. Lin, N. Ho, and M. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *ICML*, pages 3982–3991, 2019a.

T. Lin, Z. Hu, and X. Guo. Sparsemax and relaxed Wasserstein for topic sparsity. In *WSDM*, pages 141–149. ACM, 2019b.

H. Lu, R. Freund, and V. Mirrokni. Accelerating greedy coordinate descent methods. In *ICML*, pages 3263–3272, 2018.

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *NeurIPS*, pages 4541–4551, 2019.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

K. Nguyen, N. Ho, T. Pham, and H. Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *ICLR*, 2021. URL `https://openreview.net/forum?id=QYjO70ACDK`.

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

X. Nguyen. Borrowing strength in hierarchical Bayes: posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016.

J. B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.

J. B. Orlin and R. K. Ahuja. New scaling algorithms for the assignment and minimum mean cycle problems. *Mathematical Programming*, 54(1):41–56, 1992.

J. B. Orlin, S. A. Plotkin, and E. Tardos. Polynomial dual network simplex algorithms. *Mathematical Programming*, 60(1):255–276, 1993.

F-P. Paty and M. Cuturi. Subspace robust Wasserstein distances. In *ICML*, pages 5072–5081. PMLR, 2019.

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

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

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *ICML*, pages 2664–2672, 2016.

M. S. Pydi and V. Jog. Adversarial risk via optimal transport and optimal couplings. In *ICML*, pages 7814–7823. PMLR, 2020.

K. Quanrud. Approximating optimal transport with linear programs. In *SOSA*, pages 61–69, 2019.

A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *AISTATS*, pages 630–638, 2016.

M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

M. H. Schneider and S. A. Zenios. A comparative study of algorithms for matrix balancing. *Operations Research*, 38(3):439–455, 1990.

J. Sherman. Area-convexity, ℓ_∞ regularization, and undirected multicommodity flow. In *STOC*, pages 452–460. ACM, 2017.

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

M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

M. Sommerfeld, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20:1–23, 2019.

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *AISTATS*, pages 912–920, 2015.

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

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *ICLR*, 2018.

J. van den Brand, Y. T. Lee, Y. P. Liu, T. Saranurak, A. Sidford, Z. Song, and D. Wang. Minimum cost flows, MDPs, and ℓ_1-regression in nearly linear time for dense instances. In *STOC*, pages 859–869, 2021.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.