On the Efficiency of Sinkhorn and Greenkhorn and Their Acceleration for Optimal Transport

Tianyi Lin° Nhat Ho⋆ Michael I. Jordan°,†

Department of Electrical Engineering and Computer Sciences°
Department of Statistics†
University of California, Berkeley
Department of Statistics and Data Sciences, University of Texas, Austin⋆

August 3, 2021

Abstract

We present several new complexity results for the 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 the Sinkhorn algorithm, known as Greenkhorn algorithm, from \( \tilde{O}(n^2\varepsilon^{-3}) \) to \( \tilde{O}(n^2\varepsilon^{-2}) \). Notably, this matches the best known complexity bound of the Sinkhorn algorithm and sheds light to superior practical performance of the Greenkhorn algorithm. Second, we generalize an adaptive primal-dual accelerated gradient descent (APDAGD) algorithm [Dvurechensky et al., 2018] with mirror mapping \( \phi \) and prove that the resulting APDAMD algorithm achieves the complexity bound of \( \tilde{O}(n^2\sqrt{\delta}\varepsilon^{-1}) \) where \( \delta > 0 \) refers to the regularity of \( \phi \). We demonstrate that the complexity bound of \( \tilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\}) \) is invalid for the APDAGD algorithm and establish a new complexity bound of \( \tilde{O}(n^{5/2}\varepsilon^{-1}) \). Moreover, we propose a deterministic accelerated Sinkhorn algorithm and prove that it achieves the complexity bound of \( \tilde{O}(n^{7/3}\varepsilon^{-4/3}) \) by incorporating an estimate sequence. Therefore, the accelerated Sinkhorn algorithm outperforms the Sinkhorn and Greenkhorn algorithms in terms of \( 1/\varepsilon \) and the APDAGD and accelerated alternating minimization [Guminov et al., 2021] algorithms in terms of \( n \). Finally, we conduct experiments on synthetic data and real images with the proposed algorithms in the paper and demonstrate their efficiency via numerical results.

1 Introduction

From its origins in the seminal works by Monge and Kantorovich in the eighteenth and twentieth centuries, respectively, 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, Courty et al., 2017, Srivastava et al., 2018, Dvurechenski et al., 2018, Tolstikhin et al., 2018, Sommerfeld et al., 2019, Lin et al., 2019b, Ho et al., 2019]. In these applications, the focus is on the probability distributions underlying the OT formulation. These distributions are generally either empirical distributions, obtained by placing unit masses at data points, or are probability models of a putative underlying data-generating process. The OT problem accordingly often has a direct inferential meaning —
as the definition of an estimator, the definition of a likelihood, or as the robust variant of an estimator [Fournier and Guillin, 2015, Weed and Bach, 2019]. The key challenge is computational [Peyré and Cuturi, 2019]. Indeed, in machine learning applications the underlying distributions generally involve high-dimensional data sets and complex probability models.

We study the OT problem in a discrete setting where we assume that the target and source probability distributions each have at most \( n \) atoms. In this setting, one of the state-of-the-art approaches for solving OT problems are interior-point methods, reflecting the linear-programming formulation of the OT problem. For example, a specialized interior-point method delivers a complexity bound of \( O(n^3) \) [Pele and Werman, 2009] which has been recently improved to \( O(n^{5/2}) \) [Lee and Sidford, 2014] via an appeal to Laplacian linear system algorithms. Neither method, however, provides an effective practical solution to large-scale machine learning problems; the former because of scalability issues and the latter because practical implementations of Laplacian approach are yet unknown.

Cuturi [2013] initiated a productive line of research in which an entropic regularization was imposed to approximate the non-negative constraints in the transportation plan. This OT problem is referred to as entropic regularized OT or regularized OT. The key advantage of regularized OT is that its dual representation has structures that can be exploited computationally. Indeed, Cuturi [2013] demonstrated that a dual coordinate ascent algorithm for solving regularized OT is equivalent to Sinkhorn in the literature [Sinkhorn, 1974, Kalantari and Khachiyan, 1996, Knight, 2008, Kalantari et al., 2008, Chakrabarty and Khanna, 2018]. Motivated by the emerging OT applications, various new algorithms were proposed and shown to outperform Sinkhorn empirically, including Greenkhorn [Altschuler et al., 2017, Abid and Gower, 2018]. In this regard, Altschuler et al. [2017] have shown that both Sinkhorn and Greenkhorn achieve the near-linear time complexity bound of \( O(n^2 \varepsilon^{-3}) \) and Dvurechensky et al. [2018] improved the complexity bound for Sinkhorn to \( O(n^2 \varepsilon^{-2}) \).

Further progress has been made by considering other algorithmic procedures for the OT problems [Cuturi and Peyré, 2016, Genevay et al., 2016, Blondel et al., 2018, Dvurechensky et al., 2018, Altschuler et al., 2018, 2019]. One typical example is an adaptive primal-dual accelerated gradient descent (APDAGD) [Dvurechensky et al., 2018], which is claimed to achieve the complexity bound of \( \widetilde{O}(\min\{n^{9/4} \varepsilon^{-1}, n^2 \varepsilon^{-2}\}) \). There are also several second-order algorithms [Allen-Zhu et al., 2017, Cohen et al., 2017] adapted by Blanchet et al. [2018] and Quanrud [2019] with the complexity bound of \( \widetilde{O}(n^2 \varepsilon^{-1}) \). Unfortunately, these algorithms are complicated such that efficient implementations have not been available yet. Nonetheless, this complexity bound can be viewed as a theoretical benchmark for the algorithms that we consider in this paper.

We summarize the main contributions of this paper as follows:

1. We prove that the complexity bound of Greenkhorn is \( \widetilde{O}(n^2 \varepsilon^{-2}) \), which matches the best existing bound of the Sinkhorn algorithm. The proof techniques are new and different from that in Dvurechensky et al. [2018] for analyzing Sinkhorn. In particular, the Greenkhorn algorithm only updates a single row or column at a time and quantifying the per-iteration progress is accordingly more difficult than the measurement in the Sinkhorn algorithm.

2. We propose an adaptive primal-dual accelerated mirror descent (APDAMD) algorithm by generalizing the APDAGD algorithm with a mirror mapping \( \phi \). The APDAMD algorithm achieves the complexity bound of \( \widetilde{O}(n^2 \sqrt{\delta} \varepsilon^{-1}) \) where \( \delta > 0 \) refers to the regularity of \( \phi \) with respect to \( \ell_\infty \) norm. We demonstrate that the complexity bound of
We note in passing that a preliminary version with only the analysis for the Greenkhorn and APDAGD algorithms. In Section 3, we provide the complexity analysis for Greenkhorn. In Section 3, this complexity bound is believed to be optimal. [Dvurechensky et al., 2018] is invalid for the APDAGD algorithm and establish a new complexity bound of $\tilde{O}(n^{5/2}\varepsilon^{-1})$ which is slightly worse than the aforementioned bound in terms of $n$.

3. We propose a deterministic algorithm with the complexity bound of $\tilde{O}(n^{7/3}\varepsilon^{-4/3})$ which can be interpreted as an accelerated variant of Sinkhorn via appeal to an estimated sequence. The resulting accelerated algorithm involves exact minimization for the main iterates accompanied by an auxiliary sequence of iterates based on a coordinate gradient update and a monotone scheme. The new accelerated algorithm outperforms the Sinkhorn and Greenkhorn algorithms in terms of $1/\varepsilon$ and the APDAGD and accelerated alternating minimization [Guminov et al., 2021] algorithms in terms of $n$.

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 the APDAMD algorithm for solving regularized OT and provide several results on the complexity bound of APDAGD and APDAMD algorithms. In Section 5, we design an accelerated Sinkhorn with solid theoretical guarantee. In Section 6, we conduct the experiments on synthetic data and real images in which the numerical results demonstrate 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 $\ell_p$-norm and $\|x\|$ indicates an $\ell_2$-norm. diag$(x)$ is a diagonal matrix which has the vector $x$ on its diagonal. $\mathbf{1}_n$ is a $n$-dimensional vector with all components being 1. For a matrix $A \in \mathbb{R}^{n \times n}$, we denote vec$(A)$ as the vector in $\mathbb{R}^{n^2}$ obtained from concatenating the rows and columns of $A$. The notation $\|A\|_{1\rightarrow1}$ stands for $\sup_{\|x\|_1=1}\|Ax\|_1$ and the notations $r(A) = A\mathbf{1}_n$ and $c(A) = A^T\mathbf{1}_n$ 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 $\varepsilon$. We also denote $a = \Theta(b(n,\varepsilon))$ iff $a = O(b(n,\varepsilon)) = \Omega(b(n,\varepsilon))$. Similarly, we denote $a = \tilde{O}(b(n,\varepsilon))$ to indicate the previous inequality where $C_1$ depends on some logarithmic function of $n$ and $\varepsilon$. 

3
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:

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

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 \(\varepsilon\)-approximate transportation plan \(\hat{X} \in \mathbb{R}^{n \times n}_+\) satisfying the marginal distribution constraints \(\hat{X} \mathbf{1}_n = r\) and \(\hat{X}^\top \mathbf{1}_n = c\) and the inequality given by

\[
\langle C, \hat{X} \rangle \leq \langle C, X^\star \rangle + \varepsilon.
\]

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 \(\varepsilon\)-approximate transportation cost and \(\hat{X}\) an \(\varepsilon\)-approximate transportation plan for the original problem. Formally, we have the following definition of \(\varepsilon\)-approximate transportation plan.

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

\[
\langle C, \hat{X} \rangle \leq \langle C, X^\star \rangle + \varepsilon.
\]

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 \(\varepsilon\)-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 (\(\varepsilon\) 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

\[
\min_{X \in \mathbb{R}^{n \times n}} \langle C, X \rangle - \eta H(X), \quad \text{s.t.} \quad X \mathbf{1}_n = r, X^\top \mathbf{1}_n = c,
\] (2)
where $\eta > 0$ denotes the regularization parameter and $H(X)$ denotes the entropic regularization term, which is given by:

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

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 $\{ X \in \mathbb{R}^{n \times n} : X \geq 0, X1_n = r, X^\top 1_n = c \}$ is bounded. 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 $\ell_1$-norm. However, the optimal value of the entropic regularized OT problem (cf. Eq (2)) yields a poor approximation to the unregularized OT problem if $\eta$ is large. An additional issue of entropic regularization is that the sparsity of the solution is lost. Although an $\varepsilon$-approximate transportation plan can be found efficiently, it is not clear how different the resulting sparsity pattern of the obtained 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 e.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 $\eta$ 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:

$$\mathcal{L}(X, \lambda_1, \ldots, \lambda_m) = \langle C, X \rangle - \eta H(X) - \alpha^\top (X1_n - r) - \beta^\top (X^\top 1_n - c).$$

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

$$\min_{X : \|X\|_1 = 1} \langle C, X \rangle - \eta H(X) - \alpha^\top (X1_n - r) - \beta^\top (X^\top 1_n - c).$$

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

$$\bar{X}_{ij} = \frac{e^{\eta^2(\alpha_i + \beta_j - C_{ij})}}{\sum_{1 \leq i, j \leq n} e^{\eta^2(\alpha_i + \beta_j - C_{ij})}}.$$  

(4)

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

$$\max_{\alpha, \beta} \left\{ -\eta \log \left( \sum_{1 \leq i, j \leq n} e^{\eta^2(\alpha_i + \beta_j - C_{ij})} \right) + \alpha^\top r + \beta^\top c \right\}.$$  

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

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

5
To further simplify the notation, we define $B(u, v) := (B_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$ 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 2.1.** The first part of the objective function $\varphi$ 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 $\varphi$ 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 the accelerated alternating minimization algorithm.

In the remainder of the paper, we also denote $(u^*, v^*) \in \mathbb{R}^{2n}$ 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 $(u^*, v^*) \in \mathbb{R}^{2n}$ such that it has an upper bound in terms of the $\ell_\infty$-norm.

**Lemma 2.2.** For the dual entropic regularized OT problem in Eq. (5), there exists an optimal solution $(u^*, v^*)$ such that

$$\|u^*\|_\infty \leq R, \quad \|v^*\|_\infty \leq R,$$

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

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

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

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

$$\tilde{\Delta}_u = \frac{\max_{1 \leq i \leq n} \tilde{u}_i^* + \min_{1 \leq i \leq n} \tilde{u}_i^*}{2},$$

$$\tilde{\Delta}_v = \frac{\max_{1 \leq i \leq n} \tilde{v}_i^* + \min_{1 \leq i \leq n} \tilde{v}_i^*}{2},$$

and define $(u^*, v^*)$ by

$$u^* = \tilde{u}^* - \tilde{\Delta}_u 1_n, \quad v^* = \tilde{v}^* - \tilde{\Delta}_v 1_n.$$ 

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

\[
\begin{align*}
\max_{1 \leq i \leq n} u_i^* - \min_{1 \leq i \leq n} u_i^* & \leq \frac{\|C\|_{\infty}}{\eta} - \log \left( \min_{1 \leq i, j \leq n} \{r_i, c_j\} \right), \\
\max_{1 \leq i \leq n} v_i^* - \min_{1 \leq i \leq n} v_i^* & \leq \frac{\|C\|_{\infty}}{\eta} - \log \left( \min_{1 \leq i, j \leq n} \{r_i, c_j\} \right).
\end{align*}
\]

(7)

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

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

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

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

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

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

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

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

and

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

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

The upper bound for the \(\ell_{\infty}\)-norm of an optimal solution of dual entropic-regularized OT in Lemma 2.2 directly leads to the following direct bound for the \(\ell_2\)-norm.

**Corollary 2.3.** For the dual entropic regularized OT problem in Eq. (5), there exists an optimal solution \((u^*, v^*)\) such that

\[\|u^*\| \leq \sqrt{n}R, \quad \|v^*\| \leq \sqrt{n}R,\]

where \(R > 0\) is defined in Lemma 2.2.

Since the function \(-H(X)\) is strongly convex with respect to the \(\ell_1\)-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:

\[
\min_{x \in Q} f(x), \quad \text{s.t. } Ax = b,
\]
We notice that the function $\phi$ where

$O(\sqrt{n}\log(n)\varepsilon^{-2})$ to $O(n^2\|C\|_2^2 \log(n)\varepsilon^{-2})$, which matches the current state-of-the-art complexity bound for the Sinkhorn algorithm [Dvurechensky et al., 2018].
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 \((\tilde{r}, \tilde{c}) = (1 - \varepsilon' \frac{\varepsilon}{8}) (r, c) + \varepsilon' \frac{\varepsilon}{8n} (1_n, 1_n)\).

Step 2: Compute \( \tilde{X} = \text{GREENKHORN}(C, \eta, \tilde{r}, \tilde{c}, \varepsilon' \frac{\varepsilon}{2}) \).

Step 3: Round \( \tilde{X} \) to \( \hat{X} \) using Altschuler et al. [2017, Algorithm 2] such that \( \hat{X} 1_n = r \) and \( \hat{X}^\top 1_n = c \).

Output: \( \hat{X} \).

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 \( r(B(u, v)) - r = 0 \) and \( c(B(u, v)) - c = 0 \). 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 algorithms 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]:

\[ f(u, v) := \|B(u, v)\|_1 - u^\top r - v^\top c. \]

Comparing with Sinkhorn, which performs alternating updates of all rows and columns, Greenkhorn updates a single row or column at each step. Therefore, the Greenkhorn algorithm updates only \( O(n) \) entries per iteration rather than \( O(n^2) \) in Sinkhorn. It is also worth noting that Greenkhorn can be easily implemented such that each iteration runs in only \( O(n) \) arithmetic operations [Altschuler et al., 2017].

Despite significantly cheap per-iteration computational cost, it is difficult to quantify the per-iteration progress of the Greenkhorn algorithm and the proof techniques for the Sinkhorn algorithm in Dvurechensky et al. [2018] are not applicable to the Greenkhorn. This then motivates us to explore a different proof strategy which will be elaborated in the sequel.

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

\[ f(u^t, v^t) - f(u^*, v^*) \leq 2E_t(\|u^*\|_\infty + \|v^*\|_\infty), \]

where \((u^*, v^*)\) is a point which 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 $\nabla_u f(u^t, v^t) = B(u^t, v^t)1_n - r$ and $\nabla_v f(u^t, v^t) = B(u^t, v^t)^\top 1_n - c$. 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 $(u^*, v^*)$, we have

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

Combining Hölder’s inequality and the definition of $E_t$ yields

$$f(u^t, v^t) - f(u^*, v^*) \leq E_t(\|u^t - u^*\|_\infty + \|v^t - v^*\|_\infty). \quad (11)$$

Thus, it suffices to show that

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

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

$$\max\{\|u^t - u^*\|_\infty, \|v^t - v^*\|_\infty\} \leq \max\{\|u^0 - u^*\|_\infty, \|v^0 - v^*\|_\infty\}. \quad (12)$$

It is clear that Eq. (12) 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

$$\|u^{k_0+1} - u^*\|_\infty \leq \max\{\|u^{k_0} - u^*\|_\infty, \|u_I^{k_0+1} - u_I^*\|\}, \quad (13)$$

$$\|v^{k_0+1} - v^*\|_\infty = \|v^{k_0} - v^*\|_\infty. \quad (14)$$

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

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

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

$$|u_I^{k_0+1} - u_I^*| = \log \left( \frac{\sum_{j=1}^n e^{-\eta^{-1} C_{ij} + v_j}}{\sum_{j=1}^n e^{-\eta^{-1} C_{ij} + v_j^0}} \right) \leq \|v^{k_0} - v^*\|_\infty. \quad (15)$$

Combining Eq. (13) and Eq. (15) yields

$$\|u^{k_0+1} - u^*\|_\infty \leq \max\{\|u^{k_0} - u^*\|_\infty, \|v^{k_0} - v^*\|_\infty\}. \quad (16)$$

Combining Eq. (14) and Eq. (16) further implies Eq. (12). This together with $u^0 = v^0 = 0$ implies

$$\|u^t - u^*\|_\infty + \|v^t - v^*\|_\infty \leq 2(\|u^0 - u^*\|_\infty + \|v^0 - v^*\|_\infty) = 2\|u^*\|_\infty + 2\|v^*\|_\infty. \quad (17)$$

Putting Eq. (11) and Eq. (17) together yields the desired result. □

Our second lemma shows that at least one optimal solution $(u^*, v^*)$ 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 $\ell_\infty$-norm. This result is stronger than Dvurechensky et al. [2018, Lemma 1] and generalizes Blanchet et al. [2018, Lemma 8].
Lemma 3.2. There exists an optimal solution pair \((u^*, v^*)\) of \(f\) such that the following inequality holds true,
\[ \|u^*\|_\infty \leq R, \quad \|v^*\|_\infty \leq R, \]
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. First, we claim that there exists an optimal solution pair \((u^*, v^*)\) such that
\[ \max_{1 \leq i \leq n} u_i^* \geq \min_{1 \leq i \leq n} u_i^*. \] (18)
Indeed, by the convexity of \(f\) with respect to \((u, v)\), this function has at least one optimal solution. Suppose that \((\tilde{u}^*, \tilde{v}^*)\) is an optimal solution where
\[ +\infty > \max_{1 \leq i \leq n} \tilde{u}_i^* \geq \min_{1 \leq i \leq n} \tilde{u}_i^* > -\infty, \quad +\infty > \max_{1 \leq i \leq n} \tilde{v}_i^* \geq \min_{1 \leq i \leq n} \tilde{v}_i^* > -\infty. \]
Then, we can define \((u^*, v^*)\) by
\[ u^* = \tilde{u}^* - \frac{\max_{1 \leq i \leq n} u_i^* + \min_{1 \leq i \leq n} u_i^*}{2} 1_n, \quad v^* = \tilde{v}^* + \frac{\max_{1 \leq i \leq n} \tilde{v}_i^* + \min_{1 \leq i \leq n} \tilde{v}_i^*}{2} 1_n, \]
and see that \((u^*, v^*)\) satisfies Eq. (18). Then, it suffices to show that \((u^*, v^*)\) is optimal; i.e., \(f(u^*, v^*) = f(\tilde{u}^*, \tilde{v}^*)\). Since \(1_n^T r = 1_n^T c = 1\), we have \((u^*)^T r = (\tilde{u}^*)^T r\) and \((v^*)^T c = (\tilde{v}^*)^T c\).

Then, we proceed to establish the following bounds:
\[ \max_{1 \leq i \leq n} u_i^* - \min_{1 \leq i \leq n} u_i^* \leq \frac{\|C\|_\infty}{\eta} - \log \left( \frac{\min_{1 \leq i,j \leq n} \{r_i, c_j\}}{\eta} \right), \] (19)
\[ \max_{1 \leq i \leq n} v_i^* - \min_{1 \leq i \leq n} v_i^* \leq \frac{\|C\|_\infty}{\eta} - \log \left( \frac{\min_{1 \leq i,j \leq n} \{r_i, c_j\}}{\eta} \right). \] (20)
Indeed, for each \(1 \leq i \leq n\), we have
\[ e^{-\eta^{-1}\|C\|_\infty + u_i^*} \left( \sum_{j=1}^n e^{v_j^*} \right) \leq \sum_{j=1}^n e^{-\eta^{-1}C_{ij} + u_i^* + v_j^*} = [B(u^*, v^*) 1_n]_i = r_i \leq 1, \]
which implies \(u_i^* \leq \eta^{-1}\|C\|_\infty - \log(\sum_{j=1}^n e^{v_j^*})\). Furthermore, we have
\[ e^{u_i^*} \left( \sum_{j=1}^n e^{v_j^*} \right) \geq \sum_{j=1}^n e^{-\eta^{-1}C_{ij} + u_i^* + v_j^*} = [B(u^*, v^*) 1_n]_i = r_i \geq \min_{1 \leq i,j \leq n} \{r_i, c_j\}, \]
which implies \(u_i^* \geq \log(\min_{1 \leq i,j \leq n} \{r_i, c_j\}) - \log(\sum_{j=1}^n e^{v_j^*})\). Putting these pieces together yields Eq. (19). Using the similar argument, we can prove Eq. (20) holds true.

Finally, we prove our main results. Since \(\max_{1 \leq i \leq n} u_i^* \geq \min_{1 \leq i \leq n} u_i^*\), we derive from Eq. (19) that
\[ -\frac{\|C\|_\infty}{\eta} + \log \left( \frac{\min_{1 \leq i,j \leq n} \{r_i, c_j\}}{\eta} \right) \leq \min_{1 \leq i \leq n} u_i^* \leq \max_{1 \leq i \leq n} u_i^* \leq \frac{\|C\|_\infty}{\eta} - \log \left( \frac{\min_{1 \leq i,j \leq n} \{r_i, c_j\}}{\eta} \right). \]
This implies that \(\|u^*\|_\infty \leq R\). Then, we bound \(\|v^*\|_\infty\) by considering two different cases.
For the former case, we assume that $\max_{1 \leq i \leq n} v^*_{i} \geq 0$. Note that the optimality condition is $\sum_{i,j=1}^{n} e^{-\eta^{-1}C_{ij} + u^*_i + v^*_j} = 1$ and further implies that

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

Since $\max_{1 \leq i \leq n} u^*_{i} \geq 0$ and $\max_{1 \leq i \leq n} v^*_{i} \geq 0$, we have $0 \leq \max_{1 \leq i \leq n} v^*_{i} \leq \frac{\|C\|_{\infty}}{\eta}$. Combining $\max_{1 \leq i \leq n} v^*_{i} \geq 0$ with Eq. (20) yields that

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

which implies that $\|v^*\|_{\infty} \leq R$.

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

$$
\min_{1 \leq i \leq n} v^*_{i} \geq \log \left( \min_{1 \leq i,j \leq n} \{r_i, c_j\} \right) - \log \left( \sum_{i=1}^{n} e^{u^*_i} \right).
$$

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

Putting Lemma 3.1 and 3.2 together, we have the following straightforward consequence:

**Corollary 3.3.** Letting $\{(u^t, v^t)\}_{t \geq 0}$ be the iterates generated by Algorithm 1, we have

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

**Remark 3.4.** 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 a term $\log(n) - \log(\min_{1 \leq i,j \leq n}\{r_i, c_j\})$.

**Remark 3.5.** 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 $B(\overline{u}, \overline{v})\mathbf{1}_n = r$ or $B(\overline{u}, \overline{v})^\top \mathbf{1}_n = c$ where $(\overline{u}, \overline{v})$ 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 $f(u^t, v^t) - f(u^*, v^*)$ via appeal to the $\ell_{\infty}$-norm of the solution $(u^*, v^*)$.

### 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 $\varepsilon'$ in Algorithm 1. First, we start with a lower bound for the difference of function values between two consecutive iterates of Algorithm 1:

**Lemma 3.6.** 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
\[
\rho(r, r(B(u^t, v^t))) + \rho(c, c(B(u^t, v^t))) \geq \frac{1}{t} \left( \|r - r(B(u^t, v^t))\|_1^2 + \|c - c(B(u^t, v^t))\|_1^2 \right).
\]
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 3.7.** Letting \( \{(u^t, v^t)\}_{t \geq 0} \) be the iterates generated by Algorithm 1, the number of iterations required to satisfy \( E_t \leq \epsilon' \) is upper bounded by
\[
t \leq 2 + \frac{112nR}{\epsilon'} \quad \text{where} \quad R > 0 \quad \text{is defined in Lemma 3.2}.
\]

**Proof.** Letting \( \delta_t = f(u^t, v^t) - f(u^*, v^*) \), we derive from Corollary 3.3 and Lemma 3.6 that
\[
\delta_t - \delta_{t+1} \geq \max \left\{ \frac{\delta_t^2}{448nR^2} \frac{(\epsilon')^2}{28n} \right\},
\]
where \( E_t \geq \epsilon' \) 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
\[
\frac{\delta_{t+1}}{448nR^2} \leq \frac{1}{t + \frac{448nR^2}{\delta_1} \delta_t} \quad \Longrightarrow \quad t \leq 1 + \frac{448nR^2}{\delta_t} - \frac{448nR^2}{\delta_1}.
\]

(ii) Considering the process from the \((t + 1)\)-th iteration to the \((t + m)\)-th iteration for any \( m \geq 1 \), we have
\[
\delta_{t+m} \leq \delta_t - \frac{(\epsilon')^2 m}{28n} \quad \Longrightarrow \quad m \leq \frac{28n(\delta_t - \delta_{t+m})}{(\epsilon')^2}.
\]
We then minimize the sum of two estimates by an optimal choice of \( s \in (0, \delta_1) \):
\[
t \leq \min_{0 < s \leq \delta_1} \left( 2 + \frac{448nR^2}{s} - \frac{448nR^2}{\delta_1} + \frac{28ns}{(\epsilon')^2} \right) = \left\{ \begin{array}{ll}
2 + \frac{224nR^2}{\delta_1} - \frac{448nR^2}{\delta_1}, & \delta_1 \geq 4R\epsilon', \\
2 + \frac{28n\delta_1}{(\epsilon')^2}, & \delta_1 \leq 4R\epsilon'.
\end{array} \right.
\]
This implies that \( t \leq 2 + \frac{112nR}{\epsilon'} \) in both cases and completes the proof. □

Equipped with the result of Theorem 3.7 and the scheme of Algorithm 2, we are able to establish the following result for the complexity of Algorithm 2:

**Theorem 3.8.** The Greenkhorn for approximating optimal transport (Algorithm 2) returns an \( \epsilon \)-approximate transportation plan (cf. Definition 1) in
\[
O \left( \frac{n^2 \|C\|_2^2 \log(n)}{\epsilon^2} \right)
\]
arithmetic operations.
Proof. We follow the proof steps in [Altschuler et al., 2017, Theorem 1] and obtain that the transportation plan \( \tilde{X} \) returned by Algorithm 2 satisfies that

\[
\langle C, \tilde{X} \rangle - \langle C, X^* \rangle \leq 2\eta \log(n) + 4(\|\tilde{X}1_n - r\|_1 + \|\tilde{X}^T 1_n - c\|_1)\|C\|_{\infty}
\]

\[
\leq \frac{\varepsilon}{2} + 4(\|\tilde{X}1_n - r\|_1 + \|\tilde{X}^T 1_n - c\|_1)\|C\|_{\infty},
\]

where \( X^* \) is an optimal solution to the OT problem and \( \tilde{X} = \text{GREENKORN}(C, \eta, \tilde{r}, \tilde{c}, \frac{\varepsilon'}{2}) \).

The last inequality in the above display holds true since \( \eta = \frac{\varepsilon}{4 \log(n)} \). Furthermore,

\[
\|\tilde{X}1_n - r\|_1 + \|\tilde{X}^T 1_n - c\|_1 \leq \|\tilde{X}1_n - \tilde{r}\|_1 + \|\tilde{X}^T 1_n - \tilde{c}\|_1 + \|r - \tilde{r}\|_1 + \|c - \tilde{c}\|_1
\]

\[
\leq \frac{\varepsilon}{2} + \frac{\varepsilon'}{4} + \frac{\varepsilon'}{4} = \varepsilon'.
\]

Putting these pieces together with \( \varepsilon' = \frac{\varepsilon}{8\|C\|_{\infty}} \), yields that \( \langle C, \tilde{X} \rangle - \langle C, X^* \rangle \leq \varepsilon \).

The remaining step is to analyze the complexity bound. It follows from Theorem 3.7 and the definition of \( \tilde{r} \) and \( \tilde{c} \) in Algorithm 2 that

\[
t \leq 2 + \frac{112nR}{\varepsilon'} \leq 2 + \frac{96n\|C\|_{\infty}}{\varepsilon} \left( \frac{\|C\|_{\infty}}{\eta} + \log(n) - 2 \log \left( \min_{1 \leq i,j \leq n} \{r_i, c_j\} \right) \right)
\]

\[
\leq 2 + \frac{96n\|C\|_{\infty}}{\varepsilon} \left( \frac{4\|C\|_{\infty} \log(n)}{\varepsilon} + \log(n) - 2 \log \left( \frac{\varepsilon}{64n\|C\|_{\infty}} \right) \right)
\]

\[
= O \left( \frac{n\|C\|_{\infty}^2 \log(n)}{\varepsilon^2} \right).
\]

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

The complexity results presented in Theorem 3.8 improve the best known complexity bound \( \tilde{O}(n^2\varepsilon^{-3}) \) 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 the Greenkhorn algorithm can be lower bounded by a new quantity (cf. Lemmas 3.1 and 3.2). It allows us to apply the switching strategy in Theorem 3.7 to improve the complexity upper bound of the Greenkhorn algorithm.

In practice, the Greenkhorn algorithm has been reported to consistently outperform the Sinkhorn algorithm in terms of iteration number and our improved complexity result sheds light on the superior performance of Greenkhorn in practice.

4 Adaptive Primal-Dual Accelerated Mirror Descent

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

$$\min_{x \in Q} f(x), \quad \text{s.t. } Ax = b, \quad (21)$$

where $f$ is strongly convex with respect to the $\ell_1$-norm on the set $Q$:

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

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 $j \in \{i-n+l-1 : 1 \leq l \leq n\}$ and 0 otherwise. To be consistent with the notations in Algorithms 4 and 5, we specifically denote $A_{\alpha \ell}$ as the matrix $A$ corresponding to the entropic regularized OT problem.

After some calculations with the general problem (21), we obtain that the dual problem is as follows:

$$\min_{\lambda \in \mathbb{R}^{2n}} \bar{\varphi}(\lambda) := \{(\lambda, b) + \max_{x \in \mathbb{R}^{n^2}} \{-f(x) - \langle A^\top \lambda, x \rangle\}\}, \quad (22)$$

and $\nabla \bar{\varphi}(\lambda) = b - Ax(\lambda)$ where $x(\lambda) = \arg\max_{x \in \mathbb{R}^n} \{-f(x) - \langle A^\top \lambda, x \rangle\}$; see the explicit form in Eq. (9) with $\lambda = (\alpha, \beta)$. By Nesterov [2005, Theorem 1] with $\ell_1$-norm for the dual space of the Lagrange multipliers, the dual objective function $\bar{\varphi}$ satisfies the following inequality:

$$\bar{\varphi}(\lambda') - \bar{\varphi}(\lambda) - (\lambda' - \lambda)^\top \nabla \bar{\varphi}(\lambda) \leq \frac{\|A\|_{\ell_1 \rightarrow \ell_1}^2}{2\eta} \|\lambda' - \lambda\|_\infty^2. \quad (23)$$
Algorithm 4: Approximating OT by Algorithm 3

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

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

\textbf{Step 2:} Let \( A_{ot} \in \mathbb{R}^{2n \times n^2} \) and \( b \in \mathbb{R}^{2n} \) be defined by \( A_{ot} \text{vec}(X) = \left( X1_n, X^\top 1_n \right) \) and \( b = \left( \tilde{r} \tilde{c} \right) \).

\textbf{Step 3:} Compute \( \tilde{X} = \text{APDAMD}(\tilde{r}, A_{ot}, b, \frac{\varepsilon'}{2}) \) where \( \tilde{r} \) is defined by Eq. (22).

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

\textbf{Output:} \( \hat{X} \).

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 \( \|A_{ot}\|_{1 \rightarrow 1} \) is equal to maximum \( \ell_1 \)-norm of the column of this matrix, we have \( \|A_{ot}\|_{1 \rightarrow 1} = 2 \). Thus, the dual objective function \( \tilde{\varphi} \) is \( \frac{4}{\eta} \)-gradient Lipschitz with respect to the \( \ell_\infty \)-norm.

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

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

where the mirror mapping \( \phi \) is \( \frac{1}{\delta} \)-strongly convex and 1-smooth on \( \mathbb{R}^{2n} \) with respect to \( \ell_\infty \)-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 \( \phi(\lambda) = \frac{1}{2\delta n} \|\lambda\|^2 \) and \( B_\phi(\lambda', \lambda) = \frac{1}{2\delta n} \|\lambda' - \lambda\|^2 \) in the APDAMD algorithm where \( \delta = n \). As such, \( \delta > 0 \) is a function of \( n \) in general and it will appear in the complexity bound of the APDAMD algorithm for approximating the OT problem (cf. Theorem 4.5). 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 \( \tilde{O}(n^2\varepsilon^{-1}) \) using a more advanced area-convex mirror mapping [Sherman, 2017].

4.2 Properties of the APDAMD algorithm

We present several important properties of Algorithm 3 that can be used later for 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 4.1.** 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 \( M' \) by two will not stop until the line search stopping criterion is satisfied. Then, Eq. (23) implies that the number of line search iterations in the
line search strategy is finite and \( M^t \leq \frac{2\|A_{ot}\|_1^2}{\eta} \) holds true for all \( t \geq 0 \). Otherwise, the line search stopping criterion is satisfied with \( \frac{M^t}{2} \) since \( \frac{M^t}{2} \geq \frac{\|A_{ot}\|_1^2}{\eta} \).

Letting \( i_j \) denote the total number of multiplication at the \( j \)-th iteration, we have
\[
i_0 \leq 1 + \frac{\log(M^0)}{\log 2}, \quad i_j \leq 2 + \frac{\log(M^t)}{\log 2}.
\]

Then, the total number of line search iterations is bounded by
\[
\sum_{j=0}^t i_j \leq 1 + \frac{\log(M^0)}{\log 2} + \sum_{j=1}^t \left( 2 + \frac{\log(M^t)}{\log 2} \right) \leq 2t + 1 + \frac{\log(\frac{2\|A_{ot}\|_1^2}{\eta}) - \log(L^0)}{\log 2}.
\]

Since each line search contains two gradient oracle calls and \( \|A_{ot}\|_1 = 2 \), 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 \( \tilde{\varphi} \) in Algorithm 3.

**Lemma 4.2.** For each iteration \( t \) of Algorithm 3 and any \( z \in \mathbb{R}^{2n} \), we have
\[
\alpha^t \tilde{\varphi}(\lambda^t) \leq \sum_{j=0}^t (\alpha^j (\tilde{\varphi}(\mu^j) + (z - \mu^j)^\top \nabla \tilde{\varphi}(\mu^j)) + \|z\|^2_\infty.
\]

**Proof.** First, we claim that it holds for any \( z \in \mathbb{R}^n \):
\[
\alpha^{t+1}(z^t - z)^\top \nabla \tilde{\varphi}(\mu^{t+1}) \leq \alpha^{t+1}(\tilde{\varphi}(\mu^{t+1}) - \tilde{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}). \tag{24}
\]

Indeed, the optimality condition in mirror descent implies that, for any \( z \in \mathbb{R}^{2n} \), we have
\[
(z - z^{t+1})^\top \left( \nabla \tilde{\varphi}(\mu^{t+1}) + \frac{\nabla \phi(z^{t+1}) - \nabla \phi(z^t)}{\alpha^{t+1}} \right) \geq 0.
\]

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 \( B_\phi(z^{t+1}, z^{t+1}) \geq \frac{1}{2\eta} \|z^t - z^{t+1}\|^2_\infty \). Putting these pieces together yields that
\[
\begin{align*}
\alpha^{t+1}(z^t - z)^\top \nabla \tilde{\varphi}(\mu^{t+1}) &= \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \tilde{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^{t+1} - z)^\top \nabla \tilde{\varphi}(\mu^{t+1}) \\
&\leq \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \tilde{\varphi}(\mu^{t+1}) + (z - z^{t+1})^\top (\nabla \phi(z^{t+1}) - \nabla \phi(z^t)) \\
&= \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \tilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - B_\phi(z^{t+1}, z^t) \\
&\leq \alpha^{t+1}(z^t - z^{t+1})^\top \nabla \tilde{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - \frac{\|z^{t+1} - z^t\|^2_\infty}{2\eta}. \tag{25}
\end{align*}
\]

The update formulas of \( \mu^{t+1}, \lambda^{t+1}, \alpha^{t+1} \) and \( \bar{\alpha}^{t+1} \) imply that
\[
\lambda^{t+1} - \mu^{t+1} = \frac{\alpha^{t+1}}{\bar{\alpha}^{t+1}} (z^{t+1} - z^t), \quad \delta M^t (\alpha^{t+1})^2 = \bar{\alpha}^{t+1}.
\]

Therefore, we have
\[
\alpha^{t+1}(z^t - z^{t+1})^\top \nabla \tilde{\varphi}(\mu^{t+1}) = \bar{\alpha}^{t+1}(\mu^{t+1} - \lambda^{t+1})^\top \nabla \tilde{\varphi}(\mu^{t+1}),
\]
and
\[
\|z^{t+1} - z^t\|^2_\infty = (\frac{\bar{\alpha}^{t+1}}{\bar{\alpha}^{t+1}})^2 \|\mu^{t+1} - \lambda^{t+1}\|^2_\infty = \delta M^t \|\mu^{t+1} - \lambda^{t+1}\|^2_\infty.
\]
Putting these pieces together with Eq. (25) yields that
\[
\alpha^{t+1}(z^t - z)\nabla \bar{\varphi}(\mu^{t+1}) \\
\leq \alpha^{t+1}(\mu^{t+1} - \lambda^{t+1})\nabla \bar{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) - \frac{\alpha^{t+1}M_t}{2} \|\mu^{t+1} - \lambda^{t+1}\|_\infty^2 \\
= \alpha^{t+1}\left((\mu^{t+1} - \lambda^{t+1})\nabla \bar{\varphi}(\mu^{t+1}) - \frac{M_t}{2} \|\mu^{t+1} - \lambda^{t+1}\|_\infty^2\right) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) \\
\leq \alpha^{t+1}(\bar{\varphi}(\mu^{t+1}) - \bar{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}),
\]
where the last inequality comes from the stopping criterion in the line search. This implies that Eq. (24) holds true.

The next step is to bound the iterative objective gap given by
\[
\alpha^{t+1}\bar{\varphi}(\lambda^{t+1}) - \alpha^t\bar{\varphi}(\lambda^t) \\
\leq \alpha^{t+1}(\bar{\varphi}(\mu^{t+1}) + (z - \mu^{t+1})\nabla \bar{\varphi}(\mu^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}).
\]

Indeed, by combining \(\alpha^{t+1} = \alpha^t + \alpha^{t+1}\) and the update formula of \(\mu^{t+1}\), we have
\[
\alpha^{t+1}(\mu^{t+1} - z^t) = (\alpha^{t+1} - \alpha^t)\mu^{t+1} - \alpha^t\lambda^t - \alpha^t\mu^{t+1} - \alpha^{t+1}z^t = \alpha^t(\lambda^t - \mu^{t+1}).
\]

This together with the convexity of \(\bar{\varphi}\) implies that
\[
\alpha^{t+1}(\mu^{t+1} - z)\nabla \bar{\varphi}(\mu^{t+1}) \\
= \alpha^{t+1}(\mu^{t+1} - z^t)\nabla \bar{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^t - z)\nabla \bar{\varphi}(\mu^{t+1}) \\
= \alpha^t(\lambda^t - \mu^{t+1})\nabla \bar{\varphi}(\mu^{t+1}) + \alpha^{t+1}(z^t - z)\nabla \bar{\varphi}(\mu^{t+1}) \\
\leq \alpha^t(\bar{\varphi}(\lambda^t) - \bar{\varphi}(\mu^{t+1})) + \alpha^{t+1}(z^t - z)\nabla \bar{\varphi}(\mu^{t+1}).
\]

Furthermore, we derive from Eq. (24) and \(\alpha^{t+1} = \alpha^t + \alpha^{t+1}\) that
\[
\alpha^t(\bar{\varphi}(\lambda^t) - \bar{\varphi}(\mu^{t+1})) + \alpha^{t+1}(z^t - z)\nabla \bar{\varphi}(\mu^{t+1}) \\
\leq \alpha^t(\bar{\varphi}(\lambda^t) - \bar{\varphi}(\mu^{t+1})) + \alpha^{t+1}(\bar{\varphi}(\mu^{t+1}) - \bar{\varphi}(\lambda^{t+1})) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}) \\
= \alpha^t(\bar{\varphi}(\lambda^t) - \alpha^t\bar{\varphi}(\lambda^t)) + \alpha^{t+1}(\bar{\varphi}(\mu^{t+1}) + B_\phi(z, z^t) - B_\phi(z, z^{t+1}).
\]

Putting these pieces together yields that Eq. (26) holds true.

Finally, we prove our main results. By changing the index \(t\) to \(j\) in Eq. (26) and summing up the resulting inequality over \(j = 0, 1, \ldots, t - 1\), we have
\[
\alpha^t\bar{\varphi}(\lambda^t) - \alpha^0\bar{\varphi}(\lambda^0) \leq \sum_{j=0}^{t-1}(\alpha^{j+1}(\bar{\varphi}(\mu^{j+1}) + (z - \mu^{j+1})\nabla \bar{\varphi}(\mu^{j+1})) + B_\phi(z, z^j) - B_\phi(z, z^{j+1})).
\]

Since \(\alpha^0 = \alpha^0 = 0, B_\phi(z, z^j) \geq 0\) and \(\phi\) is \(1\)-smooth with respect to \(\ell_\infty\)-norm, we have
\[
\alpha^t\bar{\varphi}(\lambda^t) \leq \sum_{j=0}^{t}(\alpha^j(\bar{\varphi}(\mu^j) + (z - \mu^j)\nabla \bar{\varphi}(\mu^j))) + B_\phi(z, z^0) \\
\leq \sum_{j=0}^{t}(\alpha^j(\bar{\varphi}(\mu^j) + (z - \mu^j)\nabla \bar{\varphi}(\mu^j))) + \|z - z^0\|_\infty^2 \\
= \sum_{j=0}^{t}(\alpha^j(\bar{\varphi}(\mu^j) + (z - \mu^j)\nabla \bar{\varphi}(\mu^j))) + \|z\|_\infty^2.
\]

This completes the proof. \(\square\)

The final lemma provides us with a key lower bound for the accumulating parameter.
Lemma 4.3. 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 $\bar{\alpha}^1 = \alpha^1 = \frac{1}{M^1} \geq \frac{\eta}{8\delta}$ since $M^1 \leq \frac{8}{\eta}$ was proven in Lemma 4.1. 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

$$
\bar{\alpha}^{t+1} = \bar{\alpha}^t + \alpha^{t+1} = \bar{\alpha}^t + \frac{1}{2\delta M^t} + \sqrt{\frac{1}{4(\delta M^t)^2} + \frac{\bar{\alpha}^t}{\delta M^t}}
$$

$$
\geq \bar{\alpha}^t + \frac{1}{2\delta M^t} + \sqrt{\frac{\bar{\alpha}^t}{\delta M^t}}
$$

$$
\geq \bar{\alpha}^t + \frac{\eta}{16\delta} + \sqrt{\frac{\eta\bar{\alpha}^t}{8\delta}},
$$

where the last inequality comes from $M^t \leq \frac{8}{\eta}$ as shown in Lemma 4.1. 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. \qed

4.3 Complexity analysis for the APDAMD algorithm

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

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

The above dual form was also considered to establish the complexity bound of APDAGD algorithm [Dvurechensky et al., 2018]. Since $(\alpha, \beta)$ can be obtain by $\alpha_i = \eta u_i$ and $\beta_j = \eta v_j$, we derive from Lemma 2.2 that

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

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 $\varepsilon'$:

Theorem 4.4. Letting $\{X^t\}_{t \geq 0}$ be the iterates generated by Algorithm 3, the number of iterations required to satisfy $\|A_0 \text{vec}(X^t) - b\|_1 \leq \varepsilon'$ is upper bounded by

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

where $R > 0$ is defined in Lemma 2.2.
Proof. From Lemma 4.2, we have
\[
\bar{\alpha}^t \bar{\varphi}(\bar{\lambda}^t) \leq \min_{z \in B_{\infty}(2\eta R)} \left\{ \sum_{j=0}^{t} (\alpha^j (\bar{\varphi}(\mu^j) + (z - \mu^j)^{\top} \nabla \bar{\varphi}(\mu^j))) + \|z\|_\infty^2 \right\},
\]
where \(B_{\infty}(r) := \{ \lambda \in \mathbb{R}^n \mid \|\lambda\|_\infty \leq r \} \). This implies that
\[
\bar{\alpha}^t \bar{\varphi}(\bar{\lambda}^t) \leq \min_{z \in B_{\infty}(2\eta R)} \left\{ \sum_{j=0}^{t} (\alpha^j (\bar{\varphi}(\mu^j) + (z - \mu^j)^{\top} \nabla \bar{\varphi}(\mu^j))) \right\} + 4\eta^2 R^2.
\]
Since \(\bar{\varphi}\) is the objective function of dual regularized OT problem, we have
\[
\bar{\varphi}(\mu^j) + (z - \mu^j)^{\top} \nabla \bar{\varphi}(\mu^j) = -f(x(\mu^j)) + z^\top (b - A_{ot}x(\mu^j)).
\]
Therefore, we conclude that
\[
\bar{\alpha}^t \bar{\varphi}(\bar{\lambda}^t) \leq \min_{z \in B_{\infty}(2\eta R)} \left\{ \sum_{j=0}^{t} (\alpha^j (\bar{\varphi}(\mu^j) + (z - \mu^j)^{\top} \nabla \bar{\varphi}(\mu^j))) \right\} + 4\eta^2 R^2
\]
\[
\leq 4\eta^2 R^2 - \bar{\alpha}^t f(x^t) + \min_{z \in B_{\infty}(2\eta R)} \{ \bar{\alpha}^t z^\top (b - A_{ot}x^t) \}
\]
\[
= 4\eta^2 R^2 - \bar{\alpha}^t f(x^t) - 2\bar{\alpha}^t \eta R\|A_{ot}x^t - b\|_1,
\]
where the second inequality comes from the convexity of \(f\) and the last equality comes from the fact that \(\ell_1\)-norm is the dual norm of \(\ell_\infty\)-norm. That is to say,
\[
f(x^t) + \bar{\varphi}(\lambda^t) + 2\eta R\|A_{ot}x^t - b\|_1 \leq \frac{4\eta^2 R^2}{\bar{\alpha}^t}.
\]
Suppose that \(\lambda^*\) is an optimal solution to dual regularized OT problem satisfying \(\|\lambda^*\|_\infty \leq \eta R\), we have
\[
f(x^t) + \bar{\varphi}(\lambda^t) \geq f(x^t) + \bar{\varphi}(\lambda^*) = f(x^t) + b^\top \lambda^* + \max_{x \in \mathbb{R}^n} \left\{ -f(x) - (\lambda^*)^\top A_{ot}x \right\}
\]
\[
\geq f(x^t) + b^\top \lambda^* - f(x^t) - (\lambda^*)^\top A_{ot}x^t = (b - A_{ot}x^t)^\lambda^*
\]
\[
\geq -\eta R\|A_{ot}x^t - b\|_1,
\]
Therefore, we conclude that
\[
\|A_{ot}x^t - b\|_1 \leq \frac{4\eta R}{\bar{\alpha}^t} \leq \frac{128\delta R}{(t + 1)^2}.
\]
This completes the proof. \(\square\)

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

**Theorem 4.5.** The APDAMD algorithm for approximating optimal transport (Algorithm 4) returns an \(\varepsilon\)-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 3.8, we have

\[ \langle C, \hat{X} \rangle - \langle C, X^* \rangle \leq \frac{\varepsilon}{2} + 4(\|\hat{X}1_{n} - r\|_1 + \|\hat{X}^\top 1_{n} - c\|_1\|C\|_\infty), \]

where \( \hat{X} \) is returned by Algorithm 4, \( X^* \) is a solution to the OT problem and \( \hat{X} = \text{APDAMD}(\hat{\varphi}, A, b, \frac{\varepsilon'}{2}) \). Since \( \|\hat{X}1_{n} - r\|_1 + \|\hat{X}^\top 1_{n} - c\|_1 \leq \varepsilon' \) and \( \varepsilon' = \frac{\varepsilon}{8\|C\|_\infty} \), we have \( \langle C, \hat{X} \rangle - \langle C, X^* \rangle \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \).

The remaining step is to analyze the complexity bound. If follows from Lemma 4.1 and Theorem 4.4 that

\[
N_t \leq 4t + 4 + \frac{2 \log(\frac{\varepsilon}{\eta}) - 2 \log(L^0)}{\log 2} + \log(2) + \frac{2 \log(\frac{\varepsilon}{\eta}) - 2 \log(L^0)}{\log 2} = 8 + 256 \left[ \frac{\delta R\|C\|_\infty \log(n)}{\varepsilon} + \frac{2 \log(\frac{32 \log(n)}{\varepsilon}) - 2 \log(L^0)}{\log 2} \right].
\]

Combining the definition of \( R \) in Lemma 2.2 with the definition of \( \eta, \hat{r} \) and \( \hat{c} \) 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

\[
N_t \leq 256 \left[ \frac{\delta \|C\|_\infty \log(n)}{\varepsilon} \frac{4\|C\|_\infty \log(n)}{\varepsilon} + \log(n) - 2 \log \left( \frac{\varepsilon}{64n\|C\|_\infty} \right) \right] + 8 = O \left( \frac{\sqrt{\delta} \|C\|_\infty \log(n)}{\varepsilon} \right).
\]

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, \( \hat{r} \) and \( \hat{c} \) 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 \( O(n^2\sqrt{\delta} \|C\|_\infty \log(n)\varepsilon^{-1}) \). \( \square \)

The complexity results in Theorem 4.5 suggests an interesting feature of the (regularized) OT problem. Indeed, the dependence of that bound on \( \delta \) manifests the necessity of \( \ell_\infty \)-norm in the understanding of the complexity of the 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 \( \ell_\infty \)-norm of optimal solutions of the dual regularized OT problem to measure the progress in the objective value among the successive iterates.

4.4 Revisiting the APDAGD algorithm

We revisit the APDAGD algorithm [Dvurechensky et al., 2018] for the regularized OT problem. First, we point out that the current complexity bound of \( \tilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\}) \) is not valid by a simple counterexample. Then, we establish a new complexity bound of the APDAGD algorithm using our techniques in Section 4.3. Despite the issue with regularized
Algorithm 5: Approximating OT by Dvurechensky et al. [2018, Algorithm 3]

Input: \( \eta = \frac{1}{4 \log(n)} \) and \( \varepsilon' = \frac{\varepsilon}{8\|C\|_{\infty}} \).

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

Step 2: Let \( A_{\text{ot}} \in \mathbb{R}^{2n \times n^2} \) and \( b \in \mathbb{R}^{2n} \) be defined by \( A_{\text{ot}} \text{vec}(X) = (X1_n, X^\top 1_n) \) and \( b = (\tilde{r}, \tilde{c}) \).

Step 3: Compute \( \tilde{X} = \text{APDAGD}(\tilde{\phi}, A_{\text{ot}}, b, \varepsilon') \) where \( \tilde{\phi} \) is defined by Eq. (22).

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

OT, we wish to emphasize that the APDAGD algorithm 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 regularized OT since no dual solution exists with a constant bound in \( \ell_2 \)-norm. However, it can be used for analyzing other problems with bounded optimal dual solution.

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

\[
\min_{\alpha, \beta \in \mathbb{R}^n} \tilde{\phi}(\alpha, \beta) := \eta \left( \sum_{i,j=1}^{n} e^{-\frac{C_{ij} - \alpha_i - \beta_j}{\eta}} - 1 \right) - \alpha^\top r - \beta^\top c. \tag{27}
\]

Theorem 4.6 (Theorem 4 in Dvurechensky et al. [2018]). Let \( \overline{R} > 0 \) be defined such that there exists an optimal solution to the dual regularized OT problem in Eq. (22), denoted by \((u^*, v^*)\), satisfying \( \|(u^*, v^*)\| \leq \overline{R} \), the APDAGD algorithm for approximating optimal transport (cf. Algorithm 5) returns an \( \varepsilon \)-approximate transportation plan (cf. Definition 1) in

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

arithmetic operations.

From the above theorem, Dvurechensky et al. [2018] claims that the complexity bound for the APDAGD algorithm is \( O(\min\{n^{9/4} \varepsilon^{-1}, n^2 \varepsilon^{-2}\}) \). 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 4.7.

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 2.2 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 2.3 and Lemma 3.2 imply that a straightforward upper bound for \( \overline{R} \) is \( O(\sqrt{n}) \). Given a tolerance \( \varepsilon \in (0, 1) \), we further show that \( \overline{R} \) is indeed \( \Omega(\sqrt{n}) \) by using a specific regularized OT problem as follows.
Proposition 4.7. Suppose that $C = I_n I_n^\top$ and $r = c = \frac{1}{n} I_n$. Given a tolerance $\varepsilon \in (0, 1)$ and the regularization term $\eta = \frac{\varepsilon}{4 \log(n)}$, all the optimal solutions of the dual regularized OT problem in Eq. (27) satisfy that $\| (\alpha^*, \beta^*) \| \gtrsim \sqrt{n}$.

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

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

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

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

This implies $\alpha_i^* = \alpha_j^*$ and $\beta_i^* = \beta_j^*$ for all $i, j \in [n]$. Suppose that $A = e^{-\frac{4 \log(n)(1 - \beta_j^*)}{\varepsilon}}$ and $B = e^{-\frac{4 \log(n)(1 - \alpha_i^*)}{\varepsilon}}$, we have $ABe^{-\frac{4 \log(n)}{\varepsilon}} = \frac{\varepsilon}{n^2}$ and hence $AB = \frac{\varepsilon}{n^2}$. Putting these pieces together yields that

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

Therefore, we conclude that

$$\| (\alpha^*, \beta^*) \| \gtrsim \sqrt{n}.$$ 

As a consequence, we achieve the conclusion of the proposition. \hfill \Box

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 $\hat{\varphi}$ defined in Eq. (9). Given the corrected upper bound $\bar{F}$ and Algorithm 5 for approximating OT, we provide a new complexity bound of Algorithm 5 in the following proposition.

Proposition 4.8. The APDAGD algorithm for approximating optimal transport (Algorithm 5) returns an $\varepsilon$-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 $\varepsilon$ is

$$t \leq O \left( \max \left\{ \min \left\{ \frac{n^{1/4} \sqrt{\bar{F}} \| C \|_\infty \log(n)}{\varepsilon}, \frac{\bar{F} \| C \|_\infty \log(n)}{\varepsilon^2}, \frac{\bar{F} \sqrt{\log n}}{\varepsilon} \right\} \right\} \right). \quad (28)$$
Moreover, we have $R \leq \sqrt{n\eta}R$ 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 the APDAGD algorithm 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 $\hat{r}$ and $\hat{c}$ 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 $O(n^{5/2}\|C\|_{\infty}\sqrt{\log(n)}\varepsilon^{-1})$. \hfill \Box

Remark 4.9. As indicated in Proposition 4.8, the corrected complexity bound of the APDAGD algorithm for the regularized OT is similar to that of the APDAMD algorithm when we choose $\phi(x) = \frac{1}{2n}\|x\|^{2}$ and have $\delta = \eta$. From this perspective, our algorithm can be viewed as a generalization of the APDAGD algorithm. Since our algorithm utilizes $\ell_{\infty}$-norm in the line search criterion, it is more robust than the APDAGD algorithm in practice; see the experimental section for the details.

5 Accelerating Sinkhorn

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

5.1 Algorithmic procedure

We present the pseudocode of the 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 noting 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 $\varphi$ in a fast rate. The coordinate update guarantees that $\varphi(\tilde{u}^{t},\tilde{v}^{t}) \leq \varphi(u^{t},v^{t})$ and $\|B(\tilde{u}^{t},\tilde{v}^{t})\|_{1} = 1$. The monotone search step guarantees that $\varphi(u',v') \leq \varphi(\tilde{u}^{t},\tilde{v}^{t})$. The greedy coordinate update guarantees that $\varphi(\tilde{u}^{t+1},\tilde{v}^{t+1}) \leq \varphi(u^{t},v^{t})$ with sufficient progress.

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

$$E_{t} = \|r(B(u^{t},v^{t})) - r\|_{1} + \|c(B(u^{t},v^{t})) - c\|_{1}.$$ (29)
We first present two technical lemmas which are essential in the analysis of Algorithm 6. To facilitate the discussion, we recall Eq. (5) that satisfies Lemma 2.3. To facilitate the discussion, we recall Eq. (10) with $\|A\|_{1\to2} = \sqrt{2}$.
as follows,
\[
\varphi(u', v') - \varphi(u, v) - (u' - u) \top \nabla \varphi(u, v) \leq \left\| (u' - u) \right\|^2, \tag{31}
\]
which will be used in the proof of the first lemma.

**Lemma 5.1.** Let \( \{\tilde{u}^t, \tilde{v}^t\}_{t \geq 0} \) be the iterates generated by Algorithm 6 and \((u^*, v^*)\) be an optimal solution of the dual entropic regularized OT problem. Then, we have
\[
\delta_{t+1} \leq (1 - \theta_t) \delta_t + \theta_t^2 \left( \left\| (u^* - \tilde{u}^t) \right\| \right)^2 - \left\| (u^* - \tilde{u}^{t+1}) \right\|^2.
\]

**Proof.** Using Eq. (31) with \((u', v') = (\tilde{u}^t, \tilde{v}^t)\) and \((u, v) = (\tilde{u}^t, \tilde{v}^t)\), we have
\[
\varphi(\tilde{u}^t, \tilde{v}^t) \leq \varphi(\tilde{u}^t, \tilde{v}^t) + \theta_t \left( \tilde{u}^{t+1} - \tilde{u}^t \right) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t) + \theta_t^2 \left\| (\tilde{u}^{t+1} - \tilde{u}^t) \right\|^2.
\]
After simple calculations, we find that
\[
\varphi(\tilde{u}^t, \tilde{v}^t) = (1 - \theta_t) \varphi(\tilde{u}^t, \tilde{v}^t) + \theta_t \varphi(\tilde{u}^t, \tilde{v}^t),
\]
\[
(\tilde{u}^{t+1} - \tilde{u}^t) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t) = - (\tilde{u}^t - \tilde{u}^t) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t) + (\tilde{v}^{t+1} - \tilde{v}^t) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t).
\]
Putting these pieces together yields that
\[
\varphi(\tilde{u}^t, \tilde{v}^t) \leq \theta_t \left( \varphi(\tilde{u}^t, \tilde{v}^t) + (\tilde{u}^{t+1} - \tilde{u}^t) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t) + \theta_t \left\| (\tilde{u}^{t+1} - \tilde{u}^t) \right\|^2 \right) + (1 - \theta_t) \varphi(\tilde{u}^t, \tilde{v}^t) - \theta_t (\tilde{u}^t - \tilde{u}^t) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t).
\]

We first bound the term \(I\). Indeed, by the update formula for \((\tilde{u}^{t+1}, \tilde{v}^{t+1})\) and the definition of \(\nabla \varphi\), we have
\[
(\tilde{u}^t - \tilde{u}^{t+1}) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t) + 2 \theta_t (\tilde{u}^{t+1} - \tilde{u}^t) \top \nabla \varphi(\tilde{u}^t, \tilde{v}^t) = 0 \text{ for all } (u, v) \in \mathbb{R}^{2n}.
\]
Letting \((u, v) = (u^*, v^*)\) and rearranging the resulting equation yields that

\[
\begin{align*}
\left(\frac{\ddot{u}^{t+1} - \ddot{u}^t}{\dot{v}^{t+1} - \dot{v}^t}\right)^\top \nabla \varphi(\ddot{u}^t, \ddot{v}^t) &= \left(\frac{u^* - \ddot{u}^t}{v^* - \ddot{v}^t}\right)^\top \nabla \varphi(\ddot{u}^t, \ddot{v}^t) \\
+ \theta_t \left(\left\|\frac{u^* - \ddot{u}^t}{v^* - \ddot{v}^t}\right\|^2 - \left\|\frac{u^* - \ddot{u}^{t+1}}{v^* - \ddot{v}^{t+1}}\right\|^2 - \left\|\frac{\ddot{u}^{t+1} - \ddot{u}^t}{\ddot{v}^{t+1} - \ddot{v}^t}\right\|^2 \right).
\end{align*}
\]

Using the convexity of \(\varphi\), we have

\[
\left(\frac{u^* - \ddot{u}^t}{v^* - \ddot{v}^t}\right)^\top \nabla \varphi(\ddot{u}^t, \ddot{v}^t) \leq \varphi(u^*, v^*) - \varphi(\ddot{u}^t, \ddot{v}^t).
\]

Putting these pieces together yields that

\[
\mathbf{I} \leq \varphi(u^*, v^*) + \theta_t \left(\left\|\frac{u^* - \ddot{u}^t}{v^* - \ddot{v}^t}\right\|^2 - \left\|\frac{u^* - \ddot{u}^{t+1}}{v^* - \ddot{v}^{t+1}}\right\|^2 \right). \tag{33}
\]

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

\[
-\theta_t \left(\frac{\ddot{u}^t - \ddot{u}^t}{\ddot{v}^t - \ddot{v}^t}\right) = \theta_t \left(\frac{\ddot{u}^t}{\ddot{v}^t}\right) + (1 - \theta_t) \left(\frac{\ddot{u}^t}{\ddot{v}^t}\right) - \left(\frac{\ddot{u}^t}{\ddot{v}^t}\right) = (1 - \theta_t) \left(\frac{\ddot{u}^t}{\ddot{v}^t} - \ddot{u}^t\right).
\]

Combining the above equation with the convexity of \(\varphi\), we have

\[
\mathbf{II} = (1 - \theta_t) \left(\varphi(\ddot{u}^t, \ddot{v}^t) + \left(\frac{\ddot{u}^t - \ddot{u}^t}{\ddot{v}^t - \ddot{v}^t}\right)^\top \nabla \varphi(\ddot{u}^t, \ddot{v}^t) \right) \leq (1 - \theta_t) \varphi(\ddot{u}^t, \ddot{v}^t). \tag{34}
\]

Plugging Eq. (33) and Eq. (34) into Eq. (32) yields that

\[
\varphi(\ddot{u}^t, \ddot{v}^t) \leq (1 - \theta_t) \varphi(\ddot{u}^t, \ddot{v}^t) + \theta_t \varphi(u^*, v^*) + \theta_t^2 \left(\left\|\frac{u^* - \ddot{u}^t}{v^* - \ddot{v}^t}\right\|^2 - \left\|\frac{u^* - \ddot{u}^{t+1}}{v^* - \ddot{v}^{t+1}}\right\|^2 \right).
\]

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

\[
\varphi(\ddot{u}^{t+1}, \ddot{v}^{t+1}) - \varphi(u^*, v^*) \leq (1 - \theta_t)(\varphi(\ddot{u}^t, \ddot{v}^t) - \varphi(u^*, v^*)) + \theta_t^2 \left(\left\|\frac{u^* - \ddot{u}^t}{v^* - \ddot{v}^t}\right\|^2 - \left\|\frac{u^* - \ddot{u}^{t+1}}{v^* - \ddot{v}^{t+1}}\right\|^2 \right).
\]

This completes the proof. \(\square\)

The second lemma provides an upper bound for \(\delta_t\) defined by Eq. (30) where \(\{(\ddot{u}^t, \ddot{v}^t)\}_{t \geq 0}\) are generated by Algorithm 6 and \((u^*, v^*)\) is an optimal solution defined by Corollary 2.3.

**Lemma 5.2.** Let \(\{(\ddot{u}^t, \ddot{v}^t)\}_{t \geq 0}\) be the iterates generated by Algorithm 6 and \((u^*, v^*)\) be an optimal solution of the dual entropic regularized OT problem satisfying that \(\|u^*, v^*\| \leq \sqrt{2nR}\) where \(R\) is defined in Corollary 2.3. Then, we have

\[
\delta_t \leq \frac{8n R^2}{(t + 1)^2}.
\]
Proof. By simple calculations, we derive from the definition of \( \theta_t \) that \( \frac{\theta_{t+1}}{\theta_t} = \sqrt{1 - \theta_t} \). Therefore, we conclude from Lemma 5.1 that

\[
\left(1 - \frac{\theta_{t+1}}{\theta_t^2}\right) \delta_{t+1} - \left(1 - \frac{\theta_t}{\theta_t^2}\right) \delta_t \leq \left\| \left( u^* - \tilde{u}^{t+1} \right) \right\|_2^2 - \left\| \left( u^* - \tilde{u}^{t+1} \right) \right\|_2^2.
\]

Equivalently, we have

\[
\left(1 - \frac{\theta_t}{\theta_t^2}\right) \delta_t + \left\| \left( u^* - \tilde{u}^t \right) \right\|_2^2 \leq \left(1 - \frac{\theta_0}{\theta_0^2}\right) \delta_0 + \left\| \left( u^* - \tilde{u}^0 \right) \right\|_2^2.
\]

Since \( \theta_0 = 1 \) and \( \tilde{u}^0 = v^0 = 0_n \), we have \( \delta_t \leq \theta_0^2 t \left(\|u^*\|_2^2\right) \leq 2nR^2 \theta_0^2 t^{-1} \).

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

\[
\theta_{t_0+1} = \frac{2}{1 + \sqrt{1 + \frac{2}{t_0}}} \leq 2 \frac{t_0}{t_0 + 3}.
\]

Putting these pieces together yields the desired inequality for \( \delta_t \). \( \square \)

### 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 \( \rho : \mathbb{R}_+^n \times \mathbb{R}_+^n \to \mathbb{R}_+ \) given by: \( \rho(a,b) := 1^n (b - a) + \sum_{i=1}^n a_i \log \left( \frac{a_i}{b_i} \right) \).

**Theorem 5.3.** 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 2.2.

**Proof.** We first claim that

\[
\varphi(u^t, v^t) - \varphi(\tilde{u}^{t+1}, \tilde{v}^{t+1}) \geq \frac{1}{2} \left( \|r(B(u^t, v^t)) - r\|_1^2 + \|c(B(u^t, v^t)) - c\|_1^2 \right). \tag{35}
\]

By the definition of \( \varphi \), we have

\[
\varphi(u^t, v^t) - \varphi(\tilde{u}^{t+1}, \tilde{v}^{t+1}) = \|B(u^t, v^t)\|_1 - \|B(\tilde{u}^{t+1}, \tilde{v}^{t+1})\|_1 - (u^t - \tilde{u}^{t+1})^\top r - (v^t - \tilde{v}^{t+1})^\top c.
\]

From the update formula for \( (\tilde{u}^t, \tilde{v}^t) \) and \( (\tilde{u}^{t+1}, \tilde{v}^{t+1}) \), it is clear that \( \|B(\tilde{u}^t, \tilde{v}^t)\|_1 = 1 \) and \( \|B(\tilde{u}^{t+1}, \tilde{v}^{t+1})\|_1 = 1 \) 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

\[
\varphi(u^t, v^t) - \varphi(\tilde{u}^{t+1}, \tilde{v}^{t+1}) = -(u^t - \tilde{u}^{t+1})^\top r - (v^t - \tilde{v}^{t+1})^\top c = (\log(r) - \log(r(B(u^t, v^t))))^\top r + (\log(c) - \log(c(B(u^t, v^t))))^\top c.
\]

28
Since $1_n^T r = 1_n^T r(B(u^t, v^t)) = 1_n^T c = 1_n^T c(B(u^t, v^t)) = 1$, we have

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

Using Altschuler et al. [2017, Lemma 4], we derive Eq. (35) as desired.

By the definition of $(u^t, v^t)$, we have $\varphi(\tilde{u}^t, \tilde{v}^t) \geq \varphi(u^t, v^t)$. Plugging this inequality into Eq. (35) together with the Cauchy-Schwarz inequality yields

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

Therefore, we conclude that

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

Since $\varphi(\tilde{u}^{t+1}, \tilde{v}^{t+1}) \geq \varphi(u^*, v^*)$ for all $t \geq 1$, we have $\varphi(\tilde{u}^j, \tilde{v}^j) - \varphi(\tilde{u}^{t+1}, \tilde{v}^{t+1}) \leq \delta_j$. Then, it follows from Lemma 5.2 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

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

This completes the proof of the theorem. $\square$

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 5.4.** The accelerated Sinkhorn for approximating optimal transport (Algorithm 7) returns an $\varepsilon$-approximate transportation plan (cf. Definition 1) in

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

arithmetic operations.

**Proof.** Applying the same argument which is used in Theorem 3.8, we obtain that $\langle C, \tilde{X} \rangle - \langle C, X^* \rangle \leq \varepsilon$ 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 5.3, we have

$$t \leq 1 + \left( \frac{32\sqrt{nR}}{\varepsilon'} \right)^{2/3}.$$
By the definition of $R$ (cf. Lemma 2.2), $\eta = \frac{\varepsilon}{4 \log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8 \|C\|_{\infty}}$, we have

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

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 $O(n^{7/3}\|C\|_{\infty}^{1/3}(\log(n))^{1/3} \varepsilon^{-4/3})$. 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 $O(n^{7/3}\|C\|_{\infty}^{1/3}(\log(n))^{1/3} \varepsilon^{-4/3})$. \qed

Remark 5.5. Theorem 5.4 shows that the complexity bound of the accelerated Sinkhorn algorithm is better than that of the Sinkhorn and Greenkhorn algorithms 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, the Sinkhorn and Greenkhorn algorithms 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 with the Greenkhorn, the accelerated Sinkhorn and the APDAMD algorithms on synthetic data and real images from the MNIST Digits dataset\footnote{http://yann.lecun.com/exdb/mnist/}. We use the Sinkhorn [Cuturi, 2013], the APDAGD [Dvurechensky et al., 2018] and the GCPB algorithms [Genevay et al., 2016] as the baseline approaches. For the GCPB algorithm, we use stochastic averaged gradient (SAG) on the dual entropic regularized OT problem [Schmidt et al., 2017]. Early works suggest that the Greenkhorn and the APDAGD algorithms outperform the Sinkhorn algorithm in terms of iteration numbers [Altschuler et al., 2017, Dvurechensky et al., 2018]. We repeat these comparisons for completeness. To obtain an optimal value of the unregularized OT problem, we employ the default linear programming 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 $\ell_1$ 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 $\ell_1$ 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
Figure 2: Comparative performance of Sinkhorn v.s. Greenkhorn, APDAGD v.s. APDAMD and Sinkhorn v.s. accelerated Sinkhorn on the MNIST real images.

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 the Sinkhorn and Greenkhorn algorithms in terms of the iteration counts. In the leftmost image, the comparison uses distance to transportation polytope $d(A)$ where $A$ is either the Sinkhorn or Greenkhorn algorithm. In the middle image, the maximum, median and minimum values of the competitive ratios $\log(d(A_1)/d(A_2))$ on 10 images are utilized for the comparison where $A_1$ is the Sinkhorn algorithm and $A_2$ is the Greenkhorn algorithm. 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 observe that (i) the Greenkhorn algorithm outperforms the Sinkhorn algorithm in terms of iteration counts, illustrating the improvement from greedy coordinate descent; (ii) the APDAMD algorithm with $\delta = n$ and $\phi = (1/2n)\|\cdot\|^2$ is not faster but more robust than the APDAGD algorithm, illustrating the advantage of using mirror descent and line search with $\|\cdot\|_\infty$; (iii) the accelerated Sinkhorn algorithm outperforms the Sinkhorn algorithm in
Figure 3: Performance of the GCPB, APDAGD and APDAMD algorithms in terms of time on the MNIST real images. These three images specify the values of regularized OT with varying regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$, showing that the APDAMD algorithm is faster and more robust than the APDAGD and GCPB algorithms.

terms of iteration counts, 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 such that their sum becomes one. The maximum number of iterations is $T = 5$.

Figures 2 and 3 summarize the experimental results on MNIST. In the first row of Figure 2, we compare the Sinkhorn and Greenkhorn algorithms in terms of iteration counts. The leftmost image specifies the distances $d(A)$ to the transportation polytope for the algorithm $A$, which is either the Sinkhorn or Greenkhorn algorithm; the middle image specifies the maximum, median and minimum of competitive ratios $\log(d(A_1)/d(A_2))$ on ten random pairs of MNIST images, where $A_1$ and $A_2$ respectively correspond to the Sinkhorn and Greenkhorn algorithms; the rightmost image specifies the values of the regularized OT problem with varying regularization parameter $\eta \in \{1, \frac{1}{5}, \frac{1}{9}\}$. 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) the performance of the accelerated Sinkhorn algorithm deteriorates but remains better than that of the Sinkhorn algorithm; (iii) the APDAMD algorithm is faster and more robust than the APDAGD and GCPB algorithms. We believe there is much room for improving the accelerated Sinkhorn algorithm with parameter fine-tuning and leave it to future work.

7 Conclusion

In this paper, we first show that the complexity bound of the Greenkhorn algorithm can be improved to $\tilde{O}(n^2\varepsilon^{-2})$, which matches the best known bound of the Sinkhorn algorithm. Then, we generalize the APDAGD algorithm with mirror mapping $\phi$ and show that the resulting
APDAMD algorithm achieves the complexity bound of $\tilde{O}(n^2\sqrt{\delta}\varepsilon^{-1})$ where $\delta > 0$ refers to the regularity of $\phi$. We demonstrate that the complexity bound of $\tilde{O}(\min\{n^{9/4}\varepsilon^{-1}, n^2\varepsilon^{-2}\})$ is invalid for the APDAGD algorithm and establish a new complexity bound of $\tilde{O}(n^{5/2}\varepsilon^{-1})$. Moreover, we propose a deterministic accelerated Sinkhorn algorithm and prove that it achieves the complexity bound of $\tilde{O}(n^{7/3}\varepsilon^{-4/3})$ by incorporating an estimate sequence. Therefore, the new accelerated algorithm outperforms the Sinkhorn and Greenkhorn algorithms in terms of $1/\varepsilon$ and the APDAGD and accelerated alternating minimization [Guminov et al., 2021] algorithms in terms of $n$. Experiments on synthetic data and real images demonstrate the efficiency of the algorithms considered in this paper.

There are a few promising future directions arising from this work. First, it is important to investigate fast algorithms to compute dimension-reduced versions of OT distances. In particular, the sample complexity of OT distance can grow exponentially in dimension $d$ of the probability measures [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 distance between them. This result can be mitigated when data live on lower dimensional manifolds as shown in Weed and Bach [2019] and Paty and Cuturi [2019], but sample complexity bounds still remain pessimistic even in that case. This motivates recent works on efficient dimension-reduced Wasserstein distances, e.g., the sliced Wasserstein distance [Bonneel et al., 2015], generalized sliced Wasserstein distance [Kolouri et al., 2019], distributional sliced Wasserstein distance [Nguyen et al., 2021], further inspiring us to explore the application of our algorithms to carry out the computations of these distances 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] in which 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 probability measures. Similar to OT, an important practical question with robust OT is computational. As such, developing efficient algorithms with near-optimal complexity for approximately solving robust OT is our focus in the future work.

8 Acknowledgments

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.

References

B. K. Abid and R. M. Gower. Greedy stochastic algorithms for entropy-regularized optimal transport problems. In AISTATS, 2018. (Cited on pages 2 and 14.)

Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson. Much faster algorithms for matrix scaling. In FOCS, pages 890–901. IEEE, 2017. (Cited on page 2.)

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In NeurIPS, pages 1964–1974, 2017. (Cited on pages 2, 6, 8, 9, 12, 13, 14, 16, 21, 22, 24, 25, 26, 29, and 30.)
J. Altschuler, F. Bach, A. Rudi, and J. Weed. Approximating the quadratic transportation metric in near-linear time. *ArXiv Preprint: 1810.10046*, 2018. (Cited on page 2.)

J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *NeurIPS*, pages 4429–4439, 2019. (Cited on page 2.)

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

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In *NeurIPS*, 2020. (Cited on page 34.)

A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. In *NeurIPS*, pages 7498–7510, 2019. (Cited on page 34.)

J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. Towards optimal running times for optimal transport. *ArXiv Preprint: 1810.07717*, 2018. (Cited on pages 2, 3, and 10.)

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

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. (Cited on page 34.)

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

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. (Cited on page 2.)

L. Chizat, P. Roussillon, F. Léger, F-X. Vialard, and G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In *NeurIPS*, 2020. (Cited on page 5.)

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. (Cited on page 2.)

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

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013. (Cited on pages 2, 4, 6, 9, 25, and 30.)

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *ICML*, pages 685–693, 2014. (Cited on page 1.)

M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016. (Cited on page 2.)

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969. (Cited on pages 5 and 34.)
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. (Cited on page 1.)

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. (Cited on pages 1, 2, 3, 6, 8, 9, 10, 12, 13, 14, 16, 19, 21, 22, 23, 25, and 30.)

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. SIAM Journal on Optimization, 25(4):1997–2023, 2015. (Cited on page 24.)

R. Flamary and N. Courty. POT: Python optimal transport library, 2017. URL https://github.com/rflamary/POT. (Cited on page 3.)

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. (Cited on pages 2, 5, and 34.)

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In NeurIPS, pages 3440–3448, 2016. (Cited on pages 2 and 30.)

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In AISTATS, pages 1574–1583. PMLR, 2019. (Cited on page 5.)

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In NeurIPS, pages 5767–5777, 2017. (Cited on page 1.)

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. (Cited on pages 3, 6, 16, 24, 25, and 34.)

W. Guo, N. Ho, and M. Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In AISTATS, pages 2088–2097. PMLR, 2020. (Cited on pages 3 and 16.)

N. Ho, V. Huynh, D. Phung, and M. I. Jordan. Probabilistic multilevel clustering via composite transportation distance. AISTATS, pages 3149–3157, 2019. (Cited on page 1.)

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. (Cited on pages 3 and 16.)

B. Kalantari and L. Khachiyan. On the complexity of nonnegative matrix scaling. Linear Algebra and its Applications, 240:87–103, 1996. (Cited on page 2.)

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

L. V. Kantorovich. On the translocation of masses. In Dokl. Akad. Nauk. USSR (NS), volume 37, pages 199–201, 1942. (Cited on page 4.)

P. A. Knight. The Sinkhorn-Knopp algorithm: convergence and applications. SIAM Journal on Matrix Analysis and Applications, 30(1):261–275, 2008. (Cited on page 2.)
S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *NeurIPS*, pages 261–272, 2019. (Cited on page 34.)

N. Lahn, D. Mulchandani, and S. Raghvendra. A graph theoretic additive approximation of optimal transport. In *NeurIPS*, pages 13813–13823, 2019. (Cited on page 3.)

Y. T. Lee and A. Sidford. Path finding methods for linear programming: Solving linear programs in $\tilde{O}(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow. In *FOCS*, pages 424–433. IEEE, 2014. (Cited on page 2.)

Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015. (Cited on page 24.)

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. (Cited on pages 3, 6, and 9.)

T. Lin, Z. Hu, and X. Guo. Sparsemax and relaxed Wasserstein for topic sparsity. In *WSDM*, pages 141–149. ACM, 2019b. (Cited on page 1.)

H. Lu, R. Freund, and V. Mirrokni. Accelerating greedy coordinate descent methods. In *ICML*, pages 3263–3272, 2018. (Cited on page 24.)

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. (Cited on page 5.)

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005. (Cited on pages 6, 8, and 15.)

Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018. (Cited on page 24.)

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. (Cited on page 34.)

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

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

F-P. Paty and M. Cuturi. Subspace robust Wasserstein distances. In *ICML*, pages 5072–5081, 2019. (Cited on page 34.)

O. Pele and M. Werman. Fast and robust earth mover’s distance. In *ICCV*. IEEE, 2009. (Cited on page 2.)

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. (Cited on page 2.)

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *ICML*, pages 2664–2672, 2016. (Cited on page 1.)
M. S. Pydi and V. Jog. Adversarial risk via optimal transport and optimal couplings. In ICML, pages 7814–7823. PMLR, 2020. (Cited on page 34.)

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

A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In AISTATS, pages 630–638, 2016. (Cited on page 1.)

M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. Mathematical Programming, 162:83–112, 2017. (Cited on page 30.)

J. Sherman. Area-convexity, $\ell_\infty$ regularization, and undirected multicommodity flow. In STOC, pages 452–460. ACM, 2017. (Cited on pages 3 and 16.)

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

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

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In AISTATS, pages 912–920, 2015. (Cited on page 1.)

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

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In ICLR, 2018. (Cited on page 1.)

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009. (Cited on page 1.)

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. (Cited on pages 2, 5, and 34.)