On Efficient Optimal Transport: An Analysis of Greedy and Accelerated Mirror Descent Algorithms

Tianyi Lin [*1] **Nhat Ho** [*2] **Michael I. Jordan** [3]

Abstract

We provide theoretical analyses for two algorithms that solve the regularized optimal transport (OT) problem between two discrete probability measures with at most n atoms. We show that a greedy variant of the classical Sinkhorn algorithm, known as the *Greenkhorn algorithm*, can be improved to $\widetilde{\mathcal{O}}\left(n^2/\varepsilon^2\right)$, improving on the best known complexity bound of $\widetilde{\mathcal{O}}\left(n^2/\varepsilon^3\right)$. This matches the best known complexity bound for the Sinkhorn algorithm and helps explain why the Greenkhorn algorithm outperforms the Sinkhorn algorithm in practice. Our proof technique is based on a primal-dual formulation and provide a *tight* upper bound for the dual solution, leading to a class of *adaptive primal-dual accelerated mirror descent* (APDAMD) algorithms. We prove that the complexity of these algorithms is $\widetilde{\mathcal{O}}\left(n^2\sqrt{\gamma}/\varepsilon\right)$ in which $\gamma \in (0, n]$ refers to some constants in the Bregman divergence. Experimental results on synthetic and real datasets demonstrate the favorable performance of the Greenkhorn and APDAMD algorithms in practice.

1. Introduction

Optimal transport—the problem of finding minimal cost couplings between pairs of probability measures—has a long history in mathematics and operations research (Villani, 2003). In recent years, it has been the inspiration for numerous applications in machine learning and statistics, including posterior contraction of parameter estimation in Bayesian nonparametrics models (Nguyen, 2013; 2016), scalable posterior sampling for large datasets (Srivastava et al., 2015; 2018), optimization models for clustering complex structured data (Ho et al., 2018), deep generative models and domain adaptation in deep learning (Arjovsky et al., 2017; Gulrajani et al., 2017; Courty et al., 2017; Tolstikhin et al., 2018), and other applications (Rolet et al., 2016; Peyré et al., 2016; Carrière et al., 2017; Lin et al., 2018). These large-scale applications have placed significant new demands on the efficiency of algorithms for solving the optimal transport problem, and a new literature has begun to emerge to provide new algorithms and complexity analyses for optimal transport.

The computation of the optimal-transport (OT) distance can be formulated as a linear programming problem and solved in principle by interior-point methods. The best known complexity bound in this formulation is $\widetilde{\mathcal{O}}\left(n^{5/2}\right)$, achieved by an interior-point algorithm due to Lee & Sidford (2014). However, Lee and Sidford's method requires as a subroutine a practical implementation of the Laplacian linear system solver, which is not yet available in the literature. Pele & Werman (2009) proposed an alternative, implementable interior-point method for OT with a complexity bound is $\widetilde{\mathcal{O}}(n^3)$. Another prevalent approach for computing OT distance between two discrete probability measures involves regularizing the objective function by the entropy of the transportation plan. The resulting problem, referred to as *entropic regularized OT* or simply *regularized OT* (Cuturi, 2013; Benamou et al., 2015), is more readily solved than the original problem since the objective is strongly convex with respect to $\|\cdot\|_1$. The longstanding state-of-the-art algorithm for solving regularized OT is the Sinkhorn algorithm (Sinkhorn, 1974; Knight, 2008; Kalantari et al., 2008). Inspired by the growing scope of applications for optimal transport, several new algorithms have emerged in recent years that have been shown empirically to have superior performance when compared to the Sinkhorn algorithm. An example includes the Greenkhorn algorithm (Altschuler et al., 2017; Chakrabarty & Khanna, 2018; Abid & Gower, 2018), which is a greedy version of Sinkhorn algorithm. A variety of standard optimization algorithms have also been adapted to the OT setting, including accelerated gradient descent (Dvurechensky et al., 2018), quasi-Newton methods (Cuturi & Peyré, 2016; Blondel et al., 2018) and

[*]Equal contribution [1]Department of IEOR, University of California, Berkeley [2]Department of EECS, University of California, Berkeley [3]Department of Statistics and EECS, University of California, Berkeley. Correspondence to: Nhat Ho <minhnhat@berkeley.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

stochastic average gradient (Genevay et al., 2016). The theoretical analysis of these algorithms is still nascent.

Very recently, Altschuler et al. (2017) have shown that both the Sinkhorn and Greenkhorn algorithm can achieve the near-linear time complexity for regularized OT. More specifically, they proved that the complexity bounds for both algorithms are $\widetilde{\mathcal{O}}\left(n^2/\varepsilon^3\right)$, where n is the number of atoms (or equivalently dimension) of each probability measure and ε is a desired tolerance. Later, Dvurechensky et al. (2018) improved the complexity bound for the Sinkhorn algorithm to $\widetilde{\mathcal{O}}\left(n^2/\varepsilon^2\right)$ and further proposed an adaptive primal-dual accelerated gradient descent (APDAGD), asserting a complexity bound of $\widetilde{\mathcal{O}}\left(\min\left\{n^{9/4}/\varepsilon,\, n^2/\varepsilon^2\right\}\right)$ for this algorithm. It is also possible to use a carefully designed Newton-type algorithm to solve the OT problem (Allen-Zhu et al., 2017; Cohen et al., 2017), by making use of a connection to matrix-scaling problems. Blanchet et al. (2018) and Quanrud (2018) provided a complexity bound of $\widetilde{\mathcal{O}}\left(n^2/\varepsilon\right)$ for Newton-type algorithms. Unfortunately, these Newton-type methods are complicated and efficient implementations are not yet available. Nonetheless, this complexity bound can be viewed as a theoretical benchmark for the algorithms that we consider in this paper.

Our Contributions. The contribution of this work is threefold and can be summarized as follows:

1. We improve the complexity bound for the Greenkhorn algorithm from $\widetilde{\mathcal{O}}\left(n^2/\varepsilon^3\right)$ to $\widetilde{\mathcal{O}}\left(n^2/\varepsilon^2\right)$, matching the best known complexity bound for the Sinkhorn algorithm. This analysis requires a new proof technique—the technique used in Dvurechensky et al. (2018) for analyzing the complexity of Sinkhorn algorithm is not applicable to the Greenkhorn algorithm. In particular, the Greenkhorn algorithm only updates a single row or column at a time and its per-iteration progress is accordingly more difficult to quantify than that of the Sinkhorn algorithm. In contrast, we employ a novel proof technique with a novel tight ℓ_∞-bound for the dual optimal solution. Our results shed light on the better practical performance of the Greenkhorn algorithm compared the Sinkhorn algorithm.

2. The smoothness of the dual regularized OT with respect to $\|\cdot\|_\infty$ allows us to formulate a novel *adaptive primal-dual accelerated mirror descent* (APDAMD) algorithm for the OT problem. Here the Bregman divergence is strongly convex and smooth with respect to $\|\cdot\|_\infty$. The resulting method involves an efficient line-search strategy (Nesterov & Polyak, 2006) that is readily analyzed and is inspired by a few recently proposed mirror descent algorithms (Zhou et al., 2017a;b). It can be adapted to problems even more general than regularized OT. It can be viewed as a primal-dual extension of Algorithm 1 in Tseng (2008) and a mirror descent extension of the APDAGD algorithm (Dvurechensky et al., 2018). We establish a complexity bound for the APDAMD algorithm of $\widetilde{\mathcal{O}}\left(n^2\sqrt{\gamma}/\varepsilon\right)$ in which $\gamma \in (0, n]$ refers to some constants in the Bregman divergence. In particular, $\gamma = n$ if the Bregman divergence is simply chosen as $\frac{1}{2n}\|\cdot\|_2^2$. We provide some numerical results to show that APDAMD is more robust than APDAGD (see Section 6).

3. We show that there is a limitation in the derivation by Dvurechensky et al. (2018) of the complexity bound $\widetilde{\mathcal{O}}\left(\min\left\{n^{9/4}/\varepsilon,\, n^2/\varepsilon^2\right\}\right)$. More specifically, the complexity bound in Dvurechensky et al. (2018) depends on a parameter which is not estimated explicitly. We provide a sharp lower bound for this parameter by a simple example (Proposition 5.2), demonstrating that this parameter depends on n. Due to the dependence on n of that parameter, we demonstrate that the complexity bound of APDAGD algorithm is indeed $\widetilde{\mathcal{O}}(n^{2.5}/\varepsilon)$. This is slightly worse than the asserted complexity bound of $\widetilde{\mathcal{O}}\left(\min\left\{n^{9/4}/\varepsilon,\, n^2/\varepsilon^2\right\}\right)$ in terms of dimension n. Finally, our APDAMD algorithm potentially provides an improvement for the complexity of APDAGD algorithm as its complexity bound is $\widetilde{\mathcal{O}}(n^2\sqrt{\gamma}/\varepsilon)$ and γ can be much smaller than n.

Organization. The remainder of the paper is organized as follows. In Section 2, we provide the basic setup for regularized OT. We analyze the worst-case complexity of the Greenkhorn algorithm in Section 3. In Section 4, we propose the APDAMD algorithm for solving regularized OT and provide a theoretical complexity analysis. In Section 5, we provide detail argument with the limitation of complexity of APDAGD algorithm (Dvurechensky et al., 2018). Section 6 presents experiments that illustrate the favorable performance of the Greenkhorn and APDAMD algorithms. Proofs for all the results as well as additional experiments are presented in the Supplementary material.

Notation. We let Δ^n denote the probability simplex in $n-1$ dimensions, for $n \geq 2$: $\Delta^n = \{u = (u_1, \ldots, u_n) \in \mathbb{R}^n : \sum_{i=1}^n u_i = 1,\ u \geq 0\}$. Furthermore, $[n]$ stands for the set $\{1, 2, \ldots, n\}$ while $\mathbb{R}^n_+$ stands for the set of all vectors in $\mathbb{R}^n$ with nonnegative components for any $n \geq 1$. For a vector $x \in \mathbb{R}^n$ and $1 \leq p \leq \infty$, we denote $\|x\|_p$ as its ℓ_p-norm and $\mathrm{diag}(x)$ as the diagonal matrix with x on the diagonal. For a matrix $A \in \mathbb{R}^{n\times n}$, the notation $\mathrm{vec}(A)$ stands for the vector in $\mathbb{R}^{n^2}$ obtained from concatenating the rows and columns of A. $\mathbf{1}$ stands for a vector with all of its components equal to 1. $\partial_x f$ refers to a partial gradient of f with respect to x. Lastly, given the dimension n and accuracy ε, the notation $a = \mathcal{O}\left(b(n, \varepsilon)\right)$ stands for the upper bound $a \leq C \cdot b(n, \varepsilon)$ where C is independent of n and ε. Similarly, the notation $a = \widetilde{\mathcal{O}}(b(n, \varepsilon))$ indicates the previous inequality may depend on the logarithmic function

of n and ε, and where $C > 0$.

2. Problem Setup

In this section, we review the formal problem of computing the OT distance between two discrete probability measures with at most n atoms. We also discuss its regularized version, the *entropic regularized OT* problem. We then proceed to present the formulation of the dual regularized OT problem, which is vital for our theoretical analysis in the sequel.

2.1. (Regularized) OT

Approximating the OT distance amounts to solving a linear problem given by Kantorovich (1942):

$$\min \langle C, X \rangle \quad \text{s.t. } X\mathbf{1} = r,\ X^\top \mathbf{1} = l,\ X \geq 0, \tag{1}$$

where $X \in \mathbb{R}^{n\times n}$ is called the *transportation plan* while $C \in \mathbb{R}_+^{n\times n}$ is a cost matrix comprised of nonnegative elements. The vectors r and l are fixed vectors in the probability simplex Δ^n. The regularized version of problem (1) is proposed by Cuturi (2013) in which the entropy of X is used instead of the nonnegative constraints. The resulting problem is formulated as follows:

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

where $\eta > 0$ is the *regularization parameter* and $H(X)$ is the entropic regularization given by

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

The computational problem is to find $\hat{X} \in \mathbb{R}_+^{n\times n}$ such that $\hat{X}\mathbf{1} = r$ and $\hat{X}^\top\mathbf{1} = l$ and

$$\langle C, \hat{X}\rangle \ \leq \ \langle C, X^*\rangle + \varepsilon, \tag{4}$$

where X^* is an optimal transportation plan, i.e., an optimal solution to problem (1). In this formulation, $\langle C, \hat{X}\rangle$ is referred to an *ε-approximation* for the OT distance and $\hat{X}$ is an *ε-approximate transportation plan*.

2.2. Dual regularized OT

While problem (2) involves optimizing a convex objective with several affine constraints, its dual problem is a unconstrained optimization problem, which simplifies both algorithm design and the complexity analysis. To derive the dual, we begin with the Lagrangian $\mathcal{L}(X, \alpha, \beta) = \langle C, X\rangle - \eta H(X) - \langle \alpha, X\mathbf{1} - r\rangle - \langle \beta, X^\top\mathbf{1} - l\rangle$. The dual regularized OT is obtained by solving $\min_X \mathcal{L}(X, \alpha, \beta)$. Since $\mathcal{L}(\cdot, \alpha, \beta)$ is strictly convex and differentiable, we can solve it by setting $\partial_X \mathcal{L}(X, \alpha, \beta) = 0$. More specifically, we have

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

implying that

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

To simplify the notation, we perform a change of variables, setting $u_i = \frac{\alpha_i}{\eta} - \frac{1}{2}$ and $v_j = \frac{\beta_j}{\eta} - \frac{1}{2}$ from which we obtain $X^*_{ij} = e^{-\frac{C_{ij}}{\eta}+u_i+v_j}$. With this solution, we have

$$\mathcal{L}(X^*, \alpha, \beta) \ = \ \eta\Bigg(-\sum_{i,j=1}^{n} e^{-\frac{C_{ij}}{\eta}+u_i+v_j} + \langle u, r\rangle + \langle v, l\rangle + 1\Bigg).$$

Putting these pieces together yields the dual problem of problem (2) as follows,

$$\max_{u,v\in\mathbb{R}^n} \ -\sum_{i,j=1}^{n} e^{-\frac{C_{ij}}{\eta}+u_i+v_j} + \langle u, r\rangle + \langle v, l\rangle\,. \tag{5}$$

Letting $B(u, v) \ := \ \mathrm{diag}(e^u)\ e^{-\frac{C}{\eta}}\ \mathrm{diag}(e^v)$. Then problem (5) is simplified as follows,

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

We refer to problem (6) to the *dual regularized OT* problem.

3. The Greenkhorn Algorithm

In this section, we analyze the Greenkhorn algorithm, which stands for a "greedy Sinkhorn" algorithm (Altschuler et al., 2017). In particular, we improve the existing best known complexity bound $\mathcal{O}\left(\frac{n^2\|C\|_\infty^3\log(n)}{\varepsilon^3}\right)$ in Altschuler et al. (2017) to $\mathcal{O}\left(\frac{n^2\|C\|_\infty^2\log(n)}{\varepsilon^2}\right)$, which matches the best known complexity bound for the Sinkhorn algorithm (Dvurechensky et al., 2018). To facilitate the discussion later, we present the Greenkhorn algorithm in pseudocode form in Algorithm 1 and its application to regularized OT in Algorithm 2.

Both the Sinkhorn and Greenkhorn procedures are coordinate descent algorithms for the dual regularized OT problem (6). However, while the Greenkhorn algorithm is a greedy coordinate descent algorithm, the Sinkhorn algorithm is block coordinate descent with only two blocks. It turns out to be easier to quantify the per-iteration progress of the Sinkhorn algorithm than that of the Greenkhorn algorithm, as suggested by the fact that the proof techniques in Dvurechensky et al. (2018) are not applicable to the Greenkhorn algorithm. We thus explore a different strategy which will be elaborated in the sequel.

3.1. Algorithm scheme

The Greenkhorn algorithm is presented in Algorithm 1 with the function $\rho : \mathbb{R}_+ \times \mathbb{R}_+ \to [0, +\infty]$ (Altschuler et al.,

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

Input: $k = 0$ and $u^0 = v^0 = 0$.
while $E^k > \varepsilon'$ **do**
  $r(u^k, v^k) = B(u^k, v^k)\mathbf{1}$.
  $l(u^k, v^k) = B(u^k, v^k)^\top \mathbf{1}$.
  $I = \operatorname{argmax}_{1\le i\le n} \rho\left(r_i, r_i(u^k, v^k)\right)$.
  $J = \operatorname{argmax}_{1\le j\le n} \rho\left(l_j, l_j(u^k, v^k)\right)$.
  if $\rho\left(r_i, r_i(u^k, v^k)\right) > \rho\left(l_j, l_j(u^k, v^k)\right)$ **then**
    $u_I^{k+1} = u_I^k + \log(r_I) - \log\left(r_I(u^k, v^k)\right)$.
  else
    $v_J^{k+1} = v_J^k + \log(l_J) - \log\left(l_J(u^k, v^k)\right)$.
  end if
  $k = k + 1$.
end while
Output: $B(u^k, v^k)$.

Algorithm 2 Approximating OT by GREENKHORN

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{l} \in \Delta_n$ be defined as

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

Step 2: Compute

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

Step 3: Round $\tilde{X}$ to $\hat{X}$ by Algorithm 2 (Altschuler et al., 2017) such that $\hat{X}\mathbf{1} = r$ and $\hat{X}^\top \mathbf{1} = l$.
Output: $\hat{X}$.

2017) given by $\rho(a, b) := b - a + a\log\left(\frac{a}{b}\right)$. Note that ρ measures the progress in the dual objective value between two consecutive iterates of the Greenkhorn algorithm. In particular, $\rho(a, b) \ge 0$ for any $a, b \in \mathbb{R}_+$ and the equality holds if and only if $a = b$. On the other hand, we observe that the optimality condition of the dual regularized OT problem (6) is $B(u, v)\mathbf{1} - r = 0$ and $B(u, v)^\top \mathbf{1} - l = 0$. This brings us to the following quantity which measures the error of the k-th iterate of the Greenkhorn algorithm:

$$E^k := \|B(u^k, v^k)\mathbf{1} - r\|_1 + \|B(u^k, v^k)^\top \mathbf{1} - l\|_1.$$

3.2. Complexity analysis—bounding dual objective values

Given the definition of E^k, we first have the following lemma which yields an upper bound for the objective values of the iterates.

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

$$f(u^k, v^k) - f(u^*, v^*) \le \left(2\|u^*\|_\infty + 2\|v^*\|_\infty\right) E^k, \quad (7)$$

where (u^*, v^*) denotes an optimal solution pair for the dual regularized OT problem (6).

Our second lemma provides an upper bound for the ℓ_∞-norm of the optimal solution pair (u^*, v^*) of the dual regularized OT problem. Note that this result is stronger than Lemma 1 in Dvurechensky et al. (2018) and generalizes Lemma 8 in Blanchet et al. (2018) with fewer assumptions.

Lemma 3.2. *For the dual regularized OT problem (6), there exists an optimal solution (u^*, v^*) such that*

$$\|u^*\|_\infty \le R, \quad \|v^*\|_\infty \le R, \quad (8)$$

where $R > 0$ is defined as

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

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

Corollary 3.3. *Letting $\{(u^k, v^k)\}_{k\ge 0}$ denote the iterates returned by the Greenkhorn algorithm, we have*

$$f(u^k, v^k) - f(u^*, v^*) \le 4RE^k. \quad (9)$$

Remark 3.4. *The constant R provides an upper bound both in this paper and in Dvurechensky et al. (2018), where the same notation is used. The values in the two papers are of the same order since R in our paper only involves an additional term $\log(n) - \log\left(\min_{1\le i,j\le n}\{r_i, l_j\}\right)$.*

Remark 3.5. *We further comment on the proof techniques in this paper and Dvurechensky et al. (2018). The proof for Lemma 2 in Dvurechensky et al. (2018) depends on taking full advantage of the shift property of the Sinkhorn algorithm; namely, either $B(\overline{u}^k, \overline{v}^k)\mathbf{1} = r$ or $B(\overline{u}^k, \overline{v}^k)^\top \mathbf{1} = l$, where $(\overline{u}^k, \overline{v}^k)$ stands for the iterates of the Sinkhorn algorithm. Unfortunately, the Greenkhorn algorithm does not enjoy such a shift property. We have thus proposed a different approach for bounding $f(u^k, v^k) - f(u^*, v^*)$, based on the ℓ_∞-norm of the optimal solution (u^*, v^*) of the dual regularized OT problem.*

3.3. Complexity analysis—bounding the number of iterations

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

Lemma 3.6. *Let $\{(u^k, v^k)\}_{k\ge 0}$ be the iterates returned by the Greenkhorn algorithm, we have*

$$f(u^k, v^k) - f(u^{k+1}, v^{k+1}) \ge \frac{(E^k)^2}{28n}. \quad (10)$$

We are now able to derive the iteration complexity of Greenkhorn algorithm based on Corollary 3.3 and Lemma 3.6.

Theorem 3.7. *The Greenkhorn algorithm returns a matrix $B(u^k, v^k)$ that satisfies $E_k \leq \varepsilon'$ in the number of iterations k satisfying*

$$k \;\leq\; 2 + \frac{112nR}{\varepsilon'}, \tag{11}$$

where R is defined in Lemma 3.2.

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 the Greenkhorn algorithm:

Theorem 3.8. *The Greenkhorn algorithm for approximating optimal transport (Algorithm 2) returns $\hat{X} \in \mathbb{R}^{n\times n}$ satisfying $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$ and (4) in*

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

arithmetic operations.

The result of Theorem 3.8 improves the best known complexity bound $\widetilde{\mathcal{O}}\left(\frac{n^2}{\varepsilon^3}\right)$ for the Greenkhorn algorithm (Altschuler et al., 2017; Abid & Gower, 2018), and further matches the best known complexity bound for the Sinkhorn algorithm (Dvurechensky et al., 2018). This sheds light on the superior performance of the Greenkhorn algorithm in practice.

4. Adaptive Primal-Dual Accelerated Mirror Descent

In this section, we propose and analyze an adaptive primal-dual accelerated mirror descent (APDAMD) algorithm for a general class of problems that specializes to the regularized OT problem in (2). The APDAMD algorithm is an adaptive primal-dual optimization algorithm for finding a primal-dual optimal solution pair for a broad class of OT problems. The pseudocode for the APDAMD algorithm and its specialization to the regularized OT problem (2) are presented in Algorithm 4 and Algorithm 3, respectively. In Section 4.2 we show that the complexity of APDAMD is $\mathcal{O}\left(\frac{n^2\sqrt{\gamma}\|C\|_\infty \log(n)}{\varepsilon}\right)$ in which $\gamma \in (0, n]$ refers to some constants in the Bregman divergence.

4.1. General setup

We consider the following generalization of the regularized OT problem:

$$\min_{x\in\mathbb{R}^n} f(x), \quad \text{s.t. } Ax = b, \tag{12}$$

where $A \in \mathbb{R}^{n\times n}$ is a matrix and $b \in \mathbb{R}^n$. Here f is assumed to be strongly convex with respect to the ℓ_1-norm:

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

The Lagrangian dual problem for (12) can be written as the following minimization problem:

$$\min_{\lambda\in\mathbb{R}^n} \varphi(\lambda) \;:=\; \langle \lambda, b\rangle + \max_{x\in\mathbb{R}^n}\left\{-f(x) - \langle A^\top\lambda, x\rangle\right\}. \tag{13}$$

A direct computation leads to $\nabla\varphi(\lambda) = b - Ax(\lambda)$ where

$$x(\lambda) \;:=\; \operatorname*{argmax}_{x\in\mathbb{R}^n}\left\{-f(x) - \langle A^\top\lambda, x\rangle\right\}.$$

To analyze the complexity of the APDAMD algorithm, we first establishes the smoothness of the dual objective function φ with respect to the ℓ_∞-norm.

Lemma 4.1. *The dual objective φ satisfies that*

$$\varphi(\lambda_1) - \varphi(\lambda_2) - \langle \nabla\varphi(\lambda_2), \lambda_1 - \lambda_2\rangle \;\leq\; \frac{\|A\|_1^2}{2\eta}\left\|\lambda_1 - \lambda_2\right\|_\infty^2.$$

The Bregman divergence $B_\phi : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty]$ is

$$B_\phi(z, z') \;:=\; \phi(z) - \phi(z') - \langle \phi(z'), z - z'\rangle, \quad \forall z, z' \in \mathbb{R}^n$$

Let ϕ be a $\frac{1}{\gamma}$-strongly convex and 1-smooth on $\mathbb{R}^n$ with respect to the ℓ_∞-norm; i.e., for any $z \neq z'$,

$$\frac{1}{2\gamma} \;\leq\; \frac{\phi(z) - \phi(z') - \langle \nabla\phi(z'), z - z'\rangle}{\left\|z - z'\right\|_\infty^2} \;\leq\; \frac{1}{2}. \tag{14}$$

Note that one typical choice of ϕ is $\frac{1}{2n}\left\|z\right\|_2^2$, implying that

$$B_\phi(z, z') \;=\; \frac{1}{2n}\left\|z - z'\right\|_2^2.$$

In this extreme case, $\gamma = n$. In general, γ is a function of n. It is worth noting that the value of γ will affect the complexity bound of the APDAMD algorithm for approximating optimal transport problem (see Theorem 4.3). We make no attempt to optimize the value of γ as a function of n in the current paper. To analyze the complexity of the APDAMD algorithm for solving the regularized OT problem, we establish several key properties of APDAMD algorithm for the general setup in (12) in Section A in the Supplementary material.

4.2. Complexity analysis for the APDAMD algorithm

We start with the setting of the dual objective φ for regularized OT problem. Different from dual problem (5), we set $\varphi(\lambda)$ as the objective in the original dual OT problem in which $\lambda := (\alpha, \beta)$. The resulting problem is

$$\min_{\alpha,\beta\in\mathbb{R}^n} \varphi(\alpha, \beta) \;:=\; \eta\sum_{i,j=1}^n e^{-\frac{C_{ij} - \alpha_i - \beta_j}{\eta} - 1} - \langle\alpha, r\rangle - \langle\beta, l\rangle. \tag{15}$$

Algorithm 3 Approximating OT by APDAMD

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{l} \in \Delta_n$ be defined as

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

Step 2: Let $A \in \mathbb{R}^{2n \times n^2}$ and $b \in \mathbb{R}^{2n}$ be defined by

$$A\text{vec}(X) = \begin{pmatrix} X\mathbf{1} \\ X^\top \mathbf{1} \end{pmatrix}, \quad b = \begin{pmatrix} \tilde{r} \\ \tilde{l} \end{pmatrix}$$

Step 3: Compute $\tilde{X} = \text{APDAMD}\left(\varphi, A, b, \varepsilon'/2\right)$ with φ defined in (13) with $f(x) = \text{vec}(C)^\top \text{vec}(X) - \eta H(X)$.
Step 4: Round $\tilde{X}$ to $\hat{X}$ by Algorithm 2 (Altschuler et al., 2017) such that $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$.
Output: $\hat{X}$.

Algorithm 4 APDAMD$(\varphi, A, b, \varepsilon')$

Input: $k = 0$.
Initialization: $\bar{\alpha}^0 = \alpha^0 = 0$, $z^0 = \mu^0 = \lambda^0$ and $L^0 = 1$
repeat
  Set $M^k = \frac{L^k}{2}$.
  repeat
    Set $M^k = 2M^k$.
    Compute step size $\alpha^{k+1} = \frac{1+\sqrt{1+4\gamma M^k \bar{\alpha}^k}}{2\gamma M^k}$.
    Compute accumulating parameter $\bar{\alpha}^{k+1} = \bar{\alpha}^k + \alpha^{k+1}$.
    Compute averaging step $\mu^{k+1} = \frac{\alpha^{k+1} z^k + \bar{\alpha}^k \lambda^k}{\bar{\alpha}^{k+1}}$.
    Compute mirror descent update:

$$z^{k+1} = \underset{z \in \mathbb{R}^n}{\text{argmin}} \left\{ \left\langle \nabla\varphi(\mu^{k+1}), z - \mu^{k+1} \right\rangle + \frac{B_\phi(z, z^k)}{\alpha^{k+1}} \right\}.$$

    Compute averaging step $\lambda^{k+1} = \frac{\alpha^{k+1} z^{k+1} + \bar{\alpha}^k \lambda^k}{\bar{\alpha}^{k+1}}$.
  until stopping criterion $\varphi(\lambda^{k+1}) - \varphi(\mu^{k+1}) - \left\langle \nabla\varphi(\mu^{k+1}), \lambda^{k+1} - \mu^{k+1} \right\rangle \leq \frac{M^k}{2}\left\|\lambda^{k+1} - \mu^{k+1}\right\|_\infty^2$.
  Set averaging step $x^{k+1} = \frac{\alpha^{k+1} x(\mu^{k+1}) + \bar{\alpha}^k x^k}{\bar{\alpha}^{k+1}}$.
  Set $L^{k+1} = \frac{M^k}{2}$.
  Set $k = k + 1$.
until $\left\|Ax^k - b\right\|_1 \leq \varepsilon'$
Output: X^k where $x^k = \text{vec}(X^k)$.

This problem was also considered in Dvurechensky et al. (2018) to establish the complexity bound of APDAGD algorithm. By means of transformations $u_i = \frac{\alpha_i}{\eta} - \frac{1}{2}$ and $v_j = \frac{\beta_j}{\eta} - \frac{1}{2}$, we follow from Lemma 3.2 that

$$\|\alpha^*\|_\infty \leq \eta\left(R + \frac{1}{2}\right), \quad \|\beta^*\|_\infty \leq \eta\left(R + \frac{1}{2}\right), \quad (16)$$

We are ready to derive an upper bound for the iteration number of Algorithm 3 to reach a desired accuracy ε':

Theorem 4.2. *The APDAMD algorithm for approximating optimal transport (Algorithm 3) returns an output X^k that satisfies $\|A\,vec(X^k) - b\|_1 \leq \varepsilon'$ in a number of iterations k bounded as follows:*

$$k \leq 1 + 4\sqrt{2}\|A\|_1 \sqrt{\frac{\gamma(R + 1/2)}{\varepsilon'}}$$

where R is defined in Lemma 3.2.

Equipped with the result of Theorem 4.2, we proceed to present the complexity bound of APDAMD algorithm for approximating the OT problem.

Theorem 4.3. *The APDAMD algorithm for approximating optimal transport (Algorithm 3) returns $\hat{X} \in \mathbb{R}^{n \times n}$ satisfying $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$ and (4) in a total of*

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

arithmetic operations.

The complexity bound of the APDAMD algorithm in Theorem 4.3 suggests an interesting feature of the (regularized) OT problem. Indeed, the dependence of that bound on γ manifests the necessity of using $\|.\|_\infty$ 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 the Greenkhorn algorithm in Section 3, where we rely on the $\|.\|_\infty$ of optimal solutions of the dual regularized OT problem to measure the progress in the objective value among the successive iterates (See Section B in the Supplementary material).

5. Revisiting the APDAGD algorithm

In this section, we first point out that the complexity bound of the APDAGD algorithm (Dvurechensky et al., 2018) for regularized OT is not $\widetilde{\mathcal{O}}\left(\min\left\{\frac{n^{9/4}}{\varepsilon}, \frac{n^2}{\varepsilon^2}\right\}\right)$. Then, we provide a new complexity bound of the APDAGD algorithm based on our results in Section 4.2. Despite the issue with regularized OT, we wish to emphasize that the APDAGD algorithm is still an interesting and efficient accelerated method for general setup (12) with theoretical guarantee under the certain conditions.

To facilitate the ensuing discussion, we first present the complexity bound for regularized OT in (Dvurechensky et al., 2018) using the notation from the current paper.

Theorem 5.1 (Theorem 4 in Dvurechensky et al. (2018))**.** *The APDAGD algorithm for approximating optimal transport returns $\hat{X} \in \mathbb{R}^{n \times n}$ satisfying $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$ and (4) in a number of arithmetic operations bounded as*

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

Algorithm 5 Approximating OT by APDAGD

Input: $\eta = \frac{\varepsilon}{4\log(n)}$ and $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
Step 1: Let $\tilde{r} \in \Delta_n$ and $\tilde{l} \in \Delta_n$ be defined as

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

Step 2: Let $A \in \mathbb{R}^{2n \times n^2}$ and $b \in \mathbb{R}^{2n}$ be defined by

$$A\text{vec}(X) = \begin{pmatrix} X\mathbf{1} \\ X^\top \mathbf{1} \end{pmatrix}, \quad b = \begin{pmatrix} \tilde{r} \\ \tilde{l} \end{pmatrix}.$$

Step 3: Compute $\tilde{X} = \text{APDAGD}(\varphi, A, b, \varepsilon'/2)$ with φ defined in (13) with $f(x) = \text{vec}(C)^\top \text{vec}(X) - \eta H(X)$.
Step 4: Round $\tilde{X}$ to $\hat{X}$ by Algorithm in (Altschuler et al., 2017) such that $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$.
Output: $\hat{X}$.

where $\|(u^*, v^*)\|_2 \leq \overline{R}$ and (u^*, v^*) denotes an optimal solution pair for the dual regularized OT problem (6).

This theorem suggests that the complexity bound is at the order $\widetilde{\mathcal{O}}\left(\min\left\{\frac{n^{9/4}}{\varepsilon}, \frac{n^2}{\varepsilon^2}\right\}\right)$. However, there are two issues:

1. The upper bound $\overline{R}$ is assumed to be bounded and independent of n, which is not correct; see our counterexample in Proposition 5.2.

2. The known upper bound $\overline{R}$ is based on $\min_{1 \leq i,j \leq n}\{r_i, l_j\}$ (cf. Lemma 3.2 or Lemma 1 in (Dvurechensky et al., 2018)). This implies that the valid algorithm needs to take the rounding error with weight vectors r and l into account.

Corrected upper bound $\overline{R}$. The upper bounds from (16) imply that an upper bound for $\overline{R}$ is $\widetilde{\mathcal{O}}\left(n^{1/2}\right)$. Now we show that $\overline{R}$ is indeed $\Omega\left(n^{1/2}\right)$ for any $\varepsilon \in (0, 1)$.

Proposition 5.2. *Assume that all the entries of the ground cost matrix $C \in \mathbb{R}^{n \times n}$ are 1 and the weight vectors $r = l = \mathbf{1}/n$. Given $\varepsilon \in (0, 1)$ and the regularization term $\eta = \frac{\varepsilon}{4\log(n)}$, the optimal solution (α^*, β^*) of the dual regularized OT problem (15) satisfies $\|(\alpha^*, \beta^*)\|_2 \gtrsim n^{1/2}$.*

Approximation algorithm for OT by APDAGD. Algorithm 4 in Dvurechensky et al. (2018) can be improved by incorporating the rounding procedure, which is summarized in Algorithm 5. Here, the APDAGD algorithm used in Algorithm 5 stands for Algorithm 3 in Dvurechensky et al. (2018). Given the corrected upper bound $\overline{R}$ and Algorithm 5 for approximating OT, we provide a new complexity bound of the APDAGD algorithm in the following proposition.

Proposition 5.3. *The APDAGD algorithm for approximating optimal transport (Algorithm 5) returns $\hat{X} \in \mathbb{R}^{n \times n}$ satisfying $\hat{X}\mathbf{1} = r$, $\hat{X}^\top \mathbf{1} = l$ and (4) in a total of*

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

arithmetic operations.

The proof of Propositions 5.2 and 5.3 are provided in Sections B.9 and B.10. As indicated in Proposition 5.3, the complexity bound of APDAGD and APDAMD algorithm for the regularized OT are comparable if we choose the Bregman divergence to be $\frac{1}{2n}\|.\|_2^2$. It is still unclear whether the upper bound n of γ can be further improved (Nemirovsky, 1983). From this perspective, our APDAMD algorithm can be viewed as a generalization of the APDAGD algorithm. Finally, since our APDAMD algorithm utilizes ℓ_∞-norm in its line search criterion, it will be more robust than the APDAGD algorithm (see the experimental results in Section 6 in the main text).

6. Experiments

In this section, we conduct the extensive comparative experiments with the Greenkhorn and APDAMD algorithms on synthetic and real images[1]. Note that some results are deferred to Section C in the Supplementary material. The baseline algorithms contain the Sinkhorn (Altschuler et al., 2017), APDAGD (Dvurechensky et al., 2018) and GCPB algorithms (Genevay et al., 2016). The Greenkhorn and APDAMD algorithms outperform the Sinkhorn and APDAGD algorithms and the APDAMD algorithm is faster and more robust than APDAGD and GCPB algorithms.

We follow the setup in Altschuler et al. (2017) in order to compare different algorithms on the synthetic images. In particular, the transportation distance is defined between a pair of randomly generated synthetic images and the cost matrix is comprised of ℓ_1 distances among pixel locations in the images.

Image generation: Each of the images is of size 20 by 20 pixels and is generated based on randomly positioning a foreground square in otherwise black background. We utilize a uniform distribution on $[0, 1]$ for the intensities of the background pixels and a uniform distribution on $[0, 50]$ for the foreground pixels.

Evaluation metrics: The first metric is the distance between the output of the algorithm, X, and the transportation polytope, i.e., $d(X) = \|r(X) - r\|_1 + \|l(X) - l\|_1$ in which $r(X)$ and $l(X)$ are the row and column marginal vectors of the output X while r and l stand for the true row and

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

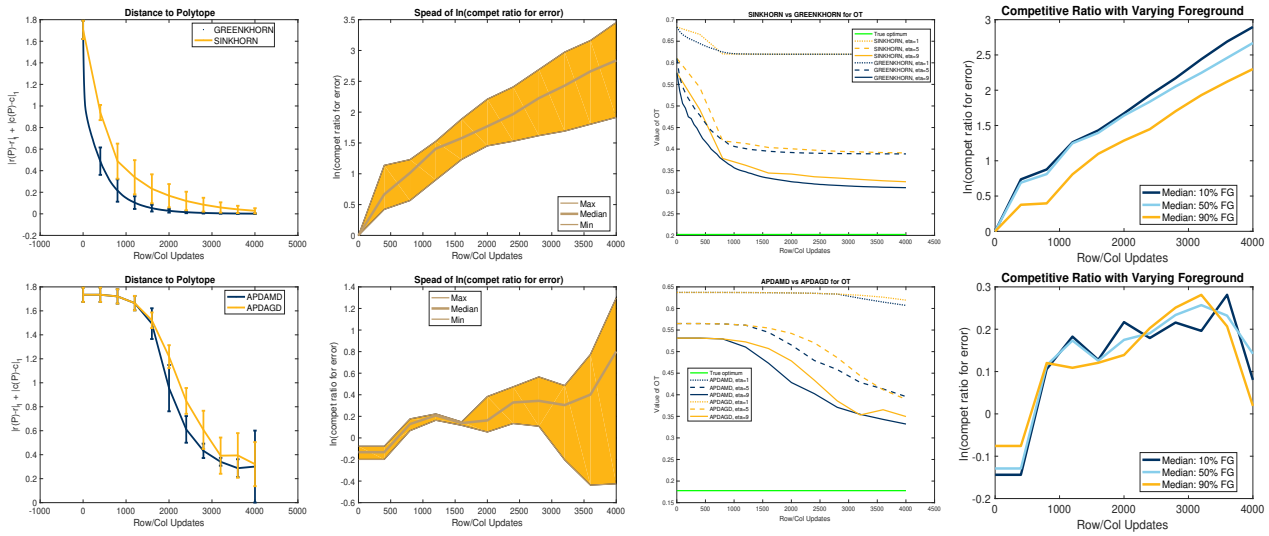

Figure 1. Performance of the Sinkhorn v.s. Greenkhorn, APDAGD v.s. APDAMD on the synthetic images. In the top two images, the comparison is based on using the distance $d(P)$ to the transportation polytope, and the maximum, median and minimum of competitive ratios $\log(d(P_S)/d(P_G))$ and $\log(d(P_{GD})/d(P_{MD}))$ on ten random pairs of images. Here, $d(P_S)$, $d(P_G)$, $d(P_{GD})$ and $d(P_{MD})$ refer to the Sinkhorn, Greenkhorn, APDAGD and APDAMD algorithms, respectively. In the bottom left image, the comparison is based on varying the regularization parameter $\eta \in \{1, 5, 9\}$ and reporting the optimal value of the original optimal transport problem without entropic regularization. Note that the foreground covers 10% of the synthetic images here. In the bottom right image, we compare by using the median of competitive ratios with varying coverage ratio of foreground in the range of 10%, 50%, and 90% of the images.

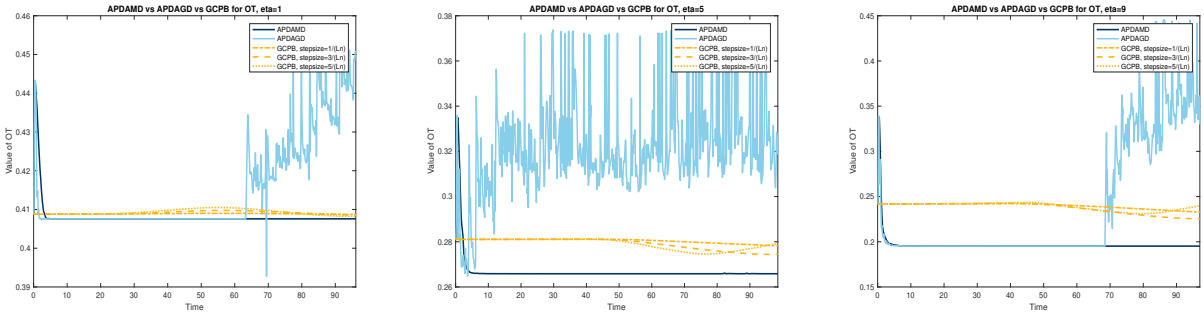

Figure 2. Performance of the GCPB, APDAGD and APDAMD algorithms in term of time on the MNIST real images. These three images specify the values of regularized OT with varying regularization parameter $\eta \in \{1, 5, 9\}$, showing that the APDAMD algorithm is faster and more robust than the APDAGD and GCPB algorithms.

column marginal vectors. The second metric is the competitive ratio, defined by $\log(d(X_1)/d(X_2))$ where $d(X_1)$ and $d(X_2)$ refer to the distance between the outputs of two algorithms and the transportation polytope.

Experimental setting: We perform two pairwise comparative experiments: Sinkhorn versus Greenkhorn and APDAGD versus APDAMD by running these algorithms with ten randomly selected pairs of synthetic images. We also evaluate all the algorithms with varying regularization parameter $\eta \in \{1, 5, 9\}$ and the optimal value of the original optimal transport problem without entropic regularization, as suggested by Altschuler et al. (2017).

Experimental results: Figure 1 shows that the Greenkhorn algorithm outperforms the Sinkhorn algorithm in terms of iteration numbers, supporting our theoretical assertion that the Greenkhorn algorithm has the complexity bound as good as the Sinkhorn algorithm (cf. Theorem 3.7). The APDAMD algorithm with the Bregman divergence $\frac{1}{2n}\|\cdot\|_2^2$ is slightly faster than the APDAGD algorithm.

Figure 2 provides the performance of the APDAMD, APDAGD and GCPB algorithms on real images. The APDAMD algorithm is not faster but more robust than the APDAGD and GCPB algorithms (Genevay et al., 2016). This makes sense since their complexity bounds are the same in terms of n and ε (cf. Theorem 4.3 and Proposition 5.3). On the other hand, the robustness comes from the fact that the APDAMD algorithm can stabilize the training by using ℓ_∞-norm in the line search criterion.

References

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

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

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

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

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

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

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

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

Chakrabarty, D. and Khanna, S. Better and simpler error analysis of the Sinkhorn-Knopp algorithm for matrix scaling. *ArXiv Preprint: 1801.02790*, 2018.

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

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

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

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

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

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

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

Ho, N., Huynh, V., Phung, D., and Jordan, M. I. Probabilistic multilevel clustering via composite transportation distance. *arXiv preprint arXiv:1810.11911*, 2018.

Jiang, B., Lin, T., and Zhang, S. A unified adaptive tensor approximation scheme to accelerate composite convex optimization. *ArXiv Preprint: 1811.02427*, 2018.

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

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

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

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

Lin, T., Hu, Z., and Guo, X. Sparsemax and relaxed Wasserstein for topic sparsity. *ArXiv Preprint: 1810.09079*, 2018.

Nemirovsky, A. S. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

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

Nesterov, Y. and Polyak, B. T. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

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

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

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

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

Quanrud, K. Approximating optimal transport with linear programs. *ArXiv Preprint: 1810.05957*, 2018.

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

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

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

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

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

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *Technical Report*, 2008. URL http://www.mit.edu/œdimitrib/PTseng/papers/apgm.pdf.

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

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., and Glynn, P. W. Stochastic mirror descent in variationally coherent optimization problems. In *NIPS*, pp. 7040–7049, 2017a.

Zhou, Z., Mertikopoulos, P., Moustakas, A. L., Bambos, N., and Glynn, P. Mirror descent learning in continuous games. In *CDC*, pp. 5776–5783. IEEE, 2017b.