Screening Sinkhorn Algorithm via Dual Projections

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper deals with the problem of approximating optimal transport (OT) distance between two discrete measures. Our proposed approach involves a convex projection of the *dual of Sinkhorn divergence*, allowing to define two appropriate active indices sets for the potential variables. These indices sets depend on two parameters acting like a threshold and a scaling factor and they are both directly linked to a priori fixed number budget of points from the supports of the given discrete measures. This new analysis induces a screened version of the dual of Sinkhorn divergence and suggests the *Screenkhorn* algorithm. We illustrate the favorable performance of Screenkhorn in practice with numerical experiments on synthetic and real datasets.

1 Introduction

Computing OT distances between pairs of probability measures or histograms, such as the earth mover's distance [31, 27] and Monge-Kantorovich or Wasserstein distance [8], are currently generating an increasing attraction in different machine learning tasks [30, 22, 4, 17], statistics [15, 24, 12, 5, 14], and computer vision [7, 27, 29], among other applications [21, 26]. In many of these problems, OT exploits the geometric features of the objects at hand in the underlying spaces to be leveraged in comparing probability measures. This effectively leads to improve performance of methods that are oblivious to the geometry, for example the chi-squared distances or the Kullback-Leibler divergence. Unfortunately, this advantage comes at the price of an enormous computational cost of solving the OT problem, that can be prohibitive in large scale applications. For instance, the OT between two histograms with supports of equal size n can be formulated as a linear programming problem that requires generally $\mathcal{O}(n^3 \log n)$ [25] arithmetic operations, which is problematic when n is larger than 10^3.

A remedy to the heavy computation burden of OT lies in a prevalent approach referred to as regularized OT [9] and operates by adding an entropic regularization penalty to the original problem. Such a regularization guarantees a unique solution, since the objective function is strongly convex, and a greater computational stability. Furthermore, [9] proposed the so-called dual of Sinkhorn divergence as the dual of the entropic problem and noticed that finding the dual solution was equivalent to finding two diagonal matrices that made a full matrix bistochastic. Therefore, the OT can be solved efficiently with celebrated matrix scaling algorithms, such as Sinkhorn's fixed point iteration method [28, 20, 18].

Sinkhorn scaling for computing OT distances is a well studied problem in many recent works. The main idea is to improve the matrix-vector operations that are the true computational bottleneck of Sinkhorn's algorithm. [3] proposed the Greenkhorn algorithm, a greedy version of Sinkhorn algorithm that selects columns and rows to be updated that most violate the constraints. [2] provided the Nys-Sink algorithm which is based on low-rank approximation of the cost matrix using Nystrom method. Other classical optimization algorithms have been considered to approximate the OT, for instance accelerated gradient descent [11, 23], quasi Newton methods [6, 10] and stochastic gradient descent [16, 1].

We give a new algorithm to approximate the regularized OT distance between discrete measures. Our algorithmic analysis is based on an approximate of the dual of Sinkhorn divergence by adding new constraints feasibility. These constraints are defined through a convex set which depends on two parameters, acting like threshold and scaling factor. We prove that dual solution of this approximation guarantees the existence of two active indices sets for the potential variables. These active sets are both directly linked to a priori fixed number budget of points from the supports of the given discrete measures. We then restrict the constraints feasibility with respect to the active sets to get a "screened" version of the dual of Sinkhorn divergence, and hence we develop the Screenkhorn algorithm.

The remainder of the paper is organized as follow. In Section 2 we briefly review the basic setup of regularized discrete OT. Section 3 contains our main contribution, that is, the Screenkhorn algorithm. Section 4 devotes to theoretical guarantees for the marginal violations of Screenkhorn. In Section 5 we present numerical results for the proposed algorithm, compared with the state-of-art Sinkhorn algorithm as implemented in [13]. The proofs of theoretical results are postponed to the supplementary material.

Notation. For any positive matrix $T \in \mathbb{R}^{n\times m}$, we define its negative entropy as $H(T) = -\sum_{i,j} T_{ij}\log(T_{ij})$. Let $r(T) = T\mathbf{1}_m \in \mathbb{R}^n$ and $c(T) = T^\top \mathbf{1}_n \in \mathbb{R}^m$ denote the rows and columns sums of T respectively. The coordinates $r_i(T)$ and $c_j(T)$ denote the i-th row sum and the j-th column sum of T, respectively. The scalar product between two matrices denotes the usual inner product, that is $\langle T, W\rangle = \mathrm{tr}(T^\top W) = \sum_{i,j} T_{ij}W_{ij}$, where $T^\top$ is the transpose of T. We write $\mathbf{1}$ (resp. $\mathbf{0}$) the vector having all coordinates equal to one (resp. zero). $\Delta(w)$ denotes the diag operator, such that if $w \in \mathbb{R}^n$, then $\Delta(w) = \mathrm{diag}(w_1, \ldots, w_n) \in \mathbb{R}^{n\times n}$. For a set of indices $L = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ satisfying $i_1 < \cdots < i_k$, we denote the complementary set of L by $L^\complement = \{1, \ldots, n\}\backslash L$. We also denote $|L|$ the cardinality of L. Given a vector $w \in \mathbb{R}^n$, we denote $w_L = (w_{i_1}, \ldots, w_{i_k})^\top \in \mathbb{R}^k$ and its complementary $w_{L^\complement} \in \mathbb{R}^{n-k}$. The notation is similar for matrices; given another subset of indices $S = \{j_1, \ldots, j_l\} \subseteq \{1, \ldots, m\}$ with $j_1 < \cdots < j_l$, and a matrix $T \in \mathbb{R}^{n\times m}$, we use $T_{(L,S)}$, to denote the submatrix of T, namely the rows and columns of $T_{(L,S)}$ are indexed by L and S respectively. When applied to matrices and vectors, $\odot$ and $\oslash$ (Hadamard product and division) and exponential notations refer to elementwise operators. Given two real numbers a and b, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

2 Regularized discrete OT

We briefly present in this section the setup of OT between two discrete measures. We then consider the case when those distributions are only available through a finite number of samples, that is $\mu = \sum_{i=1}^n \mu_i\delta_{x_i} \in \Sigma_n$ and $\nu = \sum_{j=1}^m \nu_i\delta_{y_j} \in \Sigma_m$, where Σ_n is the probability simplex with n bins, namely the set of probability vectors in $\mathbb{R}^n_+$, i.e., $\Sigma_n = \{w \in \mathbb{R}^n_+ : \sum_{i=1}^n w_i = 1\}$. We denote their probabilistic couplings set as $\Pi(\mu, \nu) = \{P \in \mathbb{R}^{n\times m}_+, P\mathbf{1}_m = \mu, P^\top\mathbf{1}_n = \nu\}$.

Sinkhorn divergence. Approximating OT distance between the two discrete measures μ and ν amounts to solving a linear problem [19] given by

$$\mathcal{S}(\mu, \nu) = \min_{P\in\Pi(\mu,\nu)} \langle C, P\rangle, \tag{1}$$

where $P = (P_{ij}) \in \mathbb{R}^{n\times m}$ is called the transportation plan, namely each entry P_{ij} represents the fraction of mass moving from x_i to y_j, and $C = (C_{ij}) \in \mathbb{R}^{n\times m}$ is a cost matrix comprised of nonnegative elements and related to the energy needed to move a probability mass from x_i to y_j. The entropic regularization of OT distances [9] relies on the addition of a penalty term as follows:

$$\mathcal{S}_\eta(\mu, \nu) = \min_{P\in\Pi(\mu,\nu)} \{\langle C, P\rangle - \eta H(P)\}, \tag{2}$$

where $\eta > 0$ is a regularization parameter. We refer to $\mathcal{S}_\eta(\mu, \nu)$ as the *Sinkhorn divergence* [9].

Dual of Sinkhorn divergence. Below we provide the derivation of the dual problem for the regularized OT problem (2). Towards this end, we begin with writing its Lagrangian dual function :

$$\mathscr{L}(P, y, z) = \langle C, P\rangle + \eta\langle \log P, P\rangle + \langle y, P\mathbf{1}_m - \mu\rangle + \langle z, P^\top\mathbf{1}_n - \nu\rangle.$$

The dual of Sinkhorn divergence can be derived by solving $\min_{P\in\mathbb{R}_+^{n\times m}} \mathscr{L}(P,y,z)$. It is easy to check that objective function $P \mapsto \mathscr{L}(P,y,z)$ is strongly convex and differentiable. Hence, one can solve the latter minimum by setting $\nabla_P \mathscr{L}(P,y,z)$ to $\mathbf{0}_{n\times m}$. Therefore, we get

$$P^\star_{ij} = \exp\Big(-\frac{1}{\eta}(y_i + z_j + C_{ij}) - 1\Big), \tag{3}$$

for all $i = 1,\dots,n$ and $j = 1,\dots,m$. Plugging this solution, and setting the change of variables $u = -y/\eta - 1/2$ and $v = -z/\eta - 1/2$, the dual problem is given by

$$\min_{u\in\mathbb{R}^n, v\in\mathbb{R}^m} \big\{\Psi(u,v) := \mathbf{1}_n^\top B(u,v)\mathbf{1}_m - \langle u,\mu\rangle - \langle v,\nu\rangle\big\}, \tag{4}$$

where $B(u,v) = \Delta(e^u)K\Delta(e^v)$ and $K = e^{-C/\eta}$ stands for the Gibbs kernel associated to the cost matrix C. We refer to problem (4) to the *dual of Sinkhorn divergence*. Therefore, the optimal solution of Sinkhorn divergence is given by $P^\star = \Delta(e^{u^\star})K\Delta(e^{v^\star})$ where the couple $(u^\star, v^\star)$ satisfies:

$$(u^\star, v^\star) = \operatorname*{argmin}_{u\in\mathbb{R}^n, v\in\mathbb{R}^m} \{\Psi(u,v)\}.$$

Note that the matrices $\Delta(e^{u^\star})$ and $\Delta(e^{v^\star})$ are unique up to a constant factor [28].

3 Screened dual of Sinkhorn divergence

For a fixed $\varepsilon > 0$ and $\kappa > 0$ we define an *approximate dual of Sinkhorn divergence* as follows:

$$\min_{u\in\mathcal{C}^n_{\frac{\varepsilon}{\kappa}}, v\in\mathcal{C}^m_{\varepsilon\kappa}} \big\{\Psi_\kappa(u,v) := \mathbf{1}_n^\top B(u,v)\mathbf{1}_m - \langle \kappa u,\mu\rangle - \langle \frac{v}{\kappa},\nu\rangle\big\}, \tag{5}$$

where $\mathcal{C}^r_\alpha \subseteq \mathbb{R}^r$, for $r\in\mathbb{N}$ and $\alpha > 0$, is a convex set given by $\mathcal{C}^r_\alpha = \{w\in\mathbb{R}^r : \min_{1\le i\le r} e^{w_i} \ge \alpha\}$. The objective function Ψ_κ is convex with respect to (u,v), then the set of optima of problem (5) is non empty. The κ-parameter plays a role of scaling factor, namely it allows to get a closed order of the potential variables e^u and e^v, while the ε-parameter acts like a threshold for e^u and e^v. Note that the approximate dual of Sinkhorn divergence coincides with the dual of Sinkhorn divergence (4) in the setting of $\varepsilon = 0$ and $\kappa = 1$. The screening procedure presented in this work is based on constructing two active sets $I_{\varepsilon,\kappa}$ and $J_{\varepsilon,\kappa}$ throughout the dual problem of (5) in the following way:

Lemma 1. *Let (u^*, v^*) be an optimal solution of the problem* (5). *Define*

$$I_{\varepsilon,\kappa} = \big\{i = 1,\dots,n : \mu_i \ge \frac{\varepsilon^2}{\kappa} r_i(K)\big\}, J_{\varepsilon,\kappa} = \big\{j = 1,\dots,m : \nu_j \ge \kappa\varepsilon^2 c_j(K)\big\} \tag{6}$$

Then one has $e^{u_i^*} = \varepsilon\kappa^{-1}$ and $e^{v_j^*} = \varepsilon\kappa$ for all $i\in I^\complement_{\varepsilon,\kappa}$ and $j \in J^\complement_{\varepsilon,\kappa}$.

First order conditions applied to (u^*, v^*) ensure that if $e^{u_i^*} > \varepsilon\kappa^{-1}$ then $e^{u_i^*}(Ke^{v^*})_i = \kappa\mu_i$ and if $e^{v_j^*} > \varepsilon\kappa$ then $e^{v_j^*}(K^\top e^{u^*})_j = \kappa^{-1}\nu_j$ which correspond to the Sinkhorn marginal conditions up to the scaling factor κ.

Screening with a fixed number budget of points. Recall that the approximate dual of Sinkhorn divergence is defined with respect to ε and κ. The explicit determination of its values depends on a priori *fixed number budget of points* from the supports of μ and ν. In the sequel of the paper, we denote by $n_b \in \{1,\dots,n\}$ and the $m_b\in\{1,\dots,m\}$ the number budget of points to be given for resolving problem (5).

Let us define $\xi\in\mathbb{R}^n$ and $\zeta\in\mathbb{R}^m$ to be the ordered decreasing vectors of $\mu \oslash r(K)$ and $\nu\oslash c(K)$ respectively, that is $\xi_1 \ge \xi_2 \ge \cdots \ge \xi_n$ and $\zeta_1 \ge \zeta_2 \ge \cdots \ge \zeta_m$. To keep only n_b-budget and m_b-budget of points, the parameters κ and ε satisfy $\varepsilon^2\kappa^{-1} = \xi_{n_b}$ and $\varepsilon^2\kappa = \zeta_{m_b}$. Hence

$$\varepsilon = (\xi_{n_b}\zeta_{m_b})^{1/4} \text{ and } \kappa = \sqrt{\frac{\zeta_{m_b}}{\xi_{n_b}}}. \tag{7}$$

Note that $|I_{\varepsilon,\kappa}| = n_b$ and $|J_{\varepsilon,\kappa}| = m_b$. Using the previous analysis, any solution (u^*, v^*) of problem (5) satisfy $e^{u_i^*} \ge \varepsilon\kappa^{-1}$ and $e^{v_j^*} \ge \varepsilon\kappa$ for all $(i,j)\in(I_{\varepsilon,\kappa}\times J_{\varepsilon,\kappa})$, and $e^{u_i^*} = \varepsilon\kappa^{-1}$ and $e^{v_j^*} = \varepsilon\kappa$ for all $(i,j)\in(I^\complement_{\varepsilon,\kappa}\times J^\complement_{\varepsilon,\kappa})$.

Basing on that facts we restrict the constraints feasibility $\mathcal{C}^n_{\frac{\varepsilon}{\kappa}} \cap \mathcal{C}^m_{\varepsilon\kappa}$ in problem (5) to the screened domain $\mathcal{U}_{\text{sc}} \cap \mathcal{V}_{\text{sc}}$ where

$$\mathcal{U}_{\text{sc}} = \{u \in \mathbb{R}^n : e^{u_{I_{\varepsilon,\kappa}}} \succeq \frac{\varepsilon}{\kappa}\mathbf{1}_{n_b}, \text{ and } e^{u_{I^{\complement}_{\varepsilon,\kappa}}} = \frac{\varepsilon}{\kappa}\mathbf{1}_{n-n_b}\},$$

and

$$\mathcal{V}_{\text{sc}} = \{v \in \mathbb{R}^m : e^{v_{J_{\varepsilon,\kappa}}} \succeq \varepsilon\kappa\mathbf{1}_{m_b}, \text{ and } e^{v_{J^{\complement}_{\varepsilon,\kappa}}} = \varepsilon\kappa\mathbf{1}_{m-m_b}\}.$$

where the vector comparison $\succeq$ has to be understood elementwise. Now, we are ready to define the *screened dual of Sinkhorn divergence* as

$$\min_{u\in\mathcal{U}_{\text{sc}}, v\in\mathcal{V}_{\text{sc}}} \{\Psi_{\varepsilon,\kappa}(u,v)\} \tag{8}$$

where

$$\begin{aligned}\Psi_{\varepsilon,\kappa}(u,v) = (e^{u_{I_{\varepsilon,\kappa}}})^\top K_{(I_{\varepsilon,\kappa},J_{\varepsilon,\kappa})} e^{v_{J_{\varepsilon,\kappa}}} + \varepsilon\kappa (e^{u_{I_{\varepsilon,\kappa}}})^\top K_{(I_{\varepsilon,\kappa},J^{\complement}_{\varepsilon,\kappa})}\mathbf{1}_{m_b} + \varepsilon\kappa^{-1}\mathbf{1}_{n_b}^\top K_{(I^{\complement}_{\varepsilon,\kappa},J_{\varepsilon,\kappa})} e^{v_{J_{\varepsilon,\kappa}}} \\ - \kappa\mu^\top_{I_{\varepsilon,\kappa}} u_{I_{\varepsilon,\kappa}} - \kappa^{-1}\nu^\top_{J_{\varepsilon,\kappa}} v_{J_{\varepsilon,\kappa}} + \Xi\end{aligned}$$

with $\Xi = \varepsilon^2 \sum_{i\in I^{\complement}_{\varepsilon,\kappa}, j\in J^{\complement}_{\varepsilon,\kappa}} K_{ij} - \kappa\log(\varepsilon\kappa^{-1})\sum_{i\in I^{\complement}_{\varepsilon,\kappa}} \mu_i - \kappa^{-1}\log(\varepsilon\kappa)\sum_{j\in J^{\complement}_{\varepsilon,\kappa}} \nu_j$.

The Screenkhorn algorithm, presented in Algorithm 1, consists of two steps: the first one is an initialization where we calculate the active sets $I_{\varepsilon,\kappa}$, $J_{\varepsilon,\kappa}$. The second is a constrained L-BFGS solver [32] for the stacked vector $\theta = (u_{I_{\varepsilon,\kappa}}, v_{J_{\varepsilon,\kappa}}) \in \mathbb{R}^{n_b\times m_b}$. It is worth to note that the couple variables (u,v) to be optimized in Screenkhorn belongs to $\mathbb{R}^{n_b\times m_b}$. Furthermore, it Screenkhorn uses only the restricted parts $K_{(I_{\varepsilon,\kappa},J_{\varepsilon,\kappa})}$, $K_{(I_{\varepsilon,\kappa},J^{\complement}_{\varepsilon,\kappa})}$, and $K_{(I^{\complement}_{\varepsilon,\kappa},J_{\varepsilon,\kappa})}$, from the Gibbs matrix K, in contrast to Sinkhorn that performs alternating updates of all rows and columns of K.

The following lemma expresses upper and lower bounds to be respected in Screenkhorn.

Lemma 2. *Let (u^{sc}, v^{sc}) be an optimal solution of problem* (8)*. Then, one has*

$$\frac{\varepsilon}{\kappa} \vee \frac{\min_{i\in I_{\varepsilon,\kappa}}\mu_i}{\varepsilon(m-m_b) + \varepsilon \vee \frac{\max_{j\in J_{\varepsilon,\kappa}}\nu_j}{n\varepsilon\kappa\min_{i,j}K_{ij}} m_b} \le e^{u_i^{sc}} \le \frac{\varepsilon}{\kappa} \vee \frac{\max_{i\in I_{\varepsilon,\kappa}}\mu_i}{m\varepsilon\min_{i,j}K_{ij}}, \tag{9}$$

and

$$\varepsilon\kappa \vee \frac{\min_{j\in J_{\varepsilon,\kappa}}\nu_j}{\varepsilon(n-n_b) + \varepsilon \vee \frac{\kappa\max_{i\in I_{\varepsilon,\kappa}}\mu_i}{m\varepsilon\min_{i,j}K_{ij}} n_b} \le e^{v_j^{sc}} \le \varepsilon\kappa \vee \frac{\max_{j\in J_{\varepsilon,\kappa}}\nu_j}{n\varepsilon\min_{i,j}K_{ij}} \tag{10}$$

for all $i \in I_{\varepsilon,\kappa}$ and $j \in J_{\varepsilon,\kappa}$.

4 Analysis of marginal violations

This section is devoted to study the marginal violations of Screenkhorn. Towards this end, let us define the screened marginals $\mu^{\text{sc}} = B(u^{\text{sc}}, v^{\text{sc}})\mathbf{1}_m$ and $\nu^{\text{sc}} = B(u^{\text{sc}}, v^{\text{sc}})^\top\mathbf{1}_n$. Lemma 3 expresses an upper bound with respect to ℓ_1-norm of $\mu^{\text{sc}}$and ν^{sc}.

Lemma 3. *Let (u^{sc}, v^{sc}) be an optimal solution of problem* (8)*. Then one has*

$$\|\mu^{sc}\|_1 \le \kappa\|\mu_{I_{\varepsilon,\kappa}}\|_1 + (n-n_b)\Big(\frac{m_b\max_{j\in J_{\varepsilon,\kappa}}\nu_j}{n\kappa\min_{i,j}K_{ij}} + (m-m_b)\varepsilon^2\Big)$$

and

$$\|\nu^{sc}\|_1 \le \kappa^{-1}\|\nu_{J_{\varepsilon,\kappa}}\|_1 + (m-m_b)\Big(\frac{n_b\kappa\max_{i\in I_{\varepsilon,\kappa}}\mu_i}{m\min_{i,j}K_{ij}} + (n-n_b)\varepsilon^2\Big).$$

The following Proposition gives also an upper bound of the marginal errors.

Proposition 1. *One has*

$$\begin{aligned}\|\mu-\mu^{sc}\|_1^2 \le & 7n_b(\kappa - \log(\kappa) - 1)\max_i \mu_i + 7(n-n_b)\Big(\frac{m_b\max_j\nu_j}{n\kappa\min_{i,j}K_{ij}} + (m-m_b)\varepsilon^2 - \min_i \mu_i \\ & + \max_i \mu_i \log\Big(\frac{\kappa(n-n_b+1)\max_i\mu_i}{m_b\min_{i,j}K_{ij}\min_{j\in J_{\varepsilon,\kappa}}\nu_j} + \frac{\kappa^2\max_i\mu_i}{mm_b\varepsilon^2(\min_{i,j}K_{ij})^2\min_{j\in J_{\varepsilon,\kappa}}\nu_j}\Big)\end{aligned}$$

Algorithm 1: Screenkhorn

input: $C, \eta, \mu \in \Sigma_n, \nu \in \Sigma_m, n_b$ and m_b;

step 1: Initialization

$\xi \leftarrow \mu \oslash r(K)$;

$\zeta \leftarrow \nu \oslash c(K)$;

$\xi \leftarrow \texttt{sort}(\xi)$; //(decreasing order)

$\zeta \leftarrow \texttt{sort}(\zeta)$; //(decreasing order)

$\varepsilon \leftarrow (\xi_{n_b}\zeta_{m_b})^{1/4},\ \kappa \leftarrow \sqrt{\zeta_{m_b}/\xi_{n_b}}$;

$I_{\varepsilon,\kappa} \leftarrow \{i = 1, \ldots, n : \mu_i \geq \varepsilon^2\kappa^{-1} r_i(K)\}$;

$J_{\varepsilon,\kappa} \leftarrow \{j = 1, \ldots, m : \nu_j \geq \varepsilon^2\kappa c_j(K)\}$;

$K_{\min} \leftarrow \min_{i \in I_{\varepsilon,\kappa}, j \in J_{\varepsilon,\kappa}} K_{ij}$;

$\underline{\mu} \leftarrow \min_{i \in I_{\varepsilon,\kappa}} \mu_i, \bar{\mu} \leftarrow \max_{i \in I_{\varepsilon,\kappa}} \mu_i$;

$\underline{\nu} \leftarrow \min_{j \in J_{\varepsilon,\kappa}} \mu_i, \bar{\nu} \leftarrow \max_{j \in J_{\varepsilon,\kappa}} \mu_i$;

$\underline{u} \leftarrow \log\big(\frac{\varepsilon}{\kappa} \vee \frac{\underline{\mu}}{\varepsilon(m-m_b) + \varepsilon \vee \frac{\bar{\nu}}{n\varepsilon\kappa K_{\min}} m_b}\big), \bar{u} \leftarrow \log\big(\frac{\varepsilon}{\kappa} \vee \frac{\bar{\mu}}{m\varepsilon K_{\min}}\big)$;

$\underline{v} \leftarrow \log\big(\varepsilon\kappa \vee \frac{\underline{\nu}}{\varepsilon(n-n_b) + \varepsilon \vee \frac{\kappa\bar{\mu}}{m\varepsilon K_{\min}} n_b}\big), \bar{v} \leftarrow \log\big(\varepsilon\kappa \vee \frac{\bar{\nu}}{n\varepsilon K_{\min}}\big)$;

$\bar{\theta} \leftarrow \text{stack}(\bar{u}\mathbf{1}_{n_b}, \bar{v}\mathbf{1}_{m_b})$;

$\underline{\theta} \leftarrow \text{stack}(\underline{u}\mathbf{1}_{n_b}, \underline{v}\mathbf{1}_{m_b})$;

step 2: L-BFGS

$u^{(0)} \leftarrow \log(\varepsilon\kappa^{-1})\mathbf{1}_{n_b}$;

$v^{(0)} \leftarrow \log(\varepsilon\kappa)\mathbf{1}_{m_b}$;

$\theta^{(0)} \leftarrow \text{stack}[u^{(0)}, v^{(0)}]$;

$\theta \leftarrow \text{L-BFGS}(\theta^{(0)}, \underline{\theta}, \bar{\theta})$;

$\theta_u \leftarrow (\theta_1, \ldots, \theta_{n_b})^\top, \theta_v \leftarrow (\theta_{n_b+1}, \ldots, \theta_{n_b+m_b})^\top$;

$u_i^{sc} \leftarrow (\theta_u)_i$ if $i \in I_{\varepsilon,\kappa}$ and $u_i \leftarrow \log(\varepsilon\kappa^{-1})$ if $i \in I^{\complement}_{\varepsilon,\kappa}$;

$v_j^{sc} \leftarrow (\theta_v)_j$ if $j \in J_{\varepsilon,\kappa}$ and $v_j \leftarrow \log(\varepsilon\kappa)$ if $j \in J^{\complement}_{\varepsilon,\kappa}$;

return $B(u^{sc}, v^{sc})$.

and

$$\begin{aligned}\|\nu - \nu^{sc}\|_1^2 \leq & 7m_b(\kappa - \log(\kappa) - 1)\max_j \nu_j + 7(m - m_b)\Big(\frac{n_b\kappa \max_i \mu_i}{n \min_{i,j} K_{ij}} + (n - n_b)\varepsilon^2 - \min_j \nu_j \\ & + \max_j \nu_j \log\Big(\frac{\kappa(m - m_b + 1)\max_j \nu_j}{n_b \min_{i,j} K_{ij} \min_{i \in I_{\varepsilon,\kappa}} \mu_i} + \frac{\kappa^2 \max_j \nu_j}{n n_b \varepsilon^2 (\min_{i,j} K_{ij})^2 \min_{i \in I_{\varepsilon,\kappa}} \mu_i}\Big)\end{aligned}$$

5 Numerical experiments

References

[1] B. K. Abid and R. Gower. Stochastic algorithms for entropy-regularized optimal transport problems. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1505–1512, Playa Blanca, Lanzarote, Canary Islands, 2018. PMLR.

[2] J. Altschuler, F. Bach, A. Rudi, and J. Weed. Massively scalable sinkhorn distances via the nyström method, 2018.

[3] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1964–1974. Curran Associates, Inc., 2017.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 2017. PMLR.

[5] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic pca in the wasserstein space by convex pca. *Ann. Inst. H. Poincaré Probab. Statist.*, 53(1):1–26, 2017.

[6] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 880–889, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[7] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Trans. Graph.*, 30(6):158:1–158:12, 2011.

[8] Villani C. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009.

[9] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

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

[11] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[12] J. Ebert, V. Spokoiny, and A. Suvorikova. Construction of non-asymptotic confidence sets in 2-wasserstein space, 2017.

[13] R. Flamary and N. Courty. Pot python optimal transport library, 2017.

[14] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

[15] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wasserstein loss. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2053–2061. Curran Associates, Inc., 2015.

[16] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

[17] N. Ho, X. L. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1501–1509. JMLR.org, 2017.

[18] 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, Apr 2008.

[19] L. Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 2:227–229, 1942.

[20] P. Knight. The sinkhorn–knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[21] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, July 2017.

[22] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France, 07–09 Jul 2015. PMLR.

[23] T. Lin, N. Ho, and M. I. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. *CoRR*, abs/1901.06482, 2019.

[24] V. M. Panaretos and Y. Zemel. Amplitude and phase variation of point processes. *Ann. Statist.*, 44(2):771–812, 04 2016.

[25] O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, 2009.

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

[27] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[28] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[29] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, 2015.

[30] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 306–314, Bejing, China, 22–24 Jun 2014. PMLR.

[31] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multidimensional histograms. *Computer Vision, Graphics, and Image Processing*, 32(3):328 – 336, 1985.

[32] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, 1997.