
A Simulated Annealing Based Inexact Oracle for Wasserstein Loss Minimization

Jianbo Ye¹ James Z. Wang¹ Jia Li²

Abstract

Learning under a Wasserstein loss, a.k.a. Wasserstein loss minimization (WLM), is an emerging research topic for gaining insights from a large set of structured objects. Despite being conceptually simple, WLM problems are computationally challenging because they involve minimizing over functions of quantities (*i.e.* Wasserstein distances) that themselves require numerical algorithms to compute. In this paper, we introduce a stochastic approach based on simulated annealing for solving WLMs. Particularly, we have developed a Gibbs sampler to approximate effectively and efficiently the partial gradients of a sequence of Wasserstein losses. Our new approach has the advantages of numerical stability and readiness for warm starts. These characteristics are valuable for WLM problems that often require multiple levels of iterations in which the oracle for computing the value and gradient of a loss function is embedded. We applied the method to optimal transport with Coulomb cost and the Wasserstein non-negative matrix factorization problem, and made comparisons with the existing method of entropy regularization.

1. Introduction

An oracle is a computational module in an optimization procedure that is applied iteratively to obtain certain characteristics of the function being optimized. Typically, it calculates the value and gradient of loss function $l(\mathbf{x}, \mathbf{y})$. In the vast majority of machine learning models, where those loss functions are decomposable along each dimension (*e.g.*, L_p norm, KL divergence, or hinge loss), $\nabla_{\mathbf{x}}l(\cdot, \mathbf{y})$ or

$\nabla_{\mathbf{y}}l(\mathbf{x}, \cdot)$ is computed in $O(m)$ time, m being the complexity of outcome variables \mathbf{x} or \mathbf{y} . This part of calculation is often negligible compared with the calculation of full gradient with respect to the model parameters. But this is no longer the case in learning problems based on Wasserstein distance due to the intrinsic complexity of the distance. We will call such problems *Wasserstein loss minimization* (WLM). Examples of WLMs include Wasserstein barycenters (Li & Wang, 2008; Agueh & Carlier, 2011; Cuturi & Doucet, 2014; Benamou et al., 2015; Ye & Li, 2014; Ye et al., 2017b), principal geodesics (Seguy & Cuturi, 2015), nonnegative matrix factorization (Rolet et al., 2016; Sandler & Lindenbaum, 2009), barycentric coordinate (Bonneel et al., 2016), and multi-label classification (Frogner et al., 2015).

Wasserstein distance is defined as the cost of matching two probability measures, originated from the literature of optimal transport (OT) (Monge, 1781). It takes into account the cross-term similarity between different support points of the distributions, a level of complexity beyond the usual vector data treatment, *i.e.*, to convert the distribution into a vector of frequencies. It has been promoted for comparing sets of vectors (*e.g.* bag-of-words models) by researchers in computer vision, multimedia and more recently natural language processing (Kusner et al., 2015; Ye et al., 2017a). However, its potential as a powerful loss function for machine learning has been underexplored. The major obstacle is a lack of standardized and robust numerical methods to solve WLMs. Even to empirically better understand the advantages of the distance is of interest.

As a long-standing consensus, solving WLMs is challenging (Cuturi & Doucet, 2014). Unlike the usual optimization in machine learning where the loss and the (partial) gradient can be calculated in linear time, these quantities are non-smooth and hard to obtain in WLMs, requiring solution of a costly network transportation problem (a.k.a. OT). The time complexity, $O(m^3 \log m)$, is prohibitively high (Orlin, 1993). In contrast to the L_p or KL counterparts, this step of calculation elevates from a negligible fraction of the overall learning problem to a dominant portion, preventing the scaling of WLMs to large data. Recently, iterative approximation techniques have been developed to compute the loss and the (partial) gradient at com-

¹College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA. ²Department of Statistics, The Pennsylvania State University, University Park, PA.. Correspondence to: Jianbo Ye <jxy198@ist.psu.edu>.

plexity $O(m^2/\varepsilon)$ (Cuturi, 2013; Wang & Banerjee, 2014). However, nontrivial algorithmic efforts are needed to incorporate these methods into WLMs because WLMs often require multi-level loops (Cuturi & Doucet, 2014; Frogner et al., 2015). Specifically, one must re-calculate through many iterations the loss and its partial gradient in order to update other model dependent parameters.

We are thus motivated to seek for a fast *inexact* oracle that (i) runs at lower time complexity per iteration, and (ii) accommodates warm starts and meaningful early stops. These two properties are equally important for efficiently obtaining adequate approximation to the solutions of a sequence of slowly changing OTs. The second property ensures that the subsequent OTs can effectively leverage the solutions of the earlier OTs so that the total computational time is low. Approximation techniques with low complexity per iteration already exist for solving a single OT, but they do not possess the second property. In this paper, we introduce a method that uses a time-inhomogeneous Gibbs sampler as an inexact oracle for Wasserstein losses. The Markov chain Monte Carlo (MCMC) based method naturally satisfies the second property, as reflected by the intuition of physicists that MCMC samples can efficiently “remix from a previous equilibrium.”

We propose a new optimization approach based on Simulated Annealing (SA) (Kirkpatrick et al., 1983; Corana et al., 1987) for WLMs where the outcome variables are treated as probability measures. SA is especially suitable for the dual OT problem, where the usual Metropolis sampler can be simplified to a Gibbs sampler. To our knowledge, existing optimization techniques used on WLMs are different from MCMC. In practice, MCMC is known to easily accommodate warm start, which is particularly useful in the context of WLMs. We name this approach *Gibbs-OT* for short. The algorithm of Gibbs-OT is as simple and efficient as the Sinkhorn’s algorithm — a widely accepted method to approximately solve OT (Cuturi, 2013). We show that Gibbs-OT enjoys improved numerical stability and several algorithmic characteristics valuable for general WLMs. By experiments, we demonstrate the effectiveness of Gibbs-OT for solving optimal transport with Coulomb cost (Benamou et al., 2016) and the Wasserstein non-negative matrix factorization (NMF) problem (Sandler & Lindenbaum, 2009; Rolet et al., 2016).

2. Related Work

Recently, several methods have been proposed to overcome the aforementioned difficulties in solving WLMs. Representatives include entropic regularization (Cuturi, 2013; Cuturi & Doucet, 2014; Benamou et al., 2015) and Bregman ADMM (Wang & Banerjee, 2014; Ye et al., 2017b). The main idea is to augment the original optimization prob-

lem with a strongly convex term such that the regularized objective becomes a smooth function of all its coordinating parameters. Neither the Sinkhorn’s algorithm nor Bregman ADMM can be readily integrated into a general WLM. Based on the entropic regularization of primal OT, Cuturi & Peyré (2016) recently showed that the Legendre transform of the entropy regularized Wasserstein loss and its gradient can be computed in closed form, which appear in the first-order condition of some complex WLM problems. Using this technique, the regularized primal problem can be converted to an equivalent Fenchel-type dual problem that has a faster numerical solver in the Euclidean space (Rolet et al., 2016). But this methodology can only be applied to a certain class of WLM problems of which the Fenchel-type dual has closed forms of objective and full gradient. In contrast, the proposed SA-based approach directly deals with the dual OT problem without assuming any particular mathematical structure of the WLM problem, and hence is more flexible to apply.

More recent approaches base on solving the dual OT problems have been proposed to calculate and optimize the Wasserstein distance between a *single pair* of distributions with very large support sets — often as large as the size of an entire machine learning dataset (Montavon et al., 2016; Genevay et al., 2016; Arjovsky et al., 2017). For these methods, scalability is achieved in terms of the support size. Our proposed method has a different focus on calculating and optimizing Wasserstein distances between *many pairs* all together in WLMs, with each distribution having a moderate support size (*e.g.*, dozens or hundreds). We aim at scalability for the scenarios when a large set of distributions have to be handled simultaneously, that is, the optimization cannot be decoupled on the distributions. In addition, existing methods have no on-the-fly mechanism to control the approximation quality at a limited number of iterations.

3. Preliminaries of Optimal Transport

In this section, we present notations, mathematical backgrounds, and set up the problem of interest.

Definition 3.1 (Optimal Transportation, OT). Let $\mathbf{p} \in \Delta_{m_1}$, $\mathbf{q} \in \Delta_{m_2}$, where Δ_m is the set of m -dimensional simplex: $\Delta_m \stackrel{\text{def.}}{=} \{\mathbf{q} \in \mathbb{R}_+^m : \langle \mathbf{q}, \mathbb{1} \rangle = 1\}$. The set of transportation plans between \mathbf{p} and \mathbf{q} is defined as $\Pi(\mathbf{p}, \mathbf{q}) \stackrel{\text{def.}}{=} \{Z \in \mathbb{R}^{m_1 \times m_2} : Z \cdot \mathbb{1}_{m_2} = \mathbf{p}; Z^T \cdot \mathbb{1}_{m_1} = \mathbf{q}\}$. Let $M \in \mathbb{R}_+^{m_1 \times m_2}$ be the matrix of costs. The optimal transport cost between \mathbf{p} and \mathbf{q} with respect to M is

$$W(\mathbf{p}, \mathbf{q}) \stackrel{\text{def.}}{=} \min_{Z \in \Pi(\mathbf{p}, \mathbf{q})} \langle Z, M \rangle. \quad (1)$$

In particular, $\Pi(\cdot, \cdot)$ is often called the coupling set.

Now we relate primal version of (discrete) OT to a variant of its dual version. One may refer to Villani (2003) for the general background of the Kantorovich-Rubenstein duality. In particular, our formulation introduces an auxiliary parameter C_M for the sake of mathematical soundness in defining Boltzmann distributions.

Definition 3.2 (Dual Formulation of OT). Let $C_M > 0$, denote vector $[g_1, \dots, g_{m_1}]^T$ by \mathbf{g} , and vector $[h_1, \dots, h_{m_2}]^T$ by \mathbf{h} . We define the dual domain of OT by

$$\Omega(M) \stackrel{\text{def.}}{=} \left\{ \mathbf{f} = [\mathbf{g}; \mathbf{h}] \in \mathbb{R}^{m_1+m_2} \mid -C_M < g_i - h_j \leq M_{i,j}, 1 \leq i \leq m_1, 1 \leq j \leq m_2 \right\}. \quad (2)$$

Informally, for a sufficiently large C_M (subject to $\mathbf{p}, \mathbf{q}, M$), the LP problem Eq. (1) can be reformulated as¹

$$W(\mathbf{p}, \mathbf{q}) = \max_{\mathbf{f} \in \Omega(M)} \langle \mathbf{p}, \mathbf{g} \rangle - \langle \mathbf{q}, \mathbf{h} \rangle. \quad (3)$$

Let the optimum set be $\Omega^*(M)$. Then any optimal point $\mathbf{f}^* = (\mathbf{g}^*, \mathbf{h}^*) \in \Omega^*(M)$ constructs a (projected) subgradient such that $\mathbf{g}^* \in \partial W / \partial \mathbf{p}$ and $-\mathbf{h}^* \in \partial W / \partial \mathbf{q}$. The main computational difficulty of WLMs comes from the fact that (projected) subgradient \mathbf{f}^* is not efficiently solvable.

Note that $\Omega(M)$ is an unbound set in $\mathbb{R}^{m_1+m_2}$. In order to constrain the feasible region to be bounded, we alternatively define

$$\Omega_0(M) = \{ \mathbf{f} = [\mathbf{g}; \mathbf{h}] \in \Omega(M) \mid g_1 = 0 \}. \quad (4)$$

One can show that the maximization in $\Omega(M)$ as Eq. (3) is equivalent to the maximization in $\Omega_0(M)$ because $\langle \mathbf{p}, \mathbb{1}_{m_1} \rangle = \langle \mathbf{q}, \mathbb{1}_{m_2} \rangle$.

4. Simulated Annealing for Optimal Transport via Gibbs Sampling

Following the basic strategy outlined in the seminal paper of simulated annealing (Kirkpatrick et al., 1983), we present the definition of Boltzmann distribution supported on $\Omega_0(M)$ below which, as we will elaborate, links the dual formulation of OT to a Gibbs sampling scheme (Algorithm 1 below).

Definition 4.1 (Boltzmann Distribution of OT). Given a temperature parameter $T > 0$, the Boltzmann distribution

¹However, for any proper M and strictly positive \mathbf{p}, \mathbf{q} , there exists C_M such that the optimal value of primal problem is equal to the optimal value of the dual problem. This modification is solely for an ad-hoc treatment of a single OT problem. In general cases of $(\mathbf{p}, \mathbf{q}, M)$, when C_M is pre-fixed, the solution of Eq. (3) may be suboptimal.

of OT is a probability measure on $\Omega_0(M) \subseteq \mathbb{R}^{m_1+m_2-1}$ such that

$$p(\mathbf{f}; \mathbf{p}, \mathbf{q}) \propto \exp \left[\frac{1}{T} (\langle \mathbf{p}, \mathbf{g} \rangle - \langle \mathbf{q}, \mathbf{h} \rangle) \right]. \quad (5)$$

It is a well-defined probability measure for an arbitrary finite $C_M > 0$.

The basic concept behind SA states that the samples from the Boltzmann distribution will eventually concentrate at the optimum set of its deriving problem (e.g. $W(\mathbf{p}, \mathbf{q})$) as $T \rightarrow 0$. However, since the Boltzmann distribution is often difficult to sample, a practical convergence rate remains mostly unsettled for specific MCMC methods.

Because $\Omega(M)$ defined by Eq. (2) (also Ω_0) has a conditional independence structure among variables, a Gibbs sampler can be naturally applied to the Boltzmann distribution defined by Eq. (5). We summarize this result below.

Proposition 4.1. Given any $\mathbf{f} = (\mathbf{g}; \mathbf{h}) \in \Omega_0(M)$ and any $C_M > 0$, we have for any i and j ,

$$g_i \leq U_i(\mathbf{h}) \stackrel{\text{def.}}{=} \min_{1 \leq j \leq m_2} (M_{i,j} + h_j), \quad (6)$$

$$h_j \geq L_j(\mathbf{g}) \stackrel{\text{def.}}{=} \max_{1 \leq i \leq m_1} (g_i - M_{i,j}). \quad (7)$$

and

$$g_i > \widehat{L}_i(\mathbf{h}) \stackrel{\text{def.}}{=} \max_{1 \leq j \leq m_2} (-C_M + h_j), \quad (8)$$

$$h_j < \widehat{U}_j(\mathbf{g}) \stackrel{\text{def.}}{=} \max_{1 \leq i \leq m_1} (C_M + g_i). \quad (9)$$

Here $U_i = U_i(\mathbf{h})$ and $L_j = L_j(\mathbf{g})$ are auxiliary variables. Suppose \mathbf{f} follows the Boltzmann distribution by Eq. (5), g_i 's are conditionally independent given \mathbf{h} , and likewise h_j 's are also conditionally independent given \mathbf{g} . Furthermore, it is immediate from Eq. (5) that each of their conditional probabilities within its feasible region (subject to C_M) satisfies

$$p(g_i | \mathbf{h}) \propto \exp \left(\frac{g_i p_i}{T} \right), \quad \widehat{L}_i(\mathbf{h}) < g_i \leq U_i(\mathbf{h}), \quad (10)$$

$$p(h_j | \mathbf{g}) \propto \exp \left(-\frac{h_j q_j}{T} \right), \quad L_j(\mathbf{g}) \leq h_j < \widehat{U}_j(\mathbf{g}), \quad (11)$$

where $2 \leq i \leq m_1$ and $1 \leq j \leq m_2$.

Remark 1. As $C_M \rightarrow +\infty$, $\widehat{U}_j(\mathbf{g}) \rightarrow +\infty$ and $\widehat{L}_i(\mathbf{h}) \rightarrow -\infty$. For $2 \leq i \leq m_1$ and $1 \leq j \leq m_2$, one can approximate the conditional probability $p(g_i | \mathbf{h})$ and $p(h_j | \mathbf{g})$ by exponential distributions.

By Proposition 4.1, our proposed time-inhomogeneous Gibbs sampler is given in Algorithm 1. Specifically in Algorithm 1, the variable g_1 is fixed to zero by the definition

Algorithm 1 Gibbs Sampling for Optimal Transport

Given $\mathbf{f}^{(0)} \in \Omega_0(M)$, $\mathbf{p} \in \Delta_{m_1}$ and $\mathbf{q} \in \Delta_{m_2}$, and $T^{(1)}, \dots, T^{(2N)} > 0$, for $t = 1, \dots, N$, we define the following Markov chain

1. Randomly sample

$$\theta_1, \dots, \theta_{m_2} \stackrel{i.i.d.}{\sim} \text{Exponential}(1).$$

For $j = 1, 2, \dots, m_2$, let

$$\begin{cases} L_j^{(t)} := \max_{1 \leq i \leq m_1} (g_i^{(t-1)} - M_{i,j}) \\ h_j^{(t)} := L_j^{(t)} + \theta_j \cdot T^{(2t-1)} / q_j \end{cases} \quad (12)$$

2. Randomly sample

$$\theta_2, \dots, \theta_{m_1} \stackrel{i.i.d.}{\sim} \text{Exponential}(1).$$

For $i = (1), 2, \dots, m_1$, let

$$\begin{cases} U_i^{(t)} := \min_{1 \leq j \leq m_2} (M_{i,j} + h_j^{(t)}) \\ g_i^{(t)} := U_i^{(t)} - \theta_i \cdot T^{(2t)} / p_i \end{cases} \quad (13)$$

of $\Omega_0(M)$. But we have found in experiments that by calculating $U_1^{(t)}$ and sampling $g_1^{(t)}$ in Algorithm 1 according to Eq. (13), one can still generate MCMC samples from $\Omega(M)$ such that the energy quantity $\langle \mathbf{p}, \mathbf{g} \rangle - \langle \mathbf{q}, \mathbf{h} \rangle$ converges to the same distribution as that of MCMC samples from $\Omega_0(M)$. Therefore, we will not assume $g_1 = 0$ from now on and develop analysis solely for the unconstrained version of Gibbs-OT.

Figure 1 illustrates the behavior of the proposed Gibbs sampler with a cooling schedule at different temperatures. As T decreases along iterations, the 95% percentile band for sample \mathbf{f} becomes thinner and thinner.

Remark 2. Algorithm 1 does not specify the actual cooling schedule, nor does the analysis of the proposed Gibbs sampler in Theorem A.2. We have been agnostic here for a reason. In the SA literature, cooling schedules with guaranteed optimality are often too slow to be useful in practice. To our knowledge, the guaranteed rate of SA approach is worse than the combinatorial solver for OT. As a result, a well-accepted practice of SA for many complicated optimization problems is to empirically adjust cooling schedules, a strategy we take for our experiments.

Remark 3. Although the exact cooling schedule is not specified, we still provide a quantitative upper bound of the chosen temperature T at different iterations in Appendix A Eq. (24). One can calculate such bound at the cost of $m \log m$ at certain iterations to check whether the current temperature is too high for the used Gibbs-OT to accurately

approximate the Wasserstein gradient. In practice, we find this bound helps one quickly select the beginning temperature of Gibbs-OT algorithm.

Definition 4.2 (Notations for Auxiliary Statistics). Besides the Gibbs coordinates \mathbf{g} and \mathbf{h} , the Gibbs-OT sampler naturally introduces two auxiliary variables, \mathbf{U} and \mathbf{L} . Let $\mathbf{L}^{(t)} = [L_1^{(t)}, \dots, L_{m_2}^{(t)}]^T$ and $\mathbf{U}^{(t)} = [U_1^{(t)}, \dots, U_{m_1}^{(t)}]^T$. Likewise, denote the collection of $g_i^{(t)}$ and $h_j^{(t)}$ by vectors $\mathbf{g}^{(t)}$ and $\mathbf{h}^{(t)}$ respectively. The following sequence of auxiliary statistics

$$\begin{aligned} & [\dots, \mathbf{z}^{2t-1}, \mathbf{z}^{2t}, \mathbf{z}^{2t+1}, \dots] \stackrel{\text{def.}}{=} \\ & \left[\dots, \begin{bmatrix} \mathbf{L}^{(t)} \\ \mathbf{U}^{(t-1)} \end{bmatrix}, \begin{bmatrix} \mathbf{L}^{(t)} \\ \mathbf{U}^{(t)} \end{bmatrix}, \begin{bmatrix} \mathbf{L}^{(t+1)} \\ \mathbf{U}^{(t)} \end{bmatrix}, \dots \right] \end{aligned} \quad (14)$$

for $t = 1, \dots, N$ is also a Markov chain. They can be re-defined equivalently by specifying the transition probabilities $p(\mathbf{z}^{n+1} | \mathbf{z}^n)$ for $n = 1, \dots, 2N$, a.k.a., the conditional p.d.f. $p(\mathbf{U}^{(t)} | \mathbf{L}^{(t)})$ for $t = 1, \dots, N$ and $p(\mathbf{L}^{(t+1)} | \mathbf{U}^{(t)})$ for $t = 1, \dots, N - 1$.

One may notice that the alternative representation converts the Gibbs sampler to one whose structure is similar to a hidden Markov model, where the \mathbf{g}, \mathbf{h} chain is conditional independent given the \mathbf{U}, \mathbf{L} chain and has (factored) exponential emission distributions. We will use this equivalent representation in Appendix A and develop analysis based on the \mathbf{U}, \mathbf{L} chain accordingly.

Remark 4. We now consider the function

$$V(\mathbf{x}, \mathbf{y}) \stackrel{\text{def.}}{=} \langle \mathbf{p}, \mathbf{x} \rangle - \langle \mathbf{q}, \mathbf{y} \rangle,$$

and define a few additional notations. Let $V(\mathbf{U}^{t'}, \mathbf{L}^t)$ be denoted by $V(\mathbf{z}^{t+t'})$, where $t' = t$ or $t - 1$. If \mathbf{g}, \mathbf{h} are independently resampled according to Eq. (12) and (13), we will have the inequalities that

$$\mathbb{E}[V(\mathbf{g}, \mathbf{h}) | \mathbf{z}^n] \leq V(\mathbf{z}^n).$$

Both $V(\mathbf{z})$ and $V(\mathbf{g}, \mathbf{h})$ converges to the exact loss $W(\mathbf{p}, \mathbf{q})$ at the equilibrium of Boltzmann distribution $p(\mathbf{f}; \mathbf{p}, \mathbf{q})$ as $T \rightarrow 0$.²

5. Gibbs-OT: An Inexact Oracle for WLMs

In this section, we introduce a non-standard SA approach for the general WLM problems. The main idea is to replace the standard Boltzmann energy with an asymptotic consistent upper bound, outlined in our previous section. Let

$$\mathfrak{R}(\theta) := \sum_{i=1}^{|\mathcal{D}|} W(\mathbf{p}_i(\theta), \mathbf{q}_i(\theta))$$

²The conditional quantity $V(\mathbf{z}^n) - V(\mathbf{g}, \mathbf{h}) | \mathbf{z}^n$ is the sum of two Gamma random variables: $\text{Gamma}(m_1, 1/T^{(2t)}) + \text{Gamma}(m_2, 1/T^{(2t'+1)})$ where $t' = t$ or $t' = t - 1$.

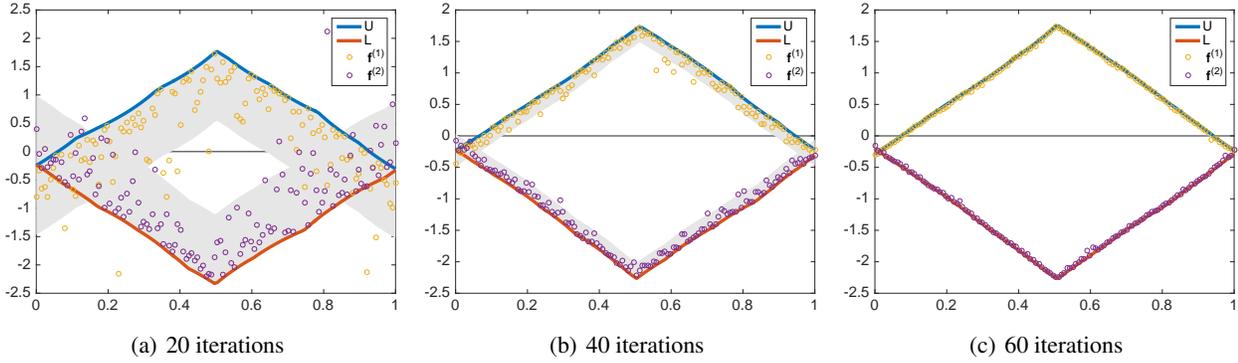


Figure 1. The Gibbs sampling of the proposed SA method. From left to right is an illustrative example of a simple 1D optimal transportation problem with Coulomb cost and plots of variables for solving this problem at different number of iterations $\in \{20, 40, 60\}$ using the inhomogeneous Gibbs sampler. Particularly, the 95% percentile of the exponential distributions are marked by the gray area.

be our prototyped objective function, where \mathcal{D} represents a dataset, $\mathbf{p}_i, \mathbf{q}_i$ are prototyped probability densities for representing the i -th instance. We now discuss how to solve $\min_{\theta \in \Theta} \mathfrak{R}(\theta)$.

To minimize the Wasserstein losses $W(\mathbf{p}, \mathbf{q})$ approximately in such WLMs, we propose to instead optimize its asymptotic consistent upper bound $\mathbb{E}[V(\mathbf{z})]$ at equilibrium of Boltzmann distribution $p(\mathbf{f}; \mathbf{p}, \mathbf{q})$ using its stochastic gradients: $\mathbf{U} \in \partial V(\mathbf{z})/\partial \mathbf{p}$ and $-\mathbf{L} \in \partial V(\mathbf{z})/\partial \mathbf{q}$. Therefore, one can calculate the gradient approximately:

$$\nabla_{\theta} \mathfrak{R} \approx \sum_{i=1}^{|\mathcal{D}|} [J_{\theta}(\mathbf{p}_i(\theta)) \mathbf{U}_i - J_{\theta}(\mathbf{q}_i(\theta)) \mathbf{L}_i]$$

where $J_{\theta}(\cdot)$ is the Jacobian, $\mathbf{U}_i, \mathbf{L}_i$ are computed from Algorithm 1 for the problem $W(\mathbf{p}_i, \mathbf{q}_i)$ respectively. Together with the iterative updates of model parameters θ , one gradually anneals the temperature T . The equilibrium of $p(\mathbf{f}; \mathbf{p}, \mathbf{q})$ becomes more and more concentrated. We assume the inexact oracle at a relatively higher temperature is adequate for early updates of the model parameters, but sooner or later it becomes necessary to set T smaller to better approximate the exact loss.

It is well known that the variance of stochastic gradient usually affects the rate of convergence. The reason to replace $V(\mathbf{g}, \mathbf{h})$ with $V(\mathbf{z})$ as the inexact oracle (for some $T > 0$) is motivated by the same intuition. The variances of MCMC samples $g_i^{(t)}, h_j^{(t)}$ of Algorithm 1 can be very large if p_i/T and q_j/T are small, making the embedded first-order method inaccurate unavoidably. But we find the variances of max/min statistics $U_i^{(t)}, L_j^{(t)}$ are much smaller. Fig. 1 shows an example. The bias introduced in the replacement is also well controlled by decreasing the temperature parameter T . For the sake of efficiency, we use a very simple convergence diagnostics in the practice of Gibbs-OT. We check the values of $V(\mathbf{z}^{(2t)})$ such that the

Markov chain is roughly considered mixed if every τ iteration the quantity $V(\mathbf{z}^{(2t)})$ (almost) stops increasing ($\tau=5$ by default), say, for some t ,

$$V(\mathbf{z}^{(2t)}) - V(\mathbf{z}^{(2(t-\tau))}) < 0.01\tau T \cdot V(\mathbf{z}^{(2t)}),$$

we terminate the Gibbs iterations.

6. Applications of Gibbs-OT

6.1. Toy OT Examples

1D Case with Euclidean Cost. We first illustrate the differences between the approximate primal solutions computed by different methods by replicating a toy example in (Benamou et al., 2015). The toy example calculates the OT between two 1D two-mode distributions. We visualize their solved coupling as a 2D image in Fig. 2 at the budgets in terms of different number of iterations. Given their different convergence behaviors, when one wants to compromise with using pre-converged primal solutions in WLMs, he or she has to account for the different results computed by different numerical methods, even though they all aim at the Wasserstein loss.

As a note, Sinkhorn, B-ADMM and Gibbs-OT share the same computational complexity per iteration. The difference in their actual CPU time comes from the different arithmetic operations used. B-ADMM may be the slowest because it requires $\log(\cdot)$ and $\exp(\cdot)$ operations. When memory efficiency is of concern, both the implementations of Sinkhorn and Gibbs-OT can be modified to take only $O(m_1 + m_2)$ additional memory besides the space for caching the cost matrix M .

Two Electrons with Coulomb Cost in DFT. In quantum mechanics, Coulomb cost (or electron-electron Coulomb repulsion) is an important energy functional in Density Functional Theory (DFT). Numerical methods that solve

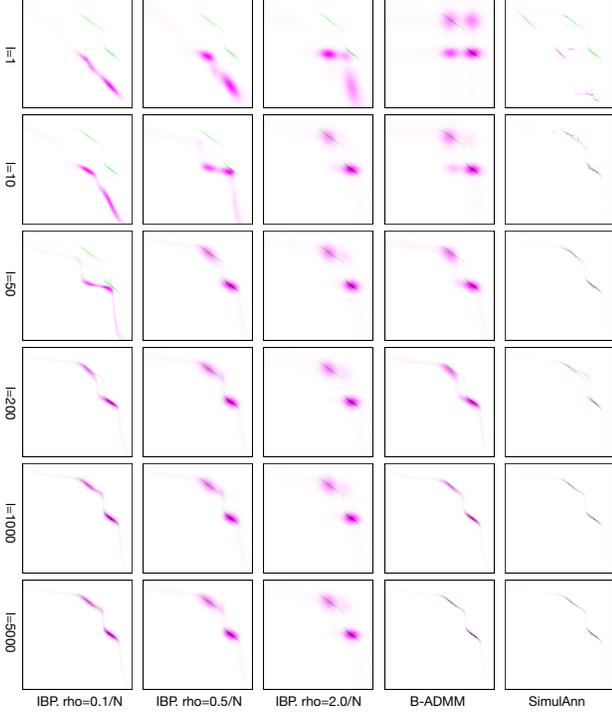


Figure 2. A simple example for OT between two 1D distribution: The solutions by Iterative Bregman Projection, B-ADMM, and Gibbs-OT are shown in pink, while the exact solution by linear programming is shown in green. Images in the rows from top to bottom present results at different iterations $\{1, 10, 50, 200, 1000, 5000\}$; The left three columns are by IBP with $\varepsilon = \{0.1/N, 0.5/N, 2/N\}$, where $[0, 1]$ is discretized with $N = 128$ uniformly spaced points. The fourth column is by B-ADMM (with default parameter $\tau_0 = 2.0$). The last column is the proposed Gibbs-OT, with a geometric cooling schedule. With a properly selected cooling schedule, one can achieve fast convergence of OT solution without comprising much solution quality.

the multi-marginal OT problem with unbounded costs remains an open challenge in DFT (Benamou et al., 2016). We consider two uniform densities on 1D domain $[0, 1]$ with Coulomb cost $c(x, y) = 1/|x - y|$ which has analytic solutions. Coulomb cost is different from the usual metric cost in the OT literature, which is unbounded and singular at $x = y$. As observed in (Benamou et al., 2016), the entropic regularized primal solution becomes more concentrated at boundaries, which is not physically plausible. This effect is not observed in the Gibbs-OT solution as shown in Appendix Fig. 3. As shown by Fig 1, the variables \mathbf{U}, \mathbf{V} in computation are always in bounded range (with an overwhelming probability), thus the algorithm does not endure any numerical difficulties.

For entropic regularization (Benamou et al., 2015; 2016), we empirically select the minimal ε which does not cause numerical overflow before 5000 iterations (in which $\varepsilon =$

$0.5/N$). For Gibbs-OT, we use a geometric temperature scheme such that $T = 2.0(1/l^4)^{n/l}/N$ at the n -th iteration, where l is the max iteration number. For the unbounded Coulomb cost, Bregman ADMM (Wang & Banerjee, 2014) does not converge to a solution close to the true optimum.

6.2. Wasserstein NMF

We now illustrate how the proposed Gibbs-OT can be used as a ready-to-plugin inexact oracle for a typical WLM — Wasserstein NMF (Sandler & Lindenbaum, 2009; Rolet et al., 2016). The data parallelization of this framework is natural because the Gibbs-OT samplers subject to different instances are independent.

Problem Formulation. Given a set of discrete probability measures $\{\Phi_i\}_{i=1}^n$ (data) over \mathbb{R}^d , we want to estimate a model $\Theta = \{\Psi_k\}_{k=1}^K$, such that for each Φ_i , there exists a membership vector $\beta^{(i)} \in \Delta_K$: $\Phi_i \approx \sum_{k=1}^K \beta_k^{(i)} \Psi_k$, where each Ψ_k is again a discrete probability measure to be estimated. Therefore, Wasserstein NMF reads $\min_{\Theta, \Xi} \sum_{i=1}^n W(\Phi_i, \sum_{k=1}^K \beta_k^{(i)} \Psi_k)$, where $\Xi = (\beta^{(1)}, \dots, \beta^{(n)})$ is the collection of membership vectors, and W is the Wasserstein distance. One can write the problem by plugging Eq. (3) in the dual formulation:

$$\min_{\Theta, \Xi} \max_{F=\{\mathbf{f}_i\}_{i=1}^n} \sum_{i=1}^n \left[\langle \widehat{\mathbf{w}}^{(i)}, \mathbf{g}_i \rangle - \langle \mathbf{w}^{(i)}, \mathbf{h}_i \rangle \right] \quad (15)$$

$$\text{s.t. } \Psi_k = \sum_{i=1}^m v_i^{(k)} \delta_{\mathbf{x}_i}, \quad (16)$$

$$\widehat{\Phi}^{(i)} = \sum_{k=1}^K \beta_k^{(i)} \Psi_k, \quad (17)$$

$$\mathbf{f}_i \in \Omega \left(M(\widehat{\Phi}^{(i)}, \Phi_i) \right), \quad (18)$$

where $\widehat{\mathbf{w}}^{(i)} \in \Delta_m$ is the weight vector of discrete probability measure $\widehat{\Phi}^{(i)}$ and $\mathbf{w}^{(i)} \in \Delta_{m_i}$ is the weight vector of $\Phi^{(i)}$. $M(\cdot, \cdot)$ denotes the transportation cost matrix between the supports of two measures. The global optimization solves all three sets of variables (Θ, Ξ, F) . In the sequel, we assume support points of $\{\Psi_k\}_{k=1}^m - \{\mathbf{x}_i\}_{i=1}^m$ are shared and pre-fixed.

Algorithm. At every epoch, one updates variables either sequentially (indexed by i) or all together. It is done by first executing the Gibbs-OT oracle subject to the i -th instance and then updating $\mathbf{v}^{(k)}$ and the membership vector $\beta^{(i)}$ accordingly at a chosen step size $\gamma > 0$. At the end of each epoch, the temperature parameter T is adjusted $T := T \left(1 - \sqrt{\frac{1}{m+\bar{m}}} \right)$, where $\bar{m} = \frac{1}{n} \sum_{i=1}^n m_i$. For each instance i , the algorithm proceeds with the following steps iteratively:

1. Initiate from the last computed \mathbf{U}/\mathbf{V} sample subject to instance i , execute the Gibbs-OT Gibbs sampler at

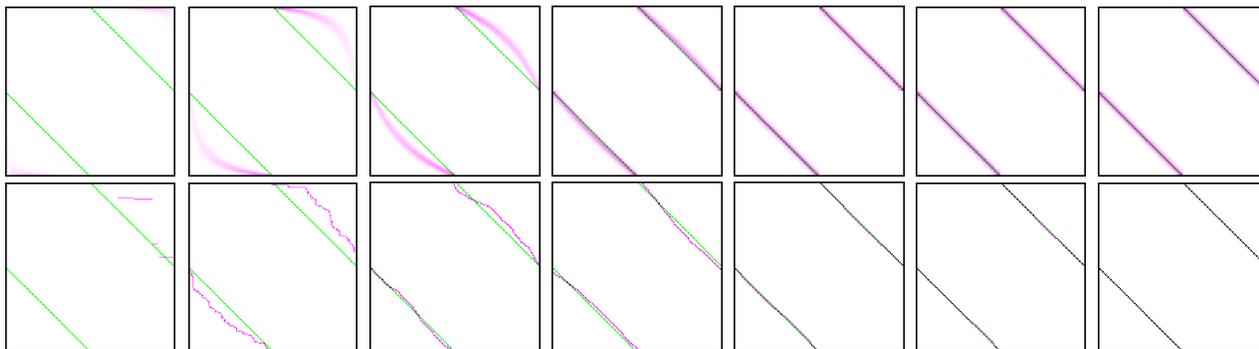


Figure 3. The recovered primal solutions for two uniform 1D distribution with Coulomb cost. The approximate solutions are shown in pink, while the exact solution by linear programming is shown in green. Top row: entropic regularization with $\varepsilon = 0.5/N$. Bottom row: Gibbs-OT. Images in the rows from left to right present results at different max iterations $\{1, 10, 50, 200, 1000, 2000, 5000\}$.

constant temperature T until a mixing criterion is met, and get \mathbf{U}_i .

2. For $k = 1, \dots, K$, update $\mathbf{v}^{(k)} \in \Delta_m$ based on gradient $\beta_k^{(i)} \mathbf{U}_i$ using the iterates of online mirror descent (MD) subject to the step-size γ (Beck & Teboulle, 2003).
3. Also update the membership vector $\beta^{(i)} \in \Delta_K$ based on gradient $(\langle \mathbf{v}^{(1)}, \mathbf{U}_i \rangle, \dots, \langle \mathbf{v}^{(K)}, \mathbf{U}_i \rangle)^T$ using the iterates of accelerated mirror descent (AMD) with restarts subject to the same step-size γ (Krichene et al., 2015).

We note that the practical speed-ups we achieved via the above procedure is the warm-start feature in Step 1. If one uses a black-box OT solver, this dimension of speed-ups is not viable.

Results. We investigate the empirical convergence of the proposed Wasserstein NMF method by two datasets: one is a subset of MNIST handwritten digit images which contains 200 digits of “5”, and the other is the ORL 400-face dataset. Our results are based on a C/C++ implementation with vectorization. In particular, we set $K = 40, \gamma = 2.0$ for both datasets. The learned components are visualized together with alternative approaches (smoothed W-NMF (Rolet et al., 2016) and regular NMF) in Appendix Figs. 4 and 5. From these figures, we observe that our learned components using Gibbs-OT are shaper than the smoothed W-NMF. This can be explained by the fact that Gibbs-OT can potentially push for higher quality of approximation by gradually annealing the temperature. We also observe that the learned components might possess some salt-and-pepper noise. This is because the Wasserstein distance by definition is not very sensitive to the sub-pixel displacements. On a single-core of a 3.3 GHz Intel

Core i5 CPU, the average time spent for each epoch for these two datasets are 0.84 seconds and 16.8 seconds, respectively. It is about two magnitude faster than fully solving all OTs via a commercial LP solver³.

7. Discussions

The solution of primal OT (Monge-Kantorovich problem) have many direct interpretations, where the solved transport is a coupling between two measures. Hence, it could be well motivated to consider regularizing the solution on the primal domain in those problems (Cuturi, 2013). Meanwhile, the solution of dual OT can be meaningful in its own right. For instance, in finance, the dual solution is directly interpreted as the vanilla prices implementing robust static super-hedging strategies. The entropy regularized OT, under the Fenchel-type dual, provides a smoothed unconstrained dual problem as shown in (Cuturi & Peyré, 2016). In this paper, we develop Gibbs-OT, whose solutions respect the dual feasibility of OT and are subject to a different regularization effect as explained by (Abernethy & Hazan, 2015). It is a numerical stable and computational suitable oracle to handle WLM.

Acknowledgement. This material is based upon work supported by the National Science Foundation under Grant Nos. ECCS-1462230 and DMS-1521092. The authors would also like to thank anonymous reviewers for their valuable comments.

³We use the specialized network flow solver in Mosek (<https://www.mosek.com>) for the computation, which is found faster than general simplex or IPM solver at moderate problem scale.

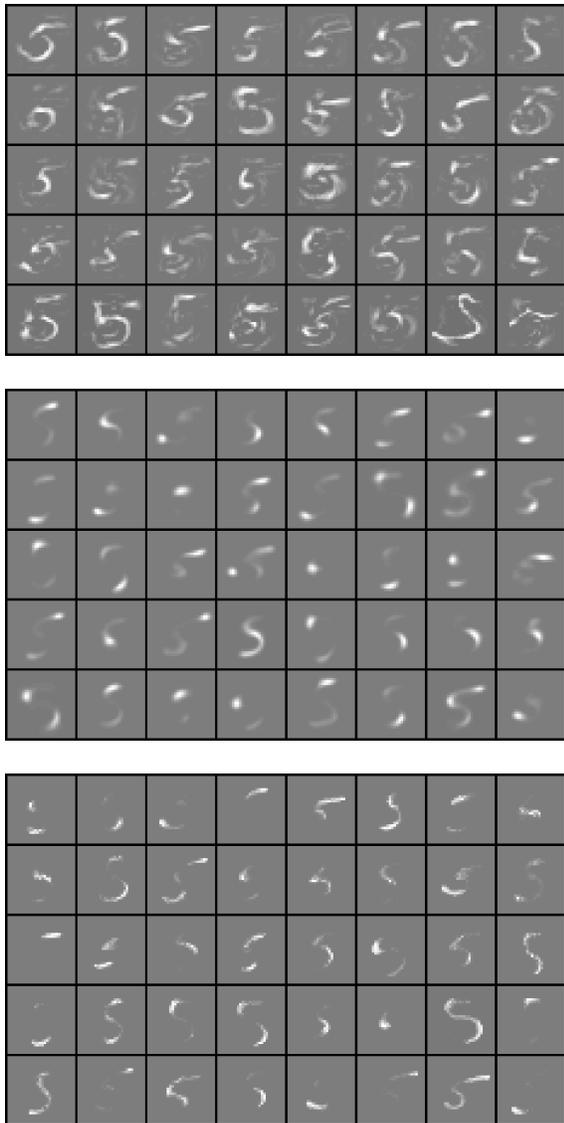


Figure 4. NMF components learned by different methods ($K = 40$) on the 200 digit “5” images. Top: regular NMF; Middle: W-NMF with entropic regularization ($\varepsilon = 1/100$, $\rho_1 = \rho_2 = 1/200$); Bottom: W-NMF using Gibbs-OT. It is observed that the components of W-NMF with entropic regularization are smoother than those optimized with Gibbs-OT.

References

- Abernethy, Jacob and Hazan, Elad. Faster convex optimization: Simulated annealing with an efficient universal barrier. *arXiv preprint arXiv:1507.02528*, 2015.
- Agueh, Martial and Carlier, Guillaume. Barycenters in the Wasserstein space. *SIAM J. Math. Analysis*, 43(2):904–924, 2011.
- Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon.

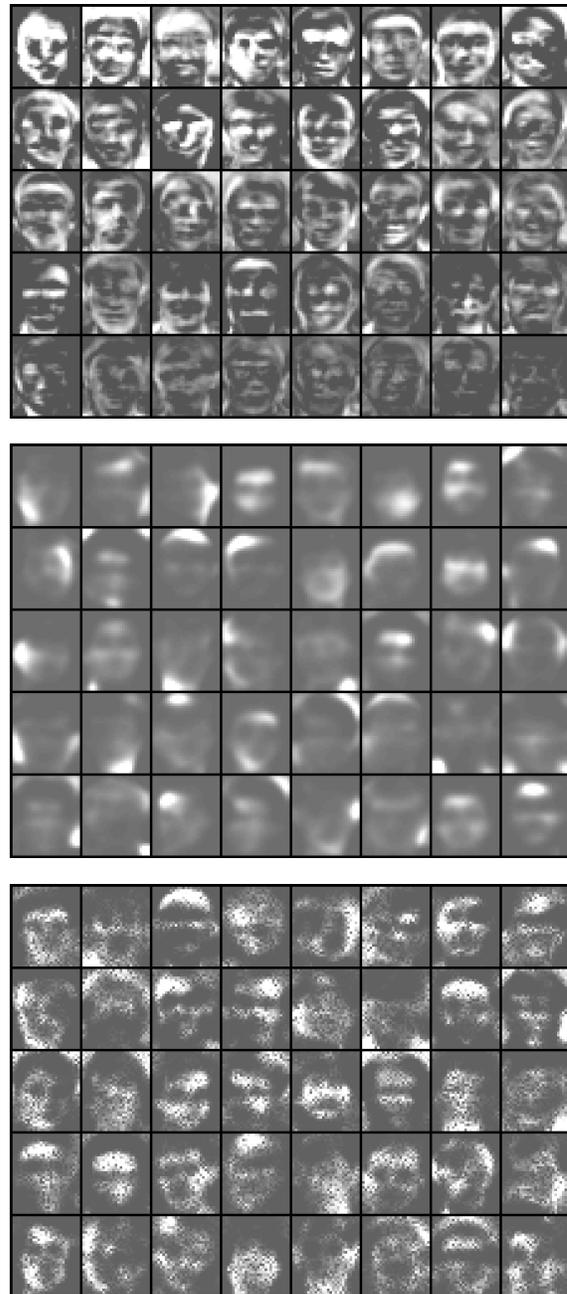


Figure 5. NMF components learned by different methods ($K = 40$) on the ORL face images. Top: regular NMF; Middle: W-NMF with entropic regularization ($\varepsilon = 1/100$, $\rho_1 = \rho_2 = 1/200$); Bottom: W-NMF using Gibbs-OT, in which the salt and pepper noises are observed due to the fact that Wasserstein distance is insensitive to the subpixel mass displacement (Cuturi & Peyré, 2016).

Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Beck, Amir and Teboulle, Marc. Mirror descent and non-

- linear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- Benamou, Jean-David, Carlier, Guillaume, Cuturi, Marco, Nenna, Luca, and Peyré, Gabriel. Iterative Bregman projections for regularized transportation problems. *SIAM J. on Scientific Computing*, 37(2):A1111–A1138, 2015.
- Benamou, Jean-David, Carlier, Guillaume, and Nenna, Luca. A numerical method to solve multi-marginal optimal transport problems with Coulomb cost. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pp. 577–601. Springer, 2016.
- Bonneel, Nicolas, Peyré, Gabriel, and Cuturi, Marco. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Trans. on Graphics*, 35(4), 2016.
- Corana, Angelo, Marchesi, Michele, Martini, Claudio, and Ridella, Sandro. Minimizing multimodal functions of continuous variables with the simulated annealing algorithm corrigenda for this article is available here. *ACM Trans. on Mathematical Software*, 13(3):262–280, 1987.
- Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.
- Cuturi, Marco and Doucet, Arnaud. Fast computation of Wasserstein barycenters. In *Proc. Int. Conf. Machine Learning*, pp. 685–693, 2014.
- Cuturi, Marco and Peyré, Gabriel. A smoothed dual approach for variational Wasserstein problems. *SIAM J. on Imaging Sciences*, 9(1):320–343, 2016.
- Frogner, Charlie, Zhang, Chiyuan, Mobahi, Hossein, Araya, Mauricio, and Poggio, Tomaso A. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2044–2052, 2015.
- Genevay, Aude, Cuturi, Marco, Peyré, Gabriel, and Bach, Francis. Stochastic optimization for large-scale optimal transport. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3440–3448. 2016.
- Kirkpatrick, Scott, Gelatt, C. Daniel, Jr, and Vecchi, Mario P. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- Krichene, Walid, Bayen, Alexandre, and Bartlett, Peter L. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, pp. 2827–2835, 2015.
- Kusner, Matt, Sun, Yu, Kolkin, Nicholas, and Weinberger, Kilian. From word embeddings to document distances. In *Proc. of the Int. Conf. on Machine Learning*, pp. 957–966, 2015.
- Li, Jia and Wang, James Z. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.
- Monge, Gaspard. *Mémoire sur la théorie des déblais et des remblais*. De l’Imprimerie Royale, 1781.
- Montavon, Grégoire, Müller, Klaus-Robert, and Cuturi, Marco. Wasserstein training of restricted boltzmann machines. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3711–3719. 2016.
- Orlin, James B. A faster strongly polynomial minimum cost flow algorithm. *Operations Research*, 41(2):338–350, 1993.
- Rolet, Antoine, Cuturi, Marco, and Peyré, Gabriel. Fast dictionary learning with a smoothed Wasserstein loss. In *AISTAT*, 2016.
- Sandler, Roman and Lindenbaum, Michael. Nonnegative matrix factorization with earth mover’s distance metric. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, pp. 1873–1880. IEEE, 2009.
- Seguy, Vivien and Cuturi, Marco. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pp. 3294–3302, 2015.
- Villani, Cédric. *Topics in Optimal Transportation*. Number 58. American Mathematical Soc., 2003.
- Wang, Huahua and Banerjee, Arindam. Bregman alternating direction method of multipliers. In *Advances in Neural Information Processing Systems*, pp. 2816–2824, 2014.
- Ye, Jianbo and Li, Jia. Scaling up discrete distribution clustering using admm. In *Proc. of the Int. Conf. on Image Processing*, pp. 5267–5271. IEEE, 2014.
- Ye, Jianbo, Li, Yanran, Wu, Zhaohui, Wang, James Z, Li, Wenjie, and Li, Jia. Determining gains acquired from word embedding quantitatively using discrete distribution clustering. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 2017a.
- Ye, Jianbo, Wu, Panruo, Wang, James Z, and Li, Jia. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Trans. on Signal Processing*, 65(9):2317–2332, 2017b.

A. Theoretical Properties of Gibbs-OT

We develop quantitative concentration bounds for Gibbs-OT in a finite number of iterations in order to understand the relationship between the temperature schedule and the concentration progress. The analysis also guides us to adjust cooling schedule on-the-fly, as will be shown. Proofs are provided in Supplement.

Preliminaries. Before characterizing the properties of Gibbs-OT by Definition 1, we first give the analytic expression for $p(\mathbf{z}^{n+1}|\mathbf{z}^n)$. Let $G(\cdot) : [-\infty, \infty] \mapsto [0, 1]$ be the c.d.f. of standard exponential distribution. Because $L_j^{(t+1)} < x$ by definition $\Leftrightarrow \forall i, g_i^{(t)} - M_{i,j} < x$, the c.d.f. of $L_j^{(t+1)}|\mathbf{U}^{(t)}$ reads

$$\Pr\left(L_j^{(t+1)} < x \mid \mathbf{U}^{(t)}\right) = \prod_{i=1}^{m_1} \left(1 - G\left(\frac{-x - M_{i,j} + U_i^{(t)}}{T^{(2t)}/p_i}\right)\right).$$

Likewise, the c.d.f. of $U_i^{(t)}|\mathbf{L}^{(t)}$ reads

$$\Pr\left(U_i^{(t)} < x \mid \mathbf{L}^{(t)}\right) = \prod_{j=1}^{m_2} G\left(\frac{x - M_{i,j} - L_j^{(t)}}{T^{(2t-1)}/q_j}\right).$$

With some calculation, the following can be shown. As a note, this lemma provides an intermediate result whose main purpose is to lay down the definition of $\phi_j^{(t)}$ and $\varphi_i^{(t)}$, which are then used in defining $O(z, T)$ (Eq. (21)) and r^n (Eq. (23)) and in Theorem A.2.

Lemma A.1. (i) Given $1 \leq j \leq m_2$ and $1 \leq t \leq N$, let the sorted index of $\{U_i^{(t)} - M_{i,j}\}_{i=1}^{m_1}$ be permutation $\{\sigma(i)\}_{i=1}^{m_1}$ such that sequence $\{U_{\sigma(i)}^{(t)} - M_{\sigma(i),j}\}_{i=1}^{m_1}$ are monotonically non-increasing. Define the auxiliary quantity

$$\phi_j^{(t)} \stackrel{\text{def.}}{=} \sum_{k=1}^{m_1} \frac{(1 - \mu_k) \prod_{i=1}^{k-1} \mu_i}{\sum_{i=1}^k p_{\sigma(i)}}, \quad (19)$$

where

$$1 \geq \mu_i \stackrel{\text{def.}}{=} \exp\left\{\frac{\sum_{i=1}^k p_{\sigma(i)}}{T^{(2t)}} \left[(U_{\sigma(i+1)} - M_{\sigma(i+1),j}) - (U_{\sigma(i)} - M_{\sigma(i),j}) \right]\right\}$$

for $i = 1, \dots, m_1 - 1$, and $\mu_{m_1} \stackrel{\text{def.}}{=} 0$. Then, the conditional expectation

$$\mathbb{E}\left[L_j^{(t+1)} \mid \mathbf{U}^{(t)}\right] = U_{\sigma(1)}^{(t)} - M_{\sigma(1),j} - \phi_j^{(t)} T^{(2t)}.$$

In particular, we denote $\sigma(1)$ by I_j^t or $I(j, t)$.

(ii) Given $1 \leq i \leq m_1$ and $1 \leq t \leq N$, let the sorted index of $\{M_{i,j} + L_j\}_{j=1}^{m_2}$ be permutation $\{\sigma(j)\}_{j=1}^{m_2}$ such

that the sequence $\{M_{i,\sigma(j)} + L_{\sigma(j)}^{(t)}\}_{j=1}^{m_2}$ are monotonically non-decreasing. Define the auxiliary quantity

$$\psi_i^{(t)} \stackrel{\text{def.}}{=} \sum_{k=1}^{m_2} \frac{(1 - \lambda_k) \prod_{j=1}^{k-1} \lambda_k}{\sum_{j=1}^k q_{\sigma(j)}}, \quad (20)$$

where

$$1 \geq \lambda_j \stackrel{\text{def.}}{=} \exp\left\{\frac{\sum_{j=1}^k q_{\sigma(j)}}{T^{(2t-1)}} \left[(M_{i,\sigma(j)} + L_{\sigma(j)}^{(t)}) - (M_{i,\sigma(j+1)} + L_{\sigma(j+1)}^{(t+1)}) \right]\right\}$$

for $i = 1, \dots, m_2 - 1$ and $\lambda_{m_2} = 0$. Then, the conditional expectation

$$\mathbb{E}\left[U_i^{(t)} \mid \mathbf{L}^{(t)}\right] = M_{i,\sigma(1)} + L_{\sigma(1)}^{(t)} + \psi_i^{(t)} T^{(2t-1)}.$$

In particular, we denote $\sigma(1)$ by J_i^t or $J(i, t)$.

We note that the calculation of Eq. (19) and Eq. (20) needs $O(m_1 \log m_1)$ and $O(m_2 \log m_2)$ time respectively. By a few additional calculations, we introduce the notation $\mathcal{O}(\cdot, \cdot)$:

$$\begin{aligned} & \mathcal{O}(\mathbf{z}^{2t}, T^{(2t)}) \\ & \stackrel{\text{def.}}{=} \mathbb{E}\left[\langle \mathbf{q}, \mathbf{L}^{(t)} \rangle - \langle \mathbf{q}, \mathbf{L}^{(t+1)} \rangle \mid \mathbf{U}^{(t)}, \mathbf{L}^{(t)}\right] \\ & = \sum_{j=1}^{m_2} \left(M_{I_j^t, j} + L_j^{(t)} - U_{I_j^t}^{(t)} + \phi_j^{(t)} T^{(2t)} \right) q_j \\ & \mathcal{O}(\mathbf{z}^{2t-1}, T^{(2t-1)}) \\ & \stackrel{\text{def.}}{=} \mathbb{E}\left[\langle \mathbf{p}, \mathbf{U}^{(t)} \rangle - \langle \mathbf{p}, \mathbf{U}^{(t-1)} \rangle \mid \mathbf{U}^{(t-1)}, \mathbf{L}^{(t)}\right] \\ & = \sum_{i=1}^{m_1} \left(M_{i, J_i^t} + L_{J_i^t}^{(t)} - U_i^{(t-1)} + \psi_i^{(t)} T^{(2t-1)} \right) p_i \end{aligned} \quad (21)$$

Note that $\mathcal{O}(\mathbf{z}^n, T^n) = \mathbb{E}\left[V(\mathbf{z}^{n+1}) - V(\mathbf{z}^n) \mid \mathbf{z}^n\right]$.

Recovery of Approximate Primal Solution. An approximate $(m_1 + m_2)$ -sparse primal solution⁴ can be recovered from \mathbf{z}^n at $n = 2t$ by

$$\begin{aligned} Z & \approx \frac{1}{2} \text{sparse}(1 : m_1, J(1 : m_1, t), \mathbf{p}) + \\ & \frac{1}{2} \text{sparse}(I(1 : m_2, t), 1 : m_2, \mathbf{q}) \in \mathbb{R}^{m_1 \times m_2}. \end{aligned} \quad (22)$$

Concentration Bounds. We are interested in the concentration bound related to $V(\mathbf{z}^n)$ because it replaces the true

⁴The notation of $\text{sparse}(\cdot, \cdot, \cdot)$ function is introduced under the syntax of MATLAB: <http://www.mathworks.com/help/matlab/ref/sparse.html>

Wasserstein loss in WLMs. Given $\mathbf{U}^{(0)}$ (i.e., \mathbf{z}^1 is implied), for $n = 1, \dots, 2N$, we let

$$r^n = V(\mathbf{z}^n) - \sum_{s=1}^{n-1} \mathcal{O}(\mathbf{z}^s, T^{(s)}). \quad (23)$$

This is crucial for one who wants to know whether the cooling schedule is too fast to secure the suboptimality within a finite budget of iterations. The following Theorem A.2 gives a possible route to approximately realize this goal. It bounds the difference between

$$V(\mathbf{z}^n) - V(\mathbf{z}^1) \text{ and } \sum_{s=1}^{n-1} \mathbb{E} [V(\mathbf{z}^{s+1}) - V(\mathbf{z}^s) | \mathbf{z}^s],$$

the second of which is a quantitative term representing sum of a sequence. We see that $\mathcal{O}(\mathbf{z}^s, T^{(s)}) = \mathbb{E} [V(\mathbf{z}^{s+1}) - V(\mathbf{z}^s) | \mathbf{z}^s] = 0$ if and only if $T^{(s)} = \mathcal{T}(\mathbf{z}^s) \stackrel{\text{def.}}{=}$

$$\begin{cases} -\frac{1}{\langle \phi^{(t)}, \mathbf{q} \rangle} \sum_{j=1}^{m_2} q_j [M_{I_j^t, j} + L_j^{(t)} - U_{I_j^t}^{(t)}] & \text{if } s=2t \\ -\frac{1}{\langle \psi^{(t)}, \mathbf{p} \rangle} \sum_{i=1}^{m_1} p_i [M_{i, J_i^t} + L_{J_i^t}^{(t)} - U_i^{(t-1)}] & \text{if } s=2t-1 \end{cases} \quad (24)$$

In the practice of Gibbs-OT, choosing the proper cooling schedule for a specific WLM needs trial-and-error. Here we present a heuristics that the temperature $T^{(s)}$ is often chosen and adapted around $\eta \mathcal{T}(\mathbf{z}^s)$, where $\eta \in [0.1, 0.9]$. We have two concerns regarding the choice of temperature T : First, in a WLM, the cost $V(\mathbf{z})$ is to be gradually minimized, hence a temperature T smaller than $\mathcal{T}(\mathbf{z}^s)$ at every iteration ensures that the cost is actually decreased by expectation, i.e., $\mathbb{E}[V(\mathbf{z}^n) - V(\mathbf{z}^1)] < 0$; second, if T is too small, it takes many iterations to reach a highly accurate equilibrium, which might not be necessary for a single outer level step of parameter update.

Theorem A.2 (Concentration bounds for finite time Gibbs-OT). *First, r^n (by definition) is a martingale subject to the filtration of $\mathbf{z}_1, \dots, \mathbf{z}_n$. Second, given a $\varepsilon \in (0, 1)$, for $n = 1, \dots, 2N - 1$ if we choose the temperature schedule $T^{(1)}, \dots, T^{(2N)}$ such that (i) $C^n \cdot T^{(n)} \leq a_n$, or (ii) $\exists \gamma > 0, \log\left(\frac{2N \max\{m_1, m_2\}}{\varepsilon}\right) \cdot T^{(n)} + D^n \leq \gamma a_n$, where $\{a_n \geq 0\}$ is a pre-determined array. Here for $t = 1, \dots, N$,*

$$\begin{aligned} C^{2t-1} &\stackrel{\text{def.}}{=} \langle \psi^{(t)}, \mathbf{p} \rangle, \\ C^{2t} &\stackrel{\text{def.}}{=} \langle \phi^{(t)}, \mathbf{q} \rangle, \\ D^{2t-1} &\stackrel{\text{def.}}{=} \sum_{i=1}^{m_1} p_i \mathcal{R}(M_{i, \cdot}^T + \mathbf{L}^{(t)}; \mathbf{q}), \\ D^{2t} &\stackrel{\text{def.}}{=} \sum_{i=1}^{m_2} q_j \mathcal{R}(M_{\cdot, j} - \mathbf{U}^{(t)}; \mathbf{p}), \end{aligned}$$

where $M_{i, \cdot}$ and $M_{\cdot, j}$ represents the i -th rows and j -th columns of matrix M respectively, $\psi^{(t)}$ and $\phi^{(t)}$ are defined in Lemma A.1, and regret function $\mathcal{R}(\mathbf{x}; \mathbf{w}) \stackrel{\text{def.}}{=} \sum_{i=1}^m w_i x_i - \min_{1 \leq i \leq m} x_i$ for any $\mathbf{w} \in \Delta_m$ and $\mathbf{x} \in \mathbb{R}^m$. Then for any $K > 0$, we have

$$\Pr(r^{2N} < r^1 - K) \leq \exp\left[-\frac{K^2}{2 \sum_{i=1}^{2N-1} a_n^2}\right], \quad (25)$$

or

$$\Pr(r^{2N} > r^1 + \gamma K) \leq \exp\left[-\frac{K^2}{2 \sum_{i=1}^{2N-1} a_n^2}\right] + \varepsilon. \quad (26)$$

Remark 5. The bound obtained is a quantitative Hoeffding bound, not a bound that guarantees contraction around the true solution of dual OT. Nevertheless, we argue that this bound is still useful in investigating the proposed Gibbs sampler when the temperature is not annealed to zero. Particularly, the bound is for cooling schedules in general, i.e., it is more applicable than a bound for a specific schedule. There has long been a gap between the practice and theory of SA despite of its wide usage. Our result likewise falls short of firm theoretical guarantee from the optimization perspective, as with the usual application of SA.

B. Proof of Lemmas and Theorem

The minimum of n independent exponential random variables with different parameters has computatable formula for its expectation. The result immediately lays out the proof of Lemma A.1.

Lemma B.1. *Suppose we have n independent exponential random variables e_i whose c.d.f. is by $f_i(x) = \min\{\exp(\omega_i(x - z_i)), 1\}$. Without lose of generality, we assume $z_1 \geq z_2 \geq \dots \geq z_n$, then let $z_{n+1} = -\infty, h_i = \exp\left[\sum_{j=1}^i \omega_j(z_{i+1} - z_i)\right] \leq 1$ (with $h_n = 0, z_{n+1}h_n = 0$), we have*

$$\mathbb{E}[\max\{e_1, \dots, e_n\}] = z_1 - \sum_{i=1}^n \frac{(1 - h_i) \prod_{j=1}^{i-1} h_j}{\sum_{j=1}^i \omega_j}.$$

Proof. The c.d.f. of $\max\{e_1, \dots, e_n\}$ is $F(x) = \prod_{i=1}^n f_i(x)$ which is piece-wise smooth with interval

(z_{i+1}, z_i) , we want to calculate $\int_{-\infty}^{\infty} x dF(x)$.

$$\begin{aligned}
 & \int_{-\infty}^{\infty} x dF(x) \\
 = & \sum_{i=1}^n \int_{z_{i+1}}^{z_i} x dF(x) + 0 \\
 = & \sum_{i=1}^n \int_{z_{i+1}}^{z_i} x d \exp \left[\sum_{j=1}^i \omega_j (x - z_j) \right] \\
 = & \sum_{i=1}^n \int_{z_{i+1}}^{z_i} \left[\sum_{j=1}^i \omega_j \right] x \exp \left[\sum_{j=1}^i \omega_j (x - z_j) \right] dx \\
 = & \sum_{i=1}^n \left\{ \left(z_i - \frac{1}{\sum_{j=1}^i \omega_j} \right) \exp \left[\sum_{j=1}^i \omega_j (z_i - z_j) \right] \right. \\
 & \left. - \left(z_{i+1} - \frac{1}{\sum_{j=1}^i \omega_j} \right) \exp \left[\sum_{j=1}^i \omega_j (z_{i+1} - z_j) \right] \right\} \\
 = & \sum_{i=1}^n \left[\left(z_i - z_{i+1} h_i \right) - \frac{1 - h_i}{\sum_{j=1}^i \omega_j} \prod_{j=1}^{i-1} h_j \right] \\
 = & \sum_{i=1}^n \left[z_i \prod_{j=1}^{i-1} h_j - z_{i+1} \prod_{j=1}^i h_j \right] \\
 & - \sum_{i=1}^n \frac{(1 - h_i) \prod_{j=1}^{i-1} h_j}{\sum_{j=1}^i \omega_j} \\
 = & z_1 - \sum_{i=1}^n \frac{(1 - h_i) \prod_{j=1}^{i-1} h_j}{\sum_{j=1}^i \omega_j}.
 \end{aligned}$$

□

Therefore Lemma A.1 is proved up to trivial calculation using the above Lemma B.1. In order to further prove Lemma B.3, we also have (by definition of $F(x)$).

Lemma B.2. *Subject to the setup of Lemma B.1, we also have*

$$\max\{e_1, \dots, e_n\} \leq z_1,$$

and

$$F(x) \leq \min \left\{ \exp \left[\sum_{i=1}^n \omega_i (x - z^*) \right], 1 \right\}, \quad -\infty < x < \infty,$$

$$\text{where } z^* = \frac{\sum_{i=1}^n \omega_i z_i}{\sum_{i=1}^n \omega_i}.$$

Therefore, based on the observation of Lemma B.2, the tail probability $Pr(\max\{e_1, \dots, e_n\} < x)$ is upper bounded by the probability of an exponential random variable, which lead us to the proof of Lemma B.3.

Lemma B.3. *Note that Eq. (21) implies $\mathbb{E}[r^{n+1} - r^n | \mathbf{z}^1, \dots, \mathbf{z}^n] = 0$ for $t = 1, \dots, 2N$. Therefore, $\{r^n\}$ is a (discrete time) martingale subject to the filtration of $\{\mathbf{z}^n\}$. (Recall the notation by Eq. (14).) Moreover, we have the following two bounds. First, we can establish the left hand side bound for $\{r^{n+1} - r^n\}_{n=1}^{2N-1}$:*

$$r^n - r^{n+1} \leq C^n \cdot T^{(n)},$$

where for $t = 1, \dots, N$

$$C^{2t-1} \stackrel{\text{def.}}{=} \langle \psi^{(t)}, \mathbf{p} \rangle \text{ and } C^{2t} \stackrel{\text{def.}}{=} \langle \phi^{(t)}, \mathbf{q} \rangle. \quad (27)$$

Second, we also bound on the right hand side. That said, for any $1 > \varepsilon > 0$, we have

$$\begin{aligned}
 & Pr(\exists n \in \{1, \dots, 2N\}, \text{ s.t. } r^{n+1} - r^n \\
 & \geq \log \left(\frac{2N \max\{m_1, m_2\}}{\varepsilon} \right) \cdot T^{(n)} + D^n | \mathbf{z}^1, \dots, \mathbf{z}^n) \leq \varepsilon,
 \end{aligned} \quad (28)$$

where for $t = 1, \dots, N$

$$D^{2t-1} \stackrel{\text{def.}}{=} \sum_{i=1}^{m_1} p_i \mathcal{R} \left(M_{i,\cdot}^T + \mathbf{L}^{(t)}; \mathbf{q} \right) \quad (29)$$

$$D^{2t} \stackrel{\text{def.}}{=} \sum_{i=1}^{m_2} q_j \mathcal{R} \left(M_{\cdot,j} - \mathbf{U}^{(t)}; \mathbf{p} \right), \quad (30)$$

where $M_{i,\cdot}$ and $M_{\cdot,j}$ represents the i -th rows and j -th columns of matrix M respectively.

Proof. On one hand, because for each $i \in \{1, \dots, m_1\}$, $U_i^{(t)} | \mathbf{L}^{(t)}$ is lower bounded by $M_{i,J(i,t)} + L_{J(i,t)}^{(t)}$ (Lemma B.2), and for each $j \in \{1, \dots, m_2\}$, $L_j^{(t)} | \mathbf{U}^{(t-1)}$ is upper bounded by $U_{I(j,t)}^{(t-1)} - M_{I(j,t),j}$ (Lemma B.2), we easily (by definition) have $r^{n+1} | \mathbf{z}^1, \dots, \mathbf{z}^n$ is lower bounded by $r^n - C^n \cdot T^{(n)}$.

On the other hand, we have if $r^{n+1} - r^n \geq \log(1/\varepsilon_0) \cdot T^{(n)} + D^n | \mathbf{z}^1, \dots, \mathbf{z}^n$ for some $\varepsilon_0 > 0$, then at least one of $U_i^{(t)}$ (or $L_j^{(t)}$) violates the bound $\log(1/\varepsilon_0) \cdot T^{(n)} + \mathcal{R}(M_{i,\cdot}^T + \mathbf{L}^{(t)}; \mathbf{q})$ (or $\log(1/\varepsilon_0) \cdot T^{(n)} + \mathcal{R}(M_{\cdot,j} - \mathbf{U}^{(t)}; \mathbf{p})$), whose probability using Lemma B.2 is shown to be less than ε_0 . Therefore, we have for each n

$$\begin{aligned}
 & Pr(r^{n+1} - r^n \geq \log(1/\varepsilon_0) \cdot T^{(n)} + D^n | \mathbf{z}^1, \dots, \mathbf{z}^n) \\
 & \leq \max\{m_1, m_2\} \varepsilon_0, \quad (31)
 \end{aligned}$$

and

$$\begin{aligned}
 & Pr(\exists n, r^{n+1} - r^n \geq \log(1/\varepsilon_0) \cdot T^{(n)} + D^n | \mathbf{z}^1, \dots, \mathbf{z}^n) \\
 & \leq 2N \max\{m_1, m_2\} \varepsilon_0, \quad (32)
 \end{aligned}$$

Let $\varepsilon = 2N \max\{m_1, m_2\} \varepsilon_0$, which concludes our result. □

Given Lemma B.3, we can prove Theorem A.2 by applying the classical Azuma's inequality for the left-hand side bound, and applying one of its extensions (Proposition 34 in (Tao and Vu, 2015)) for the right-hand side bound. Remark that Theorem A.2 is about a single OT. For multiple different OTs, which share the same temperature schedule, one can have asymptotic bounds using the Law of Large Numbers due to the fact that their Gibbs samplers are independent with each other. Let $R^n = \frac{1}{S} \sum_{k=1}^S r_k^n$, where r_k^n is defined by Eq. (23) for sample k . Since for any $\varepsilon > 0$, one has $P(|R^{n+1} - R^n| > \varepsilon) \rightarrow 0$, as $S \rightarrow \infty$, one can have the asymptotic concentration bound for R^{2N} that for any $\varepsilon_1, \varepsilon_2 > 0$, there exists S such that $P(|R^{2N} - R^1| > \varepsilon_1) \leq \exp\left(-\frac{1}{2N\varepsilon_2}\right)$.

Tao, Terence and Vu, Van. Random matrices: Universality of local spectral statistics of non-Hermitian matrices. *The Annals of Probability*, 43(2):782-874, 2015.