A Survey On Regularized Wasserstein Barycenters

Le Tran Ngoc Tran * [1]

[1] *Université de Paris, 45 Rue des Saints-Pères, 75006 Paris, France*

January 8, 2021

Abstract

We present the algorithms to compute the Wasserstein barycenters of a set of probability distributions. It is in fact the measure that minimize the sum of its Wasserstein distance to the individual input distributions. The computation of this barycenters has received a great attentions and it is very challenging especially in high dimension. Unlike the previous approaches, the supports of barycenters are finite set of points. In this work, we focus on the stochastic algorithm introduced in [1] which produces the continuous Wasserstein barycenters. We perform some numerical test cases to study this approach and compare against the previous works.

Contents

*tran.le-tran-ngoc@etu.u-paris.fr

1 Introduction

Aggregating or summarizing a collections of probability measures is play an important role in statistic and machine learning. Optimal transport (OT) theory provides the space of measures with the distance metric, known as Wasserstein disance. This kind of distance allows to us to average distribution over a geometric domain. This leads to the barycenter of the input distribution which is the is the measure that minimizes the sum of its Wasserstein distances to each element in that set. In fact, Wasserstein barycenters are very different from the usual linear averaging (barycenter in L_2 sense) which is completely blind to the geometry of the domain.

Wasserstein barycenters nowadays has a wide range of application from economics to statistic and computer sciences. For instance, one can use Wasserstein barycenters as an approach to solve the inverse problem is investigated in [2] and [3]. The idea is to find the optimal vector weights such that the barycenter of the database associated to those weights well approximate the new input (Wasserstein barycentric coordinates).

The detail of Wasserstein barycenters is investigated in [4] as Frechet means in the space of probability measures with the Wasserstein metric. They studied several properties of Wasserstein barycenters such as existence, uniqueness and the link with the multi-marginal optimal transport problem.

The exact continuous computation of Wasserstein barycenters is tractable in only a small number of specialized cases. Therefore, many works devoted to the computational Wasserstein barycenters. In the discrete setting, one can obtain it by solving a large linear program [5]. Using entropic regularization in [6, 7] is an effective way to approximate the barycenter. More importantly, the iterative Bregman projections which is normally used for the computation of the Wasserstein distances can be extend to barycenter problem. The detail of this generalized Sinkhorn algorithm is presented in [8]. However, regularization method suffer few drawbacks such as the poor behavior with decreasing regularization.

On the other hand, the previous method consider the barycenter as a discrete measure on a fixed number of support points. These fixed support methods become really cost in higher dimensions. Stochastic Wasserstein barycenters is introduced in [9] to allow the support of the barycenter to be adjusted at each iteration. Unlike most existing optimal transport algorithm, the author tackle the problem without regularization. This allows to recover much sharper result. Base on the Frank-Wolfe optimization strategy and Sinkhorn divergence , the work [10] presents a method to compute the barycenter of a set of probability distributions that does not require to fix the support beforehand. This algorithm constructs the support by incrementally adding new points and updating their weights at each iteration. These works focus on the way to handle the support of the barycenter. However, if the support is no longer to be discrete, we have to deal with continuous (or infinitely supported) probability measures. In this work, we focus on the continuous Wasserstein barycenters introduced in [1]. In this work, the author investigate the primal-dual relationship to parametrize the barycenter implicitly using the dual potentials of regularized transport problems.

The rest of this work is organized as follows: Section 2 reviews the background of the optimal transport and we also recall the generalized Sinkhorn algorithm to compute the Wasserstein barycenters. The algorithm of continuous Wasserstein barycenters and some methods to recover barycenter is mentioned in Section 3. Finally we present in Section 4 some numerical results to illustrate our purpose and Section 5 provides concluding remarks.

2 Background on Optimal Transport

2.1 Kantorovich formulation

2.1.1 Discrete Relaxation

For two histograms a and b in the probability simplex $\sum_n \stackrel{\text{def}}{=} \{x \in \mathbb{R}_+^n : x^T \mathbf{1}_n = 1\}$. We denote $U(a,b)$ is the set of admissible couplings between a and b, that is, the set of all nonnegative $n \times n$ matrices with row and column sums a and b respectively:

$$U(a,b) \stackrel{\text{def}}{=} \{P \in \mathbb{R}_+^{n \times n} | P\mathbf{1}_n = a, P^T \mathbf{1}_n = b\}, \tag{1}$$

Given a $n \times n$ cost matrix C, the cost of mapping a to b using a transportation matrix P can be quantified as $\langle P, C \rangle$. The Kantorovich optimal transport problem is defined as:

$$W_C(a,b) \stackrel{\text{def}}{=} \min_{P \in U(a,b)} \langle C, P \rangle, \tag{2}$$

2.1.2 Relaxation for Arbitrary Measures

In general case, the mass conservation constraint 1 should rewritten as a marginal constraint on joint probability distributions

$$\mathcal{U}(\alpha,\beta) \stackrel{\text{def}}{=} \{\pi \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y}) \,|\, (\mathbf{P}_1)_\# \pi = \alpha, (\mathbf{P}_1)_\# \pi = \beta)\}, \tag{3}$$

where $\mathbf{P}_1(x,y) \stackrel{\text{def}}{=} x$ and $\mathbf{P}_2(x,y) \stackrel{\text{def}}{=} y$ are the projections onto the first and second coordinate respectively, and $\mathbf{T}_\#(\mu)$ denotes the pushforward of the measure μ by a function $\mathbf{T}$.

Continuous Kantorovich problem. We consider compact spaces $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^n$ along with a symmetric cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$. We denote $\mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$ the space of probability Radon measures. For any $\alpha \in \mathcal{M}_1^+(\mathcal{X}), \beta \in \mathcal{M}_1^+(\mathcal{Y})$, the Kantorovich problem 2 is generalized as:

$$\mathcal{W}(\alpha,\beta) \stackrel{\text{def}}{=} \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) d\pi(x,y), \tag{4}$$

When $c(x,y) = \|x - y\|_2^2$, the quantity $\mathcal{W}(\alpha,\beta)$ yields the squared 2-Wasserstein distance.

2.2 Entropic Regularization

Relative entropy. The Kullback-Leibler divergence is defined as:

$$\mathbf{KL}(P|Q) \stackrel{\text{def}}{=} \sum_{i,j} \log\left(\frac{P_{i,j}}{Q_{i,j}} - P_{i,j} + Q_{i,j}\right), \tag{5}$$

with the convention $0 \log(0) = 0$ and $\mathbf{KL}(P|Q) = \infty$ if there is some (i,j) such that $Q_{i,j} = 0$ but $P_{i,j} \neq 0$. the special case $\mathbf{KL}(P|\mathbb{1})$ corresponds to minus the Shannon-Boltzmann entropy. The function $\mathbf{KL}(|Q)$ is strongly convex, because its hessian is $\partial^2 \mathbf{KL}(P|Q) = \text{diag}\left(\frac{1}{P_{i,j}}\right)$ and $P_{i,j} \leqslant 1$. The relative entropy is always non-negative, $\mathbf{KL}(P|Q) \geqslant 0$, a result known as Gibb's inequality and $\mathbf{KL}(P|Q) = 0$.

2.2.1 For Discrete Measures

The idea of the entropic regularization of optimal transport is to use $\mathbf{KL}$ as a regularizing function to obtain approximate solutions to the original transport problem 2:

$$W_C^\varepsilon(a,b) \stackrel{\text{def}}{=} \min_{P \in U(a,b)} \langle P, C \rangle + \varepsilon \mathbf{KL}(P|a \otimes b). \tag{6}$$

Here we used as a reference measure for the relative entropy $a \otimes b = (a_i b_j)_{i,j}$. This choice of normalization, specially in this discrete setting, has no importance for the selection of the optimal P since it only affects the objective by a constant, indeed for $P \in U(a,b)$, one has

$$\mathbf{KL}(P|a \otimes b) = \mathbf{KL}(P|a' \otimes b') + \mathbf{KL}(a' \otimes b'|a \otimes b)$$

2.2.2 General Formulation

One can consider arbitrary measures by replacing the discrete entropy by the relative entropy with respect to the product measure $\mathrm{d}\alpha \otimes \mathrm{d}\beta(x,y) \stackrel{\text{def}}{=} \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$, and propose a regularized counterpart to 4 using

$$\mathcal{W}_c^\varepsilon(\alpha,\beta) \stackrel{\text{def}}{=} \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta) \tag{7}$$

where the relative entropy is a generalization of the discrete Kullback-Leibler divergence 5

$$\mathrm{KL}(\pi|\xi) \stackrel{\text{def}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\right) + \int_{\mathcal{X}\times\mathcal{Y}} (\mathrm{d}\xi(x,y) - \mathrm{d}\pi(x,y))\,\mathrm{d}\pi(x,y), \tag{8}$$

where π is absolutely continuous with respect to ξ.

2.3 Wasserstein barycenter

2.3.1 Sinkhorn algorithm

The factorization of the optimal solution of 6 can be rewritte in matrix form as $P = \mathrm{diag}(\mathbf{u})\mathbf{K}\mathrm{diag}(\mathbf{v})$ where $\mathbf{K} = \exp\left(-\frac{C}{\varepsilon}\right)$ is the elementwise exponential of the negative of the rescaled cost matrix.One thus need to find two positive scaling vectors $u, v \in \mathbb{R}_+^n$ such that the following equalities hold

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = a \text{ and } \mathbf{v} \odot (\mathbf{K}^T\mathbf{u}) = b \tag{9}$$

The Sinkhorn algorithm consists in using the following iterations for $l \geqslant 1$

$$\mathbf{u}^{(l+1)} = \frac{a}{\mathbf{K}\mathbf{v}^{(l)}} \text{ and } \mathbf{v}^{(l+1)} = \frac{b}{\mathbf{K}^T\mathbf{u}^{(l+1)}} \tag{10}$$

initialized with an arbitrary positive vector, for instance $\mathbf{v}^{(0)} = \mathbb{1}_n$

2.3.2 Wasserstein barycenter

The Wasserstein barycenter of a family of measures is defined as the minimizer of the weighted sum of squared Wasserstein distances from the variable to each of the measures. For measures with the same discrete support,

we define, using entropic regularization, the barycenter of histograms $(a_1, \cdots, a_d) \in (\Sigma_n)^d$ with barycentric weights $\lambda = (\lambda_1, \cdots, \lambda_d) \in \Sigma_d$ as

$$B(A, \lambda) \stackrel{\text{def}}{=} \underset{b \in \Sigma_n}{\operatorname{argmin}} \sum_{s=1}^{d} \lambda_s W_C^{\varepsilon}(a_s, b) \tag{11}$$

where $A = (a_1^T, \cdots, a_d^T)^T \in \mathbb{R}^{nd}$. The addition of the entropy term ensures strict convexity and thus that the Wasserstein barycenter is uniquely defined. We can compute approximate Wasserstein barycenters by iterating scheme, a generalization of the Sinkhorn algorithm, once again relies on two scaling vectors:

$$u_s^{(l)} = \frac{a_s}{\mathbf{K} v_s^{(l-1)}}, \tag{12}$$

$$B^{(l)}(A, \lambda) = \prod_{s=1}^{d} \left(\mathbf{K}^T u_s^{(l)} \right)^{\lambda_s}, \tag{13}$$

$$v_s^{(l)} = \frac{B^{(l)}(A, \lambda)}{\mathbf{K}^T a_s^{(l)}}. \tag{14}$$

where $\mathbf{K} = \exp\left(-\frac{C}{\varepsilon}\right)$. In this case, the scaling vectors are of size nd, such that $u^{(l)} = \left(u_1^{(l)T}, \cdots, u_d^{(l)T} \right)^T$, $v^{(l)} = \left(v_1^{(l)T}, \cdots, v_d^{(l)T} \right)^T$ and $v^{(0)} = \mathbb{1}_{nd}$.

3 Continuous Regularized Wasserstein Barycenters

In the paper [1], authors consider two type of regularization: entropic and quadratic. Giving a convex regularizer $R : \mathbb{R} \longrightarrow \mathbb{R}$, we rewrite 7 with respect to ξ, R:

$$\mathcal{W}_R^{\xi}(\alpha, \beta) \stackrel{\text{def}}{=} \inf_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) + \int_{\mathcal{X} \times \mathcal{Y}} R\left(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x, y) \right) \mathrm{d}\xi(x, y), \tag{15}$$

where

$$\forall t \geqslant 0, \quad R(t) \stackrel{\text{def}}{=} \begin{cases} \varepsilon(t \ln t - t) & \text{entropic} \\ \frac{\varepsilon}{2} t^2 & \text{quadratic.} \end{cases} \tag{16}$$

The primal problem 15 admits an equivalent dual formulation for entropic and quadratic regularization:

$$\mathcal{W}_R^{\xi}(\alpha, \beta) = \sup_{f \in C(\mathcal{X}) g \in C(\mathcal{Y})} \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y) \mathrm{d}\beta(y) - \int_{\mathcal{X} \times \mathcal{Y}} R^{\star}(f(x) + g(y) - c(x, y)) \mathrm{d}\xi(x, y), \tag{17}$$

where $C(\mathcal{X})$ is the space of continuous real-valued functions on $\mathcal{X}$ and the regularizer $R^{\star}$ on the dual problem is determined as

$$\forall t \in \mathbb{R}, \quad R^{\star}(t) = \begin{cases} \varepsilon \exp(\frac{t}{\varepsilon}) & \text{entropic} \\ \frac{1}{2\varepsilon}(t_+)^2 & \text{quadratic.} \end{cases} \tag{18}$$

3.1 The primal problem and its dual

Given iput distributions $\alpha_1, \cdots, \alpha_n \in \mathcal{M}_1^+(\mathcal{X})$ and weights $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$, the authors in paper (??) consider the following *regularized Wasserstein barycenter* problem with respect to the regularized Wassertein distance 15 for some $\nu \in \mathcal{M}_1^+(\mathcal{X})$, where R refers to either quadratic or entropic regularization 16:

$$\inf_{\beta \in \mathcal{M}_1^+(\mathcal{X})} \sum_{i=1}^{n} \lambda_i \mathcal{W}_R^{\alpha_i \otimes \nu}(\alpha_i, \beta). \tag{19}$$

Theorem 1. *The dual problem of (19) is*

$$\sup_{\substack{\{(f_i,g_i)\}_{i=1}^n \subset C(\mathcal{X}) \times C(\mathcal{Y}) \\ \sum_{i=1}^n \lambda_i g_i = 0}} \sum_{i=1}^{n} \lambda_i \left(\int f_i \mathrm{d}\alpha_i - \iint R^\star(f_i(x) + g_i(y) - c(x,y)) \mathrm{d}\alpha_i(x) \mathrm{d}\nu(y). \right) \tag{20}$$

Moreover, strong duality holds in the sense that the infimum of (19) equals the supremum of (20), and a solution of (19) exists. If $\{(f_i, g_i)\}_{i=1}^n$ solves (20), then each (f_i, g_i) is a solution to the dual formulation (17) of $\mathcal{W}_R^{\alpha_i \otimes \nu}(\alpha_i, \beta)$.

Remark 1. *Based on Theorem 1., we can recover the optimal transport plan π_i between α_i and the barycenter β from the pair (f_i, g_i) solving (20) via the primal-dual relationship:*

$$\mathrm{d}\pi(x,y) = H(x,y)\mathrm{d}\xi(x,y), \text{ where } H(x,y) = \begin{cases} \exp\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right) & \text{entropic} \\ \left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right)_+ & \text{quadratic} \end{cases} \tag{21}$$

3.2 The Stochastic Gradient Descent (SGD) solves the regularized barycenter problem

Rewriting integrals as expectations, we onbtain the following formulation equivalent to (20):

$$\sup_{\substack{\{f_i\}_{i=1}^n \subset C(\mathcal{X}) \\ \{g_i\}_{i=1}^n \subset C(\mathcal{Y})}} \mathop{\mathbb{E}}_{\substack{X_i \sim \alpha_i \\ Y \sim \beta}} \left[\sum_{i=1}^{n} \lambda_i \left(f_i(X_i) - R^\star \left(f_i(X_i) + g_i(Y) - \sum_{j=1}^{n} \lambda_j g_j(Y) - c(X_i, Y) \right) \right) \right] \tag{22}$$

In paper [1], authors give stochastic gradient descent (Algorithm 1) to solve (20)

Algorithm 1: Stochastic gradient descent to solve the regularized barycenter problem (22)

Input: distributions $\alpha_1, \cdots, \alpha_n$ with sample access, weights $(\lambda_1, \cdots, \lambda_n$, dual regularizer $R^\star$, regularizing measures ν, cost function c, gradient update function `ApplyGradient`

Initialize parameterizations $\{(f_{\theta_i}, g_{\phi_i}\}_{i=1}^n$;

for $l \leftarrow 1$ ***to*** n_{epochs} **do**

$\forall i \in \{1, \cdots, n\}$: sample $x^{(i)} \sim \alpha_i$; sample $y \sim \nu$;

$\overline{g} \leftarrow \sum_{i=1}^{n} \lambda_i g_{\phi_i}(y)$;

$F \leftarrow \sum_{i=1}^{n} \lambda_i \left(f_{\theta_i} x^{(i)} - R^\star \left(f_{\theta_i} x^{(i)} + g_{\phi_i}(y) - \overline{g} - c(x^{(i)}, y) \right) \right)$;

for $i = 1, \cdots, n$: `ApplyGradient`(F, θ_i); `ApplyGradient`(F, ϕ_i) ;

end

return dual potentials $\{(f_{\theta_i}, g_{\phi_i}\}_{i=1}^n$.

Once we approximate the optimal potentials $\{f_i\}_{i=1}^n, \{g_i\}_{i=1}^n$, we can recover the corresponding transport plan π_i via the primal-dual relationship (21)

3.3 Some methods recover the barycenter

In paper [1], authors introduce five methods to recover barycenter, specifically, there are two ways a) and b) to recover from the transport plans and three approaches c), d), e) is to estimate a *Monge map* approximating each π_i:

(a) Use numerical integration to approximate $(P_y)_{\#}\pi_i(x) = \int \pi_i(x, y)\mathrm{d}y$ with proper discretization of the space $\mathcal{X}$, if π_i has density.

(b) Use Markov chain Monte Carlo (MCMC) methods to sample according to π_i, again assuming it has density, and then take the second components of all the samples.

(c) Compute pointwise barycentric projection. If $c(x, y) = \|x - y\|_2^2$, then the barycentric projection takes the simplified form

$$T_i(x) = \mathop{\mathbb{E}}_{Y\sim\pi_i(\cdot|x)}[Y] \tag{23}$$

(d) Recover an approximation of the Monge map usinhg the gradient of the dual potentials. For the case when $c(x, y) = \|x-y\|_2^2$ and the densities of the source distributions exist, there exists a unique Monge map realizing the (unregularized) optimal transport plan π_i:

$$T_i(x) = x - \frac{1}{2}\nabla f_i(x). \tag{24}$$

While this does not strictly hold for the regularized case, it gives a cheap approximation of T_i

(e) Find T_i as a solution to the following optimization problem, where H is defined in (21):

$$T_i \overset{\text{def}}{=} \mathop{\operatorname{argmin}}_{T:\mathcal{X}\to\mathcal{X}} \mathop{\mathbb{E}}_{(X,Y)\sim\pi_i}[c(T(X), Y)] = \mathop{\operatorname{argmin}}_{T\in\mathcal{X}\to\mathcal{X}} \mathop{\mathbb{E}}_{\substack{X\sim\alpha_i \\ Y\sim\nu}}[c(T(X), Y)H(X, Y)]. \tag{25}$$

Compared to (a) (b), options (c) (d) (e) do no require knowing the densities of the input distributions.

4 Numerical test cases

In this section, we do the experiment to compute the barycenter by applying Sinkhorn algorithm and Algorithm 1 along with some methods to recover barycenter (CWB) [1]. Specifically, we have three test cases: for 1D Gaussians, 2D Gaussians and for images. Moreover, the comparison among the results is also given, and with Gaussian distributions, the true barycenter is displayed to analyze with the approximated solutions. In all experiments, we use equal weights for input distributions i.e., $\lambda_i = \frac{1}{n}\ \forall i = 1, \cdots, n$ and the squared Euclidean distance as the cost function, i.e., $c(x, y) = \|x - y\|_2^2$.

In the test case for images, we use Adam optimizer with learning rate 10^{-4} and batch size 4096 or 8192 for the training potentials in Algorithm 1. The dual potentials $\{f_i, g_i\}_{i=1}^n$ in (22) are each parameterized as neural networks with two fully-connected layers $(d \to 128 \to 256 \to 1)$ using ReLU activations. Every T_i in (25) is parameterized with layers $(d \to 128 \to 256 \to d)$. Due to the fact that the test case for two Gaussians in 2D is very simple, we just use the batch size 4096 and reduce layers for the dual potentials and the transport map to $(d \to 64 \to 128 \to 1)$ and $(d \to 64 \to 128 \to d)$ respectively.

We remind the theorem giving an explicit construction of the barycenter in the Gaussian case.

In paper [4], authors consider the case where, for $i = 1, \cdots, p, \nu_i = \mathcal{N}(0, S_i)$; i.e., ν_i is a Gaussian measure with mean 0 and covariance matric S_i. Assuming that each S_i is positive define and, given weights $\lambda_i > 0$ that sum to 1, we consider the following barycenter problem

$$\inf_{\nu \in X' \cap \mathcal{M}_+^1(\mathbb{R})^d} J(\nu) = \sum_{i=1}^{p} \frac{\lambda}{2} W_2^2(\nu_1, \nu) \tag{26}$$

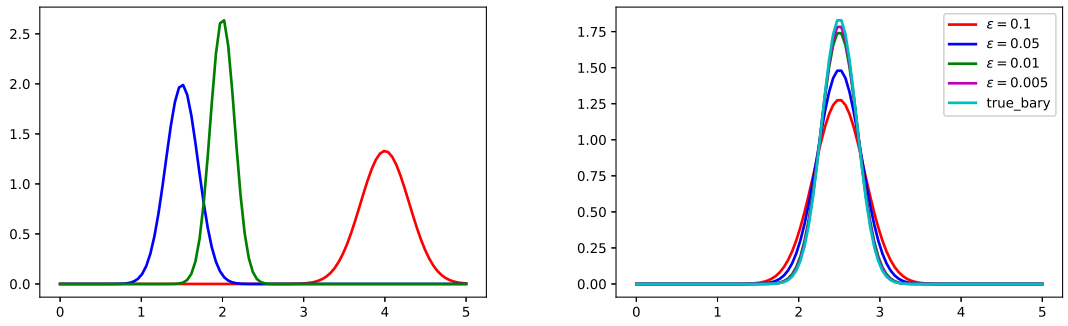
The following theorem gives a true barycenter in this case.

Theorem 2. *In the Gaussian framework, there is a unique solution $\overline{\nu}$ to (26). Moreover, $\overline{\nu} = \mathcal{N}(0, \overline{S})$, where $\overline{S}$ is the unique positive definite root of the matrix equation*

$$\sum_{i=1}^{p} \lambda_i \left(S^{1/2} S_i S^{1/2} \right)^{1/2} = S. \tag{27}$$

4.1 Compute the barycenter for 1D-Gaussians

We now begin the numerical experiment with the very simple test case where all the input distributions are Gaussians in one dimensional case. In particular, we consider here three input Gaussians g_1, g_2, and g_3, where g_1 having $\mu_1 = 4, \sigma_1 = 0.3$ (red line), g_2 having $\mu_2 = 2, \sigma_2 = 0.15$ (green line) and the parameters of g_3 is $\mu_3 = 1.5, \sigma_3 = 0.2$ (blue line). We investigate the effect of the regularized parameter ε on the generalization of Sinkhorn algorithm (from this part, we call Sinkhorn algorithm for a short). In particular, we evaluate the barycenter with various value of ε. We noticed that the smaller ε, the better approximated solution. Figure 1 illustrates that the approximated barycenter with $\varepsilon = 0.005$ has the smallest error compared to the true barycenter. However, we note that choosing the value of ε is too small makes the regularized problem becomes instability.



(a) Input distributions

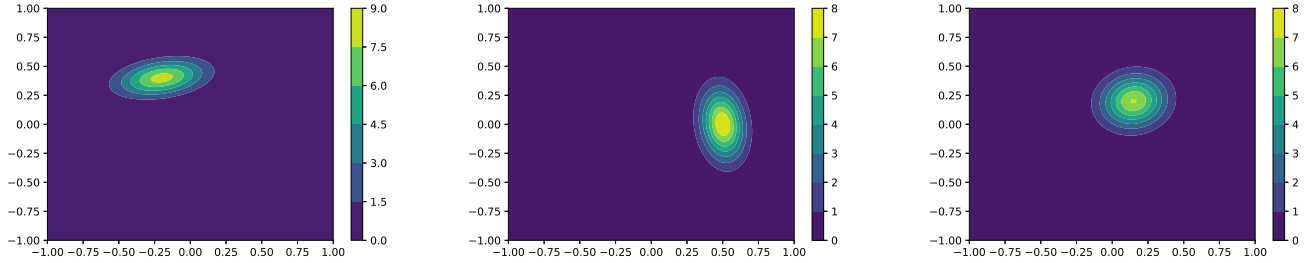
(b) The approximated and true barycenter

Figure 1: Input Gaussian distributions and barycenter.

4.2 Compute the barycenter for 2D-Gaussians

In this part, we give the approximated solution by running Sikhorn algorithm, CWB methods (using (d) to recover barycenter) with both regularization: entropic and quadratic. Specifically, we consider $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$

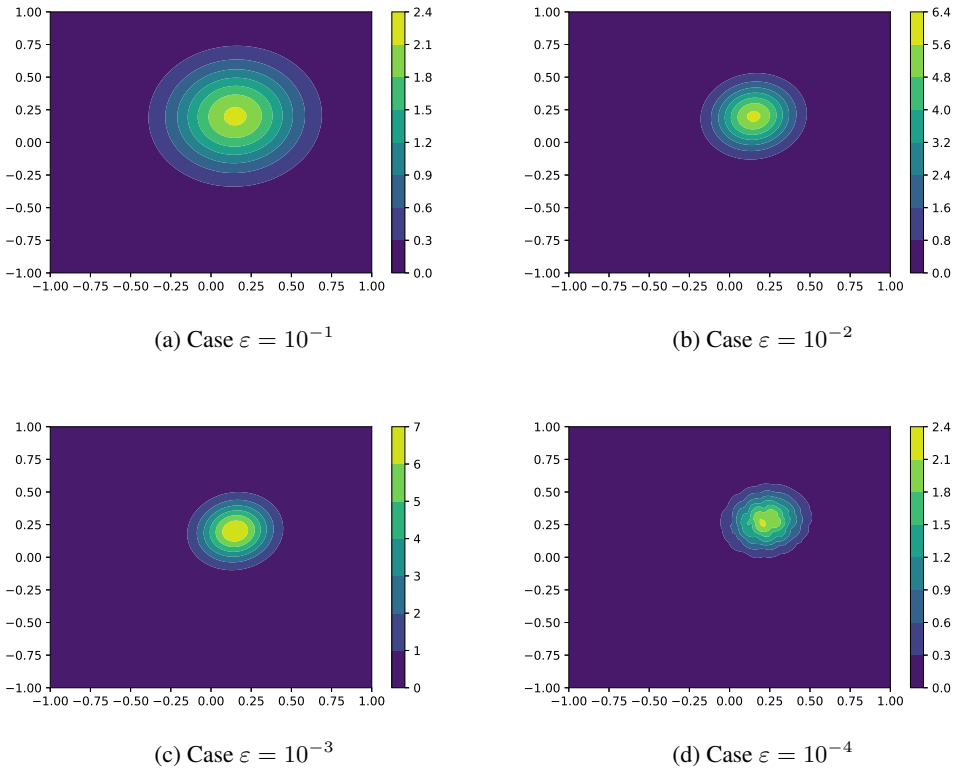
where $\mu_1 = (-0.2, 0.4), \Sigma_1 = \begin{pmatrix} 0.0404 & 0.006 \\ 0.006 & 0.0104 \end{pmatrix}$ and $X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ where $\mu_2 = (-0.5, 0), \Sigma_2 = \begin{pmatrix} 0.0101 & -0.003 \\ -0.003 & 0.0401 \end{pmatrix}$.



(a) Input 1 (b) Input 2 (c) Ground truth barycenter

Figure 2: Input Gaussian distributions and ground truth barycenter.

Sinkhorn Observing the approximated solution and ground truth barycenter, we realized that with $\varepsilon = 10^{-3}$, the result from figure 3 gives the best approximation for barycenter but the value in z-coordinate is not the same as the one of the true barycenter (we can see the color bar next to the result). Once again, we notice that the smaller ε does not give a better solution.



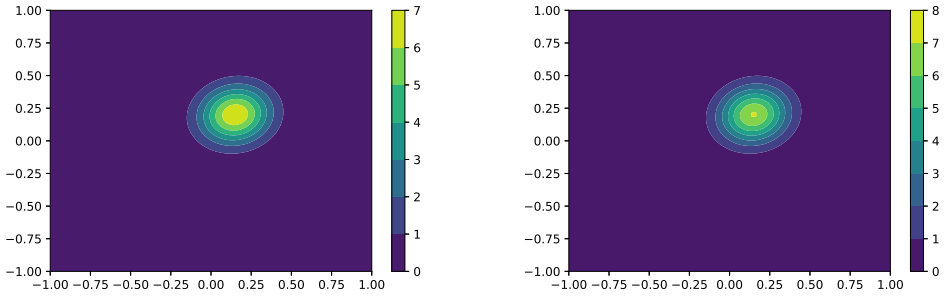
(a) Case $\varepsilon = 10^{-1}$ (b) Case $\varepsilon = 10^{-2}$

(c) Case $\varepsilon = 10^{-3}$ (d) Case $\varepsilon = 10^{-4}$

Figure 3: Compute barycenter by Sinkhorn with different values of epsilon.

CWB method With CWB method, we perform the experiment with two regularizations, each regularization is run five times and then we obtain the errors between approximated solution and true barycenter through

Table 1 and Table 2. There's no big difference among errors of the number of running times so we choose the result of fist time to visualize, which is in figure 4.



(a) Entropic regularization $\varepsilon = 10^{-3}$ (b) Quadratic regularization $\varepsilon = 10^{-4}$

Figure 4: Compute barycenter by CWB.

Based on two tables and the result from figure 4, we can observe that CWB with quadratic regularization gives the approximated solution better than this method with entropic regularization. In the other hand, we have that both solutions of Sinkhorn algorithm and CWB with entropic regularization have the same colour bar, whose the highest value is lower than of the ground truth barycenter while quadratic regularization has the colour bar same as that one of the true barycenter.

Repeat	Mean loss	Covariance loss	2-Wasserstein loss
1	$8.946900e-05$	$1.806752e-04$	$1.698161e-02$
2	$6.456989e-05$	$1.554143e-04$	$1.550135e-02$
3	$1.132215e-04$	$1.690408e-04$	$1.843264e-02$
4	$1.988828e-04$	$1.069813e-04$	$1.688803e-02$
5	$4.002236e-04$	$1.844868e-04$	$1.673603e-02$

Table 1: The error value between grouth truth and barycenter of CWB method with quadratic regularization, $\varepsilon = 10^{-4}$.

Repeat	Mean loss	Covariance loss	2-Wasserstein loss
1	2.074180e-04	8.566652e-04	1.895768e-02
2	1.730539e-04	8.633879e-04	1.923837e-02
3	2.125627e-04	8.276172e-04	1.911340e-02
4	3.437977e-04	8.391186e-04	1.693125e-02
5	1.001389e-04	8.160181e-04	1.852287e-02

Table 2: The error value between grouth truth and barycenter of CWB method with entropic regularization, $\varepsilon = 10^{-3}$.

Moreover, to illustrate the CWB method has the better performance than the method in [6], we give the table 3 showing the error between the approximated solution and true barycenter with the same setting with the CWB method.

Repeat	Mean loss	Covariance loss	2-Wasserstein loss
1	8.976861e-04	6.630438e-03	3.719102e-02
2	1.521773e-03	7.254045e-03	3.878240e-02
3	3.232934e-03	7.179928e-03	3.978970e-02
4	1.902361e-03	7.090140e-03	3.906948e-02
5	9.283114e-04	7.685556e-03	4.040739e-02

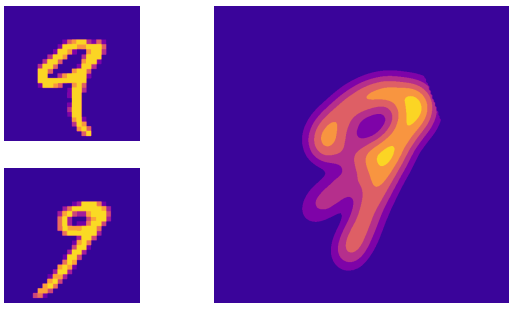
Table 3: The error value between grouth truth and barycenter of the method in [6], $\varepsilon = 10^{-2}$.

4.3 Test case for images

In the last of our experiment, we use a source including two images to test for CWB method. Furthermore, the result by applying Sinkhorn with different values of ε is also given.

Figure 5 shows the result for methods (a)-(e) from section 3.3. At first, we train the dual potential using quadratic regularization with $\varepsilon = 10^{-4}$ or $\varepsilon = 10^{-5}$. Then each method is run subsequently to obtain the barycenter. For (a) we use a discretized grid with grid size 200. For (b) we use Metropolis-Hastings to generate 10^5 samples with a symmetric Gaussian proposal. The results from (a)(b) are aggregated from all transport plans. For (c)(d)(e) we sample from each input distributions and then push the samples forward using T_i's to have 10^5 in samples total.

We have some notes on the results that: among five results, virtually, (a) has a highest performance, numerical intergration shows the transport plan π_i's computed by (21) are exact and smooth, (b) MCMC gives the result is pretty equivalent to the barycenter in (a) but this method takes a lot of time compared to other methods and can be blurry near the boundaries; (c)(d)(e) give the result with many artifacts, (c) barycentric projection has poor boundaries, (d) there is several white line in the interior, (e) the digit 9 looks pixelated.

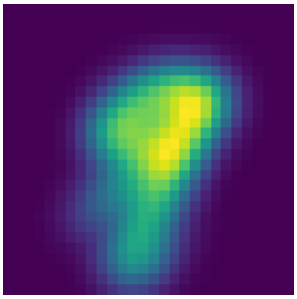


(a) Source (b) Intergrated (c) MCMC

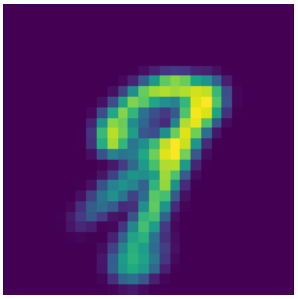
(d) Barycentric (e) Gradient map (f) Neural network

Figure 5: Several methods to recover barycenter

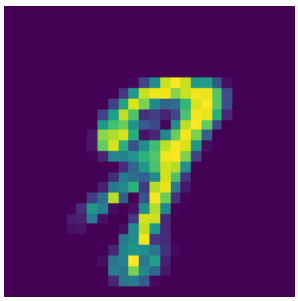
Figure 6 shows three barycenter by running Sinkhorn algorithm with different values of ε. We use entropic regularization in the setting. In general, the shape of all barycenters are similar, the case with $\varepsilon = 10^{-3}$ gives the best approximation. With this test case, we also imply that the smaller ε, the better result for barycenter. However, if we choose ε is too small, the regularized problem become instability, specifically, case (c) in figure 6 has many artifacts, looks pixelated.



(a) $\varepsilon = 10^{-2}$



(b) $\varepsilon = 10^{-3}$



(c) $\varepsilon = 10^{-4}$

Figure 6: Sinkhorn algorithm for images

5 Conclusion

This survey has motivated by the course *Computational Optimal Transport, MVA 2020-2021*, through reading the main idea in paper [1] and applying some knowledge from the course's note [11], we obtain some interesting result about *Regularized Wasserstein Barycenters*. Regulization plays important role in solving the optimal transport problem, with the discrete measures, we can apply the Sinkhorn algorithm to find the barycenters. In paper [1], authors propose a stochastic algorithm (CWB) to approximate Wasserstein barycenter without discretizing its support. We have done the experiment with 3 cases: 1D Gaussians, 2D Gaussians and for images. Our focus on 2D Gaussians case, we run the test case with both Sinkhorn and CWB method. Through the theorem proving the ground truth barycenter in Gaussian case in paper [4], we can compare the approximated solution to the true barycenter. Moreover, we also have a comparison between two regulization, when quadratic regulization gives a better result compared to entropic regularization. And with the same setting, we also have the table showing the error between the true barycenter and the approximated solution of the method was introduced in paper [6], then we conclude that, in that case, the CWB method gives the better performance. With test case for images, the CWB method gives us many results, and because, we do not have the true barycenter for this case then we cannot conclude which method gives the better approximation, but virtually, we can see the approximated barycenter with numerical integration is clearest. To sum up, compared to existing methods, the algorithm in paper [1] produces the first continuous approximation of the barycenter that allow sample access.

References

[1] L. Li, A. Genevay, M. Yurochkin, J. Solomon, Continuous regularized wasserstein barycenters, arXiv preprint arXiv:2008.12534 (2020).

[2] N. Bonneel, G. Peyré, M. Cuturi, Wasserstein barycentric coordinates: histogram regression using optimal transport., ACM Trans. Graph. 35 (4) (2016) 71–1.

[3] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, J.-L. Starck, Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning, SIAM Journal on Imaging Sciences 11 (1) (2018) 643–678.

[4] M. Agueh, G. Carlier, Barycenters in the wasserstein space, SIAM Journal on Mathematical Analysis 43 (2) (2011) 904–924.

[5] E. Anderes, S. Borgwardt, J. Miller, Discrete wasserstein barycenters: Optimal transport for discrete data, Mathematical Methods of Operations Research 84 (2) (2016) 389–409.

[6] M. Cuturi, A. Doucet, Fast computation of wasserstein barycenters (2014).

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

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

[9] S. Claici, E. Chien, J. Solomon, Stochastic wasserstein barycenters, arXiv preprint arXiv:1802.05757 (2018).

[10] G. Luise, S. Salzo, M. Pontil, C. Ciliberto, Sinkhorn barycenters with free support via frank-wolfe algorithm, in: Advances in Neural Information Processing Systems, 2019, pp. 9322–9333.

[11] G. Peyré, Course notes on computational optimal transport (2019).