Variational Wasserstein Barycenters with C-cyclical Monotonicity Regularization

Jinjin Chi[1,2], **Zhiyao Yang**[1,2], **Ximing Li**[1,2*], **Jihong Ouyang**[1,2], **Renchu Guan**[1,2]

[1] College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
chijinjin616@gmail.com, yangzy9529@gmail.com, liximing86@gmail.com, ouyj@jlu.edu.cn, guanrenchu@jlu.edu.cn

Abstract

Wasserstein barycenter, built on the theory of Optimal Transport (OT), provides a powerful framework to aggregate probability distributions, and it has increasingly attracted great attention within the machine learning community. However, it is often intractable to precisely compute, especially for high dimensional and continuous settings. To alleviate this problem, we develop a novel regularization by using the fact that c-cyclical monotonicity is often necessary and sufficient conditions for optimality in OT problems, and incorporate it into the dual formulation of Wasserstein barycenters. For efficient computations, we adopt a variational distribution as the approximation of the true continuous barycenter, so as to frame the Wasserstein barycenters problem as an optimization problem with respect to variational parameters. Upon those ideas, we propose a novel end-to-end continuous approximation method, namely Variational Wasserstein Barycenters with c-Cyclical Monotonicity Regularization (VWB-CMR), given sample access to the input distributions. We show theoretical convergence analysis and demonstrate the superior performance of VWB-CMR on synthetic data and real applications of subset posterior aggregation.

Introduction

Summarizing, combining, and comparing probability distributions defined on a metric are fundamental yet vital tasks in machine learning, statistics, and computer science. For instance, in Bayesian inference, given massive data one may perform posterior inference in parallel on different machines with data subsets, and then aggregate subset posterior distributions via their *barycenter* as an approximation to the full data posterior (Srivastava et al. 2015). Besides, the barycenter of a collection of distributions has been widely studied in various real applications, such as image processing (Rabin et al. 2011) and clustering (Ho et al. 2019).

The theory of **O**ptimal **T**ransport (**OT**) (Monge 1781; Kantorovitch 1942; Villani 2008; Le et al. 2021; Backhoff-Veraguas et al. 2022) provides a powerful framework to carry out such aggregation. OT equips the space of distributions with a distance metric known as the *Wasserstein distance*, which has gained substantial popularity in different
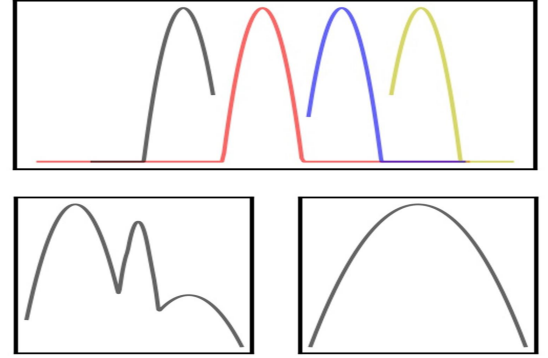

Figure 1: Top: four distributions. Bottom left: Euclidean average of distributions. Bottom right: Wasserstein barycenter.

fields. The *barycenter* of multiple given probability distributions under *Wasserstein distance* is defined as **Wasserstein barycenter**, i.e., a distribution minimizing the sum of Wasserstein distances to all input distributions. Due to geometric properties, the Wasserstein barycenter can better capture the underlying geometric structure than the barycenter with respect to other popular distances, e.g., Euclidean distance, see Figure 1. Consequently, it has a broad range of applications in text mixing (Rabin et al. 2011), imaging (Bonneel, Peyré, and Cuturi 2016), and model ensemble (Dognin et al. 2019).

Despite its impressive performance, it is often intractable to precisely compute Wasserstein barycenters, especially for high dimensional and continuous settings (Peyré, Cuturi et al. 2019; Korotin et al. 2021a; Daaloul et al. 2021). The existing studies (Agueh and Carlier 2011; Álvarez-Esteban et al. 2016; Dvurechenskii et al. 2018; Ge et al. 2019; Lin et al. 2020) on Wasserstein barycenter of continuous distributions require discretization of all input distributions or the barycenter itself, so they scale poorly to high-dimensional settings (Staib et al. 2017; Claici, Chien, and Solomon 2018; Dvurechenskii et al. 2018; Izzo, Silwal, and Zhou 2021). Moreover, the discretization techniques are commonly undesirable since they lack the inherently continuous nature of data distributions and the ability of generating new samples if needed. To the best of our knowledge, very few works

---

*Corresponding author

Literature	General cost	Barycenter	End-to-end
(Li et al. 2020)	√	Fixed prior	×
(Fan, Taghvaei, and Chen 2021)	×	Generative network	×
(Korotin et al. 2021b)	×	Fixed prior	×
(Korotin et al. 2022)	×	Generative network	×
Our method	√	Variational distribution	√

Table 1: Summary of existing Wasserstein barycenter methods.

attempt to operate directly on continuous distributions for computing a barycenter. In this work, we further contribute on this challenging topic of continuous barycenters.

Our contributions. We propose a computationally efficient and straightforward method for estimating continuous Wasserstein barycenters, given sample access to the input distributions. The key idea is to introduce a novel c-cyclical monotonicity regularization following the fact that c-cyclical monotonicity is often necessary and sufficient conditions for optimality in OT problems. We incorporate this regularization into the dual formulation of the Wasserstein barycenter problem, which guarantees the learning of optimal dual potentials, so as to obtain a good approximation of the barycenter. Our method is end-to-end without recovering the barycenters from the dual solution. This is made possible by introducing a family of continuous distributions, namely variational distribution, and using the closest one as the approximation of the barycenter. We thereby obtain a tractable objective with respect to variational parameters, which can be efficiently solved through stochastic optimization. We provide theoretical analysis on convergence as well as empirical evidence of the effectiveness of our method on both synthetic data and real applications of subset posterior aggregation.

Related work. The notion of the Wasserstein barycenter was first introduced by (Agueh and Carlier 2011) and then investigated by many works via the original Wasserstein distance as well as its regularization. Most of them mainly solve discrete Wasserstein barycenter problem, namely, averaging some discrete distributions, via linear programs (or some equivalent problem) (Anderes, Borgwardt, and Miller 2016; Cuturi and Peyré 2018; Yang et al. 2018; Ge et al. 2019) or regularized projection-based methods (Cuturi and Doucet 2014; Benamou et al. 2015; Cuturi and Peyré 2016; Ye et al. 2017; Janati, Cuturi, and Gramfort 2020). However, these methods can not solve the continuous Wasserstein barycenter problem. To address this issue, (Staib et al. 2017; Claici, Chien, and Solomon 2018; Dvurechenskii et al. 2018) assume the barycenter to be a finite set of points and rely on semi-discrete OT algorithms to compute the barycenter.

The continuous approximation methods for Wasserstein barycenters remain unexplored until recently (Li et al. 2020;

Fan, Taghvaei, and Chen 2021; Korotin et al. 2021b, 2022). Existing methods are based on dual potentials optimization or fixed point approaches. However, these methods are not *end-to-end* which consist of two sequential steps. They require recovering barycenters via an additional estimation, e.g., computing gradients of the potentials as push-forward maps or estimating Monge maps, which is itself a complex problem increasing the computational burden of the original problem. Closely related to this paper are the recent works (Li et al. 2020; Fan, Taghvaei, and Chen 2021; Korotin et al. 2021b) that rely on the dual formulation of the Wasserstein barycenter and represent the dual potentials with neural networks. In particular, the method in (Li et al. 2020) assumes a *fixed prior* as the proxy of the barycenter from the beginning, leading to inaccurate approximations especially for high-dimensional settings (Korotin et al. 2022). For the specific ground cost, i.e., *quadratic cost*, (Fan, Taghvaei, and Chen 2021; Korotin et al. 2021b) compute the barycenters under Wasserstein-2 distance. The method in (Fan, Taghvaei, and Chen 2021) employs the *generative network* to represent a barycenter, which suffers from the usual limitations of generative networks such as mode collapse. Although this method does not involve regularization terms, it ends up with a challenging min-max-min problem. Another well-known method in (Korotin et al. 2021b) constructs a new regularized formulation by adding a congruence regularization to ensure that optimal potential functions are consistent with the true barycenter. However, the congruence regularization requires the selection of a *prior* distribution that is bounded below by the barycenter, which is a non-trivial task. On the other hand, (Korotin et al. 2022) proposes an iterative algorithm for approximating the Wasserstein-2 barycenters based on the fixed point approach, where a *generative network* is used to parameterize the evolving barycenter. However, it is challenging to guarantee convergence to the true barycenter. Besides, (Cohen, Arbel, and Deisenroth 2020) proposes a generic algorithm to compute barycenters with respect to arbitrary discrepancies, which also parameterizes the barycenter by using a generative network. In contrast, our method is end-to-end and suitable for a general ground cost distance. Remarkably, we use a parameterized variational distribution as the approximation of the barycenter, allowing better flexibility. Table 1 summarizes the existing studies on continuous Wasserstein barycenter and shows our contributions.

Background and Preliminaries

In this section, we describe the Wasserstein barycenter problem, which involves all Wasserstein distances from one to many distributions.

Definition 1 (Wasserstein distance). *Let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary spaces equipped with a ground cost $c(x,y) = d(x,y)^p$ ($d(\cdot,\cdot)$ is a distance). Given two continuous probability distributions $\mu(x)$ and $\nu(y)$, for $p \in [1,\infty)$, the Wasserstein distance between μ and ν is defined as follows (Villani 2008):*

$$W_p^p(\mu,\nu) \triangleq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} \int_{\mathcal{Y}} \pi(x,y)c(x,y)dxdy, \quad (1)$$

where $\Pi(\mu, \nu)$ is the set of all joint distributions on $\mathcal{X} \times \mathcal{Y}$ with prescribed marginals $\mu(x)$ and $\nu(y)$. We also call an admissible π a transport plan.

The primal problem (1) admits an equivalent dual form (Villani 2008):

$$\sup_{\phi \oplus \psi \leq c} \int_{\mathcal{X}} \phi(x)\mu(x)dx + \int_{\mathcal{Y}} \psi(y)\nu(y)dy, \qquad (2)$$

where ϕ and ψ are dual potentials, and $(\phi \oplus \psi)(x, y) \triangleq \phi(x) + \psi(y)$. The constraint $\phi \oplus \psi \leq c$ means $\phi(x) + \psi(y) \leq c(x, y)$ for all (x, y).

Directly solving (1) and (2) is challenging, since the resulting linear problem can be too costly. In order to speed up the computation, regularized OT is introduced by (Cuturi 2013). Here, we consider the entropy and quadratic regularization:

$$\sup_{\phi, \psi} \int_{\mathcal{X}} \phi(x)\mu(x)dx + \int_{\mathcal{Y}} \psi(y)\nu(y)dy$$
$$+ \int_{\mathcal{X}} \int_{\mathcal{Y}} F(\phi(x) + \psi(y) - c(x, y))\mu(x)\nu(y)dxdy, (3)$$

where

$$\forall t \in \mathbb{R}, \quad F(t) = \begin{cases} -\varepsilon e^{\frac{t}{\varepsilon}} & \text{(entropy reg.)} \\ -\frac{1}{2\varepsilon}(t_+)^2 & \text{(quadratic reg.)} \end{cases} \quad (4)$$

Definition 2 (Wasserstein barycenter). *The Wasserstein barycenter of N ($N \geq 2$) continuous probability distributions μ_n, $n = 1, 2, \ldots, N$ with weights α_n ($\alpha_n > 0, \sum_{n=1}^{N} \alpha_n = 1$), is a solution of the following functional minimization problem (Agueh and Carlier 2011) :*

$$\inf_{\nu} \sum_{n=1}^{N} \alpha_n W_p^p(\mu_n, \nu) \qquad (5)$$

Regularized Wasserstein barycenter dual. Since the primal objective (5) is hard to compute, (Li et al. 2020) defines a new dual version of regularized Wasserstein barycenter problem by using regularized Wasserstein distance (3), which is a more popular and efficient alternative:

$$\inf_{\nu} \sup_{\substack{\phi, \psi \\ \sum_{n=1}^{N} \alpha_n \psi_n = 0}} \sum_{n=1}^{N} \alpha_n \left[\int_{\mathcal{X}} \phi_n(x_n)\mu_n(x_n)dx_n + \right. \qquad (6)$$

$$\left. \int_{\mathcal{X}} \int_{\mathcal{Y}} F(\phi_n(x_n) + \psi_n(y) - c(x_n, y))\mu_n(x_n)\nu(y)dx_n dy \right]$$

where ϕ_n and ψ_n are dual potentials, and F refers to either entropy or quadratic regularization. We notice that ν is the true barycenter, not a prior like in (Li et al. 2020). We rely on this formulation in the next section to propose a new regularized objective.

Our Method

Our goal in this paper is to compute the Wasserstein barycenter for a given set of continuous distributions $\{\mu_1, \ldots, \mu_N\}$. To effectively solve this problem, we present a novel end-to-end method, namely **V**ariational **W**asserstein **B**arycenter with c-**C**yclical **M**onotonicity **R**egularization (**VWB-CMR**). Our method is well-suited for the case where the analytic forms of input distributions are not available. Since we only have access to independent samples which can be drawn from input distributions or are provided by machine learning applications.

Imposing the C-cyclical Monotone Condition

Inspired by (Li et al. 2020), we rely on a regularized formulation to compute the Wasserstein barycenter. However, (6) cannot provide precise approximations in high dimensions (Korotin et al. 2021b). To address this issue, we enforce c-cyclical monotone condition via a regularization.

A set $\Gamma \subseteq \mathcal{X} \times \mathcal{Y}$ is said c-*cyclical monotone*, if for any $k \in \mathbb{N}$ ($k \geq 2$), any permutation $\sigma \in P(k)$ and any finite pairs $(x^1, y^1), \ldots, (x^k, y^k) \in \Gamma$, one can have

$$\sum_{i=1}^{k} c(x^i, y^i) \leq \sum_{i=1}^{k} c(x^i, y^{\sigma(i)}), \qquad (7)$$

where $P(k)$ is the set of all permutations of $\{1, \ldots, k\}$. For a finite set Γ, c-cyclical monotonicity means the points of origin x^i and destination y^i related by (x^i, y^i) have been paired so as to minimize the total transportation cost $\sum_{\Gamma} c(x, y)$. Otherwise, it would be more efficient to move mass from all x^i to y^i.

Importantly, c-cyclical monotonicity is a necessary condition for optimality in OT (Gangbo and McCann 1996), and is a sufficient condition on a weak assumption (Pratelli 2008).

Theorem 1 (Necessary and sufficient optimality conditions). *Assume a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup +\infty$ is continuous. Then a transport plan π for the Wasserstein distance is optimal (i.e., it is a solution of problem (1)) if and only if the support of π, i.e., Γ, is c-cyclical monotone.*

This theorem can be proved by contradiction. The detailed discussion can be found in (Pratelli 2008).

In order to enforce c-cyclical monotonicity, we define

$$I(x, y) \triangleq \sup_{\sigma \in P(k)} \sum_{i=1}^{k} c(x^i, y^i) - \sum_{i=1}^{k} c(x^i, y^{\sigma(i)})$$

Then $I : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$. For all $(x, y) \in \Gamma$, the optimality of π implies that $I = 0$ since Γ is c-cyclically monotone. Combing the constraint of dual formulation $\phi(x^i) + \psi(y^i) \leq c(x^i, y^i)$, we want to maximize $\sum_{i=1}^{k} \phi(x^i) + \psi(y^i) - c(x^i, y^{\sigma(i)})$. The supremum value of this difference is zero when $(x^i, y^i) \in \Gamma$. But outside Γ, we expect this supremum value is greater than zero. Hence we consider an additional convex regularization:

$$F(\sum_{i=1}^{k} c(x^i, y^{\sigma(i)}) - \phi(x^i) - \psi(y^i)), \qquad (8)$$

where F refers to either entropy or quadratic regularization defined in (4). As a result, this regularization can help to find the optimal dual potentials, so as to make more

precise approximations of barycenters, especially for high-dimensional settings, see more details in experiments.

Plugging (8) into (6), we obtain a new regularized objective of the Wasserstein barycenter problem:

$$\inf_{\substack{\nu \\ \sum_{n=1}^{N} \alpha_n \psi_n = 0}} \sup_{\phi,\psi} \sum_{n=1}^{N} \alpha_n \left[\int_{\mathcal{X}} \phi_n(x_n)\mu_n(x_n)dx_n + \right.$$

$$\left. \int_{\mathcal{X}} \int_{\mathcal{Y}} R(\phi_n(x_n),\psi_n(y))\mu_n(x_n)\nu(y)dx_n dy \right], (9)$$

where

$$R(\phi_n(x_n),\psi_n(y)) \triangleq F(\phi_n(x_n) + \psi_n(y) - c(x_n,y)) +$$

$$F(\sum_{l=1}^{k} c(x_n^l, y^{\sigma(l)}) - \phi_n(x_n^l) - \psi_n(y^l)) \quad (10)$$

For a fixed barycenter ν, (9) is concave with respect to ϕ_n and ψ_n which can be maximized through stochastic gradient methods by using deep neural networks for parameterizing ϕ_n and ψ_n.

Discussion. Though the c-cyclical monotone condition is true for any permutation σ, it is enough to check this condition by using one permutation. In practice, we have examined more permutations and found that using a single permutation can make good performances in most cases. There are cases where using two permutations is slightly better than using just one. However, the performance does not get significantly better as the permutation increases, and is even worse than using one or two permutations. We believe that more regularization terms can introduce bias. In our experiments, we only report the results of using a single permutation $\sum_{l=1}^{k} c(x^l, y^{k+1-l})$ in c-cyclical monotonicity regularization, and set k equal to the number of Monte Carlo samples as indicated in ***Algorithm 1***.

Introducing a Variational Distribution

In the context of estimating barycenters of continuous distributions, we propose to represent the barycenter by using a variational distribution $\nu'(\cdot|\lambda)$ with parameters λ. If we know the input distributions are in the same family, e.g., Gaussian, we can set the variational distribution ν' to be a Gaussian with unknown mean and covariance. Otherwise we may consider $\nu' = \Pi_n^N \nu_n'(\cdot|\lambda_n)$, where the type of ν_n' coincides with each input distribution μ_n. In the case where a fixed set of samples is provided and no information about the barycenter is known beforehand, we can assume that the variational distribution is a Gaussian mixture, which has the ability to approximate any functions (Anzai 2012).

Having specified a variational distribution $\nu'(y|\lambda)$, we transform the barycenter estimation problem into an optimization problem, where the parameters λ to be optimized adjust a variational "proxy" distribution to be similar to the barycenter. We rewrite (9) with variational distribution $\nu'(y|\lambda)$ and replace each ψ_n with $\psi_n - \sum_{j=1}^{N} \alpha_j \psi_j$ to ob-

tain an unconstrained version as follows:

$$\inf_{\lambda} \sup_{\phi,\psi} \mathcal{L}(\lambda,\phi,\psi) \triangleq \sum_{n=1}^{N} \alpha_n \left[\int_{\mathcal{X}} \phi_n(x_n)\mu_n(x_n)dx_n + \right.$$

$$\left. \int_{\mathcal{X}} \int_{\mathcal{Y}} R(\phi_n(x_n),\psi_n(y))\mu_n(x_n)\nu'(y|\lambda)dx_n dy \right], (11)$$

where

$$R(\phi_n(x_n),\psi_n(y)) =$$

$$F(\sum_{l=1}^{k} c(x_n^l, y^{\sigma(l)}) - \phi_n(x_n^l) - \psi_n(y^l) + \sum_{j=1}^{N} \alpha_j \psi_j(y^l))$$

$$+F(\phi_n(x_n) + \psi_n(y) - \sum_{j=1}^{N} \alpha_j \psi_j(y) - c(x_n,y)) \quad (12)$$

Thus we obtain a tractable objective with respect to variational parameters. We can update variational parameters through stochastic gradient methods.

The noisy gradients of $\mathcal{L}$. To optimize the objective $\mathcal{L}$ with gradient-based methods, we need to develop an unbiased estimator of its gradients which can be computed using Monte Carlo samples. To do this, we derive gradients of $\mathcal{L}$ with respect to λ as an expectation form:

$$\nabla_\lambda \mathcal{L}$$

$$= \nabla_\lambda \sum_{n=1}^{N} \alpha_n \int_{\mathcal{X}} \int_{\mathcal{Y}} R(\phi_n,\psi_n)\mu_n(x_n)\nu'(y|\lambda)dx_n dy$$

$$= \sum_{n=1}^{N} \alpha_n \int_{\mathcal{X}} \int_{\mathcal{Y}} \nabla_\lambda R(\phi_n,\psi_n)\mu_n(x_n)\nu'(y|\lambda)dx_n dy$$

$$= \sum_{n=1}^{N} \alpha_n \mathbb{E}_{\mu_n \nu'}[\nabla_\lambda \log \nu'(y|\lambda) R(\phi_n,\psi_n)] \quad (13)$$

where $\nabla_\lambda \log \nu'(y|\lambda)$ is called the score function (Cox and Hinkley 1979), and it can be calculated using automatic differentiation tools (Team 2015). Note that $\nabla_\lambda[\nu'(y|\lambda)] = \nabla_\lambda[\log \nu'(y|\lambda)]\nu'(y|\lambda)$. For brevity, we omit notations of x_n and y in R.

With this equation in hand, we can directly compute the gradient using Monte Carlo samples drawn from $\mu(x)\nu'(y|\lambda)$:

$$\nabla_\lambda \mathcal{L}$$

$$\approx \frac{1}{S} \sum_{n=1}^{N} \sum_{s=1}^{S} \alpha_n \nabla_\lambda \log \nu'(y^{(s)}|\lambda) R(\phi_n(x_n^{(s)}), \psi_n(y^{(s)}))$$

where S is the number of Monte Carlo samples. Then, at each iteration t, the parameter of interest λ can be updated as follows:

$$\lambda_t \leftarrow \lambda_{t-1} - \rho_t \nabla_\lambda \mathcal{L},$$

where ρ_t is the learning rate.

We notice that the optimization over ν is not convex. However, it guarantees to converge to a local optimum, if the learning rate satisfies the Robbins-Monro condition (Robbins and Monro 1951):

$$\sum_{t=1}^{\infty} \rho_t = \infty, \quad \sum_{t=1}^{\infty} \rho_t^2 < \infty$$

Algorithm 1: Optimization of VWB-CMR

1: **Input**: Continuous distributions $\{\mu_n\}_{n=1}^N$ with weight α_n; cost function c; batch size S; regularization F; learning rate ρ; network gradient update function *BackWard*.
2: **While** *not converged* do
3:    $\forall n \in \{1, ..., N\}$: sample x_n from μ_n
4:    sample y from $\nu'(y|\lambda)$ and obtain a permutation y^σ
5:    $\overline{\psi}_y \leftarrow \sum_{j=1}^N \alpha_j \psi_j(y)$
6:    **For** $n = 1$ to N do
7:      $R_n^1 \leftarrow F(\sum_{l=1}^S c(x_n^l, y^{\sigma(l)}) - \phi_n(x_n^l) - \psi_n(y^l) + \overline{\psi}_{y^l})$
8:      $R_n^2 \leftarrow F(\phi_n(x_n) + \psi_n(y) - \overline{\psi}_y - c(x_n, y))$
9:      $R_n \leftarrow R_n^1 + R_n^2$
10:   **End For**
11:   $\mathcal{L} \leftarrow \sum_{n=1}^N \alpha_n(\phi_n(x_n) + R_n)$
12:   **For** $n = 1$ to N do **in parallel**
13:      *BackWard($\mathcal{L}, \phi_n$)*;   *BackWard($\mathcal{L}, \psi_n$)*
14:   **End For**
15:   $f \leftarrow \sum_{n=1}^N \alpha_n \nabla_\lambda \log \nu'(y|\lambda) R_n$
16:   $h \leftarrow \nabla_\lambda \log \nu'(y|\lambda)$
17:   $a^* \leftarrow \text{Cov}(f, h)/\text{Var}(h)$
18:   update $\lambda \leftarrow \lambda - \rho(f - a^* h)$
19: **End While**
20: **Output**: The continuous Wasserstein barycenter ν' of $\{\mu_n\}_{n=1}^N$.

Variance reduction. The noisy gradients formed by using Monte Carlo samples often can be too large to be useful. In practice, the high variance would lead to slow convergence and even worse performance (Paisley, Blei, and Jordan 2012).

To alleviate this, we use the control variate to reduce variance (Ross 2002; Paisley, Blei, and Jordan 2012), which is a family of functions with equivalent expectations. The basic idea is to replace the target function with another function that has the same expectation but a smaller variance. For example, when we compute $\mathbb{E}[f]$ with Monte Carlo samples, we use the empirical average of $\hat{f}$ where $\hat{f}$ is chosen so $\mathbb{E}[f] = \mathbb{E}[\hat{f}]$ and $\text{Var}[f] > \text{Var}[\hat{f}]$. Next, we describe how to reduce the variance via easy-to-implement control variates.

Define $\hat{f}$ to be

$$\hat{f}(z) \triangleq f(z) - a(h(z) - \mathbb{E}[h(z)]),$$

where $h(z)$ is a function with a finite first moment, and a is a scalar. Note that $\mathbb{E}[f] = \mathbb{E}[\hat{f}]$ as required. The variance of $\hat{f}$ is:

$$\text{Var}(\hat{f}) = \text{Var}(f) + a^2 \text{Var}(h) - 2a\text{Cov}(f, h)$$

This equation implies that a good control variate $\hat{f}$ with a smaller variance will have high covariance with the function f. Given a function h, taking the derivative of $\text{Var}(\hat{f})$ with respect to a and setting it equal to zero, one can get the optimal scaling,

$$a^* = \text{Cov}(f, h)/\text{Var}(h),$$

which can be estimated with the ratio of empirical covariance and variance using Monte Carlo samples.

Inspired by this, in our method we choose h to be the score function of the variational distribution, i.e., $\nabla_\lambda \log \nu'(y|\lambda)$ which always has expectation zero, and re-compute the gradient of $\mathcal{L}$ with respect to λ using a new Monte Carlo method:

$$\nabla_\lambda \mathcal{L} \approx$$
$$\frac{1}{S} \sum_{n=1}^N \sum_{s=1}^S \alpha_n \nabla_\lambda \log \nu'(y^{(s)}|\lambda)(R(\phi_n(x_n^{(s)}), \psi_n(y^{(s)}) - a^*)$$
$$x_n^{(s)} \sim \mu_n(x_n), y^{(s)} \sim \nu'(y|\lambda) \quad (14)$$

Finally, we summarize the full algorithm of VWB-CMR in *Algorithm 1*.

Computation complexity. For our method, as well as the methods proposed in (Li et al. 2020; Korotin et al. 2021b; Fan, Taghvaei, and Chen 2021; Korotin et al. 2022), estimating Wasserstein barycenters includes repeated computation on N input distributions. Therefore, the computational complexity per iteration scales with $O(NSP)$ where P is the size of the networks. Besides, these previous methods require recovering the barycenter by an additional calculation, which is not an easy task scaling with square level complexity at least. However, our method is end-to-end and can easily be parallelized, where N dual variable pairs (ϕ_n, ψ_n) can be computed in a fully parallel manner.

Theoretical Results

In this section, we discuss the theoretical guarantee for the convergence of our method.

The work (Agueh and Carlier 2011) shows if at least one of the distributions μ_n is absolutely continuous, then the Wasserstein barycenter ν is unique. Moreover, ν is also absolutely continuous. This analysis ensures the existence and uniqueness of the Wasserstein barycenter for continuous distributions.

Theorem 2 (Convergence). *Let $\mu_1, ..., \mu_N$ be continuous distributions with respect to the Lebesgue measure. Assume a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup +\infty$ is continuous. If $\{\phi_n, \psi_n\}_{n=1}^N$ are the optimal dual potentials in (9), then each $\{\phi_n, \psi_n\}$ is a solution to the regularized dual formulation (3). Let ν^ε the solution of regularized Wasserstein barycenter problem (9), and let ε converge to 0 sufficiently fast. Then, ν^ε converges weakly to the solution ν of the Wasserstein barycenter problem (5).*

We include the proof of Theorem 2 in the supplementary document.

Experiments

In this section, we demonstrate the effectiveness of VWB-CMR on both synthetic and real data.

Experimental Setup

Our aim is to examine whether VWB-CMR can accurately approximate the continuous barycenter. To this end, we compute the Wasserstein barycenter with the squared Euclidean

distance as the cost function, i.e., $c(x,y) = \|x-y\|_2^2$, $x, y \in \mathbb{R}^D$. In all experiments, we use equal weights for input distributions, i.e., $\alpha_n = \frac{1}{N}$ for all $n = 1, \ldots, N$. Note that the proposed method is not limited to Euclidean distance and equal weights, it can be applied to a more general setting.

We compare VWB-CMR against the following state of the art methods: (i) continuous regularized Wasserstein barycenter (CRWB) [1] (Li et al. 2020); (ii) continuous Wasserstein barycenter without minimax optimization (CWB) [2] (Korotin et al. 2021b); (iii) scalable computations of Wasserstein barycenter via input convex neural networks (SCWB) [3](Fan, Taghvaei, and Chen 2021). These methods recover barycenter through the gradient-based method. Besides, we consider quadratic regularization in CRWB, since it performs better than entropy regularization.

In our method, we use Adam method to adjust the learning rate, where parameters β_1=0.9, β_2=0.999 and α=0.001. For the training, the number of iterations and batch size are 20000 and 1024, respectively. The dual potentials $\{\phi_n, \psi_n\}_{n=1}^N$ are parameterized as neural networks with two fully-connected hidden layers ($D \to 128 \to 256 \to D$) using ReLU activations. We consider quadratic regularization and report the average results of 5 independent runs for all experiments. The choice of regularization parameter ε depends on the examples. See the supplementary document for more details and empirical results.

Learning the Wasserstein Barycenter in 2D

Figure 2 shows the qualitative performance of our method on 2D examples. Each example is represented in a row. The first and second columns are input distributions, and the third column is the approximated barycenter by using our method. In the first example, we set the variational distribution ν' to be a Gaussian mixture with 30 components. In the second example, we set ν' to be a Gaussian mixture with 20 components. We can find that VWB-CMR learns the barycenter qualitatively well. For comparison with CRWB and SCWB, see (Li et al. 2020, Figure 1) and (Fan, Taghvaei, and Chen 2021, Figure 4).

Evaluation on Multivariate Gaussians

Though (Korotin et al. 2022) constructs a dataset for quantitative evaluation of barycenter methods, this dataset may be only suitable for methods using generative models in high-dimensional settings (e.g., ≥ 16). Therefore, for quantitative evaluation we compute the Wasserstein barycenters of N multivariate Gaussians $\mathcal{N}(m_n, \Sigma_n)$ in dimension D with mean m_n and non-diagonal covariance matrix Σ_n. When the input distributions are multivariate Gaussians, the Wasserstein barycenter is always a Gaussian $\mathcal{N}(m_*, \Sigma_*)$, where $m_* = \sum_{n=1}^N \alpha_n m_n$ and $\Sigma_* = \sum_{n=1}^N \alpha_n (\Sigma_*^{\frac{1}{2}} \Sigma_n \Sigma_*^{\frac{1}{2}})^{\frac{1}{2}}$ computed through an efficient fixed-point algorithm (Álvarez-Esteban et al. 2016).

[1]https://github.com/lingxiaoli94/CWB

[2]https://github.com/iamalexkorotin/Wasserstein2Barycenters

[3]https://github.com/sbyebss/Scalable-Wasserstein-Barycenter

To evaluate the performance of these methods, we use the **B**ures-**W**asserstein **U**nexplained **V**ariance **P**ercentage (BW$_2^2$-UVP) to measure whether the approximation of the Wasserstein barycenter, denoted by ν with mean m_ν and covariance Σ_ν, approaches the true one $\tilde{\nu}$ with mean $m_{\tilde{\nu}}$ and covariance $\Sigma_{\tilde{\nu}}$ (Korotin et al. 2021b):

$$\mathrm{BW}_2^2 - \mathrm{UVP}(\nu, \tilde{\nu}) \triangleq 100 \frac{\mathrm{BW}_2^2(\nu, \tilde{\nu})}{\frac{1}{2}\mathrm{Var}(\tilde{\nu})}\%,$$

where $\mathrm{BW}_2^2(\nu, \tilde{\nu})$ equals

$$\frac{1}{2}\|m_\nu - m_{\tilde{\nu}}\|^2 + \frac{1}{2}\mathrm{Tr}\left(\Sigma_\nu + \Sigma_{\tilde{\nu}} - 2(\Sigma_\nu^{1/2}\Sigma_{\tilde{\nu}}\Sigma_\nu^{1/2})^{1/2}\right).$$

Here $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. When $\mathrm{BW}_2^2 - \mathrm{UVP} \approx 0\%$, ν is a good approximation of $\tilde{\nu}$. In the case of $\mathrm{BW}_2^2 - \mathrm{UVP} \geq 100\%$, the approximation ν is undesirable.

The results of 3 randomly generated multivariate Gaussians in varying dimensions D are shown in Table 2. We can find that the estimation errors of VWB-CMR are lower than those of other methods in all cases. These results imply that VWB-CMR can make good approximations of continuous Wasserstein barycenters, in particular for high-dimensional cases. To further demonstrate the effectiveness of c-cyclical monotonicity regularization, we give the result of the ablative version of VWB-CMR, namely VWB, which computes the barycenters based on the original regularized objective (6) without this regularization. We see that VWB-CMR is significantly better than VWB. This result directly indicates the positive impact of c-cyclical monotonicity. On the other hand, VWB outperforms CRWB, which shows the effectiveness of the variational distribution. Because the main difference between VWB and CRWB is that VWB introduces a variational distribution to represent the barycenter instead of a fixed prior in CRWB.

Convergence. Figure 3 shows the convergence curves of VWB-CMR with quadratic and entropy regularization at D=128 and 256 within 25000 iterations. Though the optimization over ν' is not convex, we can see that VWB-CMR
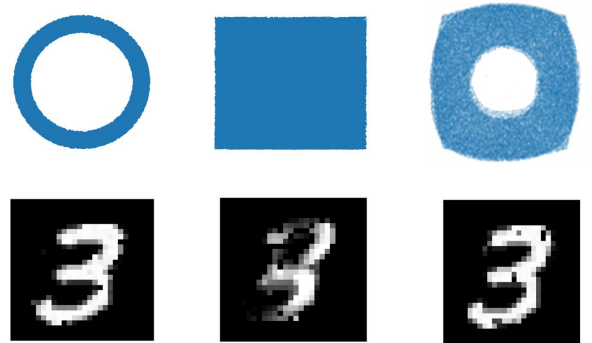

Figure 2: Qualitative results of VWB-CMR in 2D setting. The first and second columns are input distributions, and the third column is the barycenter learned by VWB-CMR.

Method	$D = 2$	4	8	16	32	64	128	256
VWB-CMR (**Ours**)	**0.02**	**0.06**	**0.11**	**0.13**	**0.17**	**0.20**	**0.40**	**1.10**
VWB	0.04	0.11	0.21	0.24	0.30	0.77	5.01	14.65
CRWB (Li et al. 2020)	0.12	0.62	0.84	9.58	25.81	26.31	51.75	>100
CWB (Korotin et al. 2021b)	0.06	0.10	0.17	0.20	0.21	0.25	1.92	4.81
SCWB (Fan, Taghvaei, and Chen 2021)	0.03	0.12	0.16	0.21	0.40	0.67	1.42	3.41

Table 2: Numerical results of the $\mathrm{BW}_2^2 - \mathrm{UVP}$ error for estimating barycenters of Gaussians of varying dimensions. Smaller is better. The second row lists the results of the ablative version of VWB-CMR without the c-cyclical monotonicity regularization (CMR).
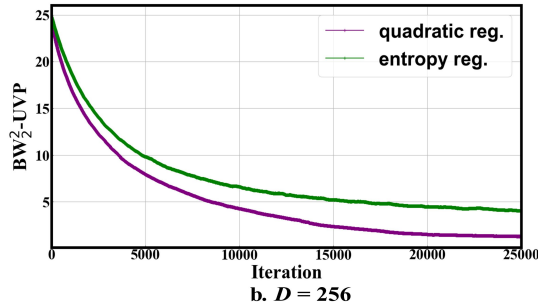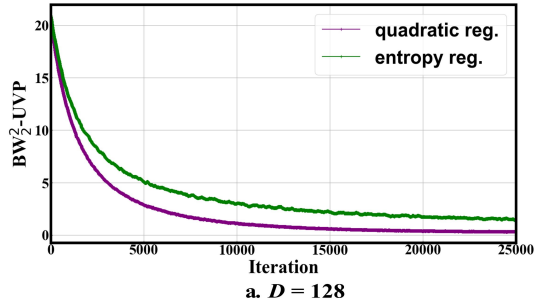

Figure 3: Convergence curves of VWB-CMR with quadratic and entropy regularization at D=128 and 256.

Method	$\mathrm{BW}_2^2 - \mathrm{UVP}$
CRWB (Li et al. 2020)	0.96
CWB (Korotin et al. 2021b)	0.13
SCWB (Fan, Taghvaei, and Chen 2021)	0.17
VWB-CMR (**Ours**)	**0.10**

Table 3: Empirical results on subset posterior aggregation. The $\mathrm{BW}_2^2 - \mathrm{UVP}$ is computed for comparison by CRWB, CWB, SCWB and VWB-CMR.

into 5 subsets with equal-size and get 10^5 samples from each subset posterior using the Stan library (Carpenter et al. 2017). Then we compute the Wasserstein barycenter of these subset posterior distributions by all methods. In our method, we set ν' to be a Gaussian. To evaluate the performance of these methods, we compute the BW_2^2-UVP between the approximated barycenter and the full data posterior.

The results are shown in Table 3. All methods approach the true Wasserstein barycenter well since the values of $\mathrm{BW}_2^2 - \mathrm{UVP}<1\%$. The performance of VWB-CMR is better than CRWB, CWB and SCWB. This further indicates that VWB-CMR is an effective alternative to compute continuous Wasserstein barycenters.

can converge to a local optimum. Besides, the empirical results show that quadratic regularization is more stable than entropy regularization, since there is no exponential term causing overflow.

Evaluation on Real Data

We now evaluate VWB-CMR in real-world applications. We apply the Wasserstein barycenter to aggregate posterior distributions of subsets in a large data set to approximate the true posterior of the full data. Following (Li et al. 2020), we use Poisson regression for the task of predicting the hourly number of bike rentals using features such as the day of the week and weather conditions [4]. We consider the posterior distribution on the 8-dimensional regression coefficients for the Poisson model. We firstly randomly split the full data

---
[4]http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Conclusion

We develop a novel VWB-CMR method to compute the Wasserstein barycenters of continuous distributions. The main idea is to explore c-cyclical monotone condition to construct a new regularized objective and introduce a variational distribution to represent the barycenter. The objective can be solved by stochastic optimization only requiring samples access to the input distributions, where variational parameters are optimized to find the closest distribution as the approximation of the barycenter. The limitation of our method is that we require to adjust the regularization parameter in different applications. One future direction is to find a way to solve this issue dynamically.

Acknowledgments

References

Agueh, M.; and Carlier, G. 2011. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924.

Álvarez-Esteban, P. C.; Del Barrio, E.; Cuesta-Albertos, J.; and Matrán, C. 2016. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2): 744–762.

Anderes, E.; Borgwardt, S.; and Miller, J. 2016. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2): 389–409.

Anzai, Y. 2012. *Pattern recognition and machine learning*. Elsevier.

Backhoff-Veraguas, J.; Fontbona, J.; Rios, G.; and Tobar, F. 2022. Stochastic gradient descent in Wasserstein space. *arXiv preprint arXiv:2201.04232*.

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

Bonneel, N.; Peyré, G.; and Cuturi, M. 2016. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4): 1–10.

Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; and Riddell, A. 2017. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1): 1–32.

Claici, S.; Chien, E.; and Solomon, J. 2018. Stochastic Wasserstein Barycenters. In *International Conference on Machine Learning*, 999–1008.

Cohen, S.; Arbel, M.; and Deisenroth, M. P. 2020. Estimating barycenters of measures in high dimensions. *arXiv preprint arXiv:2007.07105*.

Cox, D. R.; and Hinkley, D. V. 1979. *Theoretical statistics*. CRC Press.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems*, 2292–2300.

Cuturi, M.; and Doucet, A. 2014. Fast Computation of Wasserstein Barycenters. In *International Conference on Machine Learning*, 685–693.

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

Cuturi, M.; and Peyré, G. 2018. Semidual regularized optimal transport. *SIAM Review*, 60(4): 941–965.

Daaloul, C.; Gouic, T. L.; Liandrat, J.; and Tournus, M. 2021. Sampling from the wasserstein barycenter. *arXiv preprint arXiv:2105.01706*.

Dognin, P.; Melnyk, I.; Mroueh, Y.; Ross, J.; Santos, C. D.; and Sercu, T. 2019. Wasserstein barycenter model ensembling. *arXiv preprint arXiv:1902.04999*.

Dvurechenskii, P.; Dvinskikh, D.; Gasnikov, A.; Uribe, C.; and Nedich, A. 2018. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *Neural Information Processing Systems*, 10760–10770.

Fan, J.; Taghvaei, A.; and Chen, Y. 2021. Scalable computations of wasserstein barycenter via input convex neural networks. In *International Conference on Machine Learning*, 1571–1581.

Gangbo, W.; and McCann, R. J. 1996. The geometry of optimal transportation. *Acta Mathematica*, 177(2): 113–161.

Ge, D.; Wang, H.; Xiong, Z.; and Ye, Y. 2019. Interior-point methods strike back: solving the Wasserstein barycenter problem. In *Neural Information Processing Systems*, 6894–6905.

Ho, N.; Huynh, V.; Phung, D.; and Jordan, M. 2019. Probabilistic multilevel clustering via composite transportation distance. In *International Conference on Artificial Intelligence and Statistics*, 3149–3157.

Izzo, Z.; Silwal, S.; and Zhou, S. 2021. Dimensionality reduction for wasserstein barycenter. In *Neural Information Processing Systems*, 15582–15594.

Janati, H.; Cuturi, M.; and Gramfort, A. 2020. Debiased sinkhorn barycenters. In *International Conference on Machine Learning*, 4692–4701.

Kantorovitch, L. 1942. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2): 227–229.

Korotin, A.; Egiazarian, V.; Li, L.; and Burnaev, E. 2022. Wasserstein iterative networks for barycenter estimation. *arXiv preprint arXiv:2201.12245*.

Korotin, A.; Li, L.; Genevay, A.; Solomon, J.; Filippov, A.; and Burnaev, E. 2021a. Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark. In *Neural Information Processing Systems*, 14593–14605.

Korotin, A.; Li, L.; Solomon, J.; and Burnaev, E. 2021b. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*.

Le, K.; Nguyen, H.; Nguyen, Q.; Pham, T.; Bui, H.; and Ho, N. 2021. On Robust Optimal Transport: Computational Complexity and Barycenter Computation. In *Neural Information Processing Systems*, 21947–21959.

Li, L.; Genevay, A.; Yurochkin, M.; and Solomon, J. M. 2020. Continuous regularized wasserstein barycenters. In *Neural Information Processing Systems*, 17755–17765.

Lin, T.; Ho, N.; Chen, X.; Cuturi, M.; and Jordan, M. 2020. Fixed-Support Wasserstein Barycenters: Computational Hardness and Fast Algorithm. In *Neural Information Processing Systems*, 5368–5380.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Paisley, J. W.; Blei, D. M.; and Jordan, M. I. 2012. Variational Bayesian Inference with Stochastic Search. In *International Conference on Machine Learning*, 1367–1374.

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

Pratelli, A. 2008. On the sufficiency of c-cyclical monotonicity for optimality of transport plans. *Mathematische Zeitschrift*, 258(3): 677–690.

Rabin, J.; Peyré, G.; Delon, J.; and Bernot, M. 2011. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 435–446.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Ross, S. M. 2002. *Simulation*. Elsevier.

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

Staib, M.; Claici, S.; Solomon, J. M.; and Jegelka, S. 2017. Parallel streaming Wasserstein barycenters. In *Neural Information Processing Systems*, 2647–2658.

Team, S. D. 2015. Stan: A c++ library for probability and sampling, version 2.8.0. https://mc-stan.org/. Accessed: 2022-05-10.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Yang, L.; Li, J.; Sun, D.; and Toh, K.-C. 2018. A fast globally linearly convergent algorithm for the computation of Wasserstein barycenters. *arXiv preprint arXiv:1809.04249*.

Ye, J.; Wu, P.; Wang, J. Z.; and Li, J. 2017. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9): 2317–2332.