Approximate Optimal Transport for Continuous Densities with Copulas

Jinjin Chi[1,2], **Jihong Ouyang**[1,2], **Ximing Li**[1,2*], **Yang Wang**[3] and **Meng Wang**[3]
[1] College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[3] School of Computer Science and Information Engineering, Hefei University of Technology, China
chijinjin616@gmail.com, ouyj@jlu.edu.cn, liximing86@gmail.com, yangwang@hfut.edu.cn
eric.mengwang@gmail.com

Abstract

Optimal Transport (OT) formulates a powerful framework by comparing probability distributions, and it has increasingly attracted great attention within the machine learning community. However, it suffers from severe computational burden, due to the intractable objective with respect to the distributions of interest. Especially, there still exist very few attempts for continuous OT, *i.e.,* OT for comparing continuous densities. To this end, we develop a novel continuous OT method, namely Copula OT (Cop-OT). The basic idea is to transform the primal objective of continuous OT into a tractable form with respect to the copula parameter, which can be efficiently solved by stochastic optimization with less time and memory requirements. Empirical results on real applications of image retrieval and synthetic data demonstrate that our Cop-OT can gain more accurate approximations to continuous OT values than the state-of-the-art baselines.

1 Introduction

Optimal **T**ransport (**OT**), a powerful distance for comparing probability distributions, has recently attracted great attention in machine learning area [Rubner *et al.*, 2000; Pele and Werman, 2009; Ni *et al.*, 2009; Courty *et al.*, 2017; Arjovsky *et al.*, 2017; Wang *et al.*, 2018; Li *et al.*, 2019]. In contrast to other popular distances of distributions, *e.g.,* KL-divergence, (1) OT is more effective to describe the geometry of distributions by considering the spatial location of the distribution modes [Villani, 2003]; (2) it has also gained the superior performance in real applications, such as image retrieval [Pele and Werman, 2009], image segmentation [Ni *et al.*, 2009] and generative adversarial network [Arjovsky *et al.*, 2017] *etc.*

The motivation of OT begins with the Monge problem [Monge, 1781]. That is, for any two D-dimensional probability distributions, denoted as $p(x)$ and $q(y)$ over metric spaces $\mathcal{X}$ and $\mathcal{Y}$, it consists of finding an optimal map $f(x)$: $\mathcal{X} \to \mathcal{Y}$ that transports the mass of $p(x)$ to $q(y)$ with minimum cost given a predefined cost function $c(x, y)$: $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$:

$$\begin{aligned} \min_{f} \quad & \int_{\mathcal{X}} c(x, f(x)) dp(x) \\ \mathbf{s.t.} \quad & f(x) \sim y \end{aligned}$$

One concern of the Monge's problem is that the map $f(x)$ that satisfies the constraint may not always exist, *e.g.,* when $p(x)$ is a discrete distribution. To achieve a more feasible solution, the Kantorovich problem [Kantorovitch, 1942], a relaxed version of the Monge problem, has been proposed. It transforms the optimal map problem into finding an **optimal joint distribution** $\pi(x, y)$ with marginals $p(x)$ and $q(y)$, *i.e.,* an cheapest transport plan from $p(x)$ to $q(y)$:

$$\begin{aligned} \min_{\pi \in \Pi(p,q)} \quad & \int_{\mathcal{Y}} \int_{\mathcal{X}} \pi(x, y) c(x, y) dx dy \\ \mathbf{s.t.} \quad & x \sim p(x), \quad y \sim q(y), \end{aligned} \tag{1}$$

where $\Pi(p, q)$ is the set of all joint distributions with marginals $p(x)$ and $q(y)$. For our case, we focus on the Kantorovich problem due to its feasibility, and claim OT for this problem in the subsequent sections of the paper.

Generally speaking, the primal objective of OT, *i.e.,* Eq.1, is intractable to compute, since it involves the optimization about the distributions of interest. A mainstream methodology is to formulate its dual problem with regularizers, *e.g.,* entropic and ℓ^2-norm penalization [Cuturi, 2013; Ferradans *et al.*, 2014; Genevay *et al.*, 2016; Seguy *et al.*, 2018; Blondel *et al.*, 2018], leading to approximate but easier solved objectives with respect to dual variables. For example, the representative work proposed in [Cuturi, 2013] incorporates an entropic regularization term into the dual OT problem of two discrete distributions, so as to achieve a convex objective, which can be solved by the Sinkhorn's algorithm. Naturally, they have been successfully applied to approximately computing OT values in many real applications [Kusner *et al.*, 2015].

The above existing methods, however, mainly focus on the discrete distributions, and cannot cope with continuous OT, *i.e.,* OT for comparing continuous densities. To the best of our knowledge, rarely few works [Genevay *et al.*, 2016; Seguy *et al.*, 2018] attempt to fill this gap for continuous OT. A typical method [Genevay *et al.*, 2016], referred to as **K**ernel **O**ptimal **T**ransport (**K-OT**), parameterizes the dual variables by kernel

*Corresponding Author

Method	Primal objective	Time complexity	Space complexity
K-OT [Genevay *et al.*, 2016]	$\times$	$O(T^3S^3D)$	$O(TSD + f_k(D))$
NN-OT [Seguy *et al.*, 2018]	$\times$	$O(TS^2D)$	$O(SD + f_n(D))$
Cop-OT (Ours)	$\surd$	$O(TSD^2)$	$O(SD + D^2)$

Table 1: A brief summary of methods for continuous OT

functions, so as to transform the regularized dual objective into an expectation form with respect to $p(x)q(y)$, and then it can be solved through stochastic optimization by drawing Monte Carlo samples from $p(x)q(y)$.

However, since at each iteration the update process has to recycle the samples of previous iterations, K-OT suffers from higher costs of $O(T^3S^3D)$ time complexity and $O(TSD + f_k(D))$ memory space, where S and T are sample numbers of per-iteration and the total number of iterations, and $f_k(D)$ specifies the memory cost of kernel functions. [Seguy *et al.*, 2018] utilizes neural networks to parameterize the dual variables, instead of kernel functions, leading to another candidate method of continuous OT, namely **N**eural **N**etwork **O**ptimal **T**ransport (**NN-OT**). That requires $O(TS^2D)$ time and $O(SD + f_n(D))$ space, where $f_n(D)$ specifies the memory cost of neural networks. However, one salient problem remains: their target is to find the optima of the regularized dual objective, rather than the primal objective of OT, *i.e.,* Eq.1, theoretically introducing biases. Another problem is that they are empirically insensitive to the regularization parameters, making them less practical.

Our contributions. In this work, we attempt to further contribute on this challenging topic of continuous OT. Compared with the prior arts of **K-OT** and **NN-OT** under in Table 1, our motivation is to directly optimize the primal objective of continuous OT, *i.e.,* Eq.1, instead of its regularized dual versions, so as to achieve more accurate approximations. Our basic idea is to formulate the joint distribution $\pi(x, y)$ using the well-established copula [Sklar, 1959], and then transform the primal objective with respect to distributions of interest into a tractable form with respect to the copula parameter, enabling to be efficiently solved by stochastic optimization. Following this, a novel continuous OT method, namely **Cop**ula **O**ptimal **T**ransport (**Cop-OT**), was developed. We first propose the **Cop-OT** for comparing 1-D continuous densities, and then extend it to high dimensional ones, *i.e.,* $D \geq 2$, with requirements of time, *i.e.,* $O(TSD^2)$, and memory, *i.e.,* $O(SD + D^2)$. Referring to Table 1, **Cop-OT** is more efficient than **K-OT**, and practically more efficient than **NN-OT** due to the expensive computational cost on neural networks. Empirical studies on synthetic data and real applications of image retrieval demonstrate that **Cop-OT** can gain more accurate approximations to continuous OT values than the state-of-the-art baselines of **K-OT** and **NN-OT**.

Related work. The existing works on OT mainly focus on comparing discrete distributions. In this setting, OT is actually equivalent to the linear programming problem, which can be solved in cubic complexity. Some attempts investigate approximate but more efficient methods. The mainstream methodology is built on the idea of formulating regularized dual objectives [Cuturi, 2013; Benamou *et al.*, 2015; Genevay *et al.*, 2016; Seguy *et al.*, 2018], and solving them using the commonly used optimization schemes, *e.g.,* the Sinkhorn's algorithm and stochastic optimization. Besides, there are some works [Aurenhammer *et al.*, 1998; Genevay *et al.*, 2016; Lévy and Schwindt, 2018; Seguy *et al.*, 2018] on the semi-discrete OT, *i.e.,* OT for comparing a continuous density and a discrete distribution. They employ the machinery of c-transforms to eliminate one dual variable, and then obtain convex finite dimensional optimizations, solved by the Newton method [Lévy and Schwindt, 2018] and stochastic optimization [Genevay *et al.*, 2016; Seguy *et al.*, 2018]. To the best of our knowledge, only very few works on continuous-OT, which is still an open problem we attempt to tackle in this paper.

2 Proposed Method

Our goal on the problem is for the computation of OT between continuous densities $p(x)$ and $q(y)$, *i.e.,* continuous OT. However, it is intractable to compute, since its target, *i.e.,* its objective of Eq.1, is to minimize the transport cost by finding an optimal joint distribution $\pi(x, y)$ from an unknown set of distributions with marginals $p(x)$ and $q(y)$. To effectively solve the continuous OT, we propose a novel method, namely **Cop**ula **O**ptimal **T**ransport (**Cop-OT**). The basic idea is to re-parameterize the joint distribution $\pi(x, y)$ with the well-established copula [Sklar, 1959] on $p(x)q(y)$, so as to transform the primal objective of continuous OT into a tractable objective with respect to the copula parameter, referring to as the **copula objective**. We directly derive its gradient to an expectation form with respect to $p(x)q(y)$, therefore we can efficiently solve it following the spirit of stochastic optimization, where we iteratively update the copula parameter by forming noisy gradients using Monte Carlo samples drawn from $p(x)q(y)$.

For the rest of this section, we briefly review the concept of copula, and then describe the **Cop-OT** for 1-D continuous densities and the extension to high dimensional ones, *i.e.,* $D \geq 2$.

2.1 Copula

The story of copula begins with the theorem of Sklar [Sklar, 1959]: For any continuous joint distribution $\pi(x, y), x, y \in \mathbb{R}^D$, its Cumulative Distribution Function (CDF), denoted as $F(\cdot)$, can be represented by a copula with respect to the

marginal CDFs $F_i(\cdot)$ as follows:

$$F(x,y) = \mathbf{COP}\left(F_1(x_1), \cdots, F_D(x_D), F_{D+1}(y_1), \cdots, F_{2D}(y_D)|\eta\right),$$

where η is the copula parameter. Then, one can directly derive the following equation of $\pi(x,y)$:

$$\pi(x,y) = \prod_{i=1}^{D} p_i(x_i)q_i(y_i) \mathbf{cop}\left(F_1(x_1), \cdots, F_D(x_D), F_{D+1}(y_1), \cdots, F_{2D}(y_D)|\eta\right), \quad (2)$$

where $\mathbf{cop}(\cdot)$ denotes the copula density; $p_i(\cdot)$ and $q_i(\cdot)$ are the marginals of $p(x)$ and $q(y)$, respectively. For brevity, we omit the notation of CDF, *i.e.,* $F(\cdot)$, enabling to re-write Eq.2 as follows:

$$\pi(x,y) = \prod_{i=1}^{D} p_i(x_i)q_i(y_i)\mathbf{cop}(x_1, y_1, \cdots, x_D, y_D|\eta) \quad (3)$$

In practice, there are many well-established families of copulas, including Gaussian, Clayton, Frank, Gumbel, Joe and Student-t copulas [Dissmann *et al.*, 2013]. Given these copulas, one can describe any continuous joint distribution via a copula-related form parameterized by η.

2.2 Cop-OT for 1-D Continuous Densities

We are now ready to propose the Cop-OT for 1-D continuous densities, where both $p(x)$ and $q(y)$ are 1-dimensional densities. Under this situation, referring to Eq.3, the joint distribution $\pi(x,y)$ can be directly represented by the following form with the copula density:

$$\pi(x,y) = p(x)q(y)\mathbf{cop}(x,y|\eta), \quad (4)$$

Combining Eq.4 with the primal objective of Eq.1, we can equivalently transform it into a tractable copula objective with respect to η:

$$\min_{\eta} \ \mathcal{L}(\eta) \triangleq \int_{\mathcal{Y}}\int_{\mathcal{X}} p(x)q(y)\mathbf{cop}(x,y|\eta)c(x,y)dxdy \quad \mathbf{s.t.} \quad x \sim p(x), \quad y \sim q(y), \quad (5)$$

Given pre-defined copula $\mathbf{cop}(x,y|\eta)$ and cost function $c(x,y)$, we can iteratively optimize the objective $\mathcal{L}(\eta)$ by gradient-based methods, so as to derive its gradient to an expectation form with $p(x)q(y)$[1]:

$$\begin{aligned}\nabla_\eta \mathcal{L} &= \nabla_\eta \int_{\mathcal{Y}}\int_{\mathcal{X}} p(x)q(y)\mathbf{cop}(x,y|\eta)c(x,y)dxdy \\ &= \int_{\mathcal{Y}}\int_{\mathcal{X}} p(x)q(y)\nabla_\eta\mathbf{cop}(x,y|\eta)c(x,y)dxdy \\ &= \mathbb{E}_{p(x)q(y)}[\nabla_\eta\mathbf{cop}(x,y|\eta)c(x,y)] \end{aligned} \quad (6)$$

Therefore, we can directly form a noisy gradient using the Monte Carlo samples drawn from $p(x)q(y)$, to simultaneously satisfy the constraints in Eq.5:

$$\nabla_\eta \mathcal{L} \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\eta\mathbf{cop}(x^{(s)}, y^{(s)}|\eta)c(x^{(s)}, y^{(s)}) \quad x^{(s)} \sim p(x), y^{(s)} \sim q(y), \quad (7)$$

[1]There exist many off-the-shelf copula libraries, e.g., VineCopula, which can compute the gradients of copulas, *i.e.,* $\nabla_\eta\mathbf{cop}(x,y|\eta)$.

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

$$\eta_t \leftarrow \eta_{t-1} - \rho_t \nabla_\eta \mathcal{L}, \quad (8)$$

where ρ_t is the learning rate.

Note that this optimization process strictly follows the spirit of stochastic optimization [Robbins and Monro, 1951], where the expectation of the noisy gradient used, *i.e.,* Eq.7, is equivalent to the true gradient, *i.e.,* Eq.6. Therefore it guarantees to converge to a local optimum, if the learning rate satisfies the Robbins-Monro condition:

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

Variance Reduction by Reparameterization

Such noisy gradient of Eq.7, formed by using Monte Carlo samples, often suffers from high variance, resulting in slower convergence and even worse performance [Paisley *et al.*, 2012; Li *et al.*, 2018]. To alleviate this, we use the reparameterization trick [Kingma and Welling, 2013], an empirically effective method for variance reduction. The basic idea is to use a transformation of a simple random variable by a differentiable mapping function, and then form the noisy gradient using Monte Carlo samples drawn from the distribution of this simple random variable. Inspired by this, we re-write the noisy gradient of Eq.7 with reparameterization as follows:

$$\nabla_\eta \mathcal{L} \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\eta\mathbf{cop}(f_x(u^{(s)}), f_y(v^{(s)})|\eta)c(f_x(u^{(s)}), f_y(v^{(s)})) \quad u^{(s)} \sim \widehat{p}(u), v^{(s)} \sim \widehat{q}(v), \quad (9)$$

where u and v are the corresponding simple random variables for x and y, respectively, and $\widehat{p}(u)$ and $\widehat{q}(v)$ are their distributions; $f_x(\cdot)$ and $f_y(\cdot)$ are the mapping functions for x and y, respectively.

For clarity, we illustrate an example: If $p(x)$ is an Gaussian with mean μ and variance σ^2, *i.e.,* $x \sim \mathcal{N}(\mu, \sigma^2)$, the corresponding distribution $\widehat{p}(u)$ of the simple random variable can be the standard Gaussian, *i.e.,* $u \sim \mathcal{N}(0,1)$, and the mapping function is $f_x(u) = \mu + u\sigma$.

Learning Rate Setting

The stochastic optimization of Cop-OT may be sensitive to the learning rate ρ_t. To address this, we use the Adam method [Kingma and Ba, 2015] to adaptively adjust the optimization process, which guarantees the convergence. For clarity, we briefly review this method.

The Adam method is built on the first and second moments of the gradients, requiring to compute exponential moving averages of the gradient m_t and the squared gradient r_t.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\eta\mathcal{L},$$

$$r_t = \beta_2 r_{t-1} + (1-\beta_2)\nabla_\eta\mathcal{L}^2,$$

where β_1 and β_1 control the decay rates of these moving averages; and $\nabla_\eta\mathcal{L}^2$ is the elementwise square of the gradient.

Then, at each iteration t, we can update η using the following equation:

$$\eta_t \leftarrow \eta_{t-1} - \alpha\frac{m_t}{r_t}, \quad (10)$$

where α is the step size in the Adam method. Following [Kingma and Ba, 2015], in this work we empirically fixed the parameters of Adam as follows: $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\alpha = 0.001$.

2.3 Extension to High-dimensional Densities

In this section, we extend Cop-OT to high-dimensional continuous densities $p(x)$ and $q(y)$, $x, y \in \mathbb{R}^D$, $D \geq 2$.

Be analogous to the 1-D problem setting, we can also achieve a high-dimensional copula objective, however, it suffers from a severe problem: Since the copula formulation of $\pi(x, y)$, *i.e.,* Eq.3, refers to all marginals, we can only form noisy gradients using Monte Carlo samples drawn from the marginals of $p(x)$ and $q(y)$, rather than samples from $p(x)$ and $q(y)$. This doesn't satisfy the constraints of OT, *i.e.,* $x \sim p(x)$ and $y \sim q(y)$.

To resolve this, we resort to a novel form of $\pi(x, y)$ that refers to $p(x)$ and $q(y)$ using the D-vine method [Dissmann *et al.*, 2013], which decomposes a high-dimensional copula into a set of conditional bivariate copulas. After some simple algebra[2], we propose the following form with the product of D^2 conditional bivariate copulas between each component of x and y:

$$\pi(x, y) = p(x)q(y) \prod_{i,j}^{D} \mathbf{cop}_{xy}(x_i, y_j | D(e_{ij}), \eta_{ij}^{xy}), \quad (11)$$

where $D(e_{ij})$ denotes the variable conditioning set of the pair (x_i, y_j).

Applying Eq.11 to the primal objective of the above copula objective, we achieve a copula objective of OT for high-dimensional densities:

$$\begin{aligned} &\min_{\eta^{xy}} \ \mathcal{L}(\eta^{xy}) \triangleq \\ &\int_{\mathcal{Y}} \int_{\mathcal{X}} p(x)q(y) \prod_{i,j}^{D} \mathbf{cop}_{xy}(x_i, y_j | D(e_{ij}), \eta_{ij}^{xy}) c(x, y) dx dy \\ &\mathbf{s.t.} \quad x \sim p(x), \quad y \sim q(y) \end{aligned} \quad (12)$$

Be analogous to Eq.6, the gradient of Eq.12 can be also represented by an expectation form with respect to $p(x)q(y)$:

$$\nabla_{\eta_{ij}^{xy}} \mathcal{L} = \mathbb{E}_{p(x)q(y)} \left[\nabla_{\eta_{ij}^{xy}} \prod_{i,j}^{D} \mathbf{cop}_{xy}(x_i, y_j | D(e_{ij}), \eta_{ij}^{xy}) c(x, y) \right] \quad (13)$$

We propose a noisy gradient using Monte Carlo samples from $p(x)q(y)$, which naturally satisfies the constraints in Eq.12. However, we actually need a further approximation, since every conditional bivariate copula $\mathbf{cop}_{xy}(\cdot)$ refers to the conditional marginals $p_i(x_i|D(e_{ij}))$ and $q_i(y_i|D(e_{ij}))$, rather than $p(x)$ and $q(y)$, so that the samples from these conditional marginals are also needed to compute $\mathbf{cop}_{xy}(\cdot)$. Since the conditional marginals are commonly intractable to compute, we use the corresponding marginals to replace them to further simplify the computation. Then, we approximate the noisy gradient using Monte Carlo samples from both $p(x)q(y)$ and their marginals:

$$\begin{aligned} &\nabla_{\eta_{ij}^{xy}} \mathcal{L} \approx \\ &\frac{1}{S} \sum_{s=1}^{S} \nabla_{\eta_{ij}^{xy}} \prod_{i,j}^{D} \mathbf{cop}_{xy}(x_i^{(s)}, y_j^{(s)} | \eta_{ij}^{xy}) c(x^{(s)}, y^{(s)}) \\ &\qquad x^{(s)} \sim p(x), \ y^{(s)} \sim q(y) \\ &\qquad x_i^{(s)} \sim p_i(x_i), \ y_j^{(s)} \sim q_j(y_j) \end{aligned} \quad (14)$$

For variance reduction, we also apply the reparameterization trick, so as to re-write the gradient of Eq.14 as follows:

$$\begin{aligned} \nabla_{\eta_{ij}^{xy}} \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_{\eta_{ij}^{xy}} \prod_{i,j}^{D} &\mathbf{cop}_{xy}(f_{x_i}(u_i^{(s)}), f_{y_j}(v_j^{(s)}) | \eta_{ij}^{xy}) \\ &\times c(f_x(u^{(s)}), f_y(v^{(s)})) \\ &u^{(s)} \sim \widehat{p}(u), \ v^{(s)} \sim \widehat{q}(v) \\ &u_i^{(s)} \sim \widehat{p}_i(u_i), \ v_j^{(s)} \sim \widehat{q}_j(v_j), \end{aligned} \quad (15)$$

where u_i and v_j are the components of u and v, respectively; $\widehat{p}_i(u_i)$ and $\widehat{q}_j(v_j)$ are the marginals of $\widehat{p}(u)$ and $\widehat{q}(v)$, respectively. Given these noisy gradients, the parameter of interest, *i.e.,* η^{xy}, can be finally updated using the Adam method.

Algorithm 1 Optimization of Cop-OT

Input: Continuous densities $p(x), q(y),\ x, y \in \mathbb{R}^D$
Setting: (1) Choose the cost function $c(\cdot)$; (2) construct distributions $\widehat{p}(u)$ and $\widehat{q}(v)$ for reparameterization; (3) choose the bivariate copula function $\mathbf{cop}(\cdot|\eta)$, and initialize η randomly
For $s = 1$ to S
  Draw $u^{(s)} \sim \widehat{p}(u)$
  Draw $v^{(s)} \sim \widehat{p}(v)$
End For
If $D \geq 2$
  For $s = 1$ to S && $i = 1$ to D
    Draw $u_i^{(s)} \sim \widehat{p}_i(u_i)$
    Draw $v_i^{(s)} \sim \widehat{q}_i(v_i)$
  End For
Repeat
  Form the noisy gradient using Eq.9, **If** $D = 1$
  Form the noisy gradient using Eq.15, **If** $D \geq 2$
  Update η or η^{xy} using the Adam method.
Until Convergence
Output: OT between $p(x)$ and $q(y)$ approximated by Cop-OT

Full algorithm. In summary, we outline the full optimization process of Cop-OT in ***Algorithm 1***.

Discussion. We discuss some crucial details of Cop-OT: (1) Referring to Eqs.9 and 15, at each iteration, Cop-OT computes the noisy gradients using samples from the same distributions, *i.e.,* $\widehat{p}(x), \widehat{q}(y)$ and their marginals. Therefore, we only need to draw samples once and iteratively re-use them as indicated in ***Algorithm 1***. (2) For high-dimensional Cop-OT,

[2]The derivation details of Eq.11 can be found at *https://github.com/jinjinchi/Approximate-Optimal-Transport-for-Continuous-Densities-with-Copulas.*

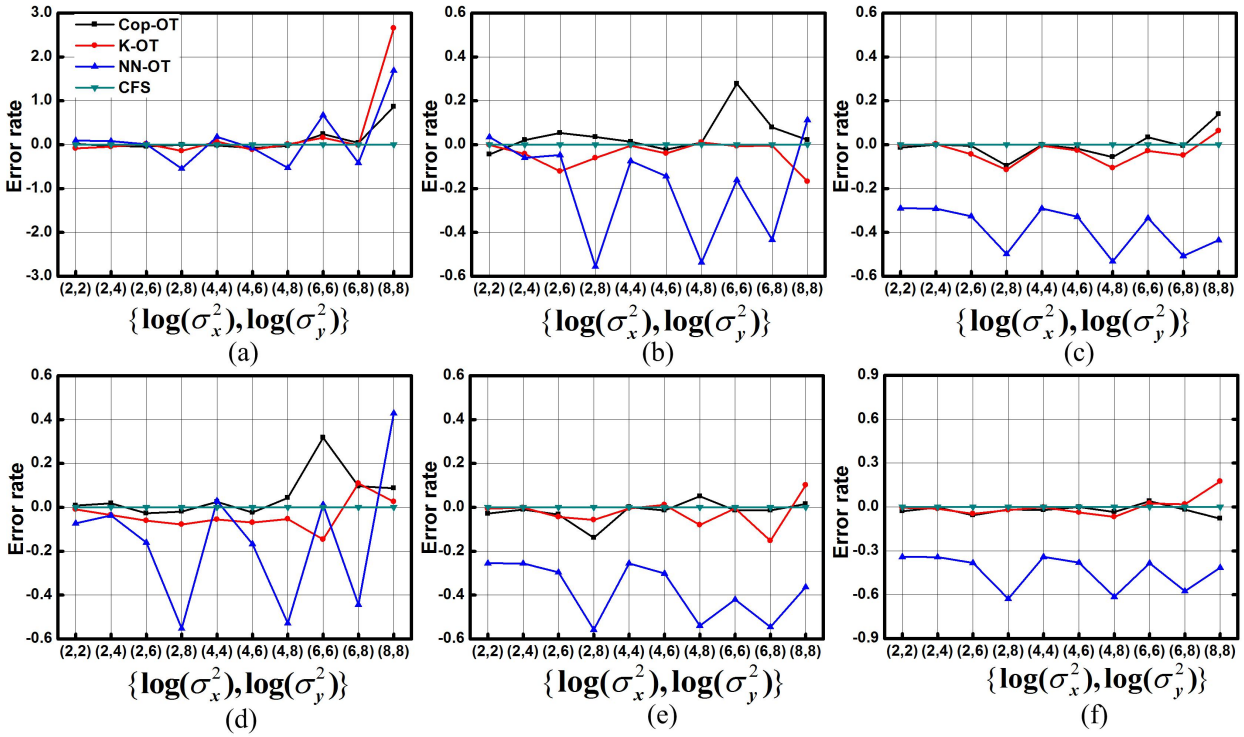

Figure 1: Error rate of 1-D Gaussians with different means and variances. Each sub-figure presents the results of different pairs of $\{\mu_x, \mu_y\}$: (a) $\{2^1, 2^3\}$; (b) $\{2^1, 2^5\}$; (c) $\{2^1, 2^7\}$; (d) $\{2^3, 2^5\}$; (e) $\{2^3, 2^7\}$; (f) $\{2^5, 2^7\}$.

its noisy gradient of Eq.15 doesn't strictly satisfy the convergence condition of stochastic optimization, since its expectation only approximates the true gradient of Eq.13. Fortunately, it can empirically converge fast to achieve competitive results. We show more details of results in the Section of experiment.

Complexity analysis. We discuss the complexity of Cop-OT. Reviewing Eq.12, the copula objective actually involves D^2 bivariate copulas between each pair of x_i and y_j, leading to D^2 copula parameters of interest. Therefore, reviewing Eq.15, Cop-OT needs to form the noisy gradient for each copula parameter using $2S$ samples, requiring $O(TSD^2)$ time and $O(SD + D^2)$ memory, where T is the total iteration number. Besides, we would like to notice that the 1-D setting can be considered as a special case of $D = 1$, requiring $O(TS)$ time and $O(S)$ memory.

3 Empirical Study

In this section, we empirically evaluate Cop-OT on both synthetic and real data.

3.1 Experimental Setup

Our aim is to examine whether Cop-OT can accurately approximate the continuous OT. To this end, we evaluate the OT with the squared Euclidean cost, *i.e.,* $c(x, y) = \|x - y\|_2^2$, $x, y \in \mathbb{R}^D$, between two Gaussian distributions with means $\mu_x, \mu_y \in \mathbb{R}^D$ and covariance matrices $\Sigma_x, \Sigma_y \in \mathbb{R}^{D\times D}$, *i.e.,* $p(x) = \mathcal{N}(\mu_x, \Sigma_x)$, $q(y) = \mathcal{N}(\mu_y, \Sigma_y)$, which has a **C**losed-**F**orm **S**olution (**CFS**) [Givens *et al.*, 1984] computed by:

$$W^*_{xy}(p(x), q(y)) = \sqrt{\|\mu_x - \mu_y\|_2^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_y - 2(\Sigma_y^{1/2}\Sigma_x\Sigma_y^{1/2})^{1/2}\right)} \quad (16)$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

We compare Cop-OT against two baseline methods of **K-OT**[3] [Genevay *et al.*, 2016] and **NN-OT**[4] [Seguy *et al.*, 2018]. For both baselines, we implemented in-house codes based on their demo codes of the semi-continuous versions provided by their authors.

For K-OT and NN-OT, we adjust the regularization parameter over the set $[10^{-3}, 10^{-2}, \cdots, 10^3]$ and report the best results. For Cop-OT, we employ the family of Gaussian copula function, and use the standard Gaussians as the mapping distribution in the reparameterization trick. For all methods. the sample number S is set to 200, and we report the average results of five independent runs.

Besides, in our early experiments, we have examined various copula families, *e.g.,* Clayton, Frank, Gumbel, Joe and Student-t copulas, and found that the Gaussian copula performed the best in most settings.

[3]https://github.com/audeg/StochasticOT

[4]https://github.com/vivienseguy/Large-Scale-OT

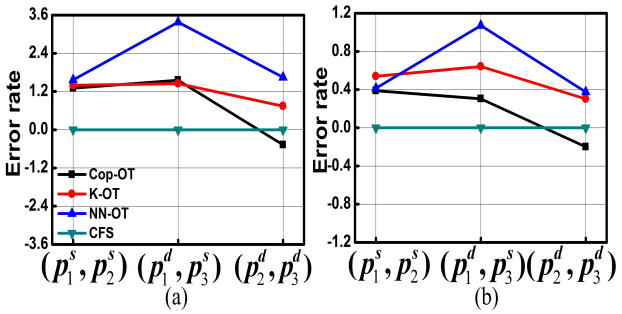

Figure 2: Error rate results of high-dimensional Gaussians: (a) $D = 2$; (b) $D = 8$.

Method	CFS	K-OT	NN-OT	Cop-OT
Accuracy	0.47	0.44	0.39	0.46

Table 2: Empirical results of image retrieval

3.2 Evaluation on Synthetic Data

Results of 1-D Gaussians

We now evaluate on 1-D Gaussians of $\mathcal{N}(\mu_x, \sigma_x^2)$ and $\mathcal{N}(\mu_y, \sigma_y^2)$. The means and variances vary over the sets $\{2^1, 2^3, 2^5, 2^7\}$ and $\{2^2, 2^4, 2^6, 2^8\}$, respectively. We examine the performance of the OT methods by measuring whether their approximations, denoted by $\widehat{W}_{xy}$, approach the **CFS** values, *i.e.,* W^*_{xy} computed by Eq.16. To this end, we define an evaluation metric of ***Error rate*** computed as follows:

$$\mathbf{ErrorRate}(\widehat{W}_{xy}) = \frac{\widehat{W}_{xy} - W^*_{xy}}{W^*_{xy}},$$

The results are shown in Fig.1, where the error rates of **Cop-OT** are significantly lower than those of **NN-OT** in most settings, and **Cop-OT** performs very competitive with **K-OT**. These results imply that **Cop-OT** can compute more accurate approximations to **CFS**.

Results of High-dimensional Gaussians

We further evaluate the high-dimensional Gaussians ($D = 2$ and 8) with mean zero and different covariance matrices drawn from inverse-Wishart distributions. For each dimension, we draw three relatively sparse Σ^s and dense Σ^d from $\mathcal{W}^{-1}(\mathbf{I}, D+2)$ and $\mathcal{W}^{-1}(10\mathbf{I}, D+2)$, leading to three pairs of Gaussians, denoted by $\{p_1^s, p_2^s\}$, $\{p_1^d, p_3^s\}$ and $\{p_2^d, p_3^d\}$, where $p_i^s = \mathcal{N}(0, \Sigma_i^s)$ and $p_i^d = \mathcal{N}(0, \Sigma_i^d)$.

We show the results in Fig.2, where **Cop-OT** outperforms **K-OT** and **NN-OT** in most cases. This empirically indicates that **Cop-OT** works well on comparing high-dimensional densities, even it follows a further approximation to its noisy gradients during optimization. Besides, we observe that the error rates of high-dimensional Gaussians seem less stable than those of 1-D densities.

Convergence

We discuss the convergence of OT methods on two pairs of 1-D Gaussians. Fig.3 presents the convergence curves of all methods within 25 seconds. We can observe that **Cop-OT** converges faster than the baseline methods of **K-OT** and **NN-OT**. Surprisingly, **NN-OT** is converged slower than **K-OT**. The possible reason is that the neural network of **NN-OT** is computationally expensive due to its relatively complex structure.

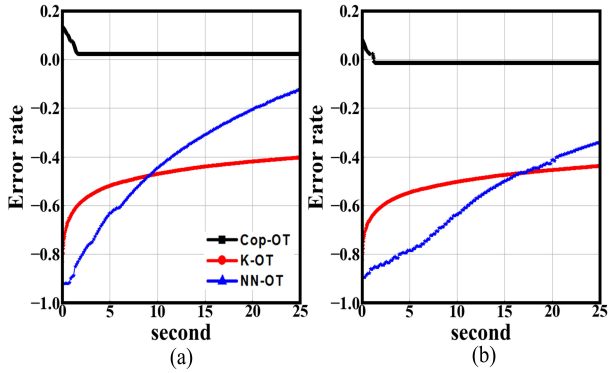

Figure 3: Convergence curves of the OT methods: (a) $\{\mathcal{N}(2^1, 2^2), \mathcal{N}(2^3, 2^8)\}$; (b) $\{\mathcal{N}(2^1, 2^2), \mathcal{N}(2^5, 2^8)\}$.

3.3 Evaluation on Real Data

We evaluate Cop-OT on image retrieval. The MINST[5], a dataset of handwritten digits from zero to nine, is used. We randomly select 10,000 images as the database and 150 images for texting. We refer to each image as a Gaussian estimated by its pixels. Following [Li *et al.*, 2013], we find the nearest 10 images of test images measured by OT, and compute the **accuracy** by counting how many nearest images belong to the same digit number of the test images.

The results are shown in Table 2. First, we observe that the accuracy scores of **Cop-OT** are higher than those of **K-OT** and **NN-OT**. Since image retrieval is factually a problem of finding nearest neighboring images for test inputs, the performance gain of **Cop-OT** indicates that it can better maintain the relative distances among images. Besides, we can see that the accuracy gap between **Cop-OT** and **CFS** is not obvious. This further indicates the effectiveness of **Cop-OT**.

4 Conclusion

We develop a novel Cop-OT method for continuous OT. It formulates the joint distribution with the copula function, and then transforms the primal objective of OT into a copula objective with respect to the copula parameter, solved by stochastic optimization with the reparameterization trick. Both Cop-OT with 1-D and high-dimensional model are proposed. Empirical studies on both real and synthetic data demonstrate the effectiveness of Cop-OT.

Acknowledgments

This research was supported the National Natural Science Foundation of China (NSFC) [No.61602204, No.61876071, No.61806035].

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

References

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[Aurenhammer *et al.*, 1998] Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[Benamou *et al.*, 2015] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[Blondel *et al.*, 2018] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 880–889, 2018.

[Courty *et al.*, 2017] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

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

[Dissmann *et al.*, 2013] Jeffrey Dissmann, Eike C Brechmann, Claudia Czado, and Dorota Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69, 2013.

[Ferradans *et al.*, 2014] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[Genevay *et al.*, 2016] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Neural Information Processing Systems*, pages 3440–3448, 2016.

[Givens *et al.*, 1984] Clark R Givens, Rae Michael Shortt, et al. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

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

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Kingma and Welling, 2013] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2013.

[Kusner *et al.*, 2015] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[Lévy and Schwindt, 2018] Bruno Lévy and Erica L Schwindt. Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148, 2018.

[Li *et al.*, 2013] Peihua Li, Qilong Wang, and Lei Zhang. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *IEEE International Conference on Computer Vision*, pages 1689–1696, 2013.

[Li *et al.*, 2018] Ximing Li, Changchun Li, Jinjin Chi, and Jihong Ouyang. Variance reduction in black-box variational inference by adaptive importance sampling. In *International Joint Conferences on Artificial Intelligence*, pages 2404–2410, 2018.

[Li *et al.*, 2019] Changchun Li, Jihong Ouyang, and Ximing Li. Classifying extremely short texts by exploiting semantic centroids in word mover's distance space. In *The World Wide Web Conference*, pages 939–949, 2019.

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

[Ni *et al.*, 2009] Kangyu Ni, Xavier Bresson, Tony Chan, and Selim Esedoglu. Local histogram based segmentation using the Wasserstein distance. *International Journal of Computer Vision*, 84(1):97–111, 2009.

[Paisley *et al.*, 2012] John William Paisley, David M. Blei, and Michael I. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, page 1367–1374, 2012.

[Pele and Werman, 2009] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *International Conference on Computer Vision*, pages 460–467, 2009.

[Robbins and Monro, 1951] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

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

[Seguy *et al.*, 2018] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018.

[Sklar, 1959] Abe Sklar. Fonstions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

[Villani, 2003] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Society, 2003.

[Wang *et al.*, 2018] Yang Wang, Lin Wu, Xuemin Lin, and Junbin Gao. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE transactions on neural networks and learning systems*, 29(10):4833 – 4843, 2018.