HHS Public Access

Author manuscript

IEEE Access. Author manuscript; available in PMC 2019 December 27.

Published in final edited form as:

IEEE Access. 2018 ; 7: 6269–6278. doi:10.1109/ACCESS.2018.2889838.

Optimal transport for Gaussian mixture models

YONGXIN CHEN[1] **[Member, IEEE]**, **TRYPHON T. GEORGIOU**[2] **[FELLOW, IEEE]**, **ALLEN TANNENBAUM**[3] **[Fellow, IEEE]**

[1]School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[2]Department of Mechanical and Aerospace Engineering, University of California at Irvine, Irvine, CA 92697, USA

[3]Departments of Computer Science and Applied Mathematics and Statistics, The State University of New York at Stony Brook, Stony Brook, NY 11794, USA

Abstract

We introduce an optimal mass transport framework on the space of Gaussian mixture models. These models are widely used in statistical inference. Specifically, we treat Gaussian mixture models as a submanifold of probability densities equipped with the Wasserstein metric. The topology induced by optimal transport is highly desirable and natural because, in contrast to total variation and other metrics, the Wasserstein metric is weakly continuous (i.e., convergence is equivalent to convergence of moments). Thus, our approach provides natural ways to compare, interpolate and average Gaussian mixture models. Moreover, the approach has low computational complexity. Different aspects of the framework are discussed and examples are presented for illustration purposes.

Keywords

Gaussian mixture models; optimal mass transport; statistical signal analysis; Wasserstein metric

I. Introduction

A mixture model is a probabilistic model reflecting the presence of subpopulations within an overall population. Formally, it is a mixture of distributions where each component represents a subpopulation. Mixture models are used in statistics for modeling subgroups and their impact on the total population, inferring properties of subpopulations from the whole, as well as inferring membership for hypothesis testing and other decision making tasks [1]. An important case of mixture models is the so-called Gaussian mixture model (GMM), which is simply a weighted average of several Gaussian distributions. Each Gaussian component stands for a subpopulation. The Gaussian mixture model is widely used because of its mathematical simplicity as well as the efficiency of pertinent inference tests (e.g., in Expectation Maximization algorithms).

Optimal mass transport (OMT) on the other hand is an active and rapidly developing area of research that deals with the geometry of probability densities. It has applications in probability theory, stochastic processes, partial differential equations, fluid mechanics, and many other areas [2], [3]. The subject began with the work of Monge [4] in 1781. The next quantum leap came more than 150 years later, in the work of Kantorovich [5], who introduced a brilliant relaxation, duality theory and linear programming to solve the hitherto intractable Monge formulation. A more recent transformative phase of development came in the works of Brenier, McCann, Otto, Benamou, Gangbo and others [6–11] which culminated in an effective form of calculus on the space of probability measures [12]. Due to these recent advances OMT has become a powerful tool in mathematics, physics, economics, medical imaging and so on [13]–[19], while algorithms developed to address OMT [20]–[28] have provided enabling tools in data science [29], [30].

Briefly, OMT deals with problems of transporting masses from an initial distribution to a terminal one in a *mass preserving* manner (continuity) incurring minimum cost. When the cost of transporting unit Dirac masses is the square of the Euclidean distance between the points of their support, OMT induces a rich Euclidian-like geometry on probability densities. It formally endows the space of probability densities with a Riemannian metric [2], [3], [31]. This geometry enables us to construct geodesic paths, to compare, interpolate, and average probability densities in a natural way, which is in line with our needs in a range of applications. Most importantly, the metric induced by the cost of transport (Wasserstein W_2-metric) is weakly continuous, i.e. convergence in the metric is equivalent to convergence of moments.

While OMT has brought about transformative advances in many fields, computational difficulties persist: solving OMT problems in high dimensional spaces may be computationally prohibitive. The starting point of the present paper is the realization that, in many applications, probability densities have specific structure and can be effectively approximated as points in a suitable manifold [32]. In particular, this work has been motivated by problems in data science where high dimensional data often has lumped structure suggesting mixture models and subgroup membership. Thus, we aim to develop a mathematical framework that takes advantage of such data structures. More specifically, we seek to develop OMT on a most basic submanifolds of probability densities the space of Gaussian mixture models. The extension to more general structured densities will be a future research topic.

Section II provides backgound on OMT. Section III introduces an OMT inspired geometry on Gaussian mixture models, defines a suitable metric, explains how to construct the shortest paths between two points (geodesics), and details some of its properties. Section IV discusses how to average and construct barycenters on the space of Gaussian mixtures. Section V concludes with numerical examples that highlight differences and similarities between the metric structures we are discussing (W_2 and our OMT-inspired metric on the manifold of Gaussian mixtures). The final section provides a summary and conclusions.

II. Background on OMT

We provide a very brief overview of OMT theory. We only cover materials that are pertinent to the present work. We refer the reader to [2] for a more detailed development of the subject and references.

Consider two nonnegative measures μ_0, μ_1 on $\mathbb{R}^n$ with equal total mass. Without loss of generality, we take μ_0 and μ_1 to be probability distributions. In the original formulation of OMT, a transport map

$$T:\mathbb{R}^n \to \mathbb{R}^n : x \mapsto T(x)$$

is sought that specifies where mass $\mu_0(dx)$ at x should be transported so as to match the final distribution in the sense that $T_{\#}\mu_0 = \mu_1$, i.e. μ_1 is the "push-forward" of μ_0 under T, meaning

$$\mu_1(B) = \mu_0(T^{-1}(B))$$

for every Borel set B in $\mathbb{R}^n$. Moreover, the map should achieve a minimum cost of transportation

$$\int_{\mathbb{R}^n} c(x, T(x))\mu_0(dx)\,. \tag{1}$$

Here, $c(x, y)$ represents the transportation cost per unit mass from point x to y. In this paper we focus on the case when $c(x, y) = \|x - y\|^2$. To ensure finite cost, it is standard to assume that μ_0 and μ_1 live in the space of probability densities with finite second moments, denoted by $P_2(\mathbb{R}^n)$.

The dependence of the transportation cost on T is highly nonlinear and a minimum may not exist in general. This fact complicated early analyses of the problem [2]. To circumvent this difficulty, Kantorovich presented a relaxed formulation in 1942. In this, instead of seeking a transport map, one seeks a joint distribution $\pi \in \Pi(\mu_0, \mu_1)$ on $\mathbb{R}^n \times \mathbb{R}^n$, referred to as "coupling" of μ_0 and μ_1, so that the marginals along the two coordinate directions coincide with μ_0 and μ_1, respectively. Thus, in the Kantorovich formulation, we solve

$$J := \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \pi(dxdy)\,. \tag{2}$$

For the case where μ_0, μ_1 are absolutely continuous with respect to the Lebesgue measure, it is a standard result that OMT (2) has a unique solution [2], [3], [6]. This is of the form

$$\pi = (\mathrm{Id} \times T)_{\#}\mu_0,$$

where Id stands for the identity map, and T is the unique minimizer of (1). Moreover, the unique optimal transport T is the gradient of a convex function ϕ, i.e.,

$$y = T(x) = \nabla\phi(x)\,. \tag{3}$$

A. Wasserstein metric

The square root of the optimal cost formally defines a Riemannian metric on $P_2(\mathbb{R}^n)$, known as the Wasserstein metric W_2 [2], [3], [9], [31], i.e.,

$$W_2(\mu_0, \mu_1) := \sqrt{J}$$

with J in (2). Naturally $P_2(\mathbb{R}^n)$ is a geodesic space: a geodesic between μ_0 and μ_1 is of the form

$$\mu_t = (T_t)_\sharp \mu_0, \quad T_t(x) = (1-t)x + tT(x)\,. \tag{4}$$

A geodesic path is also known as displacement interpolation (McCann). It holds that

$$W_2(\mu_s, \mu_t) = (t-s)W_2(\mu_0, \mu_1), \quad 0 \le s < t \le 1. \tag{5}$$

B. OMT between Gaussian distributions

When both of the marginals μ_0, μ_1 are Gaussian distributions, OMT is substantially simplified [33]. In fact, the solution exists in closed-form as explained next.

Denote the mean and covariance of μ_i, $i = 0, 1$ by m_i and Σ_i, respectively. Let X, Y be two Gaussian random vectors associated with μ_0, μ_1, respectively. Suppose the joint distribution between X and Y is π, then $\pi \in \Pi(\mu_0, \mu_1)$ and the cost in (2) becomes

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2 \pi(dxdy) = \mathbb{E}\{\|X-Y\|^2\}\,.$$

This can be further decomposed to

$$\mathbb{E}\left\{\|X-Y\|^2\right\} = \mathbb{E}\left\{\|\widetilde{X}-\widetilde{Y}\|^2\right\} + \|m_0 - m_1\|^2, \tag{6}$$

where $\widetilde{X} = X - m_0$, $\widetilde{Y} = Y - m_1$ are zero mean versions of X and Y. We minimize (6) over all the possible Gaussian joint distributions between X and Y. This gives

$$\min_S \left\{ \|m_0 - m_1\|^2 + \operatorname{trace}(\Sigma_0 + \Sigma_1 - 2S) \middle| \begin{bmatrix} \Sigma_0 & S \\ S^T & \Sigma_1 \end{bmatrix} \ge 0 \right\}, \tag{7}$$

with $S = \mathbb{E}\{\widetilde{X}\widetilde{Y}^T\}$. The constraint is semidefinite constraint, so the above problem is a semidefinite programming (SDP). It turns out that the minimum is achieved by the unique minimizer in closed-form

$$S = \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}$$

with minimum value

$$W_2(\mu_0, \mu_1)^2 = \left\|m_0 - m_1\right\|^2 + \operatorname{trace}(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}). \tag{8}$$

The consequent displacement interpolation μ_t between μ_0 and μ_1 is Gaussian with mean $m_t = (1 - t)m_0 + tm_1$ and covariance

$$\Sigma_t = \Sigma_0^{-1/2}\left((1-t)\Sigma_0 + t(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\right)^2\Sigma_0^{-1/2}. \tag{9}$$

The Wasserstein distance and interpolation can be extended to singular Gaussian distributions by replacing the inverse by the Moor-Penrose pseudoinverse. In particular, when $\Sigma_0 = \Sigma_1 = 0$ and the distributions are Dirac, we have that

$$W_2(\mu_0, \mu_1) = \left\|m_0 - m_1\right\|.$$

Thus, the Wasserstein space of Gaussian distributions, denoted by $\mathcal{G}(\mathbb{R}^n)$, can formally be seen as an extension of the Euclidean space $\mathbb{R}^n$.

III. OMT FOR GAUSSIAN MIXTURE MODELS

A Gaussian mixture model is an important instance of mixture models, which are commonly used to study properties of populations with several subgroups. Mathematically, a Gaussian mixture model is a probability density consisting of several Gaussian components. Namely, it has the form

$$\mu = p^1\nu^1 + p^2\nu^2 + \cdots + p^N\nu^N,$$

where each ν^k is a Gaussian distribution and $p = (p_1, p_2, \ldots, p^N)^T$ is a probability vector. Here the finite number N stands for the number of components of μ. We denote the space of Gaussian mixture distributions by $\mathcal{G}(\mathbb{R}^n)$.

As we have already seen in Section II-B, the displacement interpolation of two Gaussian distributions remains Gaussian. This invariance, however, no longer holds for Gaussian mixtures. Yet, the mixture models may contain some physical or statistical features that we may want to retain. This gives rise to the following question we would like to address: how

can we define a geometry that inherits the nice properties of OMT and while respects the Gaussian mixture structure? Our approach relies on viewing a Gaussian mixture as a discrete measure[1] on the space of Gaussian distributions $\mathcal{G}(\mathbb{R}^n)$ and computing OMT distances between such by taking into account respective means and variances. We explain this next.

Our motivation in developing an optimal transport framework that that is specific to Gaussian mixture models stems from i) a general interest to compare such models as these are these are inferred from data and widely used in applications, and ii) as a way to bring in measurement uncertainty into data analysis since, in general, uncertain data points can be viewed as Gaussian distributions and then, at this fine scale, empirical distributions themselves can be viewed as Gaussian mixture models.

A. A metric on $\mathcal{G}(\mathbb{R}^n)$

Let μ_0, μ_1 be two Gaussian mixture models of the form

$$\mu_i = p_i^1 v_i^1 + p_i^2 v_i^2 + \cdots + p_i^{N_i} v_i^{N_i}, \quad i = 0, 1.$$

Here N_0 maybe different from N_1. The distribution μ_i is equivalent to a discrete measure p_i with supports $v_i^1, v_i^2, \ldots, v_i^{N_i}$ for each $i = 0, 1$. Our framework is built on the discrete OMT problem

$$\min_{\pi \in \Pi(p_0, p_1)} \sum_{i,j} c(i,j)\pi(i,j) \tag{10}$$

for these two discrete measures. Here $\prod(p_0, p_1)$ denote the space of joint distributions between p_0 and p_1. The cost $c(i, j)$ is taken to be the square of the Wasserstein metric on $\mathcal{G}(\mathbb{R}^n)$, that is,

$$c(i,j) = W_2(v_0^i, v_1^j)^2.$$

By standard linear programming theory, the discrete OMT problem (10) always has at least one solution. Let π^* be a minimizer, and define

$$d(\mu_0, \mu_1) = \sqrt{\sum_{i,j} c(i,j)\pi^*(i,j)}. \tag{11}$$

Theorem 1. *$d(\cdot, \cdot)$ defines a metric on $\mathcal{G}(\mathbb{R}^n)$.*

Proof: Clearly, $d(\mu_0, \mu_1) \geq 0$ for any $\mu_0, \mu_1 \in \mathcal{G}(\mathbb{R}^n)$ and $d(\mu_0, \mu_1) = 0$ if and only if $\mu_0 = \mu_1$. We next prove the triangular inequality, namely,

[1] A similar viewpoint has been used in [34] to reduce the dimensionality of Gaussian mixture models. However, the approach in [34] is based on KL divergence instead of OMT.

$$d(\mu_0, \mu_1) + d(\mu_1, \mu_2) \geq d(\mu_0, \mu_2)$$

for any μ_0, μ_1, $\mu_2 \in \mathcal{G}(\mathbb{R}^n)$. Denote the probability vector associated with μ_0, μ_1, μ_2 by p_0, p_1, p_2 respectively. The Gaussian components of μ_i are denoted by ν^j. Let $\boldsymbol{\pi}_{01}$ ($\boldsymbol{\pi}_{12}$) be the solution to (10) with marginals μ_0, μ_1 (μ_1, μ_2). Define $\boldsymbol{\pi}_{02}$ by Denote the probability vector associated with μ_0, μ_1, μ_2 by p_0, p_1, p_2 respectively. The Gaussian components of μ_i are denoted by v_i^j. Let $\boldsymbol{\pi}_{01}$ ($\boldsymbol{\pi}_{12}$) be the solution to (10) with marginals μ_0, μ_1 (μ_1, μ_2). Define $\boldsymbol{\pi}_{02}$ by

$$\pi_{02}(i,k) = \sum_j \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}.$$

Clearly, $\boldsymbol{\pi}_{02}$ is a joint distribution between p_0 and p_2, namely, $\boldsymbol{\pi}_{02} \in \prod(p_0, p_2)$. It follows from direct calculation

$$\begin{aligned}\sum_i \pi_{02}(i,k) &= \sum_{i,j} \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j} \\ &= \sum_j \frac{p_1^j \pi_{12}(j,k)}{p_1^j} \\ &= p_2^k.\end{aligned}$$

Similarly, we have $\sum_k \pi_{02}(i,k) = p_0^i$. Therefore,

$$
\begin{aligned}
d(\mu_0,\mu_2) &\le \sqrt{\sum_{i,k}\pi_{02}(i,k)W_2(v_0^i,v_2^k)^2} \\
&= \sqrt{\sum_{i,j,k}\frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}W_2(v_0^i,v_2^k)^2} \\
&\le \sqrt{\sum_{i,j,k}\frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}(W_2(v_0^i,v_1^j)+W_2(v_1^i,v_2^k))^2} \\
&\le \sqrt{\sum_{i,j,k}\frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}W_2(v_0^i,v_1^j)^2} \\
&\quad + \sqrt{\sum_{i,j,k}\frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}W_2(v_1^j,v_2^k)^2} \\
&= \sqrt{\sum_{i,j}\pi_{01}(i,j)W_2(v_0^i,v_1^j)^2} \\
&\quad + \sqrt{\sum_{j,k}\pi_{12}(j,k)W_2(v_1^j,v_2^k)^2} \\
&= d(\mu_0,\mu_1)+d(\mu_1,\mu_2).
\end{aligned}
$$

In the above, the second inequality is due to the fact W_2 is a metric, and the third inequality is an application of the Minkowski inequality. ■

B. Geodesics on $\mathcal{G}(\mathbb{R}^n)$

A geodesic[2] on $\mathcal{G}(\mathbb{R}^n)$ connecting μ_0 and μ_1 is given by

$$\mu_t = \sum_{i\,.\,j}\pi^*(i,j)v_t^{ij}, \tag{12}$$

where v_t^{ij} is the displacement interpolation (see (9)) between v_0^i and v_1^j.

Theorem 2.

$$d(\mu_s,\mu_t) = (t-s)d(\mu_0,\mu_1), \quad 0 \le s < t \le 1. \tag{13}$$

Proof: For any $0 \le s \le t \le 1$, we have

$$
\begin{aligned}
d(\mu_s,\mu_t) &\le \sqrt{\sum_{i,j}\pi^*(i,j)W_2(v_s^{ij},v_t^{ij})^2} \\
&= (t-s)\sqrt{\sum_{i,j}\pi^*(i,j)W_2(v_0^i,v_1^j)^2} = (t-s)d(\mu_0,\mu_1)
\end{aligned}
$$

[2]Here by *geodesic* we mean the shortest path between two points.

where we have used the property (5) of W_2. It follows that

$$\begin{aligned} &d(\mu_0, \mu_s) + d(\mu_s, \mu_t) + d(\mu_t, \mu_1) \\ &\le sd(\mu_0, \mu_1) + (t-s)d(\mu_0, \mu_1) + (1-t)d(\mu_0, \mu_1) \\ &= d(\mu_0, \mu_1). \end{aligned}$$

On the other hand, by Theorem 1, we have

$$d(\mu_0, \mu_s) + d(\mu_s, \mu_t) + d(\mu_t, \mu_1) \ge d(\mu_0, \mu_1).$$

Combining these two, we obtain (13). ■

We remark that μ_t is a Gaussian mixture model since it is a weighted average of the Gaussian distributions v_t^{ij}. Even though the solution to (10) is not unique in some instances, it is unique for generic $\mu_0, \mu_1 \in \mathcal{G}(\mathbb{R}^n)$. Therefore, in most real applications, we need not worry about the uniqueness.

C. Relation between the metrics d(·, ·) and W2(·, ·)

We first note that we have

$$d(\mu_0, \mu_1) \ge W_2(\mu_0, \mu_1) \tag{14}$$

for any $\mu_0, \mu_1 \in \mathcal{G}(\mathbb{R}^n)$. To see this, note that for any $\boldsymbol{\pi} \in \prod(p_0, p_1)$, $\gamma := \sum_{i,j} \pi(i,j)\gamma^{i,j}$ is a joint distribution between μ_0 and μ_1. Here γ^{ij} is the optimal coupling of v_0^i, v_1^j solving the OMT problem with marginals v_0^i, v_1^j. Additionally,

$$\sum_{i,j} c(i,j)\pi(i,j) = \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2 \gamma(dxdy).$$

Therefore, any joint distribution $\pi \in \prod(p_0, p_1)$ corresponds to a feasible solution γ to the Kantorovich problem (2). The inequality (14) then follows.

The equality in (14) holds when both μ_0 and μ_1 have only one Gaussian component. In general, $d > W_2$. This is due to the fact that the restriction to the submanifold $\mathcal{G}(\mathbb{R}^n)$ induces sub-optimality in the transport plan. Let $\gamma(t)$, $0 \le t \le 1$ be any piecewise smooth curve on $\mathcal{G}(\mathbb{R}^n)$ connecting μ_0 and μ_1. Define the Wasserstein length of γ by

$$L_W(\gamma) = \sup_{0 = t_0 < t_1 < \cdots < t_s = 1} \sum_k W_2(\gamma_{t_k}, \gamma_{t_{k+1}}),$$

and natural length by

$$L(\gamma) = \sup_{0 = t_0 < t_1 < \cdots < t_s = 1} \sum_k d(\gamma_{t_k}, \gamma_{t_{k+1}}).$$

Then $L_W(\gamma) \leq L(\gamma)$.

Using the metric property of d we get

$$d(\mu_0, \mu_1) \leq \inf_\gamma L(\gamma),$$

where the minimization is taken over all the piecewise smooth curve on $\mathcal{G}(\mathbb{R}^n)$ connecting μ_0 and μ_1. In view of (13), we conclude

$$d(\mu_0, \mu_1) \leq \inf_\gamma L(\gamma) \geq \inf_\gamma L_W(\gamma).$$

However, it is unclear whether d is the restriction of W_2 to $\mathcal{G}(\mathbb{R}^n)$.

In general, d is a very good approximation of W_2 if the variances of the Gaussian components are small compared with the differences between the means. This may lead to an efficient algorithm to approximate Wasserstein distance between two distributions with such properties. If we want to compute the Wasserstein distance $W_2(\mu_0, \mu_1)$ between two distributions $\mu_0, \mu_1 \in \mathcal{G}(\mathbb{R}^n)$, a standard procedure is discretizing the densities first, and then solving a discrete OMT problem. Depending upon the resolution of the discretization, the second step may become very costly. In contrast, to compute our new distance $d(\mu_0, \mu_1)$, we need only to solve (10). When the number of Gaussian components of μ_0, μ_1 is small, this is very efficient.

IV. Barycenter of Gaussian mixtures

The barycenter [35] of L distributions $\mu_0, \mu_1, \ldots, \mu_L$ is defined to be the minimizer of

$$J(\mu) = \frac{1}{L} \sum_{k=1}^{L} W_2(\mu, \mu_k)^2. \tag{15}$$

This resembles the average $\frac{1}{L}(x_1 + x_2 + \cdots + x_L)$ of L points in the Euclidean space, which minimizes

$$J(x) = \frac{1}{L} \sum_{k=1} \|x - x_k\|^2.$$

The above definition can be generalized to the cost

$$\min_{\mu \in p_2(\mathbb{R}^n)} \sum_{k=1}^{L} \lambda_k W_2(\mu, \mu_k)^2 . \tag{16}$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_L]$ is a probability vector. The existence and uniqueness of (16) has been extensively studied in [35] where it is shown that under some mild assumptions, the solution exists and is unique.

In the special case when all μ_k are Gaussian distributions, the barycenter remains Gaussian. In particular, denoting the mean and covariance of μ_k as m_k, Σ_k, then the barycenter has mean

$$m = \sum_{k=1}^{L} \lambda_k m_k \tag{17}$$

and covariance Σ solving

$$\Sigma = \sum_{k=1}^{L} \lambda_k (\Sigma^{1/2} \Sigma_k \Sigma^{1/2})^{1/2} . \tag{18}$$

A fast algorithm to get the solution of (18) is through the fixed point iteration [36]

$$(\Sigma)_{\text{next}} = \Sigma^{-1/2} \left(\sum_{k=1}^{L} \lambda_k (\Sigma^{1/2} \Sigma_k \Sigma^{1/2})^{1/2} \right)^2 \Sigma^{-1/2} .$$

In practice, the iteration

$$(\Sigma)_{\text{next}} = \sum_{k=1}^{L} \lambda_k (\Sigma^{1/2} \Sigma_k \Sigma^{1/2})^{1/2}$$

appears to also work. However, no convergence proof for the latter is known at present [35], [36].

For general distributions, the barycenter problem (16) is difficult to solve. It can be reformulated as a multi-marginal optimal transport problem and is therefore convex. Recently several algorithms have been proposed to solve (16) through entropic regularization [26]. However, due to the curse of dimensionality, solving such a problem for dimension greater than 3 is still unrealistic. This is the case even for Gaussian mixture models. What's more, the Gaussian mixture structure is often lost when solving problem (16).

To overcome this issue for Gaussian mixtures, herein, we propose to solve a modified barycenter problem

$$\min_{\mu \in \mathcal{G}(\mathbb{R}^n)} \sum_{k=1}^{L} \lambda_k d(\mu, \mu_k)^2 . \tag{19}$$

The optimization variable is restricted to be Gaussian mixture distribution and the Wasserstein distance W_2 is replaced by its relaxed version (11). Let μ_k be a Gaussian mixture distribution with N_k components, namely, $\mu_k = p_k^1 v_k^1 + p_k^2 v_k^2 + \cdots + p_k^{N_k} v_k^{N_k}$. If we view μ as a discrete measure on $\mathcal{G}(\mathbb{R}^n)$, then clearly, it can only have support at the points (Gaussian distributions) of the form

$$\text{argmin}_v \sum_{k=1}^{L} \lambda_k W_2(v, v_k^{i_k})^2 \tag{20}$$

with $v_t^{i_k}$ being any component of μ_k. As we discussed before, the optimal v is Gaussian. Denote the set of all such minimizers as $\{v^1, v^2, \ldots, v^N\}$, then μ is equal to

$$\mu = p^1 v^1 + p^2 v^2 + \cdots + p^N v^N,$$

for some probability vector $p = (p_1, p_2, \ldots, p^N)^T$. The number of element, denoted by N, is bounded above by $N_1 N_2 \cdots N_L$. Finally, utilizing the definition of $d(\cdot, \cdot)$ we obtain an equivalent formulation of (19), which reads as

$$\min_{\pi_1 \geq 0, \cdots, \pi_L \geq 0} \sum_{k=1}^{L} \sum_{i=1}^{N} \sum_{j_k=1}^{N_k} \lambda_k c_k(i, j_k) \pi_k(i, j_k) \tag{21a}$$

$$\sum_{i=1}^{N} \pi_k(i, j_k) = p_k^{j_k}, \quad \forall 1 \leq k \leq L, 1 \leq j_k \leq N_k \tag{21b}$$

$$\sum_{j_1=1}^{N_1} \pi_1(i, j_1) = \cdots = \sum_{j_L=1}^{N_L} \pi_L(i, j_L), \quad \forall 1 \leq i \leq N . \tag{21c}$$

The cost

$$c_k(i, j) = W_2(v^i, v_k^j)^2 \tag{22}$$

is the optimal transport cost from v^i to $\mathbb{R}^n$. After solving the above linear programming problem (21), we get the barycenter $\mu = p^1 v^1 + p^2 v^2 + \cdots + p^N v^N$ with

$$p^i = \sum_{j=1}^{N_1} \pi_1(i,j)$$

for each $1 \le i \le N$. We remark that our formulation is independent of the dimension of the underlying space $\mathbb{R}^n$. The dimension affects only the computation of the cost function (22) where a closed-form (8) is available. The complexity of (21) relies on the numbers of components of the Gaussian mixtures distributions $\{\mu_k\}$. Therefore, our formulation is extremely efficient for high dimensional Gaussian mixtures with small number of components.

The difficulty of formulation (21) lies in the number N of components of the barycenter μ, which is usually of order $N_1 N_2 \cdots N_L$. To overcome this issue, we can consider the barycenter problem for Gaussian mixture with specified components. More specifically, given N Gaussian components $\nu^1, \nu^2, \ldots, \nu^N$, we would like to find a minimizer of the optimization problem (16) subject to the structure constraint that

$$\mu = p^1 \nu^1 + p^2 \nu^2 + \cdots + p^N \nu^N$$

for some probability vector $p = (p^1, p^2, \ldots, p^N)^T$. Note that ν^k here doesn't have to be of the form (20). It can be any Gaussian distribution. Moreover, the number N can be chosen to be small. It turns out this problem can be solved in exactly the same way. Clearly, a linear programming reformulation (21) is straightforward.

V. Numerical Examples

Several examples are provided to illustrate our framework in computing distance, computational cost, geodesic and barycenter.

A. Comparing d(·, ·) and W2(·, ·)

To demonstrate the difference between d and W_2, we choose μ_0 to be a one dimensional zero-mean Gaussian distribution with unit variance. The terminal distribution μ_1 is set to be the average of two unit variance Gaussian distributions, one with mean Δ and the other one with mean $-\Delta$. Clearly, $d(\mu_0, \mu_1) = \Delta$. Figure 1 depicts $d(\mu_0, \mu_1)$ and $W_2(\mu_0, \mu_1)$ for different Δ values. As can be seen, these two distances are not equivalent and d is always bounded below by W_2.

We also compared the computational cost of our method to two other algorithms commonly used in solving optimal transport problems. One is a simplex method developed by Rubner *et al.* [20], and the other is based on the Sinkhorn algorithm [22]. The latter is employed to solve an entropic regularized optimal transport problem. We ran the algorithms for problems in 1d and 2d with varying number of components and summarize the results in Table I. The computational cost is averaged over 3 trials for each configuration. Both algorithms in [20] and [22] require discretizing the densities on grids (here we use 100 points for 1d and 30×30 points for 2d) first, and therefore do not work for high dimensional problems due to the

"curse of dimensionality." In contrast, our framework leverages the closed-form expression for optimal transport distance between Gaussian components and because of that it scales nicely as the dimension increases. Figure 2 depicts the computational cost (again averaged over 3 trials) for different choice of dimension. The number of components is fixed to be 10 and the parameters are chosen randomly.

B. Comparing geodesics

We compare shortest path interpolation provided by standard OMT theory with our proposed geodesic interpolation (see (12)) on $\mathcal{G}(\mathbb{R}^n)$. To this end we consider the two Gaussian mixture models in Figure 3. Both have two (Gaussian) components; one shown in red and one in blue, while the mixture model is shown in black. The components as well as the mixture model are normalized to have unt integrals. The two mixture models serve as marginals, with mass equally distributed among the respective components. The corresponding means and covariances are $m_0^1 = 0.5$, $m_0^2 = 0.1$, $\Sigma_0^1 = 0.01$, $\Sigma_0^2 = 0.05$ for μ_0, and $m_1^1 = 0$, $m_1^2 = -0.35$, $\Sigma_1^1 = 0.02$, $\Sigma_1^2 = 0.02$ for μ_1 (see Section III for notations).

Figure 4 compares interpolation between the two marginals based on standard OMT and interpolation based on the geometry on $\mathcal{G}(\mathbb{R}^n)$ introduced herein. As seen in the figures, the flow of densities that is based on standard OMT interpolation loses the Gaussian mixture character. Our method is of course designed to preserve the Gaussian mixture structure as can be seen from Figure 5, which displays the two Gaussian components of the density flow interpolation using our method.

We make similar observations on a 2-dimensional example displayed in Figures 6–8. To this end, we plot the level sets of the densities so as to highlight their actual 3-dimensional shape. The two marginal distributions are the Gaussian mixtures shown in Figure 6. Figures 7 and 8 show snapshots of the two interpolation paths, the first one based on OMT (Figure 7) and the second based on our method (Figure 8), respectively. We can easily discern that the Gaussian mixture structure is not preserved along the geodesic path of Figure 7. In contrast, the snapshots in Figure 8 are seen to contain two (dominant) Gaussians that are clearly recognizable.

C. Barycenter

Three Gaussian mixture distributions are given in Figure 9. The masses are equally distributed among the respective components. The statistics for μ_1, μ_2, μ_3 are $(m_1^1 = 0, m_1^2 = 0.1, \Sigma_1^1 = 0.01, \Sigma_1^2 = 0.05)$, $(m_2^1 = 0, m_2^2 = -0.35, \Sigma_2^1 = 0.02, \Sigma_2^2 = 0.02)$ and $(m_3^1 = 0.4, m_3^2 = -0.45, \Sigma_3^1 = 0.025, \Sigma_3^2 = 0.021)$ respectively. We compare our method with the traditional OMT theory to compute the barycenter. Two sets of weights are considered and the results are displayed in Figures 10 and 11. It is quite clear that our method gives a more "desirable" average. The Gaussian mixtures character is not preserved with traditional OMT-barycenter construction while it is evident in our setting.

VI. Conclusion

In this paper, we have defined a new optimal mass transport distance for Gaussian mixture models by restricting ourselves to the submanifold of Gaussian mixture distributions. Consequently, the geodesic interpolation utilizing this metric remains on the submanifold of Gaussian mixture distributions. On the numerical side, computing this distance between two densities is equivalent to solving a linear programming problem whose number of variables grows linearly as the number of Gaussian components. This represents a huge reduction in computational cost as compared with traditional OMT.

When the covariances of the respective components in Gaussian mixture models are small, our distance is a very good approximation of the standard OMT W_2 distance, and the respective geodesics are also close. Thus, in this case the computationally more efficient framework herein represents a good compromise.

In general, our objective in this paper has been twofold. First, being interested in Gaussian mixtrure models, we set out to develop an OMT-based geometry on the respective manifold $\mathcal{G}(\mathbb{R}^n)$. Geometric constructions (geodesics, averages) on this manifold, with respect to the metric $d(\cdot, \cdot)$ retain the Gaussian mixture character. Besides computations are much more tractable even for very high dimensional spaces as compared to OMT. The computational burden is only dictated by the number of components of the Gaussian mixture model and not the size of the space, and it is quite modest. Thus, our approach can be used for approximating OMT transport as well, for cases where distributions can be reasonably well approximated by Gaussian mixtures. A subject of interest for future research is to extend our toolset so that we are able to work efficiently on more general mixture models that are not necessarily Gaussian.

Biography

Yongxin Chen received his BSc from Shanghai Jiao Tong University in 2011 and Ph.D. from University of Minnesota in 2016, both in Mechanical Engineering. He is currently an Assistant Professor in the School of Aerospace Engineering at Georgia Institute of Technology. He received the George S. Axelby Best Paper Award in 2017 for his joint work with Tryphon Georgiou and Michele Pavon. His research interests include optimal transport, control, machine learning and robotics.

Tryphon T. Georgiou is currently a Chancellor's Professor in the Department of Mechanical and Aerospace Engineering at the University of California, Irvine. He has served on the faculty at Florida Atlantic University, Iowa State University and the University of Minnesota. He is a co-recipient of the G.S. Axelby award of the IEEE Control Systems Society for the years 1992, 1999, 2003 and 2017, a Fellow of the IEEE, a Fellow of IFAC, and a Foreign Member of the Royal Swedish Academy of Engineering Sciences (IVA).

Allen Tannenbaum is presently Distinguished Professor of Computer Science and Applied Mathematics/Statistics at SUNY Stony Brook. He works in systems and control, signal processing, computer vision, and systems biology.

References

[1]. McLachlan G and Peel D, Finite mixture models John Wiley & Sons, 2004.

[2]. Villani C, Topics in Optimal Transportation American Mathematical Soc., 2003, no. 58.

[3]. Villani C, Optimal Transport: Old and New Springer, 2008, vol. 338.

[4]. Monge G, Mémoire sur la the´orie des déblais et des remblais De l'Imprimerie Royale, 1781.

[5]. Kantorovich LV, "On the transfer of masses," in Dokl. Akad. Nauk SSSR, vol. 37, no. 7–8, 1942, pp. 227–229.

[6]. Brenier Y, "Polar factorization and monotone rearrangement of vector-valued functions," Communications on pure and applied mathematics, vol. 44, no. 4, pp. 375–417, 1991.

[7]. Gangbo W and McCann RJ, "The geometry of optimal transportation," Acta Mathematica, vol. 177, no. 2, pp. 113–161, 1996.

[8]. McCann RJ, "A convexity principle for interacting gases," Advances in mathematics, vol. 128, no. 1, pp. 153–179, 1997.

[9]. Jordan R, Kinderlehrer D, and Otto F, "The variational formulation of the Fokker–Planck equation," SIAM journal on mathematical analysis, vol. 29, no. 1, pp. 1–17, 1998.

[10]. Benamou JD and Brenier Y, "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," Numerische Mathematik, vol. 84, no. 3, pp. 375–393, 2000.

[11]. Otto F and Villani C, "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality," Journal of Functional Analysis, vol. 173, no. 2, pp. 361–400, 2000.

[12]. Ambrosio L, Gigli N, and Savaré G, Gradient flows: in metric spaces and in the space of probability measures Springer, 2006.

[13]. Evans LC and Gangbo W, Differential equations methods for the Monge-Kantorovich mass transfer problem American Mathematical Soc., 1999, vol. 653.

[14]. Haker S, Zhu L, Tannenbaum A, and Angenent S, "Optimal mass transport for registration and warping," International Journal of Computer Vision, vol. 60, no. 3, pp. 225–240, 2004.

[15]. Mueller M, Karasev P, Kolesov I, and Tannenbaum A, "Optical flow estimation for flame detection in videos," IEEE Transactions on image processing, vol. 22, no. 7, pp. 2786–2797, 2013. [PubMed: 23613042]

[16]. Chen Y, Georgiou TT, and Pavon M, "On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint," Journal of Optimization Theory and Applications, vol. 169, no. 2, pp. 671–691, 2016.

[17]. Chen Y, Georgiou TT, and Pavon M, "Optimal transport over a linear dynamical system," IEEE Transactions on Automatic Control, vol. 62, no. 5, pp. 2137–2152, 2017.

[18]. Galichon A, Optimal Transport Methods in Economics Princeton University Press, 2016.

[19]. Chen Y, "Modeling and control of collective dynamics: From Schrödinger bridges to optimal mass transport," Ph.D. dissertation, University of Minnesota, 2016.

[20]. Rubner Y, Tomasi C, and Guibas LJ, "The earth mover's distance as a metric for image retrieval," International journal of computer vision, vol. 40, no. 2, pp. 99–121, 2000.

[21]. Angenent S, Haker S, and Tannenbaum A, "Minimizing flows for the Monge–Kantorovich problem," SIAM journal on mathematical analysis, vol. 35, no. 1, pp. 61–97, 2003.

[22]. Cuturi M, "Sinkhorn distances: Lightspeed computation of optimal transport," in Advances in Neural Information Processing Systems, 2013, pp. 2292–2300.

[23]. Benamou JD, Froese BD, and Oberman AM, "Numerical solution of the optimal transportation problem using the Monge-Ampere equation," Journal of Computational Physics, vol. 260, pp. 107–126, 2014.

[24]. Tabak EG and Trigila G, "Data-driven optimal transport," Commun. Pure. Appl. Math. doi, vol. 10, p. 1002, 2014.

[25]. Haber E and Horesh R, "A multilevel method for the solution of time dependent optimal transport," Numerical Mathematics: Theory, Methods and Applications, vol. 8, no. 01, pp. 97–111, 2015.

[26]. Benamou JD, Carlier G, Cuturi M, Nenna L, and Peyré G, "Iterative Bregman projections for regularized transportation problems," SIAM Journal on Scientific Computing, vol. 37, no. 2, pp. A1111–A1138, 2015.

[27]. Chen Y, Georgiou T, and Pavon M, "Entropic and displacement interpolation: a computational approach using the Hilbert metric," SIAM Journal on Applied Mathematics, vol. 76, no. 6, pp. 2375–2396, 2016.

[28]. Genevay A, Cuturi M, Peyré G, and Bach F, "Stochastic optimization for large-scale optimal transport," in Advances in Neural Information Processing Systems, 2016, pp. 3440–3448.

[29]. Montavon G, Müller K-R, and Cuturi M, "Wasserstein training of restricted Boltzmann machines," in Advances in Neural Information Processing Systems, 2016, pp. 3718–3726.

[30]. Arjovsky M, Chintala S, and Bottou L, "Wasserstein generative adversarial networks," in International Conference on Machine Learning, 2017, pp. 214–223.

[31]. Otto F, "The geometry of dissipative evolution equations: the porous medium equation," Communications in Partial Differential Equations, 2001.

[32]. Amari SI and Nagaoka H, Methods of information geometry American Mathematical Soc., 2007, vol. 191.

[33]. Takatsu A, "Wasserstein geometry of gaussian measures," Osaka Journal of Mathematics, vol. 48, no. 4, pp. 1005–1026, 2011.

[34]. Akaho S, "Dimension reduction for mixtures of exponential families," in International Conference on Artificial Neural Networks Springer, 2008, pp. 1–10.

[35]. Agueh M and Carlier G, "Barycenters in the Wasserstein space," SIAM Journal on Mathematical Analysis, vol. 43, no. 2, pp. 904–924, 2011.

[36]. Álvarez-Esteban PC, del Barrio E, Cuesta-Albertos J, and Matrán C, "A fixed-point approach to barycenters in Wasserstein space," Journal of Mathematical Analysis and Applications, vol. 441, no. 2, pp. 744–762, 2016.

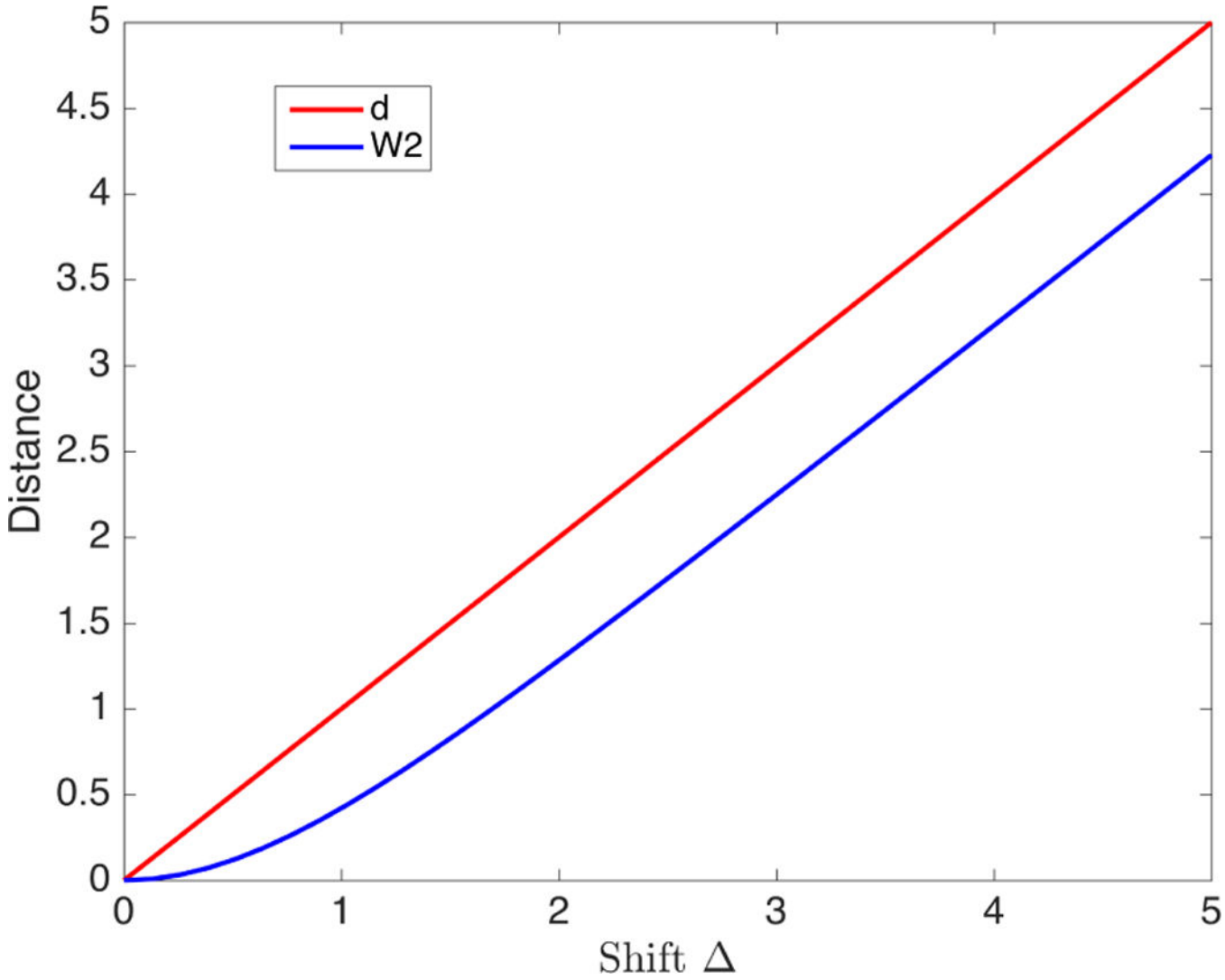


Fig. 1:
d vs W_2 for different shift Δ value

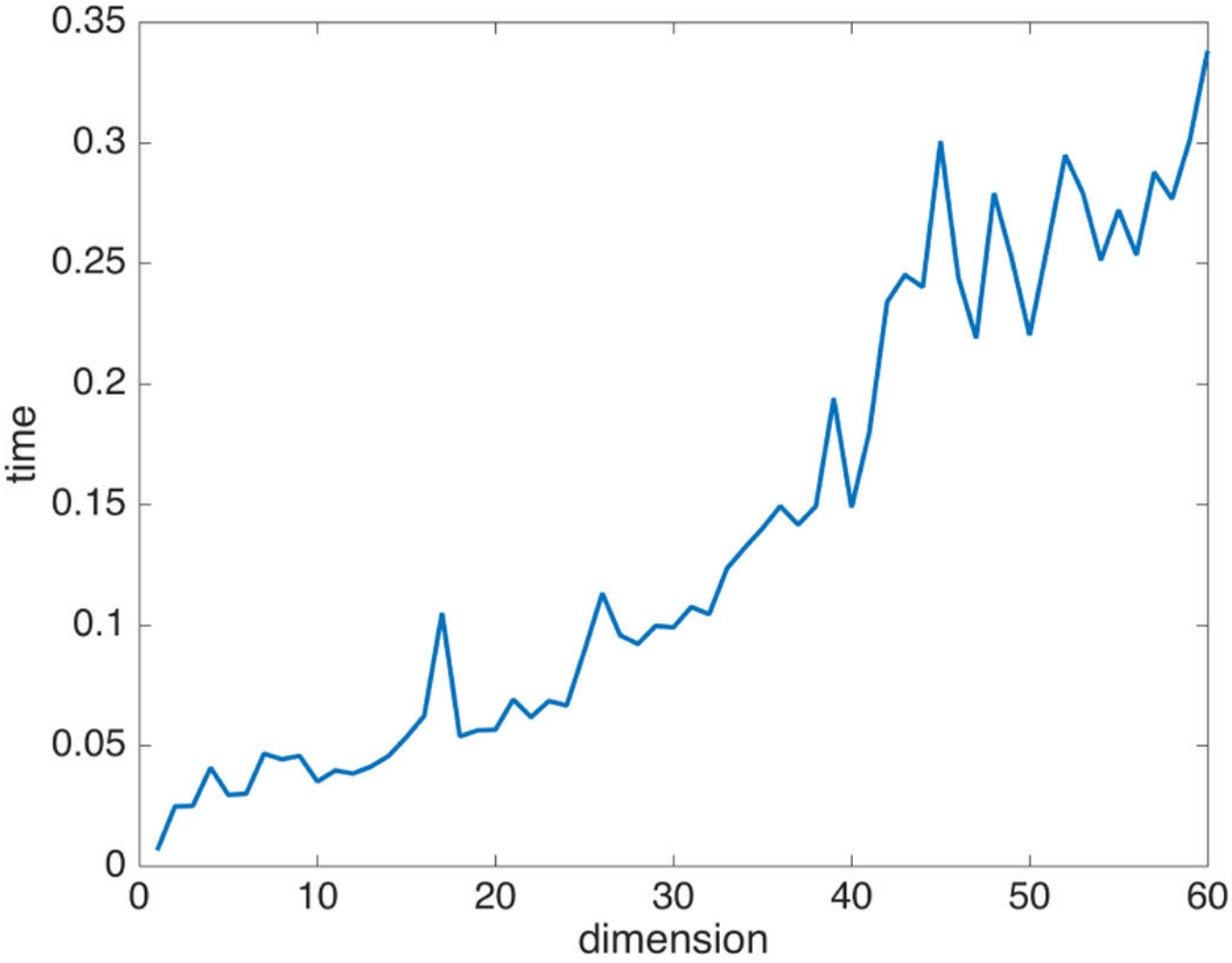


Fig. 2:
Computation cost vs. dimension (averaged over 3 experiments)

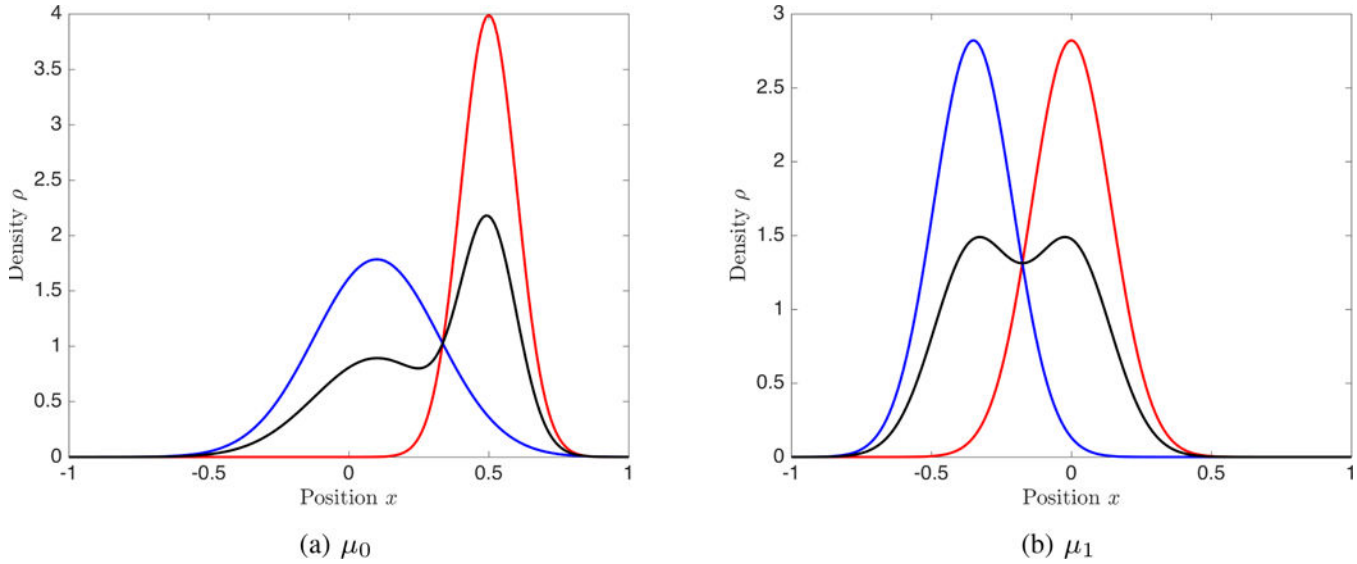
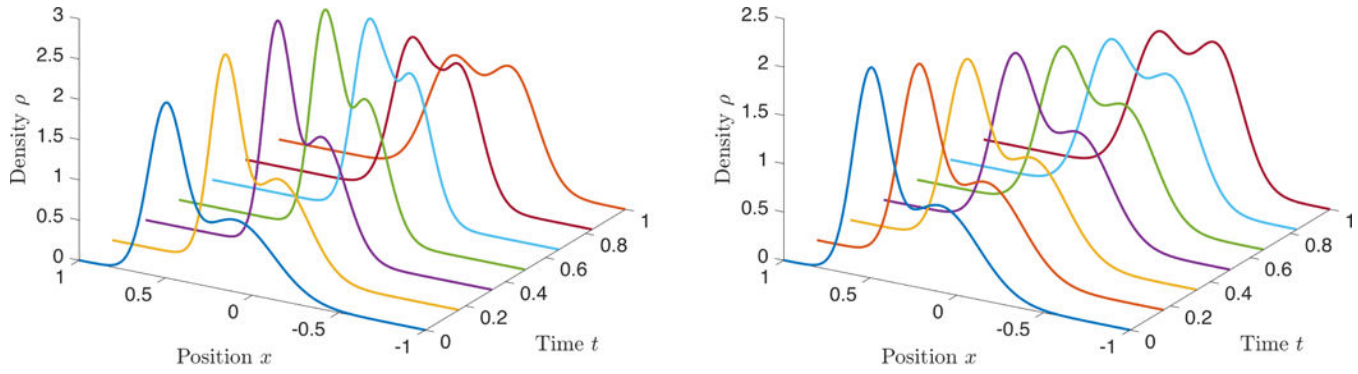


Fig. 3:
Marginal distributions (blue and red represent the two Gaussian components and black is the Gaussian mixture)



(a) OMT (b) our framework

Fig. 4:
Shortest path interpolations between μ_0 and μ_1 using different methods

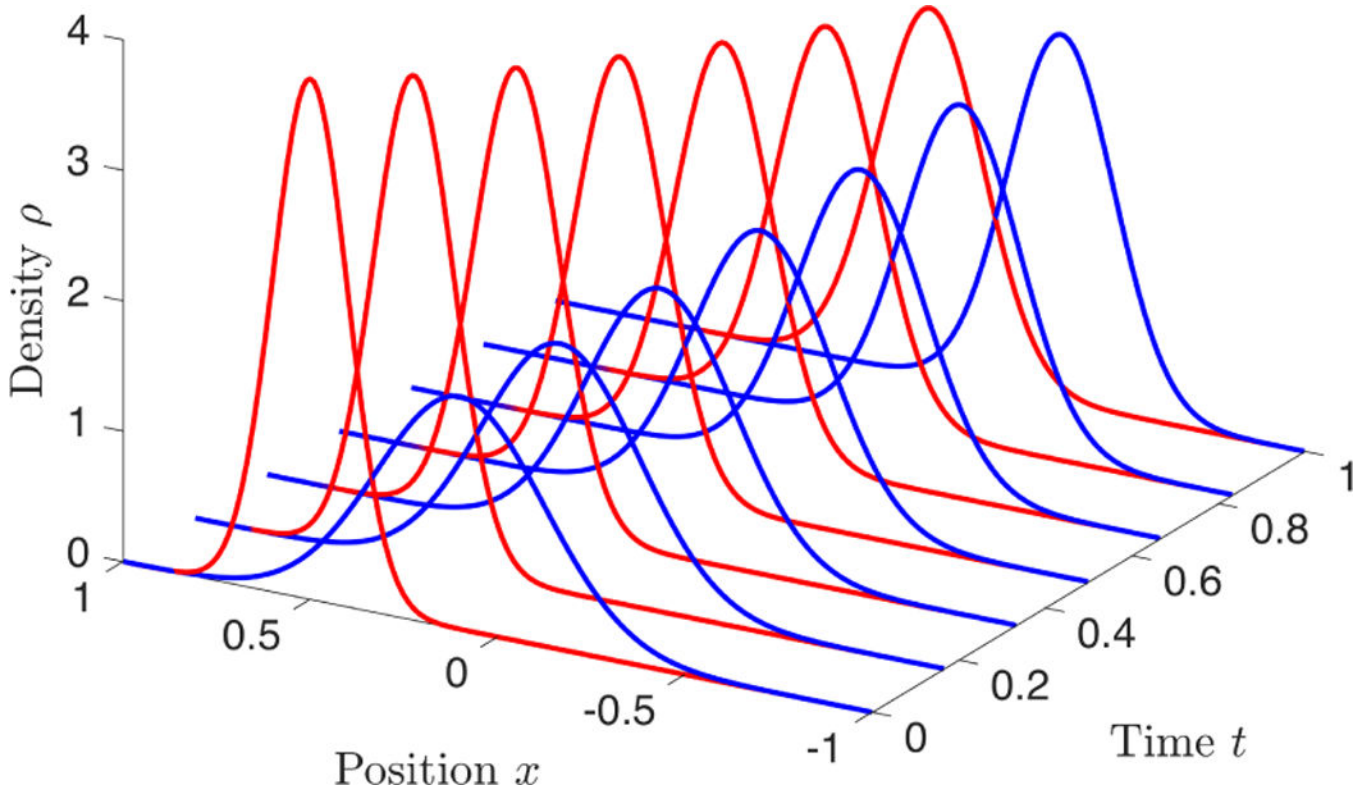
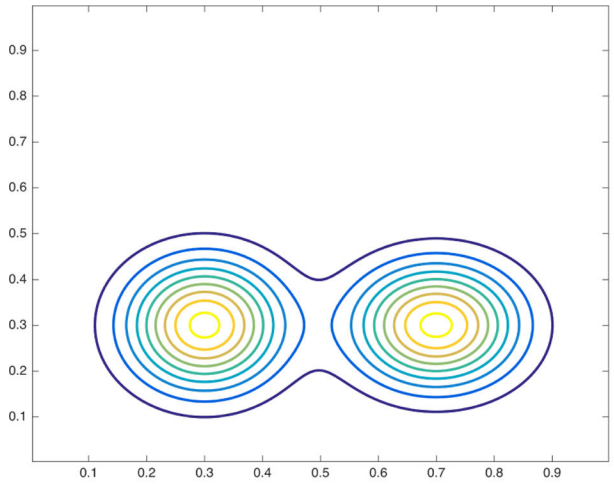
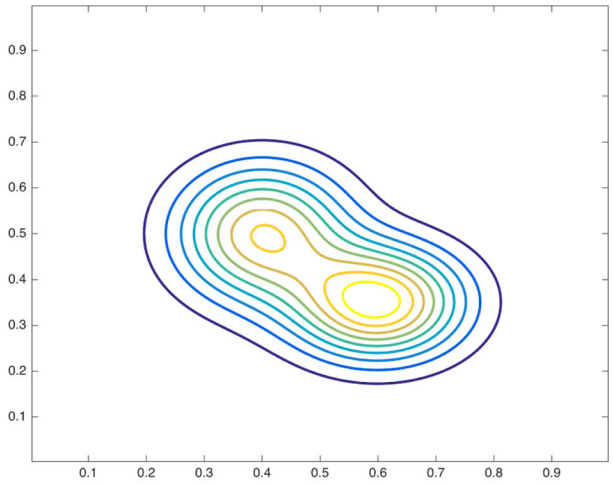


Fig. 5:
Two Gaussian components of the shortest path interpolation



(a) μ_0



(b) μ_1

Fig. 6:
Level sets of marginal distributions

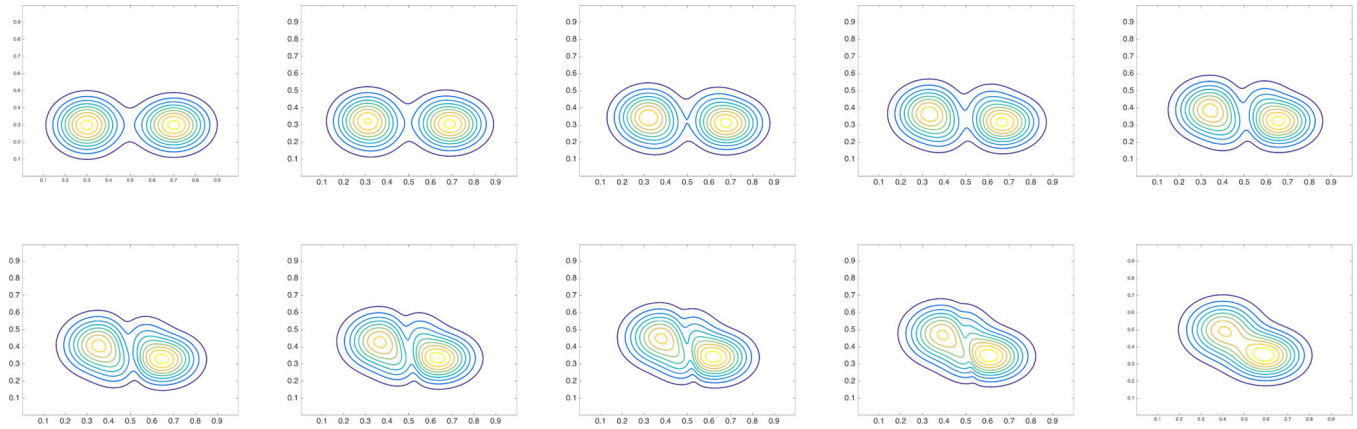


Fig. 7:
Level sets of OMT interpolation

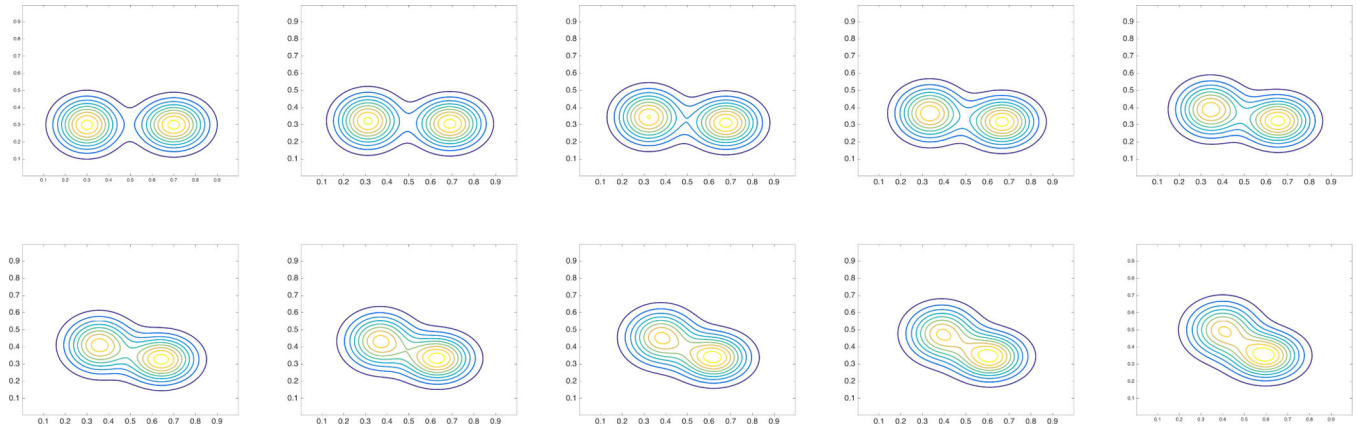


Fig. 8:
Level set of shortest path interpolation using our method

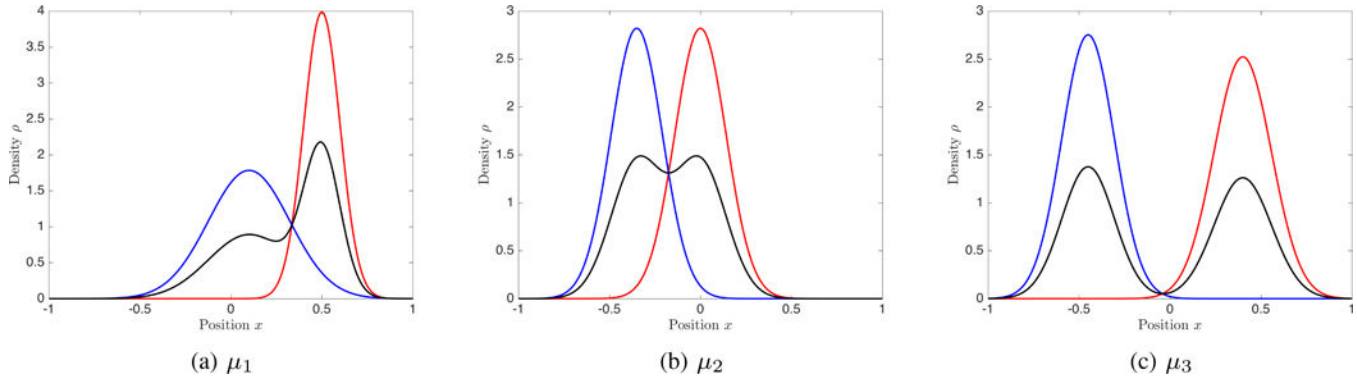
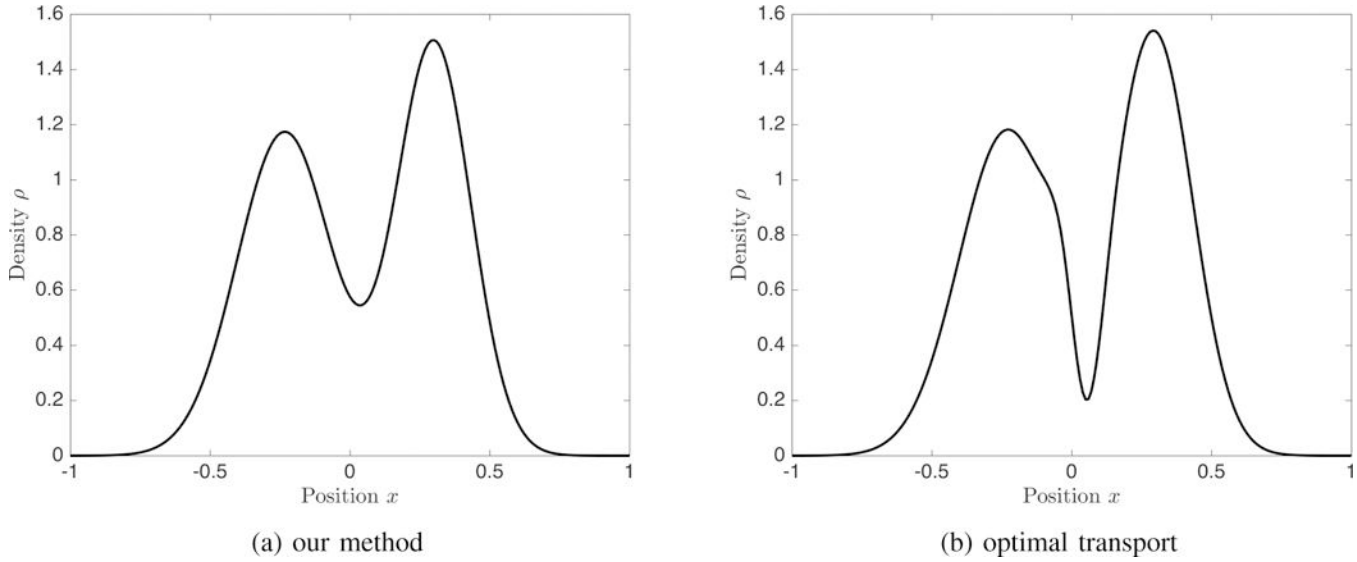


Fig. 9:
Marginal distributions (blue and red represent the two Gaussian components and black is the Gaussian mixture)



(a) our method (b) optimal transport

Fig. 10:
Barycenters of μ_1, μ_2, μ_3 with weight λ = (1/3, 1/3, 1/3)

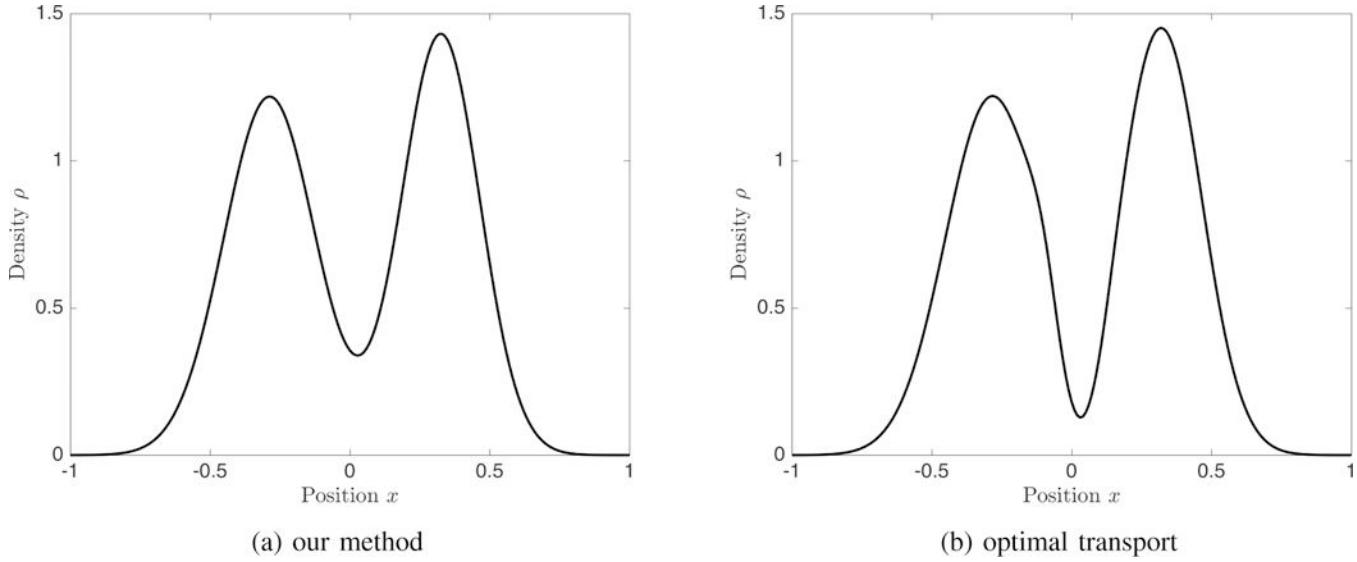


Fig. 11:
Barycenters of μ_1, μ_2, μ_3 with weight λ = (1/4, 1/4, 1/2)

TABLE I:

Computational cost comparison (in seconds)

	Our algorithm	**Rubner [20]**	**Sinkhorn [22]**
1d, N=2	0.00046	0.013	0.0046
1d, N=10	0.005	0.014	0.0048
2d, N=10	0.028	15.5	0.78
2d, N=50	0.53	15.9	0.80