23rd International Symposium on Mathematical Theory of Networks and Systems
Hong Kong University of Science and Technology, Hong Kong, July 16-20, 2018

Optimal transport for Gaussian mixture models

Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum

Abstract— **An optimal mass transport framework on the space of Gaussian mixture models is presented. These models are widely used in statistical inference. We treat such models as discrete measures on the space of Gaussian densities. Our method leads to a natural way to compare and interpolate Gaussian mixture models, with low computational cost. The method represents a first attempt to study optimal transport problems for probability densities with specific structure that can be suitably exploited.**

I. Introduction

Optimal mass transport (OMT) has become in recent years a very active research area with many deep theoretical results as well as many applications including in physics, economics, engineering, and biology [1], [2], [3], [4], [5], [6], [7]. Employing powerful numerical computational algorithms [8], [9], [10], [11], [12], [13], [14], OMT has even found applications in data science [15], [16]. OMT deals with the problem of transporting a mass from an initial distribution to a final distribution in a mass preserving manner that minimizes a given cost functional. When the unit cost is the square of the Euclidean distance, the OMT problem equips the space of probability densities with a natural Riemannian structure [17], [18], [19]. This geometry enables us to compare, interpolate and average probability densities in a very natural way, which is in line with the needs in a range of applications.

A mixture model is a probabilistic model describing properties of populations with subpopulations. Formally, it is a mixture distribution with each component representing a subpopulation. Mixture models are widely used in statistics in detecting subgroups, inferring properties of subpopulations, and many other areas [20]. An important case of mixture models is the 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 commonly used in applications due to its mathematical simplicity as well as efficient algorithms in inference (e.g., Expectation Maximization algorithm).

Computing mass transport on the entire manifold of probability densities may be computationally expensive, and so we are motivated to study OMT on certain submanifolds of probability densities. To retain the nice properties of OMT, we seek an explicit OMT framework on Gaussian mixture models. This study is partially motivated by certain problems in data analysis. As is well-known, real-world data are many times high dimensional and always have some structure. Thus, they are not densely distributed in the high dimensional space, meaning that they typically live in a low dimensional submanifold. Moreover, many times, the data are sparsely distributed among subgroups, and the difference between data within a subgroup is much less significant than that between subgroups. In such circumstances, mixture models are quite natural, and so it is of interest to develop a mathematical framework that respects such data structures.

II. Background on OMT

We now give a very brief overview of OMT theory. We only cover materials that are related to the present work. We refer the reader to [18] for more details.

Consider two 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 $T_\sharp \mu_0 = \mu_1$, 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

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

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 [18]. To circumvent this difficulty, Kantorovich presented a relaxed formulation in 1942. In this, instead of seeking a transport map, one seeks a joint distribution $\Pi(\mu_0, \mu_1)$ 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

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \pi(dxdy). \tag{1}$$

For the case where μ_0, μ_1 are absolutely continuous with corresponding densities ρ_0 and ρ_1, it is a standard result that

OMT (1) has a unique solution [21], [18], [19]. Moreover, the unique optimal transport T is the gradient of a convex function ϕ, i.e.,

$$y = T(x) = \nabla\phi(x). \tag{2}$$

Having the optimal mass transport map T, as in (2), the optimal coupling is

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

where Id stands for the identity map. The square root of the minimum of the cost defines a Riemannian metric on $P_2(\mathbb{R}^n)$, known as the Wasserstein metric W_2 [22], [17], [18], [19]. On this Riemannian-type manifold, the geodesic curve connecting μ_0 and μ_1 is given by

$$\mu_t = (T_t)_\sharp \mu_0, \quad T_t(x) = (1-t)x + tT(x), \tag{3}$$

which is called displacement interpolation. It satisfies

$$W_2(\mu_s, \mu_t) = (t-s)W_2(\mu_0, \mu_1), \quad 0 \le s < t \le 1. \tag{4}$$

A. Gaussian marginal distributions

When both of the marginals μ_0, μ_1 are Gaussian distributions, the problem can be greatly simplified [23]. In fact, a closed-form solution exists. Denote the mean and covariance of $\mu_i, i = 0, 1$ by m_i and Σ_i, respectively. Let X, Y be two Gaussian random vectors associated with μ_0, μ_1, respectively. Then the cost in (1) becomes

$$\mathbb{E}\{\|X - Y\|^2\} = \mathbb{E}\{\|\tilde{X} - \tilde{Y}\|^2\} + \|m_0 - m_1\|^2, \tag{5}$$

where $\tilde{X} = X - m_0, \tilde{Y} = Y - m_1$ are zero mean versions of X and Y. We minimize (5) over all the possible Gaussian joint distributions between X and Y. This gives

$$\min_S \left\{ \|m_0 - m_1\|^2 + \mathrm{trace}(\Sigma_0 + \Sigma_1 - 2S) \mid \begin{bmatrix} \Sigma_0 & S \\ S^T & \Sigma_1 \end{bmatrix} \ge 0 \right\} \tag{6}$$

with $S = \mathbb{E}\{\tilde{X}\tilde{Y}^T\}$. 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

$$\begin{aligned} W_2(\mu_0, \mu_1)^2 &= \|m_0 - m_1\|^2 + \mathrm{trace}(\Sigma_0 \\ &\quad + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}). \end{aligned}$$

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

$$\Sigma_t = \Sigma_0^{-1/2}\left((1-t)\Sigma_0 + t(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\right)^2\Sigma_0^{-1/2}. \tag{7}$$

The Wasserstein distance can be extended to singular Gaussian distributions by replacing the inverse by the pseudoinverse †, which leads to

$$\begin{aligned} W_2(\mu_0, \mu_1)^2 &= \|m_0 - m_1\|^2 \\ + \mathrm{trace}(\Sigma_0 + \Sigma_1 - 2\Sigma_0^{1/2}&((\Sigma_0^{1/2})^\dagger\Sigma_1(\Sigma_0^{1/2})^\dagger)^{1/2}\Sigma_0^{1/2}). \end{aligned} \tag{8}$$

In particular, when $\Sigma_0 = \Sigma_1 = 0$, we have

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

implying that the Wasserstein space of Gaussian distributions, denoted by $G(\mathbb{R}^n)$, is an extension, at least formally, 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 $M(\mathbb{R}^n)$.

As we have already seen in Section II-A, 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 do we establish a geometry that inherits the nice properties of OMT and in the meantime keeps the Gaussian mixture structure?

Our approach relies on a different way of looking at Gaussian mixture models. Instead of treating the given mixture as a distribution on the Euclidean space $\mathbb{R}^n$, we view it as a discrete distribution on the Wasserstein space of Gaussian distributions $G(\mathbb{R}^n)$. A Gaussian mixture distribution is equivalent to a discrete measure, and therefore we can apply OMT theory to such discrete measures. We will see next that this strategy retains the Gaussian mixture structure.

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

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

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

$$\min_{\pi \in \Pi(p_0, p_1)} \sum_{i,j} c(i,j)\pi(i,j) \tag{9}$$

for these two discrete measures. Here $\Pi(p_0, p_1)$ 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 $G(\mathbb{R}^n)$, that is,

$$c(i,j) = W_2(\nu_0^i, \nu_1^j)^2.$$

By standard linear programming theory, the discrete OMT problem (9) always has at least one solution. Let π^* be a minimizer, and define

$$d(\mu_0,\mu_1)=\sqrt{\sum_{i,j}c(i,j)\pi^*(i,j)}. \tag{10}$$

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

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

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

for any $\mu_0,\mu_1,\mu_2\in M(\mathbb{R}^n)$. 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 is denoted by ν_i^j. Let π_{01} (π_{12}) be the solution to (9) with marginals μ_0,μ_1 (μ_1,μ_2). Define π_{02} by

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

Clearly, π_{02} is a joint distribution between p_0 and p_2, namely, $\pi_{02}\in\Pi(p_0,p_2)$. It follows that

$$\begin{aligned}
d(\mu_0,\mu_2) &\leq \sqrt{\sum_{i,k}\pi_{02}(i,k)W_2(\nu_0^i,\nu_2^k)^2}\\
&= \sqrt{\sum_{i,j,k}\frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}W_2(\nu_0^i,\nu_2^k)^2}\\
&\leq \sqrt{\sum_{i,j,k}\frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}W_2(\nu_0^i,\nu_1^j)^2}\\
&\quad+\sqrt{\sum_{i,j,k}\frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}W_2(\nu_1^j,\nu_2^k)^2}\\
&= d(\mu_0,\mu_1)+d(\mu_1,\mu_2).
\end{aligned}$$

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. ∎

A. Geodesic

A geodesic on $M(\mathbb{R}^n)$ connecting μ_0 and μ_1 is given by

$$\mu_t=\sum_{i,j}\pi^*(i,j)\nu_t^{ij}, \tag{11}$$

where ν_t^{ij} is the displacement interpolation (see (7)) between ν_0^i and ν_1^j.

Theorem 2:

$$d(\mu_s,\mu_t)=(t-s)d(\mu_0,\mu_1),\quad 0\leq s<t\leq 1. \tag{12}$$

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

$$\begin{aligned}
d(\mu_s,\mu_t) &\leq \sqrt{\sum_{i,j}\pi^*(i,j)W_2(\nu_s^{ij},\nu_t^{ij})^2}\\
&= (t-s)\sqrt{\sum_{i,j}\pi^*(i,j)W_2(\nu_0^i,\nu_1^j)^2}\\
&= (t-s)d(\mu_0,\mu_1)
\end{aligned}$$

where we have used the property (4) of W_2. It follows that

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

On the other hand, by Theorem 1, we have

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

Combining these two, we obtain (12). ∎

We remark that μ_t is a Gaussian mixture model since it is a weighted average of the Gaussian distributions ν_t^{ij}. Even though the solution to (9) is not unique in some instances, it is unique for generic $\mu_0,\mu_1\in M(\mathbb{R}^n)$. Therefore, in most real applications, we need not worry about the uniqueness.

B. Relation between d and W_2

We first note that we have

$$d(\mu_0,\mu_1)\geq W_2(\mu_0,\mu_1)$$

for any $\mu_0,\mu_1\in M(\mathbb{R}^n)$. Equality 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 $M(\mathbb{R}^n)$ induces sub-optimality in the transport plan. Let $\gamma(t),0\leq t\leq 1$ be any piecewise smooth curve on $M(\mathbb{R}^n)$ 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 $M(\mathbb{R}^n)$ connecting μ_0 and μ_1. In view of (12), we conclude

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

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

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 $\mu_0, \mu_1 \in M(\mathbb{R}^n)$, 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 (9). When the number of Gaussian components of μ_0, μ_1 are small, this is extremely efficient.

IV. Conclusion

In this note, 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 is a huge reduction in computational cost compared with traditional OMT. Finally, when the covariances of the components are small, our distance is a very good approximation of the standard OMT distance. The extension to general mixture models or structural models will be an interesting direction in the future.

Acknowledgements

This project was supported by AFOSR grants (FA9550-15-1-0045 and FA9550-17-1-0435), grants from the National Center for Research Resources (P41- RR-013218) and the National Institute of Biomedical Imaging and Bioengineering (P41-EB-015902), NCI grant (1U24CA18092401A1,) and NIA grant (R01 AG053991), and the Breast Cancer Research Foundation.

References

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

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

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

[4] Y. Chen, T. T. Georgiou, and M. Pavon, "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.

[5] ——, "Optimal transport over a linear dynamical system," *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2137–2152, 2017.

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

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

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

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

[10] J. D. Benamou, B. D. Froese, and A. M. Oberman, "Numerical solution of the optimal transportation problem using the Monge-Ampere equation," *Journal of Computational Physics*, vol. 260, pp. 107–126, 2014.

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

[12] E. Haber and R. Horesh, "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.

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

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

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

[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

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

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

[19] ——, *Optimal Transport: Old and New*. Springer, 2008, vol. 338.

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

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

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

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