

Solving General Elliptical Mixture Models through an Approximate Wasserstein Manifold

Shengxi Li,¹ Zeyang Yu,¹ Min Xiang,¹ Danilo Mandic¹

¹Imperial College London

South Kensington Campus

London SW7 2AZ, UK

{shengxi.li17, z.yu17, m.xiang13, d.mandic}@imperial.ac.uk

Abstract

We address the estimation problem for general finite mixture models, with a particular focus on the elliptical mixture models (EMMs). Compared to the widely adopted Kullback–Leibler divergence, we show that the Wasserstein distance provides a more desirable optimisation space. We thus provide a stable solution to the EMMs that is both robust to initialisations and reaches a superior optimum by adaptively optimising along a manifold of an approximate Wasserstein distance. To this end, we first provide a unifying account of computable and identifiable EMMs, which serves as a basis to rigorously address the underpinning optimisation problem. Due to a probability constraint, solving this problem is extremely cumbersome and unstable, especially under the Wasserstein distance. To relieve this issue, we introduce an efficient optimisation method on a statistical manifold defined under an approximate Wasserstein distance, which allows for explicit metrics and computable operations, thus significantly stabilising and improving the EMM estimation. We further propose an adaptive method to accelerate the convergence. Experimental results demonstrate the excellent performance of the proposed EMM solver.

Introduction

This paper establishes a general solution to the finite mixture model problem, which has been attracting extensive research effort for decades, due to both its simple representation and potential for universal approximation on arbitrary distributions in \mathbb{R}^M . The finite mixture model also provides interpretable and statistical descriptions of data, which makes it a popular choice in a wide range of statistical learning paradigms, such as semi-supervised learning, capsule networks, and various image processing paradigms (e.g., denoising, matching and registration).

The estimation on a finite mixture model boils down to a minimisation problem which considers the mixture of distributions as a parametric model $\rho(\theta)$; it is then optimised through a minimisation of a certain discrepancy measure between $\rho(\theta)$ and empirical distributions of observed data ρ^* , namely, $\min_{\theta} d(\rho(\theta), \rho^*)$. This minimisation, although not explicitly stated, is a constrained problem because $\rho(\theta)$ must

maintain the property of a probability density throughout, to ensure that $d(\cdot, \cdot)$ is tractable.

Due to this probability constraint, various advanced numerical algorithms (solvers) have been typically restricted by either the requirement of a flexible $\rho(\theta)$ or a powerful $d(\cdot, \cdot)$. Such restrictions, for example, are one of the main rationales for using the expectation-maximisation (EM) algorithm to minimise the Kullback–Leibler (KL) divergence in Gaussian mixture model (GMM) problems (Xu and Jordan 1996). On the other hand, gradient-based numerical algorithms typically rest upon additional techniques that only work in particular situations (e.g., gradient reduction (Redner and Walker 1984), positive definite projection (Hosseini and Sra 2015), re-parametrisation (Jordan and Jacobs 1994) and Cholesky decomposition (Naim and Gildea 2012)). Besides the GMM, there exist other solutions that allow for flexible choices of $\rho(\theta)$, which still belong to the EM-type methods (e.g., mixtures of t-distributions (Peel and McLachlan 2000), Laplace distributions (Tan and Jiao 2007) and hyperbolic distributions (Browne and McNicholas 2015)). Unfortunately, given other suitable candidates of distributions, those EM-type methods cannot ensure universal convergence (Kent and Tyler 1991; Zhang, Wiesel, and Greco 2013; Sra and Hosseini 2013), which dramatically limits the power of finite mixture models.

Another issue that has been highlighted in the literature is the sensitivity to initialisations when solving GMMs (Xu, Hsu, and Maleki 2016; Jin et al. 2016). One of the main reasons is due to the use of KL divergence, which operates based on a “bin-to-bin” comparison between two density histograms. This means that mixtures which fall into a spurious local minimum cannot be corrected via the points outside. Indeed, with random initialisations for the GMM, Jin has proved that the EM method or any other first-order method which minimises the KL divergence are highly likely to result in arbitrary bad local minima (Jin et al. 2016). This can also be verified from the non-smooth optimization space with various local optimum of the KL divergence as illustrated in Fig. 1-(b), even for estimating a simple GMM (Fig. 1-(a)). The gradients on the space are highly concentrated as well, which also leads to ill-posed gradient descent.

On the other hand, by virtue of the reflection of sam-

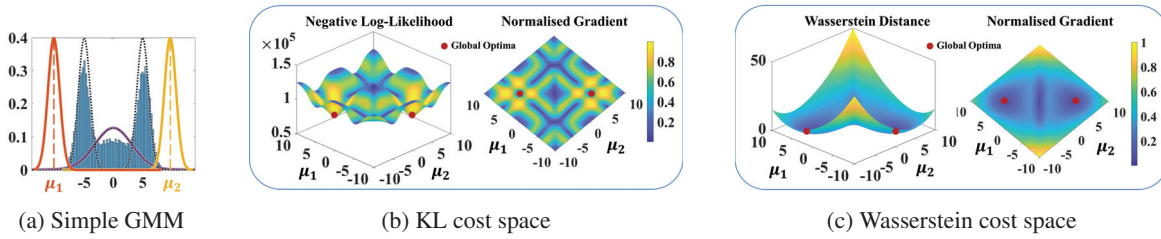


Figure 1: Illustration of a simple one-dimensional GMM with 3 clusters. (a): The ground-truth of clusters is plotted via the black dotted lines; Parametric models $\rho(\theta)$ are denoted as colourful lines; We set ALL the parameters θ of the GMM to the ground-truth EXCEPT the mean values of only two clusters, i.e., μ_1 and μ_2 . (b): The KL divergence and the normalised gradient versus the variance of μ_1 and μ_2 . (c): The Wasserstein distance and the normalised gradient versus the variance of μ_1 and μ_2 .

ple space (Kolouri et al. 2017) within the Wasserstein distance¹ (Monge 1781) which employs a “cross-bin” comparison, many practical benefits may be achieved in learning tasks (Arjovsky, Chintala, and Bottou 2017). This property is particularly appealing in mixture model problems, where it ideally provides a comprehensive distance measure over all possible transport plans. The cost space of the Wasserstein distance is also shown in Fig. 1-(c), of which the smoothness is witnessed. Basically, there only exist global optimum in this case. The gradients on this space are also well behaved, which promises to achieve superior optimisation. Most recently, Kolouri *et al.* (Kolouri, Rohde, and Hoffmann 2018) have adopted the Wasserstein distance for solving GMM problems. However, the aforementioned probability constraint enforces an extremely small stepsize (learning rate) during optimisation. Given the random projections of the sliced Wasserstein distance, this setup leads to extremely slow convergence or even non-convergence (especially in high dimensions as shown in our experiments). At the end of their optimisation, the EM algorithm may be still needed to stabilised the algorithm.

Motivations and Contributions: Despite extensive benefits, optimising the GMM under the Wasserstein distance is still not feasible due to the probability constraint. To address this from a different perspective, we resort to the statistical manifold. We should further point out that our work is different from information geometry (Amari 1998), as directly establishing a manifold in the whole density space under the Wasserstein distance is absolutely intractable and cumbersome in optimisation (Wang et al. 2015). Another problem in the Wasserstein space is that the geodesic between two mixture models may not lie in mixture models of the same type (Chen and Li 2018), which leads to non-convergence.

We propose to resolve this problem by introducing an approximation to the Wasserstein distance, followed by establishing a statistical manifold via the so induced distance, which exhibits the desirable property of being complete within mixture models. The optimisation along this manifold intrinsically satisfies the probability constraint and ensures that the solution resides in the same mixture models. More importantly, minimising the induced distance is shown

to be truly reducing the discrepancy between two mixture models, unlike most existing solutions which are based on the minimisation of the Euclidean distance to the optimal parameters. This ensures fast and stable convergence in optimisation. By realising that the existing Riemannian adaptive algorithms only make sense in updating vector parameters, we further develop a novel accelerated stochastic gradient descent method to update the positive definite matrices.

In this way, our proposed framework makes it possible to incorporate a broad family of distributions, $\rho(\theta)$, including an important class of multivariate analysis techniques called elliptical distributions (Fang, Kotz, and Ng 1990) and to further investigate the mixture family termed the elliptical mixture model (EMM). We therefore provide computable and identifiable EMMs in a unified way, which demonstrates that EMMs are quite general and flexible and include the GMMs as special cases (Fang, Kotz, and Ng 1990).

Overall, this paper proposes a complete and efficient framework for solving general EMM problems, by establishing a statistical manifold under an approximate Wasserstein distance which promotes stability and efficiency, together with an adaptive stochastic gradient algorithm to further accelerate the optimisation². Compared to the existing literature on mixture problems, the proposed solution achieves consistently superior performance not only in the GMM problems but also for general EMM problems. Our contributions can be summarised as follows:

- A unified framework for dealing with computable and identifiable EMMs, which introduces a rich choice of candidates for flexible finite mixture models.
- Establishment of the statistical manifold through the proposed approximate Wasserstein distance, which provides explicit and complete operations within the manifold.
- An adaptive accelerated Riemannian gradient descent algorithm on the established manifold, to improve the optimisation and accelerate convergence.

Computable and Identifiable EMMs

Elliptical distributions include a wide range of standard distributions, and it therefore comes as no surprise that a unified summary of computable candidates as components in

¹Throughout this paper, the term Wasserstein distance refers to the square-Wasserstein distance.

²The code of this paper is available at <https://github.com/ShengxiLi/wass.emm>

the EMMs is a prerequisite to problem definition and subsequent solutions. A classical summary can be found in Chapter 3 in (Fang, Kotz, and Ng 1990); however, despite progress this framework is not general enough as various elliptical distributions are still missing, and more importantly, it involves complicated representations for each type of elliptical distributions. The existing literature also employs different notations of particular distributions, which may lead to confusion. We thus provide a simple and unified framework for summarising the existing computable elliptical distributions via the stochastic representation, which can be used to constitute flexible and identifiable EMMs.

Preliminaries on Elliptical Distributions

A random variable, $\mathcal{X} \in \mathbb{R}^m$, is said to exhibit an elliptical distribution if and only if it admits the following stochastic representation (Fang, Kotz, and Ng 1990),

$$\mathcal{X} = {}^d \mu + \mathcal{R} \Lambda \mathcal{S}, \quad (1)$$

where $\mathcal{R} \in \mathbb{R}^+$ is a non-negative real scalar random variable which models tail properties of the elliptical distribution; $\mathcal{S} \in \mathbb{S}^{(m'-1)}$ is a random vector which is uniformly distributed on a unit spherical surface³ with the pdf within the class of $2\pi^{-m'/2}\Gamma(m'/2)$; $\mu \in \mathbb{R}^m$ is a mean (location) vector, while $\Lambda \in \mathbb{R}^{m \times m'}$ is a matrix that transforms \mathcal{S} from a sphere to an ellipse, and “ $=^d$ ” designates “the same distribution”. For a comprehensive review of elliptical distributions, we refer to (Fang, Kotz, and Ng 1990).

When $m' = m$, that is, for a non-singular scatter matrix $\Sigma = \Lambda \Lambda^T$, the pdf for elliptical distributions does exist and has the following form

$$p(\mathbf{x}) = \underbrace{2\pi^{-\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)}_{c_m} \det(\Sigma)^{-\frac{1}{2}} \underbrace{g((\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu))}_t. \quad (2)$$

In (2), the term c_m serves as a normalisation term and relates solely to m . We denote the Mahalanobis distance $(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$ by t . The density generator, $g(t)$, can be explicitly expressed as $t^{-(m-1)/2} p_{\mathcal{R}}(\sqrt{t})$, where $t > 0$ and $p_{\mathcal{R}}(t)$ denotes the pdf of \mathcal{R} . Thus, \mathcal{R} , or equivalently⁴ \mathcal{R}^2 , fully characterises $g(\cdot)$, i.e., the type of elliptical distributions. For example, when $\mathcal{R}^2 = {}^d \chi_m^2$ (χ_m^2 denotes the chi-squared distribution of dimension m), then in (2), $g(t) \propto \exp(-t/2)$, which formulates the multivariate Gaussian distribution. Therefore, the elliptical distribution can be fully characterised by μ , Σ and \mathcal{R} . For simplicity, the elliptical distribution in (2) will be denoted by $\mathcal{X} \sim \mathcal{E}(\mathbf{x}; \mu, \Sigma, \mathcal{R})$, where $\mathcal{E}(\mathbf{x}; \mu, \Sigma, \mathcal{R}) = c_m \det(\Sigma)^{-1/2} g(t)$ of (2).

Computable and Identifiable EMMs

Due to the fact that the \mathcal{R}^2 decides the type of elliptical distributions, we here provide a unified summary of elliptical distributions in Table 1; this is achieved through stochastic

representations in (1). This makes it possible to avoid complicated formulations, and to instead classify different categories simply through several typical distributions of \mathcal{R}^2 , which also allows for simple and intuitive sample generations for elliptical distributions. The proof for expressions in this table is provided in (Li et al. 2019). Uniquely, this further clarifies the commonalities between the members of the elliptical family of distributions. More importantly, constructing an EMM with the candidates in Table 1 can be easily proved to be identifiable based on Theorem 2 in (Holzmann, Munk, and Gneiting 2006). It is thus convenient and safe to establish a well-defined EMM by the candidates summarised in Table 1.

Problem Statement and Notations

For generality, we assume that there are k mixtures in the EMM model, latent variables $\mathcal{Z}_i \in \{0, 1\}$ are binary, and the probability of choosing the i -th mixture is denoted by $p(\mathcal{Z}_i = 1) = \pi_i$, so that $\sum_{i=1}^k \mathcal{Z}_i = 1$ and $\sum_{i=1}^k \pi_i = 1$. If we use the random variable \mathcal{Y} to denote the EMM, then $\mathcal{Y} = {}^d \sum_{i=1}^k \mathcal{Z}_i \mathcal{X}_i \sim \sum_{i=1}^k \pi_i \mathcal{E}(\mathbf{x} | \mu_i, \Sigma_i, \mathcal{R})$, and the pdf of \mathcal{Y} can be expressed as

$$p(\mathbf{y}) = \sum_{i=1}^k \pi_i c_m \det(\Sigma_i)^{-\frac{1}{2}} g((\mathbf{y} - \mu_i)^T \Sigma_i^{-1} (\mathbf{y} - \mu_i)), \quad (3)$$

where the density generator $g(\cdot)$ can be chosen flexibly from Table 1, for which the identifiability is ensured (Holzmann, Munk, and Gneiting 2006). Thus, without ambiguity in the context, we shall denote the EMM \mathcal{Y} through the pdf of (3), that is, as \mathcal{Y}_{θ} , where $\theta = \{\pi, \mu_i, \Sigma_i\}$ ($\pi = [\pi_1, \pi_2, \dots, \pi_k]^T$ and $i = 1, 2, \dots, k$).

Statistical Manifold towards EMMs

As the Wasserstein distance possesses a Riemannian structure (Kolouri et al. 2017), it is then natural to treat each EMM as a “point” in the manifold, whereby the metric is defined by the Hessian of the Wasserstein distance. However, as pointed out by Chen (Chen and Li 2018), in the Wasserstein space, the geodesic between two points, i.e., two GMMs, does not necessarily belong to the GMM of the same type. A way to solve this problem is to pull back from the whole density space to the parametric space (Chen and Li 2018). This treatment, nevertheless, is still a toy solution due to two main limitations. The first deficiency is that the metric does not have a closed-form representation, and needs to be numerically obtained for every possible value before the optimisation. This is highly prohibitive especially for multivariate cases due to the curse of dimensionality. Besides computational intractability, operations such as the exponential mapping and the vector transport in the manifold cannot be well defined as there is typically a second-order differential equation involved to obtain those operations.

We thus propose an approximate Wasserstein distance between two EMMs as a means to define a well-behaved manifold for the EMM problems; this is achieved by the property that the Wasserstein distance of two elliptical distributions is

³The term $\mathbb{S}^{m'-1}$ is defined as $\mathbb{S}^{m'-1} := \{\mathbf{x} \in \mathbb{R}^{m'} : \mathbf{x}^T \mathbf{x} = 1\}$.

⁴The term \mathcal{R}^2 is frequently used in practice because $\mathcal{R}^2 = {}^d (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$.

Table 1: $\mathcal{R}^2 \leftrightarrow$ Computable elliptical distributions

Types	\mathcal{R}^2	\leftrightarrow	$c_m \cdot g(t)$	Typical Multivariate Dist.
Kotz Type Ref[a]	$\mathcal{R}^2 = {}^d \mathcal{G}^{1/s}$, $\mathcal{G} \sim \text{Ga}(\frac{2a+m-2}{2s}, b)$ $a > 1 - \frac{m}{2}, b, s > 0$	$=$	$\left(\frac{\Gamma(m/2) s b^{(2a+m-2)/(2s)}}{\Gamma((2a+m-2)/2s) \pi^{m/2}} \right) t^{a-1} \exp(-bt^s)$	Gamma: $s = 1$ Weibull: $a = s$ Generalised Gaussian: $a = 1$ Gaussian: $a = 1, s = 1, b = \frac{1}{2}$
Scale Mixture of Normals	Pearson Type VII Ref[a] $\mathcal{K}^{-1} \sim \text{Ga}(s - \frac{m}{2}, 2v)$, $v > 0, s > m/2$	$=$	$\left(\frac{(\pi v)^{-m/2} \Gamma(s)}{\Gamma(s - m/2)} \right) (1 + t/v)^{-s}$	T -dist.: $s = \frac{m+v}{2}$ Cauchy: $v = 1, s = \frac{m+1}{2}$
$\mathcal{R}^2 = {}^d \mathcal{G} \cdot \mathcal{K}$, $\mathcal{G} \sim \text{Ga}(\frac{m}{2}, \frac{1}{2})$, \mathcal{K} has different dist.	Hyperbolic Type Ref[b] $\mathcal{K} \sim \text{GIG}(v, a, \lambda)$ $v, a > 0, \lambda \in \mathbb{R}$	$=$	$\left(\frac{(v/a)^{\lambda/2}}{(2\pi)^{m/2} \text{BeK}_\lambda(\sqrt{av})} \right) \frac{\text{BeK}_{(\lambda-m/2)}(\sqrt{av+vt})}{(\sqrt{a/v+tv})^{m/2-\lambda}}$	Inverse-Gaussian: $\lambda = -1/2$ K -dist. (Ollila et al. 2012): $a \rightarrow 0, \lambda > 0$ Laplace: $a \rightarrow 0, \lambda = 1, v = 2$
	Other Types Ref[c] $\sqrt{\mathcal{K}} \sim \partial \text{Kov}(\frac{\kappa}{2}) / \partial \mathcal{K}$ $\mathcal{K} \sim \text{SaS}(\frac{a}{2}), a \in (0, 2)$	$=$	$\exp(-t) / (1 + \exp(-t))^2$	Logistic
		\propto	$\text{SaS}(a)$	α -stable
Pearson Type II	$\mathcal{R}^2 \sim \text{Beta}(m/2, s), s > 1$	$=$	$\left(\frac{\Gamma(m/2+s)}{\pi^{m/2} \Gamma(s)} \right) (1-t)^{s-1}, t \in [0, 1]$	Ref[a]
Notations:	$a, b, s, v, \lambda, \alpha$ are adjustable parameters for different types of dist.; \mathcal{G} and \mathcal{K} are random variables related to \mathcal{R}^2 ; m is the dimension. Gamma dist.: $\text{Ga}(x, y) = y^x t^{x-1} \exp(-yt) / \Gamma(x)$; Inverse Gaussian dist.: $\text{GIG}(x, y, z) = \frac{(xy/y)^{z/2}}{2 \text{BeK}_z(\sqrt{xy})} t^{z-1} \exp(-\frac{xt^2+y}{2t})$; Kolmogorov-Smirnov dist.: $\text{Kov}(x) = 1 - 2 \sum_{n=1}^{\infty} (-1)^{n+1} \exp(-2n^2 x^2)$; Beta dist.: $\frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)} t^{x-1} (1-t)^{y-1}$; $\text{SaS}(a)$: the symmetric α -stable dist. with index a ; $\text{BeK}_\lambda(y)$: the Bessel function of the third kind; $\Gamma(x)$: the Gamma function.			
References:	[a]: (Fang, Kotz, and Ng 1990); [b]: (Browne and McNicholas 2015); [c]: (Andrews and Mallows 1974; Stefanski 1991)			

completely and explicitly defined. We then provide the Riemannian metric according to the defined distance.

Approximate Wasserstein Distance between EMMs

We now focus on the distance between two EMMs \mathcal{Y}_1 and \mathcal{Y}_2 , and propose an approximate Wasserstein distance by treating each distribution within an EMM as a “super-point” and defining a transport-like distance between those “super-points”. A rigorous definition is given as follows.

Definition 1. Given two EMMs \mathcal{Y}_1 and \mathcal{Y}_2 , a discrepancy measure is defined as

$$d_U(\mathcal{Y}_1, \mathcal{Y}_2) = \min_{\gamma(i,j)} \left(\sum_{i,j} \frac{\gamma(i,j)}{k} d_W^2(\mathcal{X}_{i,1}, \mathcal{X}_{j,2}) + \arccos \left(\sum_{i,j} \gamma(i,j) \sqrt{\pi_{i,1} \pi_{j,2}} \right) \right), \quad (4)$$

where $d_W^2(\mathcal{X}_{i,1}, \mathcal{X}_{j,2})$ is the Wasserstein distance between the elliptical distributions $\mathcal{X}_{i,1}$ and $\mathcal{X}_{j,2}$. $\gamma(i, j)$ is binary $\in \{0, 1\}$; for each i and j , $\gamma(i, j)$ satisfies $\sum_{i=1}^k \gamma(i, j) = 1$ and $\sum_{j=1}^k \gamma(i, j) = 1$.

Theorem 1. Given two EMMs \mathcal{Y}_1 and \mathcal{Y}_2 , the discrepancy measure $d_U(\mathcal{Y}_1, \mathcal{Y}_2)$ defines a distance.

Proof. Please refer to (Li et al. 2019). \square

By definition, the term $\arccos(\sum_{i,j} \gamma(i, j) \sqrt{\pi_{i,1} \pi_{j,2}})$ intrinsically relates to the probability constraint of $\sum_{i=1}^k \pi_{i,1} = \sum_{j=1}^k \pi_{j,2} = 1$, while $\gamma(i, j)$ operates as a bijection between mixture components in \mathcal{Y}_1 and \mathcal{Y}_2 . $\sum_{i,j} \frac{\gamma(i,j)}{k} d_W^2(\mathcal{X}_{i,1}, \mathcal{X}_{j,2})$ can thus be regarded as a discrete transport between k uniformly distributed “super-points” $\mathcal{X}_{i,1}$ and $\mathcal{X}_{j,2}$, for which their cost is defined as $d_W^2(\mathcal{X}_{i,1}, \mathcal{X}_{j,2})$. We also noticed a computational optimal

transport distance proposed by (Chen, Georgiou, and Tanenbaum 2018). Our metric can be further seen as a restriction (or upper bound) of the Chen’s metric because we operate via a one-to-one transport whilst the Chen’s metric admits arbitrary transport plans. However, the most advantageous property of our metric is that it defines an explicit manifold over EMMs, which will be introduced shortly.

More importantly, $d_U(\mathcal{Y}_1, \mathcal{Y}_2)$ comprehensively reflects the discrepancy between \mathcal{Y}_1 and \mathcal{Y}_2 via a summation operation; this means $d_U(\mathcal{Y}_1, \mathcal{Y}_2) = 0$ if and only if both the difference between mixture components $\mathcal{X}_{i,1}$ and $\mathcal{X}_{j,2}$ and the difference between latent variables $\pi_{i,1}$ and $\pi_{j,2}$ equal to 0. The following lemma further proves that for balanced EMMs, $d_U(\mathcal{Y}_1, \mathcal{Y}_2)$ is an upper bound of the Wasserstein distance $d_W^2(\mathcal{Y}_1, \mathcal{Y}_2)$.

Lemma 1. Given two balanced EMMs \mathcal{Y}_1 and \mathcal{Y}_2 (i.e., $\pi_{i,1} = \pi_{j,2} = 1/k$ for all i, j), $d_U(\mathcal{Y}_1, \mathcal{Y}_2)$ is an upper bound of the Wasserstein distance:

$$d_W^2(\mathcal{Y}_1, \mathcal{Y}_2) \leq d_U(\mathcal{Y}_1, \mathcal{Y}_2). \quad (5)$$

The equality holds when $k = 1$.

Proof. Please refer to (Li et al. 2019), where the tightness of the upper bound is also analysed. \square

Statistical Manifold for EMM Problems

Before introducing the statistical manifold for EMMs, we would like to give credit to several most recent works on Gaussian distributions and elliptical distributions, which lay a basis of our metric proposed in this section. We would like to mention (Knott and Smith 1984) of the Wasserstein distance of Gaussian measures and (Muzellec and Curi 2018) for the elliptical distributions. Most recently, the Riemannian manifold for Gaussian distributions has been established in an explicit form (Takatsu and others 2011;

Malagò, Montrucchio, and Pistone 2018). Then, on the basis of the approximated Wasserstein distance, we provide the Riemannian metric of the EMM problems by calculating the Hessian of the proposed distance in Definition 1 as follows.

Lemma 2. *The approximate Wasserstein distance $d_U(\mathcal{Y}_1, \mathcal{Y}_2)$ represents an explicit Riemannian metric in the parametric space, and the corresponding Riemannian manifold is a product manifold of $\mathbb{R}^k \times \prod_{i=1}^k (\mathbb{R}^m \times \mathbb{P})$, where \mathbb{P} is the $m \times m$ positive definite manifold:*

- 1) *The manifold for the square root of π_i , i.e., $[\sqrt{\pi_1}, \sqrt{\pi_2}, \dots, \sqrt{\pi_k}]^T$ is a sphere manifold of \mathbb{R}^k .*
- 2) *The manifold for μ_i is the Euclidean space of \mathbb{R}^m .*
- 3) *The manifold for Σ_i , i.e., \mathbb{P} , is defined by*

$$ds^2 = \frac{\mathbb{E}[\mathcal{R}^2]}{m} (\mathbb{L}_{\Sigma_i}[d\Sigma]) \Sigma_i (\mathbb{L}_{\Sigma_i}[d\Sigma]). \quad (6)$$

In (6), $\mathbb{L}_A[C] = B$ is a Lyapunov operator: $AB + BA = C$, where $A, B, C \in \mathbb{P}$. More importantly, the sectional curvature is non-negative ($= m/\mathbb{E}[\mathcal{R}^2]k_G$ where $k_G \geq 0$ is the sectional curvature for Gaussian cases (Takatsu and others 2011)). Recall that \mathcal{R} is defined in (1).

Proof. Please refer to (Li et al. 2019). \square

Remark. *The metric in (6) provides a manifold for positive definite matrices. Compared to the best known manifold belonging to the Hadamard manifold (non-positive sectional curvature) (Sra and Hosseini 2015), the newly developed manifold provides an example of non-negative manifolds. It is actually further stated that an Alexandrov space has the non-negative curvature iff it is a Wasserstein space (Sturm and others 2006). The established metric thus provides a desirable reflection on the curvature when dealing with Wasserstein related EMM problems.*

Adaptively Accelerated Optimisation

By virtue of optimising on a statistical manifold, the probability constraint can be satisfied automatically; this allows us to incorporate various numerical algorithms when solving the EMM problems. More specifically, we first show that the constrained minimisation problem can be transformed to an unconstrained one when restricted on the statistical manifold, which results in the “vanilla” gradient descent on the manifold. We then propose an adaptively accelerated solver.

Vanilla Gradient Descent on Statistical Manifold

Similar to (Martens 2014), for gradient descent methods, the way for gradient descent in our work is given by,

$$\begin{aligned} \mathcal{Y}_{\theta+\Delta\theta}^* &= \arg \min_{\mathcal{Y}_{\theta+\Delta\theta}} d_{SW}(\mathcal{Y}_{\theta+\Delta\theta}, \mathcal{Y}^*) + \frac{d_U(\mathcal{Y}_{\theta}, \mathcal{Y}_{\theta+\Delta\theta})}{c^2}, \\ &\rightarrow \Delta\theta^* = \text{Exp}(-\alpha \nabla_U(\theta)) \end{aligned} \quad (7)$$

where $\Delta\theta$ is the step size for the next iteration and c^2 denotes a sphere of all realisable distributions, which ensures searching for optimal $\Delta\theta^*$ without being slowed down by the curvature; $d_{SW}(\mathcal{Y}_{\theta+\Delta\theta}, \mathcal{Y}^*)$ denotes the sliced Wasserstein distance (Rabin et al. 2011) between the parametric

mixture model $\mathcal{Y}_{\theta+\Delta\theta}$ and the observed samples \mathcal{Y}^* . By approximating $d_U(\mathcal{Y}_{\theta}, \mathcal{Y}_{\theta+\Delta\theta})$ with its Hessian term (i.e., the inner product on the tangent space), we naturally obtain an unconstrained Riemannian gradient descent of the second row of (7), where $\nabla_U(\theta)$ is the Riemannian gradient on the statistical manifold defined by $d_U(\cdot, \cdot)$; α is the stepsize of each iteration; $\text{Exp}(-\alpha \nabla_U(\theta))$, called exponential mapping, projects the step movement $-\alpha \nabla_U(\theta)$ from the tangent space to the statistical manifold along geodesics (Absil, Mahony, and Sepulchre 2009), which ensures the probability constraint. On the other hand, when the proposed distance $d_U(\mathcal{Y}_{\theta}, \mathcal{Y}_{\theta+\Delta\theta})$ is changed to the Euclidean distance, (7) is the trivial gradient descent (Kolouri, Rohde, and Hoffmann 2018). However, as mentioned above, the trivial gradient descent does not satisfy the probability constraint nor does it reflect the probability space curvature, which leads to inefficient and unstable optimisation.

In (7), the sliced Wasserstein distance (Rabin et al. 2011) provides a feasible solution in solving the semi-discrete Wasserstein problem, i.e., $\mathcal{Y}_{\theta+\Delta\theta}$ is continuous and \mathcal{Y}^* is the sum of Dirac masses (Lévy 2015). It uses unit random projections to turn the original problem to a one-dimensional Wasserstein problem via the Radon transform (Rabin et al. 2011), so that a closed-form solution can be obtained⁵. More importantly, as the sliced Wasserstein distance is composed of a set of random projections (Rabin et al. 2011), it is then natural to implement a stochastic gradient descent on the Riemannian manifold for each random projection (denoted as $\mathbf{p} \in \mathbb{S}^{m-1}$). The random projection process also allows for parallel gradient descent. We provide the basic operations of problem in (7) in Table 2, while the details are provided in (Li et al. 2019).

Adaptively Accelerated Algorithm

On the basis of several accelerated Riemannian stochastic gradient descent methods which adopt the first-order moment information (Zhang and Sra 2016), recently Becigneul and Ganea (Becigneul and Ganea 2019) further proposed a Riemannian adaptive method by employing the second-order moments. Although this adaptive stochastic gradient descent method (Becigneul and Ganea 2019) may be incorporated to our work to further improve the convergence, we argue that it only makes sense for updating vector parameters, i.e., π^h and μ_i^h , because in this case the second-order moment can be calculated element-wise by a decomposition into product manifolds (Becigneul and Ganea 2019). This is similar to the *Adam* algorithm in the Euclidean space (Kingma and Ba 2014). When updating the matrix parameter, Σ_i^h , however, it does not capture the second-order information properly due to the direct accumulation of the second-order moments over the whole matrix manifold; this slows down convergence in the optimisation.

The key difficulty for the matrix case is that it is meaningless to accumulate second-order moments in an element-wise manner due to the structure within the matrix. We here

⁵It needs to be pointed out that other approximate distances (e.g., entropy relaxed Wasserstein distances) for the semi-discrete problem can also be seamlessly adopted in (7) of our work.

Table 2: Basic operations of the manifold in Lemma 2

For the h -th iteration		
$\sqrt{\pi}^h$	$\nabla_U(\cdot)$	$\text{Exp}(-\alpha \nabla_U(\cdot))$
μ_i^h	$\nabla_E(\sqrt{\pi}^h) - (\sqrt{\pi}^h)^T \nabla_E(\sqrt{\pi}^h) \cdot \sqrt{\pi}^h$	$\cos(\ \alpha \nabla_U(\sqrt{\pi}^h)\ _2) \sqrt{\pi}^h - \frac{\sin(\ \alpha \nabla_U(\sqrt{\pi}^h)\ _2)}{\ \nabla_U(\sqrt{\pi}^h)\ _2} \nabla_U(\sqrt{\pi}^h)$
Σ_i^h	$\nabla_E(\mu_i^h)$	$\mu_i^h - \alpha \nabla_U(\mu_i^h)$
	$\nabla_E(\Sigma_i^h) \Sigma_i^h + \Sigma_i^h \nabla_E(\Sigma_i^h)$	$(\mathbb{L}_{\Sigma_i^h}[-\alpha \nabla_U(\Sigma_i^h)] + I) \Sigma_i^h (\mathbb{L}_{\Sigma_i^h}[-\alpha \nabla_U(\Sigma_i^h)] + I)$
For the Radon transform $\mathbf{p} \in \mathbb{S}^{m-1}$, specified in the EMM problems:		
$\nabla_E(\sqrt{\pi}^h)$	The i -th dimension of $\nabla_E(\sqrt{\pi}^h)$ is $2 \int_{\mathbb{R}} c_m \phi(y) \sqrt{\pi_i}^h (\mathbf{p}^T \Sigma_i^h \mathbf{p})^{-1/2} g(\frac{(y - \mathbf{p}^T \mu_i^h)^2}{\mathbf{p}^T \Sigma_i^h \mathbf{p}}) dy$	
$\nabla_E(\mu_i^h)$	$= \left(-2 \int_{\mathbb{R}} c_m \phi(y) \pi_i^h (\mathbf{p}^T \Sigma_i^h \mathbf{p})^{-3/2} g'(\frac{(y - \mathbf{p}^T \mu_i^h)^2}{\mathbf{p}^T \Sigma_i^h \mathbf{p}}) (y - \mathbf{p}^T \mu_i^h) dy \right) \mathbf{p}$	
$\nabla_E(\Sigma_i^h)$	$= \left(- \int_{\mathbb{R}} c_m \phi(y) \pi_i^h (\mathbf{p}^T \Sigma_i^h \mathbf{p})^{-3/2} \left(\frac{1}{2} g(\frac{(y - \mathbf{p}^T \mu_i^h)^2}{\mathbf{p}^T \Sigma_i^h \mathbf{p}}) + g'(\frac{(y - \mathbf{p}^T \mu_i^h)^2}{\mathbf{p}^T \Sigma_i^h \mathbf{p}}) \frac{(y - \mathbf{p}^T \mu_i^h)^2}{\mathbf{p}^T \Sigma_i^h \mathbf{p}} \right) dy \right) \mathbf{p} \mathbf{p}^T$	
$\phi(y)$ is the Kantorovich potential (Chen and Li 2018). $\nabla_E(\cdot)$ denotes the Euclidean gradient with regard to $\sqrt{\pi}^h$, μ_i^h and Σ_i^h .		

Alg. 1: Riemannian adaptively accelerated manifold optimisation

Input: n observed samples $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$; stepsize $\{\alpha^h\}_{h=1}^H$;
hyper-parameters $\{\beta_1^h\}_{h=1}^H$ and β_2
Initialise: 1st-order moment $\{\mathbf{u}_i^0\}_{i=1}^k$; 2nd-order moment $\{\mathbf{v}_i^0\}_{i=1}^k$;
for: $h = 1$ to H **do**
 Random projection: $\mathbf{p} \in \mathbb{S}^{(m-1)}$
 Update $\sqrt{\pi}^{h+1} = \text{Exp}(-\alpha \nabla_U(\sqrt{\pi}^h))$
 for: $i = 1$ to k **do**
 Update $\mu_i^{h+1} = \text{Exp}(-\alpha \nabla_U(\mu_i^h))$
 Update Σ_i^h by the *Dadam*:
 $\mathbf{u}_i^h = \beta_1^h \varphi_{\Sigma_i^{h-1} \rightarrow \Sigma_i^h}(\mathbf{u}_i^{h-1}) + (1 - \beta_1^h) \nabla_U(\Sigma_i^h)$
 $\mathbf{v}_i^h = \beta_2 \mathbf{v}_i^{h-1} + (1 - \beta_2) \nabla_E(\Sigma_i^h) \nabla_E(\Sigma_i^h)^T$
 $\text{adp}_i^h = \max\{\mathbf{p}^T \mathbf{v}_i^h \mathbf{p}, \text{adp}_i^{h-1}\}$
 $\Sigma_i^{h+1} = \text{Exp}(-\alpha^h \mathbf{u}_i^h / \sqrt{\text{adp}_i^h})$
 end for
end for
Return: π^H , $\{\mu_i^H\}_{i=1}^k$, $\{\Sigma_i^H\}_{i=1}^k$

view the second-order moments of matrices from another perspective, by realising that a positive definite matrix can be decomposed into a set of eigenvectors and the corresponding eigenvalues. The eigenvectors can be regarded as a set of projection directions and the eigenvalues are scalars that can be connected with the accumulation in those directions. Furthermore, we can see from Table 2 that the Euclidean gradient $\nabla_E(\Sigma_i^h)$ consists of a scalar weight multiplied by a rank-1 matrix characterised by the direction $\mathbf{p} \mathbf{p}^T$. Therefore, instead of the element-wise accumulation, we propose a direction-wise accumulation of second-order moments to adaptively adjust the stepsize when updating the matrix. The details of our algorithm in Algorithm 1, with our directional adaptive accelerated method (*Dadam*) of updating Σ_i^h being achieved by \mathbf{v}_i^h and adp_i^h .

Moreover, as there is no explicit parallel transport in the Wasserstein space, we propose a new means of vector transport $\varphi_{\Sigma_i^{h-1} \rightarrow \Sigma_i^h}(\mathbf{u}_i^{h-1}) = \mathbb{L}_{\Sigma_i^{h-1}}[\mathbf{u}_i^{h-1}] \Sigma_i^h + \Sigma_i^h \mathbb{L}_{\Sigma_i^{h-1}}[\mathbf{u}_i^{h-1}]$ to accumulate the first-order moments.

As for the convergence analysis, because our goal is to minimise the Wasserstein distance, qualitatively, our manifold, defined by the Hessian of the approximate Wasserstein distance as in Lemma 2, implicitly involves the second-order information and accelerates the convergence. Furthermore,

adp_i^h denotes the accumulation in the current projection \mathbf{p} , which is a scalar (for computational ease) and does not need any eigen-decomposition operations. This, on the one hand, avoids the high computational complexity of calculating any inverse of matrices, making our method the same computational complexity as that of (Becigneul and Ganea 2019) at each iteration. On the other hand, it also makes our convergence analysis similar to that of (Becigneul and Ganea 2019), upon realising that the sectional curvature is automatically bounded from below; we omit the analysis here.

Experimental Results

We evaluated the effectiveness of our manifold and the proposed *Dadam* on both synthetic data and image data, by employing 4 EMMs, i.e., mixtures of *Gaussian*, *Logistic*, *Cauchy* and *Gamma* ($s = 1, a = 2, b = 0.5$ in Table 1).

Synthetic Data: Each synthetic dataset contains different mixtures of Gaussian distributions, with 10,000 samples in total. Specifically, we employed three types of synthetic datasets, i.e., $\{m = 2, k = 3\}$, $\{m = 8, k = 9\}$ and $\{m = 16, k = 27\}$. Each type contains 10 randomly generated mixture datasets (the eccentricity ε and separation c of (Dasgupta 1999) were equal to 10). For each method, we tested on each dataset over 10 random initialisations and recorded the mean and standard deviation. We then averaged over all datasets to obtain the performance of each method.

Image Data: We adopted the MNIST (LeCun, Cortes, and Burges 2010) dataset as well as the BSDS500 (Arbelaez et al. 2011) benchmark dataset in our evaluation. For the MNIST dataset, each image in both the training and testing sets was downsampled to 3×3 and 5×5 ; by vectorising each we obtained two test data with $(n \times m)$ being $(70,000 \times 9)$ and $(70,000 \times 25)$, respectively. The evaluation on the BSDS500 dataset is slightly different from that on the MNIST dataset, which enables to verify our method on some extra basic tasks such as the image reconstruction. Specifically, instead of modelling over the whole dataset, each image in the dataset was treated as one test data; this means we had 500 test data in the BSDS500 dataset. Furthermore, each test data in both the MNIST and BSDS500 datasets was tested for every algorithm with 10 random initialisations, where the mean and standard deviation were also recorded.

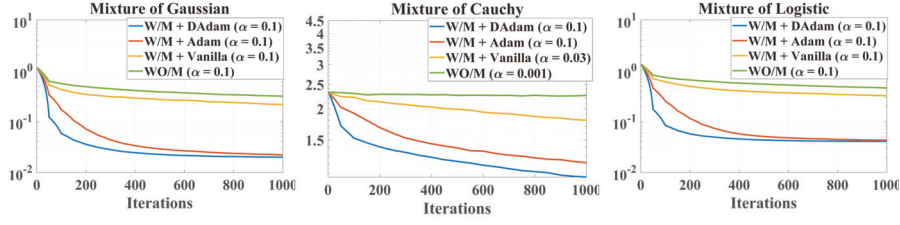


Figure 2: The Wasserstein distance against the number of iterations for the four considered algorithms, averaged over the three types of datasets ($\{m, k\} = \{2, 3\}, \{8, 9\}, \{16, 27\}$). The best learning rate is shown in the legend. 1000 iterations are for illustration purpose. The computational time per iteration is discussed in Table 3.



Figure 3: The optimised GMMs via the EM and our ($W/M + Dadam$) methods on a flower-shaped synthetic data ($n = 10,000$) with random initialisations. Each cluster in the flower-shaped data is composed by Gaussian-distributed samples.

Table 3: Comparisons among our, WO/M and EM methods for GMMs by varying m and k .

	$m = 2, k = 3$			$m = 8, k = 9$			$m = 16, k = 27$		
	Wass	NLL	Time per It.	Wass	NLL	Time per It.	Wass	NLL	Time per It.
W/M+Dadam	0.01 \pm 0.00	5.10 \pm 0.00	4.35ms	0.03 \pm 0.03	19.49 \pm 0.03	9.83ms	2.39 \pm 0.33	44.92 \pm 1.04	32.51ms
WO/M	0.03 \pm 0.00	5.12 \pm 0.00	4.23ms	4.26 \pm 3.61	22.27 \pm 1.69	9.64ms	5e3 \pm 589	Inf	29.77ms
EM	0.65 \pm 1.06	5.30 \pm 0.35	3.10ms	0.70 \pm 0.46	19.83 \pm 0.29	13.92ms	2.51 \pm 1.27	48.84 \pm 1.24	215.00ms

Table 4: Comparison performance of our, WO/M and IRA algorithms over the BSDS500 dataset via five metrics

	Gaussian Mixture			Logistic Mixture			Cauchy Mixture			Gamma Mixture		
	Our	WO/M	IRA	Our	WO/M	IRA	Our	WO/M	IRA	Our	WO/M	IRA
Wass	35.0 ± 4.68	89.5 ± 10.0	53.6 ± 21.9	45.8 ± 12.7	154 ± 10.9	68.0 ± 24.2	179 ± 15.3	211 ± 7.77	253 ± 65.4	35.5 ± 1.82	510 ± 71.2	—
NLL	11.8 ± 0.01	12.2 ± 0.11	11.8 ± 0.05	10.6 ± 0.02	12.5 ± 0.25	10.6 ± 0.06	12.1 ± 0.01	12.3 ± 0.01	12.0 ± 0.05	13.3 ± 0.07	19.1 ± 0.86	—
PSNR (dB)	18.9 ± 0.07	18.6 ± 0.18	18.3 ± 0.66	18.8 ± 0.11	18.2 ± 0.18	17.8 ± 0.67	19.3 ± 0.05	19.3 ± 0.06	18.4 ± 0.56	21.0 ± 0.08	18.0 ± 0.58	—
SSIM	0.68 ± 0.00	0.65 ± 0.01	0.66 ± 0.03	0.67 ± 0.01	0.62 ± 0.01	0.63 ± 0.03	0.70 ± 0.00	0.70 ± 0.00	0.69 ± 0.02	0.73 ± 0.00	0.59 ± 0.03	—
FailR	0.20%	53.1%	2.98%	0.46%	75.5%	2.68%	1.16%	0.82%	17.16%	0%	2.02%	100%

Parameter Settings and Metrics: We found the best learning rates α by searching from $\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3\}$. Similar to (Kingma and Ba 2014; Becigneul and Ganea 2019), $\beta_{11} = 0.9$ and $\beta_2 = 0.999$. The maximum number of iterations for testing the synthesis data was 2,000 and that for the image data was 10,000. The Wasserstein distance (simplified as *Wass*) and averaged negative log-likelihood (*NLL*) were mainly employed for comparison. The optimisation fail ratio (*FailR*) was also reported to show the stability of optimisation under initialisations. For BSDS500 data, we adopted two well-known metrics, namely, the peak signal-to-noise ratio (*PSNR*) and structural similarity index (*SSIM*) (Wang et al.

2004), to evaluate the quality of reconstructed images via maximising the posterior of optimised EMMs.

Baselines: Each main part of our method was assessed. The vanilla Riemannian gradient descent method on our manifold is denoted by *W/M + Vanilla*. The Riemannian Adam of (Becigneul and Ganea 2019) with our manifold is *W/M + Adam*. Our proposed *Dadam* is denoted by *W/M + Dadam* or *Our* interchangeably. We also performed comparisons with the trivial gradient descent over the sliced Wasserstein distance without the manifold (*WO/M*), which is the basics of (Kolouri, Rohde, and Hoffmann 2018). The EM-type methods were also compared, which are denoted as *EM* for GMMs and *IRA* for other EMMs (Kent and Tyler

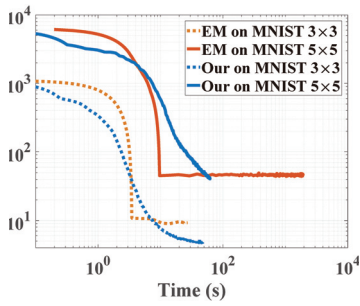


Figure 4: Wasserstein distance against the computational time for the *EM* and *Our*, on the downsampled 3×3 and 5×5 MNIST test data. The best learning rate was chosen as $\alpha = 0.1$. The cluster number was chosen as $k = 3$. The maximal iterations were set to 10,000 for both methods.

1991). All the methods were run on Matlab 2017a under Intel Core(TM) i7-6700 CPU, where the time was recorded.

Assessment over the *Dadam* on Synthesis Data

We first evaluated the effectiveness of our *Dadam* in adaptively accelerating the convergence. In this part, we only optimised the Σ_i whilst setting the π and μ_i as the ground truth of the synthesis data. Fig. 2 illustrates the convergence speed; observe that the optimisation on our established manifold exhibits much faster convergence. Furthermore, both the adaptive methods of (Becigneul and Ganea 2019) and the *Dadam* were shown to boost the speed. The proposed *Dadam* achieved the fastest convergence, which verifies the effectiveness of our *Dadam*.

Overall Comparisons on Synthesis Data

We next compared our algorithm (i.e., *W/M+Dadam*) with the *WO/M* and the *EM* methods under GMM problems. Similar results can be found for other EMMs; the result for GMMs is given in Table 3. It should be pointed out that all parameters (π , μ_i and Σ_i) were optimised in this test and each result was reported via the corresponding best learning rates. From this table, observe that our algorithm not only achieves an extremely stable estimation (the lowest deviations) but also the lowest values of both the Wasserstein distance and the NLL. An illustrative example can be found in Fig. 3, in which our algorithm consistently achieves the optimal clustering but the optimised GMMs via the *EM* are highly unstable. From Table 3, we should also point out that although our algorithm implements a set of manifold operations, the computational burden over that without manifold optimisation (i.e., *WO/M*) is in average around 5%. However, our algorithm enjoys much faster speed compared to the *EM* algorithm per iteration.

Overall Comparisons on the MNIST

We further compared the computational complexity of our algorithm (i.e., *W/M+Dadam*) with the *EM* method, besides the time per iteration reported in Table 3. Fig. 4 plots the decrease of Wasserstein loss with regard to the computational time when optimising GMMs and the trends for other EMMs

are the same. We can see clearly from this figure that our algorithm consistently achieves the lowest Wasserstein cost. The computational time of our algorithm is comparable to that of the *EM* algorithm when $m = 3 \times 3 = 9$; our algorithm, however, requires much less time when in higher dimensions such as $m = 5 \times 5 = 25$. We shall also need to point out that the random projections in our algorithm can be implemented parallelly, which can further accelerate the convergence speed of our algorithm.

Overall Comparisons on the BSDS500

Finally, we evaluated our algorithm on the BSDS500 dataset, with the results shown in Table 4. From this table, we see some improvements of using the Wasserstein distance instead of the KL divergence, by comparing the *WO/M* with the *IRA*. However, due to the probability constraint, the *WO/M* is highly unstable and its fail ratio reaches $> 50\%$ for the Gaussian mixture and Logistic mixture. More importantly, by optimising on our established manifold together with our *Dadam* method (denoted as *Our*), we can consistently achieve best performances over the five metrics. Again, our algorithm also achieve highly stable estimation, taking advantages from the Wasserstein distance, compared to the *IRA* which operates under the KL divergence. Another advantage of our algorithm is its universal convergence, while the *IRA* does not converge for the mixture of *Gamma*, which limits the flexibility of EMMs.

Conclusion

We have introduced a new and complete framework for solving general EMMs. To this end, we have first provided a unified and ready-to-use summary of the candidates for identifiable EMMs, which is achieved via a stochastic representation. Then, an approximate Wasserstein distance for EMMs has been proposed, which unlike the existing Wasserstein distances allows the corresponding metrics to be explicitly calculated. The so established manifold has been shown to consistently improve performances in terms of Wasserstein cost and even the NLL cost, and also significantly stabilise the optimisation on EMMs, making them robust to initialisations. We have further proposed a directional adaptively accelerated algorithm to enhance and stabilise the convergence, the performance of which has been validated through comprehensive experimental results.

References

- Absil, P.-A.; Mahony, R.; and Sepulchre, R. 2009. *Optimization algorithms on matrix manifolds*. Princeton University Press.
- Amari, S.-I. 1998. Natural gradient works efficiently in learning. *Neural Computation* 10(2):251–276.
- Andrews, D. F., and Mallows, C. L. 1974. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 99–102.
- Arbelaez, P.; Maire, M.; Fowlkes, C.; and Malik, J. 2011. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(5):898–916.
- Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Gan. *arXiv preprint arXiv:1701.07875*.

- Becigneul, G., and Ganea, O.-E. 2019. Riemannian adaptive optimization methods. In *Proceedings of the International Conference on Learning Representations*.
- Browne, R. P., and McNicholas, P. D. 2015. A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* 43(2):176–198.
- Chen, Y., and Li, W. 2018. Natural gradient in Wasserstein statistical manifold. *arXiv preprint arXiv:1805.08380*.
- Chen, Y.; Georgiou, T. T.; and Tannenbaum, A. 2018. Optimal transport for Gaussian mixture models. *IEEE Access* 7:6269–6278.
- Dasgupta, S. 1999. Learning mixtures of Gaussians. In *Proceedings of the 40th IEEE Annual Symposium on Foundations of Computer Science*, 634–644. IEEE.
- Fang, K. W.; Kotz, S.; and Ng, K. W. 1990. *Symmetric multivariate and related distributions*. London, U.K.: Chapman & Hall.
- Holzmänn, H.; Munk, A.; and Gneiting, T. 2006. Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics* 33(4):753–763.
- Hosseini, R., and Sra, S. 2015. Matrix manifold optimization for Gaussian mixtures. In *Advances in Neural Information Processing Systems*, 910–918.
- Jin, C.; Zhang, Y.; Balakrishnan, S.; Wainwright, M. J.; and Jordan, M. I. 2016. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, 4116–4124.
- Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6(2):181–214.
- Kent, J. T., and Tyler, D. E. 1991. Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics* 2102–2119.
- Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Knott, M., and Smith, C. S. 1984. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications* 43(1):39–49.
- Kolouri, S.; Park, S. R.; Thorpe, M.; Slepcev, D.; and Rohde, G. K. 2017. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine* 34(4):43–59.
- Kolouri, S.; Rohde, G. K.; and Hoffmann, H. 2018. Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3427–3436.
- LeCun, Y.; Cortes, C.; and Burges, C. 2010. Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist> 2.
- Lévy, B. 2015. A numerical algorithm for L2 semi-discrete optimal transport in 3d. *ESAIM: Mathematical Modelling and Numerical Analysis* 49(6):1693–1715.
- Li, S.; Yu, Z.; Xiang, M.; and Mandic, D. 2019. Solving general elliptical mixture models through an approximate wasserstein manifold. *arXiv preprint arXiv:1906.03700*.
- Malagò, L.; Montrucchio, L.; and Pistone, G. 2018. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry* 1(2):137–179.
- Martens, J. 2014. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*.
- Monge, G. 1781. *Mémoire sur la théorie des déblais et des remblais*. De l’Imprimerie Royale.
- Muzellec, B., and Cuturi, M. 2018. Generalizing point embeddings using the Wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems*, 10237–10248.
- Naim, I., and Gildea, D. 2012. Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients. *arXiv preprint arXiv:1206.6427*.
- Ollila, E.; Tyler, D. E.; Koivunen, V.; and Poor, H. V. 2012. Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on Signal Processing* 60(11):5597–5625.
- Peel, D., and McLachlan, G. J. 2000. Robust mixture modelling using the t distribution. *Statistics and Computing* 10(4):339–348.
- Rabin, J.; Peyré, G.; Delon, J.; and Bernot, M. 2011. Wasserstein barycenter and its application to texture mixing. In *Proceedings of the International Conference on Scale Space and Variational Methods in Computer Vision*, 435–446. Springer.
- Redner, R. A., and Walker, H. F. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26(2):195–239.
- Sra, S., and Hosseini, R. 2013. Geometric optimisation on positive definite matrices for elliptically contoured distributions. In *Advances in Neural Information Processing Systems*, 2562–2570.
- Sra, S., and Hosseini, R. 2015. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization* 25(1):713–739.
- Stefanski, L. A. 1991. A normal scale mixture representation of the logistic distribution. *Statistics & Probability Letters* 11(1):69–70.
- Sturm, K.-T., et al. 2006. On the geometry of metric measure spaces. *Acta Mathematica* 196(1):65–131.
- Takatsu, A., et al. 2011. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics* 48(4):1005–1026.
- Tan, S., and Jiao, L. 2007. Multivariate statistical models for image denoising in the wavelet domain. *International Journal of Computer Vision* 75(2):209–230.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4):600–612.
- Wang, W.; Wang, R.; Huang, Z.; Shan, S.; and Chen, X. 2015. Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2048–2057.
- Xu, L., and Jordan, M. I. 1996. On convergence properties of the EM algorithm for gaussian mixtures. *Neural Computation* 8(1):129–151.
- Xu, J.; Hsu, D. J.; and Maleki, A. 2016. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems*, 2676–2684.
- Zhang, H., and Sra, S. 2016. First-order methods for geodesically convex optimization. In *Proceedings of the Conference on Learning Theory*, 1617–1638.
- Zhang, T.; Wiesel, A.; and Greco, M. S. 2013. Multivariate generalized Gaussian distribution: Convexity and graphical models. *IEEE Transactions on Signal Processing* 61(16):4141–4148.