A unified framework for hard and soft clustering with regularized optimal transport

Jean-Frédéric Diebold
Safran Tech
Châteaufort, France

Nicolas Papadakis
IMB, UMR 5251
F-33400 Talence, France

Arnaud Dessein
Back Market
Bordeaux, France

Charles-Alban Deledalle
Brain Corp
San Diego, United States

Abstract—We formulate the inference of a finite mixture model from discrete data as an optimal transport problem with entropic regularization of parameter $\lambda \geq 0$. Our method unifies hard and soft clustering, the Expectation-Maximization (EM) algorithm being exactly recovered for $\lambda = 1$. The family of clustering algorithm we propose relies on the resolution of nonconvex problems using alternating minimization. We study the convergence property of our generalized $\lambda$--EM algorithms and show that each algorithmic step has a closed form solution when inferring finite mixture models of exponential families. Experiments highlight the benefits of taking a parameter $\lambda > 1$ to improve the inference performance and $\lambda \to 0$ for classification.

Index Terms—Statistical inference, Regularized optimal transport, Expectation-Maximization, K-means.

I. INTRODUCTION

The inference of a probabilistic parametric model from a set of realizations consists in estimating parameters allowing an explanation or even a generalization of these realizations. The inference of finite mixtures of Gaussian distributions, or more generally of exponential families, is thus widely used in data science task: video modeling [15], image segmentation [6], audio source separation [3] to name but a few. To estimate model parameters, a classical unsupervised approach is maximizing the likelihood, that is computed as the probability of realizations $\{x_1, \ldots, x_n\}$ along a model depending on parameters $\eta$. In many cases, such as for mixtures of exponential families, maximizing the likelihood with respect to the model parameters is not straightforward. Enriching the observations $x$ with hidden latent states $\pi$ allows to maximize the likelihood of the equivalent model in $(x, \pi)$ in a more convenient way. As an example, when considering a mixture of exponential families, if $\pi$ describes the belonging of realizations $x$ to their respective cluster, the inference of the law of each cluster (i.e. group of the mixture) can be done independently and admits a closed form expression.

Several hard and soft clustering algorithms rely on the complete information formulation in $(x, \pi)$ to iteratively infer parameters that simultaneously increase the likelihood of the observations $x$ with respect to the model. For instance, the k-means (KM) algorithm [21] assigns each realization $x_i$ to a single centroid (known as hard clustering) before updating its position. The Expectation-Maximization (EM) algorithm [10] is a soft clustering algorithm that distributes realizations $x_i$ to all clusters $j$ according to an a posteriori law $\pi_{ij}$. Such a soft strategy is meant to be more robust than a hard one, though both approaches can be competitive in practice depending on the task at hand and the dataset.

Contributions and outline.: The main contribution of this paper is to bridge the gap between hard and soft clustering by introducing a general formulation of statistical inference based on Regularized Optimal Transport (ROT). We cast the inference of finite mixture models as a ROT problem with an entropic regularization of parameter $\lambda \geq 0$. Solving the ROT problem amounts to minimizing a nonconvex functional. We propose to optimize this functional with an alternating minimization method. We show that each minimization step admits a closed form solution for exponential family models, and prove that accumulation points of the alternating algorithm are stationary points of the functional. We demonstrate that EM is recovered as a specific case of our algorithm for $\lambda = 1$ and that for $\lambda = 0$ we recover a hard clustering algorithm similar to k-means. Contrary to existing methods for statistical inference through optimal transport [23], the computational burden of our approach does not require a full OT resolution and the remains tractable with increased data dimension. We finally numerically illustrate the interest of (i) taking a parameter $\lambda > 1$ for the inference of 2D Gaussian Mixture Models and (ii) considering hard clustering with $\lambda \to 0$ for classification.

After a review of closely related works in section II, We introduce the regularized optimal transport problem for generalizing the EM algorithm in section III. In section IV, we analyze mathematically the convergence properties of the numerical algorithm. In section V, we specify the algorithm for mixtures of exponential families, with a focus on Gaussian Mixture Models (GMM). Finally, we provide in Section VI experiments that mainly illustrate the robustness properties brought by the regularization parameter in the context of GMM inference. We underline that this paper is based on unpublished preprint by the same authors [11].

II. RELATED WORK

Optimal Transport (OT) defines a family of distance between distributions. It consists in estimating a map transferring a source distribution onto a target one that minimizes a given cost of displacement. The OT distance (also known as the Wasserstein distance or the Earth Mover distance) has been shown to produce state of the art results for the comparison of
Gaussian Mixture Models (GMMs) \cite{7, 25} and generalized to exponential families in \cite{29}.

Many clustering algorithms are based on OT \cite{17, 22, 27, 30}. Since the theoretical work in \cite{2}, the statistical inference by minimizing OT distances has been increasingly studied \cite{4, 13, 20, 28}. In another line of works, Wasserstein Generative Adversarial Networks \cite{1} design generators that provide samples respecting the data distribution according to regularized or approximate OT distances \cite{5, 14, 18}.

On the other hand, there exists no general framework for the use of OT in “traditional” algorithms such as KM, or EM. In this paper, we fill this gap and study numerical algorithms considering statistical inference based on OT. We nevertheless mention the method of \cite{23} that learns GMMs by solving an OT problem. Notice that this competing approach is limited to cases where the weights of the mixture components are known and fixed. It also involves the resolution of a ROT problem on a discretized grid at each iteration, thus limiting the experiments to 1D examples or rough discretization. On the other hand, our algorithm admits closed form expressions for each step, and the computational complexity is only related to the number of clusters and independent to the data dimension.

III. PARAMETER ESTIMATION IN FINITE MIXTURE MODELS WITH REGULARIZED OPTIMAL TRANSPORT

In this section, we define our inference framework. We introduce the notations relative to clustering in section III-A. The formulation of statistical inference of a finite mixture model as a regularized optimal transport (ROT) problem is proposed in section III-B. This corresponds to estimating a low dimensional ROT map between sample points distribution and the weighted distribution of cluster parameters. In section III-C, we show that the problem can be recast as a simplified nonconvex functional to optimize.

A. Hard and soft Clustering

Hard and soft clustering methods suppose that we dispose of \( n \) observations \((x_i)_{i=1}^n\) from a space \( \mathcal{X} \) issued from a mixture of \( k \) components \((X_j)_{j=1}^k\) called clusters. The belonging of each point \( x_i \) to each cluster \( X_j \) is represented by a membership coupling \( \pi_{ij} \). The weights \( \omega_j \) of the clusters \( X_j \) are estimated through the proportion of points that are assigned to them. The clusters \( X_j \) are also represented by individual models \( p_{\eta_j} \). A generic scheme to optimize all variables is to update the memberships \( \pi_{ij} \), weights \( \omega_j \) and model parameters \( \eta_j \) in turn until convergence \cite{24}. The difference between hard and soft clustering lies in the computation of memberships \( \pi_{ij} \) for the assignment of points to the clusters. In hard clustering such as KM, the memberships \( \pi_{ij} \in \{0, 1\} \) are binary so that each point \( x_i \) belongs to exactly one cluster \( X_j \). The clustering thus produces a partition of the data. In soft clustering such as EM, the memberships \( \pi_{ij} \in [0, 1] \) are relaxed so that each point is softly affected to all clusters \( X_j \).

B. Finite Mixture Models and General Problem

Let \( \mathbb{P} = \{P_\eta\}_{\eta \in \Omega} \) be a parametric model. The finite mixture model with \( k \) components from \( \mathbb{P} \) is the model \( \mathcal{M} = \{P_\omega, \eta \in \Sigma_k, \eta \in \Gamma_k \} \) whose probability measures can be expressed as \( P_\omega, \eta = \sum_{j=1}^k \omega_j P_{\eta_j} \), where \( \eta = (\eta_j)_{j=1}^k \in \Gamma_k \) and \( \omega = (\omega_j)_{j=1}^k \in \Sigma_k := \{\omega \in \mathbb{R}_+^k : \sum_{j=1}^k \omega_j = 1\} \) are the parameters and weights of the \( k \) components.

Let \( x = (x_i)_{i=1}^n \in \mathcal{X}^n \) be composed of i.i.d. random variables that are distributed according to an unknown probability measure \( P_\omega, \eta \) from \( \mathcal{M} \). We call \( P_\omega, \eta \) the true distribution. We consider the inference of the weights \( \omega \in \Sigma_k \) and \( \eta \in \Gamma_k \) on the basis of the sample observations \( x \). The dataset \( x \) can be expressed as an empirical distribution \( p_\nu \) on \( \mathcal{X} \) made of the sum of Dirac masses located at points \( x_i \) with some arbitrary weights \( \nu = (\nu_i)_{i=1}^n \in \Sigma_n \) chosen by the user:

\[
p_\nu = \sum_{i=1}^n \nu_i \delta_{x_i}.
\]

The weights are typically taken uniform equal to \( 1/n \), though non-uniform weights \( \nu_i \) can also be used to put more or less emphasis on the respective observations \( x_i \). We then introduce a second measure for defining the OT problem. This measure is the weighted sum of Dirac localized along parameters \( \eta \):

\[
q_{\omega, \eta} = \sum_{j=1}^k \omega_j \delta_{\eta_j}.
\]

From the perspective of OT, we consider the transportation between the distribution \( p_\nu \) that is observed and the distribution of clusters \( q_{\omega, \eta} \) to fit. This gives a coupling between observations and source clusters. Introducing a cost matrix \( \gamma \in \mathbb{R}^{n \times k} \) between pairwise components and the convex and differentiable entropy function \( H : \pi \in \mathbb{R}^{n \times k} \rightarrow \sum_{ij} \pi_{ij} \log(\pi_{ij}) - 1 \in \mathbb{R} \) with penalty \( \lambda \geq 0 \) on transport plans, the ROT between the two distributions reads:

\[
d(p_\nu, q_{\omega, \eta}) = \inf_{\pi \in \Pi(\nu, \omega)} \left\{ \langle \gamma, \pi \rangle + \lambda H(\pi) \right\},
\]

where the transport polytope is defined by:

\[
\Pi(\nu, \omega) = \{\pi \in \mathbb{R}^{n \times k} : \pi 1_k = \nu, \pi^\top 1_n = \omega \}.
\]

This entropic ROT between discrete distributions is known as the Sinkhorn distance \cite{8}. We aim at identifying the coupling that maximizes the complete information log-likelihood in \( (x, \pi) \). To that end, we introduce the following cost matrix:

\[
\gamma_{ij}(\omega, \eta) = -\log(\omega_j p_{\eta_j}(x_i)).
\]

Intuitively, the more plausible the observation \( i \) is in relation to the component \( j \), the less expensive it is to transport the underlying mass. Notice that the cost matrix \( \gamma \) depends on the variables of the problem. The cost matrix will thus be updated during the optimization process. Estimating the mixture model amounts to finding the weights and parameters that minimize the total transportation cost \( d_\omega, \eta(p_\nu, q_{\omega, \eta}) \), that is:

\[
\inf_{\omega \in \Sigma_k} \inf_{\pi \in \Pi(\nu, \omega)} \lambda H(\pi) - \sum_{i=1}^n \sum_{j=1}^k \pi_{ij} \log(\omega_j p_{\eta_j}(x_i)).
\]
The objective (4) being nonconvex, reaching a global optimum is generally not tractable. We rather target a critical point with an algorithm that iteratively decreases the objective function.

C. Proposed model with Transport Plan Relaxation

Tackling problem (4) through alternate optimization on the different variables is computationally expensive. For fixed weights \( \omega \) and parameters \( \eta \), updating the transport plan \( \pi \) instead requires solving a full ROT problem, that gets costly as the optimization is repeated through iterations. We thus formulate the relaxed problem (5), where we replace the constraint \( \pi \in \Pi(\nu, \omega) \) with \( \pi \in \Pi(\nu) := \{ \pi \in \mathbb{R}^{n \times k}_+ : \mathbf{1}_k = \nu \} \):

\[
\inf_{\omega \in \mathbb{R}^k} \, \inf_{\pi \in \Pi(\nu)} \left( \sum_{i=1}^n \sum_{j=1}^k \pi_{ij} \log(\omega_j p_{ij}(x_i)) + \lambda H(\pi) \right). \tag{5}
\]

Thanks to this relaxation, we can avoid the computational burden associated to the resolution of the transport problem (4) w.r.t. \( \pi \in \Pi(\nu, \omega) \). As detailed in section IV, the expectation step of our algorithm (i.e. estimating \( \pi \)) is rather a simple closed form projection on \( \Pi(\nu) \). We now show that solutions of the relaxed problem (5) provide solutions for problem (4).

To that end, we first introduce the notion of coordinatewise minimum [26]. We say that \( x^* \in \mathbb{R}^n \) is a coordinatewise minimum of the nonconvex problem \( \min_{x \in \mathbb{R}^n} f(x_1, x_2, \ldots, x_n) \) if for all \( i = 1 \ldots n \), \( x^*_i = \arg\min_j f(x_1, x_2, \ldots, x_n) \).

**Proposition 1.** Any coordinatewise minimum of the relaxed problem (5) \((\hat{\pi}, \hat{\omega}, \hat{\eta})\) is an admissible solution and a coordinatewise minimum of problem (4).

The proof of this proposition is given in [12]. It first relies on the fact that the minimal value of the objective function in the relaxed problem is smaller than the one of the constrained problem (4). The proof then consists in showing that a minimizer of the relaxed problem necessarily belongs to set of constraints of the original problem (4).

IV. \( \lambda \)-EM algorithm

We propose to solve problem (5) using block coordinate descent. This gives an alternate optimization scheme on the transport plan \( \pi \), mixture weights \( \omega \) and mixture parameters \( \eta \). We present the optimization steps in section IV-A. In section IV-B, we analyze the convergence properties of the algorithm. In section IV-C we detail the connection of our framework with hard and soft clustering algorithms.

A. Alternate Optimization

We propose to target a local minima of the nonconvex problem (5) with an alternate procedure following the update order \((\pi, \omega, \eta)\).

**Transport plan.** For fixed weights \( \omega \in \Sigma_k \) and parameters \( \eta \in \Gamma^k \), the transport plan \( \pi \in \Pi(\nu) \) is updated by solving a relaxed ROT problem:

\[
\min_{\pi \in \Pi(\nu)} \, \sum_{i=1}^n \sum_{j=1}^k \pi_{ij} \log(\omega_j p_{ij}(x_i)) + \lambda H(\pi). \tag{6}
\]

For \( H(\pi) = \sum_{i,j} \pi_{ij} \log(\pi_{ij}) - 1 \) the entropic regularization, the global optimum of (6) for fixed \( \eta \) and \( \omega \) writes:

\[
\pi_{ij}^* = \frac{\nu_j \rho_j(x_i)^{1/\lambda}}{\sum_{l=1}^k (\omega_l p_{ij}(x_i))^{1/\lambda}}. \tag{7}
\]

**Weights.** For fixed parameters \( \eta \in \Gamma^k \) and transport plan \( \pi \in \Pi(\nu) \), we are now left out with the maximization of a concave objective: \( \max_{\omega \in \Sigma_k} \sum_{i=1}^n \sum_{j=1}^k \pi_{ij} \log(\omega_j) \). The solution is actually reached at \( \omega = \pi^1 \mathbf{1}_n \), i.e.

\[
\omega_j = \prod_{i=1}^n \pi_{ij} \tag{8}
\]

If a cluster is empty after a weight update, it is removed for the remaining steps of the algorithm.

**Parameters.** For fixed weights \( \omega \in \Sigma_k \) and transport plan \( \pi \in \Pi(\nu) \), if weights are non zero, the updates of parameters \( \eta \in \Gamma^k \) can be solved independently for each cluster:

\[
\eta_{ij} = \prod_{i=1}^n \pi_{ij} \log(p_{ij}(x_i)) \tag{9}
\]

In Sec. V, we explicit the closed form solution of this update step for the exponential family mixture model.

B. Convergence analysis

As the process alternates between minimization steps, the sequence of iterates may admit accumulation points. Using [16], we show that such accumulation points are necessarily stationary points of the nonconvex problem (5).

**Definition 1.** Let \( f(x_1, \ldots, x_n) \) be a differentiable function defined that may not be convex. A point \( \hat{x} = (\hat{x}_1, \ldots, \hat{x}_n) \) is a stationary point iff \( \forall y \in \mathbb{R}^n \), \( \nabla f(\hat{x}) \cdot y \geq 0 \).

**Theorem 1** (Proof in [12]). Let \((\hat{\pi}, \hat{\omega}, \hat{\eta})\) be an accumulation point of the algorithm given by the steps (7), (8) and (9). If (i) \( \exists B \subset \Gamma^k \) closed and convex so that all outputs from step (9) satisfy \( \eta \in B \), (ii) \( \forall (i,j) \), \( \log(p_{ij}(x_i)) \) is continuously differentiable along \( \eta \) in the neighborhood of \( \hat{\eta} \), and (iii) \( \pi_{ij} > 0, \omega_j > 0 \), then \((\hat{\pi}, \hat{\omega}, \hat{\eta})\) is a stationary point of the problem (5) in \( \Pi(\nu) \times \Sigma_k \times B \). Moreover, if \( \forall i,j - \log(p_{ij}(x_i)) \) is convex along \( \eta \), \((\hat{\pi}, \hat{\omega}, \hat{\eta})\) is a coordinatewise minimum of the function.

C. Hard and soft clustering; and beyond

Depending on the choice of the regularization parameter, the proposed framework interpolates between hard and soft clustering for \( \lambda \in [0, 1] \), and it recovers the maximum likelihood estimator (MLE) for \( \lambda \to +\infty \).

For \( \lambda = 0 \), the transport plan optimization (6) simplifies as \( \inf_{\pi \in \Pi(\nu)} - \sum_{i=1}^n \sum_{j=1}^k \pi_{ij} \log(\omega_j p_{ij}(x_i)) \). The constraint \( \pi \in \Pi(\nu) \) allows to solve the problem separately for each \( x_i \); if the value function \( \gamma_{ij} = - \log(\omega_j p_{ij}(x_i)) \) is higher for a cluster \( j = j^* \) compared to the others, the solution is attributing the point’s mass \( v_i \) to it: \( \pi_{ij} = v_i \) and \( \pi_{ij} = 0 \) for \( j \neq j^* \). The algorithm is thus a hard clustering.

In the case \( \lambda = 1 \), the transport plan (7) is the "a posteriori" distribution \( p(\pi | x, \omega, \eta) \) of the likelihood of a realization.
issued from the clusters. This is a soft clustering of data points \( x_i \) w.r.t. all components \( j \), that exactly corresponds to EM.

When \( \lambda = +\infty \), problem (6) is a maximum likelihood: the entropy is minimum for \( \omega_j = 1/k \) and all components have the same contribution in the obtained parametrization.

V. EXPONENTIAL FAMILY MIXTURE MODELS

In this section, we apply our algorithm to Exponential Family mixtures models. We show in section V-A that in this setting, all steps admit closed form solutions. The case of Gaussian Mixture Models (GMM) is detailed in section V-B.

A. Parameter Estimation for Exponential Family Components

We consider a mixture of minimal exponential family components with natural parameters \( \eta \) [19]:

\[
p_\eta(x) = h(x) \exp(T(x)^T \eta - A(\eta)) 
\]

where \( T \) is a transformation on \( x \) and the cumulant function \( A \) is strictly convex and differentiable. By introducing \( \mu = E[T(X)] \), there is a bijective mapping between \( \mu \) and \( \eta \) through \( \mu = \nabla A(\eta) \). Therefore, the law \( p_\eta \) can be parameterized along \( \mu \). In this setting, step (9) writes:

\[
\sup_{\eta_j} \sum_{i=1}^n \pi_{ij}(T(x_i)^T \eta_j - A(\eta_j)) 
\]

As parameters update follows weights update, empty clusters have been removed and \( \omega_j = \sum_{i=1}^n \pi_{ij} > 0 \). It corresponds to solving \( \sup_{\eta_j} \sum_{i=1}^n \frac{\pi_{ij}}{\omega_j} (T(x_i)^T \eta_j - A(\eta_j)) \), that involves a strictly concave function whose gradient vanishes for \( \mu_j = \nabla A(\eta_j) = \sum_{i=1}^n \frac{\pi_{ij}}{\omega_j} T(x_i) \). Then, if \( \mu_j = \sum_{i=1}^n \frac{\pi_{ij}}{\omega_j} T(x_i) \) belongs to the definition set of the family barycentric parameter, it solves the parameter optimization step (11). The whole inference process is given in Algorithm 1.

**Algorithm 1 Inference of Exponential Family Mixture Models**

Initialize \( \omega \) and \( \mu \). Do until convergence

- Expectation: \( \pi_{ij} = \frac{p_\eta(x_i)}{\sum_{i=1}^n p_\eta(x_i)^{1/\lambda}} \)
- Weight update: \( \omega_j = \sum_{i=1}^n \pi_{ij} \)
- Maximization: \( \mu_j = \sum_{i=1}^n \frac{\pi_{ij}}{\omega_j} T(x_i) \)

B. Case of Gaussian Mixture Models

Gaussian Mixture Model (GMM) is the most well-known model of the exponential family. For a 1D Gaussian of mean value \( \nu \) and variance \( \sigma^2 \), expression (10) with natural parameters is obtained with \( \eta = (\frac{\nu}{\sigma^2}, \frac{1}{2\sigma^2})^T \), \( T(x) = (x, x^2)^T \), \( h(x) = (2\pi)^{-1/2} \) and \( A(\eta) = -\frac{\nu^2}{4\sigma^2} - \frac{1}{2} \log(-2\eta_2) \), so as to recover \( p_\eta(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp(-\frac{(x-\nu)^2}{2\sigma^2}) \).

In higher dimensions, multivariate normal law parameterized with mean \( \nu \) and variance \( \Sigma \) remain within the exponential family by taking:

\[
\eta = (\Sigma^{-1}\nu, -\frac{1}{2}\Sigma^{-1})^T \quad T(x) = (x, xx^T)^T \quad \text{and} \quad A(\eta) = -\frac{1}{2} \eta_1^T \eta_2 - \frac{1}{2} \log | -2 \eta_2 |.
\]

After the maximization step (11), the update of the parameters \( (\nu_j, \Sigma_j) \) of the GMM components \( p_\eta \) are obtained at the expectation step using \( \mu_j = \sum_{i=1}^n \frac{\pi_{ij}}{\omega_j} T(x_i) = [\nu_j, \Sigma_j + \nu_j \nu_j^T] \), so that \( \nu_j = (\mu_j)_1 \) and \( \Sigma_j = (\mu_j)_2 - (\mu_j)_1 (\mu_j)_1^T \).

VI. EXPERIMENTS

A. Inference of GMMS

We first study the effect of the regularization parameter \( \lambda \) in the case of a 2D GMM (see [12] for 1D experiments). In order to quantify the inference quality, we compare the reference and inferred models with the \( MW_2 \) distance introduced in [9], that provides an upper bound of the true Wasserstein Distance between GMMs.

**Influence of \( \lambda \) and robustness to initialization.** For 20 random reference GMMs with 3 clusters, we sample \( n = 1000 \) points and infer GMMs with \( k = 5 \) components for different \( \lambda \) and 20 different random initializations for each experiment. As shown in Fig. 1, best accuracy (in terms of \( MW_2 \) distance) is reached for \( \lambda \approx 1.08 \), whereas performances deteriorate for \( \lambda < 0.9 \) or \( \lambda > 1.2 \). Taking \( \lambda = 1.1 \) improves both robustness (smaller standard deviation of the \( MW_2 \) distance) and accuracy compared to the EM algorithm \( \lambda = 1 \).

We provide in Figure 2 a qualitative illustration. For smaller values of \( \lambda \), clusters may concentrate on small group of points, at the expense of the global shape recovery of the reference GMM. Inference performance is increased in the setting \( \lambda = 1 \), which corresponds to the EM algorithm. Finally \( \lambda = 1.1 \) is the optimal value for the inference. It provides a more regular shape, closer to the reference distribution. With higher values \( \lambda > 1.5 \), the inferred distribution becomes close to the maximum likelihood estimator with only one cluster.

**Fig. 1.** Mean and variance of the inference accuracy (\( MW_2 \) distance) of 2D GMMs, for various regularization levels \( \lambda \).

![Fig. 1](image1.png)

**Fig. 2.** 2D Inferences for several values of \( \lambda \).

![Fig. 2](image2.png)

**Number of sampled points.** We consider the average inference performance over 80 reference GMMs with \( k = 4 \) clusters and a KM initialization, and we study the influence of the number of sampled points on the optimal \( \lambda \). Figure 3(top) highlights that, for a lower (resp. larger) number of samples the optimal regularization value is close to \( \lambda = 1.2 \) (resp.
We finally consider a clustering in $k = 4$ clusters. As shown in Figure 3(bottom), the best inference is obtained with $\lambda = 1.02$ and the true number of cluster $k = 4$. A small decrease of performance is observed for $k = 5$. In all cases, a parameter $\lambda \approx 1.05$ allows to be robust to the uncertainty of the number of clusters for the inferred GMM.

**B. Classification in higher dimension**

We finally present a classification experiment in higher dimension. We realize a clustering in $k = 16$ groups of 2000 samples of the latent space (dimension $n = 64$) of an autoencoder learnt on MNIST. We display in Figure 4 the accuracy of the corresponding clustering for $\lambda \in [0, 3]$. Best accuracy (0.666) is obtained with $\lambda \to 0$, compared to EM ($\lambda = 1$, accuracy 0.647, almost to identical to the scikit-learn implementation of EM that gives 0.645) or higher values (accuracy 0.605 for $\lambda = 3$). This suggests that hard clustering with $\lambda = 0$ is relevant for clustering purpose.

**VII. CONCLUSION**

In this paper we reformulate the parameter estimation problem in finite mixture models from the point of view of regularized OT. Considering Sinkhorn distance, we are able to recover standard algorithms such as Maximum Likelihood, EM or $k$-means as specific instances of our general $\lambda$-EM algorithm. Numerical results highlight the interest of taking a parameter $\lambda \approx 1.1$ to increase the robustness of the classical EM to initialization and to few data points, while considering hard clustering with $\lambda \to 0$ for classification purposes.

**REFERENCES**

[1] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML. pp. 214–223 (2017)
[2] Bassetti, F., Bodini, A., Regazzini, E.: On minimum kantorovich distance estimators. Statistics & Probability Letters 76(12), 1298 – 1302 (2006)
[3] Benaroya, L., Bimbot, F., Gribonval, R.: Audio source separation with a single sensor. IEEE Trans. on ASLP 14(1), 191–199 (2006)
[4] Blanchet, J., Kang, Y.: Sample out-of-sample inference based on wasserstein distance. Operations Research 69(3), 985–1013 (2021)
[5] Bouquuet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.J., Schoelkopf, B.: From optimal transport to generative modeling: the vegan cookbook. arXiv preprint arXiv:1705.07642 (2017)
[6] Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Trans. on PAMI 24(8), 1026–1038 (2002)
[7] Chen, Y., Georgiou, T.T., Tannenbaum, A.: Optimal transport for gaussian mixture models. IEEE Access 7, 6269–6278 (2018)
[8] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: NeurIPS. vol. 26 (2013)
[9] Delon, J., Desolneux, A.: A wasserstein-type distance in the space of gaussian mixture models. SIAM J. Imaging. Sci. 13(2), 936–970 (2020)
[10] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. J. of the royal statistical soc. series B 39(1), 1–22 (1977)
[11] Dessein, A., Papadakis, N., Deledalle, C.A.: Parameter estimation in finite mixture models by regularized optimal transport: A unified framework for hard and soft clustering. preprint arXiv:1711.04366 (2017)
[12] Diebold, J.F., Papadakis, N., Dessein, A., Deledalle, C.A.: A unified framework for hard and soft clustering with regularized optimal transport. preprint arXiv:1711.04366v2 (2024)
[13] Frogner, C., Poggio, T.: Approximate inference with wasserstein gradient flows. In: Int. Conf. Artif. Intel. and Stat. pp. 2581–2590 (2020)
[14] Genevay, A., Peyré, G., Cuturi, M.: Learning generative models with sinkhorn divergences. In: AISTATS. pp. 1608–1617 (2018)
[15] Greenspan, H., Goldberger, J., Mayer, A.: Probabilistic space-time video modeling via piecewise gmm. IEEE Trans. Pat. Analys. and Mach. Intel. 26(3), 384–396 (2004)
[16] Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear gauss–seidel method under convex constraints. Operations res. let. 26(3), 127–136 (2000)
[17] Ho, N., Nguyen, X., Yurochkin, M., Bui, H.H., Huynh, V., Phung, D.: Multilevel clustering via wasserstein means. arXiv preprint arXiv:1706.03883 (2017)
[18] Houard, A., Leclaire, A., Papadakis, N., Robin, J.: On the gradient formula for learning generative models with regularized optimal transport costs. Transactions on Machine Learning Research (2022)
[19] Jordan, M.: Exponential families : basics. Berkeley course notes (2009)
[20] Lambert, M., Chewi, S., Bach, F., Bonnabel, S., Rigollet, P.: Variational inference via wasserstein gradient flows. In: NeurIPS. vol. 35 (2022)
[21] Lloyd, S.P.: Least square quantization in PCM. IEEE Trans. on Information Theory 28(2), 129–137 (1982)
[22] Marti, G., Andler, S., Nielsen, F., Donnat, P.: Optimal transport vs. fisher-rao distance between copulas for clustering multivariate time series. In: IEEE Statistical Signal Proces. Workshop. pp. 1–5 (2016)
[23] Mena, G., Nejatbakhsh, A., Varol, E., Niles-Weed, J.: Sinkhorn em: an expectation-maximization algorithm based on entropic optimal transport. arXiv preprint arXiv:2006.16548 (2020)
[24] Nock, R., Nielsen, F.: On weighting clustering. IEEE Trans. Pat. Analys. and Mach. Intel. 28(8), 1223–1235 (2006)
[25] Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover’s distance as a metric for image retrieval. Int. J. of Computer Vis. 40(2), 99–121 (2000)
[26] Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. of Optim. Theory and Applic. 109, 475–494 (2001)
[27] Ye, J., Wu, P., Wang, J.Z., Li, J.: Fast discrete distribution clustering using Wasserstein barycenter with sparse support. IEEE Trans. Signal Process. 65(9), 2317–2332 (2017)
[28] Yi, M., Liu, S.: Sliced wasserstein variational inference. In: ACML. pp. 11382–11395 (2022)