Unsupervised Learning of Finite Mixture Models using Mean Field Games

Sergio Pequito [†◇], A. Pedro Aguiar [†], Bruno Sinopoli [◇], Diogo A. Gomes[♯]

[†] Department of Electrical and Computer Engineering
Institute for System and Robotics
Instituto Superior Tecnico
Technical University of Lisbon
Lisbon, Portugal
spequito@isr.ist.utl.pt,pedro@isr.ist.utl.pt

[◇]Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
brunos@ece.cmu.edu

[♯]Department of Mathematics
Instituto Superior Tecnico
Technical University of Lisbon
Lisbon, Portugal
dgomes@math.ist.utl.pt

***Abstract*—In this paper we develop a dynamic continuous solution to the clustering problem of data characterized by a mixture of K distributions, where K is given *a priori*. The proposed solution resorts to game theory tools, in particular mean field games and can be interpreted as the continuous version of a generalized Expectation-Maximization (GEM) algorithm. The main contributions of this paper are twofold: first, we prove that the proposed solution is a GEM algorithm; second, we derive closed-form solution for a Gaussian mixture model and show that the proposed algorithm converges exponentially fast to a maximum of the log-likelihood function, improving significantly over the state of the art. We conclude the paper by presenting simulation results for the Gaussian case that indicate better performance of the proposed algorithm in term of speed of convergence and with respect to the overlap problem.**

I. Introduction

Finite mixtures are a flexible and powerful probabilistic modeling tool for univariate and multivariate data. Statistical modeling of data, such as pattern recognition, computer vision, signal and image analysis, machine learning, system identification and estimation, constitutes a large of class of applications. Finite mixtures describe data generated by a convex combination of probability density functions (pdf). In statistical pattern recognition, finite mixtures allow a formal probabilistic model-based approach to unsupervised learning, i.e, clustering [2]. Finite mixtures naturally model observations which are assumed to have been produced by one (randomly selected and unknown) of a set of alternative random sources. Inferring (the parameters of) these sources and identifying which source produced each observation leads to a clustering problem. The usefulness of mixture models is not limited to unsupervised learning applications. Mixture models are also able to represent arbitrarily complex pdf. This makes them an excellent choice for representing complex class-conditional pdfs (i.e, likelihood functions) in Bayesian supervised learning scenarios or priors for Bayesian parameter estimation [2].

A. Standard method

Expectation-maximization (EM) is a standard algorithm [1] and one of the fundamental tools in unsupervised learning. Briefly, EM is a technique that allows to incrementally increase the maximum likelihood (ML) and converges to a maximum likelihood estimate of the mixture model's parameters. Let the probability be represented by $p(x|\Theta)$, assumed to be fully characterized by a set of parameters Θ. Let $\mathcal{X} = \left\{x_{(1)}, ..., x_{(N)}\right\}$ be a known data set of size N, drawn from this distribution, also referred as *incomplete data*. Consider that a *complete data* set exists $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$, where $\mathcal{Y}$ is the missing information, random and presumably governed by an underlying distribution. The complete data is specified by the following joint density function

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta),$$

called the *complete-data likelihood*. This function is random, due to $\mathcal{Y}$. The EM method is composed of two steps: Expectation (E-Step) and Maximization (M-Step). First, we find the expected value of the complete-data likelihood $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$ with respect to the unknown data $\mathcal{Y}$ given the observed data $\mathcal{X}$ and the current estimate $\Theta^{(i-1)}$ (E-Step). Next the following functional

$$Q(\Theta, \Theta^{(i-1)}) = E\left[\log p(\mathcal{X}, \mathcal{Y}|\Theta)|\mathcal{X}, \Theta^{(i-1)}\right], \qquad (1)$$

is maximized with respect to Θ. This maximization represents the M-Step, and defines the new estimate of parameters as follows

$$\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)}). \qquad (2)$$

These two steps are repeated as necessary and each iteration is guaranteed to increase the log-likelihood until the algorithm converges to a local maximum of the likelihood function. Instead of maximizing $Q(\Theta, \Theta^{(i-1)})$ one may be interested into modifying the M-Step to find $\Theta^{(i)}$ such that $Q(\Theta^{(i)}, \Theta^{(i-1)}) > Q(\Theta^{(i-1)}, \Theta^{(i-1)})$. This class of algorithms is known as generalized EM (GEM), and shares the same properties as EM and can be used an alternative to it. Despite being a powerful tool, EM has several drawbacks [3], [4]. The degeneracy problem and the selection of number of components are properly dealt with Bayesian inference and variational methods [2], [5]. The convergence rate can be sped up using Gauss-Newton methods [4], with some extra computation cost. In addition EM does not apply to non-parametric distributions. Finally, the interdependency between the two steps leads to discontinuities in the optimization and a low convergence rate.

B. Problem statement

In this paper we take a dynamic system point of view to perform maximum likelihood clustering adopting tools from game and control theory. We are concerned with unsupervised learning of a mixture of distributions. The main assumption is that the data set $\mathcal{X} = \{x_{(1)}, ..., x_{(N)}\}$ is generated by a mixture model where number of mixtures $K \in \mathbb{N}$ is known *a priori*, that is

$$p(x) = \sum_{k=1}^{K} \alpha_k p_k(x), \quad \sum_{k=1}^{K} \alpha_k = 1 \tag{3}$$

where $\alpha_k > 0$ is the mixture coefficient or weight relative to the distribution $p_k(x)$. This distribution is assumed to be greater than zero almost everywhere, but not necessarily characterized by a finite number of parameters. The goal is to maximize the log-likelihood of the probability of the mixture $p(x)$ given the data $\mathcal{X}$:

$$\begin{aligned} &\mathcal{L}(\alpha_1^{\xi}(t), ..., \alpha_K^{\xi}(t), p_1^{\xi}(t,\xi), ..., p_K^{\xi}(t,\xi)) = \\ &\qquad \sum_{k=1}^{K} \sum_{i=1}^{N} \log\left(\alpha_k^{\xi}(t) p_k^{\xi}(t, x_{(i)})\right) p_k^{x}(t, x_{(i)}). \end{aligned} \tag{4}$$

This is closely related to (1) in the case where $p_k^{\xi}(t,\xi)$ is parameterized and $\mathcal{L}$ does not depend upon time. In this case we obtain the same expression as in [eq. 5,19]. In (4), α_k^{ξ} are the mixing coefficients of the solution that maximizes (4), $p_k^{\xi}(t,\xi)$ is a distribution driven by the dynamics of a system to be introduced in the next section and finally $p_k^{x}(t,\xi)$ is the estimated distribution of the data at time k.

C. Proposed solution

In order to cope with the interdependency mentioned above, we use mean field games (MFG), introduced and developed by Lions and Lasry in two seminal papers [6], [7] and independently by P. Caines ([20] and references therein), to understand the economical/social behavior of a large number of players. In our case the MFG framework is especially suitable to model populations of agents that are characterized by a probability distribution $p_k^{\xi}(t,\xi)$ that evolves in time according to the following dynamics of $\xi_k(t) \in \mathbb{R}$:

$$\begin{aligned} d\xi_k &= u_k dt + \epsilon dw_t \\ \xi_k(0) &= \xi_k^0. \end{aligned} \tag{5}$$

In (5), ξ_k represents the state of each agent $\xi_k(t)$ in the population $k \in \{1, ..., K\}$. The initial condition is a random variable with distribution given by $p_k^{\xi}(0,.)$, w_t is a Wiener process with volatility $\epsilon > 0$ that codifies possible errors of the evolution and error sparseness and $u_k : [0 : \infty) \to \mathbb{R}^n$ can be viewed as a *decision* of each player (terminology from game theory) [18] or a *control* (terminology in control theory) [13] to be designed.

The main idea to solve the finite mixture model problem stated in Section I-B is to use (5), where u_k is chosen as the solution to the following finite-time optimal control problem

$$J^k(t,\xi) = \min_{u_{k:[t,T]}} \mathbb{E}\left[\int_t^T \|u_k(\tau,\xi)\|^2 + \log\left(\frac{p_k^{\xi}(\tau,\xi)}{p_k^{x}(\tau,\xi)}\right) d\tau\right], \tag{6}$$

where $0 \leq t \leq T$ and $T \in (0, \infty]$. In (6) the expectation is taken with respect to ξ_k. The running cost is composed of two terms: the first one penalizes the energy of the control of each agent in the population while the second is related to the Kullback-Leibler divergence between the underlying distribution that represents each population $p_k^{\xi}(t,\xi)$ and the estimated distribution $p_k^{x}(t,\xi)$ to be computed using the data set $\mathcal{X}$. For example, for a distribution characterized by a finite set of parameters, p_k^{x} can be computed by maximizing the likelihood of the data with respect to p_k^{ξ}, as shown by the simulation results. The rationale for the selection of this cost (6) is that it forces the system (5) and its underlying distribution p_{ξ} to approach p_x. This is enforced by the minimization of the Kullback-Leiber divergence $D_{\mathcal{KL}}(p||q) = \mathbb{E}\left[\log\left(\frac{p}{q}\right)\right]$ where the expectation is taken with respect to p.

The cost (6) and p_x depend on the evolution of p_{ξ}, which in turn depends on the cost due to the control in (5). This coupling is analogous to the interdependency between the E and the M steps.

At this point, we introduce the behavior/dynamics of each ξ_k for all the $k = 1, ..., K$ populations, and consequently the behavior of p_k^{ξ}. These populations are weakly coupled through the mixing coefficients $\alpha_k(t)$ that depend on time and can be seen in a probabilistic sense as reputation rates given by

$$\alpha_k(t) = \frac{1}{N} \sum_{i=1}^{N} p_k^{\xi}(t, x_{(i)}), \qquad \sum_{k=1}^{K} \alpha_k(t) = 1, \tag{7}$$

where $x_{(i)} \in \mathcal{X}$ and N is the number of elements of $\mathcal{X}$. This choice of the mixing coefficients α_ks both resembles the probabilistic interpretation [2] of the number of samples $x_{(i)}$ that are most likely to belong to a class k and it also maximizes the likelihood of the mixing coefficients with respect to the data.

The characterization of the solution using the MFG approach requires the simultaneous solution of two partial differential equations: a Hamilton-Jacobi-Bellman (HJB) equation characterizes the evolution of the control [12] (the M-Step), and a Fokker-Planck (FP) equation drives the evolution of the population p_k^ξ through (5) (the E-step). In this way we cope with the interdependency, since (6) depends on the evolution of p_k^ξ and the FP equation through the control presented in (5) that solve the problem in (6). The coupled HBJ and FP equations constitute the basis of the MFG framework.

The proposed approach provides a dynamical system interpretation of GEM, as it can be seen as a continuous version of GEM.

The main contributions of this paper are twofold: we first prove that the proposed solution is a GEM algorithm; we then derive a closed-form solution for Gaussian mixture model and show that the proposed algorithm converges exponentially fast to a maximum of the log-likelihood function. We conclude the paper by validating the algorithm on artificial data and showing some simulation results relative to the overlap problem.

The remainder of this article is organized as follows: In Section II we present in detail the equations that describe the MFG. In Section III we derive the dynamic expression for the mean field game and present the proof that the proposed solution is a GEM. In Section IV we provide closed form solutions for the case of Gaussian Mixture Models (GMM) and provide the proof of exponential convergence and we enunciate some stability results. Finally in Section V we provide simulation results on artificial data, and analyze the worst case scenario using stationary assumptions for the GMM case as well as some promising results concerning the *overlap problem* [8].

II. Mean Field Games

Mean field games (MFG) is a mathematical framework developed by Larsy and Lions [6], [7], and investigated by Caines [20], that is suitable to model and analyze the behavior of games among N agents [21], when $N \to \infty$. Although It uses a Hamilton-Jacobi equation [12], MFG is more general as it allows the cost function of each agent to also depend on the probability density p_ξ of all the other agents. Such generalization can be achieved adding a term to the cost function as in (6), that captures the influence of the collective, in our case the Kullback-Leiber divergence. From (6) we obtain the following Hamilton-Jacobi equation for each population k

$$J_t^k\left(t,\xi\right) + \left|J_\xi^k\left(t,\xi\right)\right|^2 + \frac{\epsilon^2}{2}\Delta J^k\left(t,\xi\right) - \log\left(\frac{p_k^\xi\left(t,\xi\right)}{p_k^x\left(t,\xi\right)}\right) = 0 \tag{8}$$

with terminal condition given by $J(T,.) = 0$, where J_t and J_ξ denote the partial derivative of J w.r.t. t and ξ, respectively. The Euclidean norm is given by $|.|$ and $\Delta = \frac{\partial^2}{\partial\xi_1^2} + ... + \frac{\partial^2}{\partial\xi_n^2}$ represents the Laplacian, where ξ_i is the i-th entry of the vector $\xi \in \mathbb{R}^n$. This last equation is coupled with a Fokker-Planck equation that drives the behavior of p_k^ξ according to the dynamics (5). The Fokker-Planck equation is given by

$$\left(p_k^\xi\right)_t - \operatorname{div}\left(p_k^\xi u_k\left(t,\xi\right)\right) = \frac{\epsilon^2}{2}\Delta p_k^\xi \tag{9}$$

with initial condition $p_k^\xi(0,.) = p_k^{\xi_0}(.)$, where $\operatorname{div}(f) = \left[\frac{\partial f_1}{\partial\xi_1} \; \cdots \; \frac{\partial f_n}{\partial\xi_n}\right]$ and $u_k = -J_\xi^k(t,\xi)$ is a function that depends on the evolution of the cost function J^k. The system of coupled equations given by (8)-(9) is a MFG. The main difficulty of working with these equations arise from the fact that they both depend on each other and (8) evolves backward in time, whereas (9) evolves forward in time.

In the Appendix we provide more details about these equations, giving a particular emphasis on (9) since it explains why the controlled evolution (i.e. introducing the control u_k in (9)) drives the underlying distribution p^ξ toward p^x, and why this increases the likelihood.

III. Unsupervised Learning of Finite Mixture Models using Mean Field Games

In this section we show that the MFG approach is indeed a GEM. Consider the game where each player acts according to (5) and (6). The expected optimal behavior can be written as

$$\hat{\xi}_k(t) = \arg\min_{z\in\mathbb{R}^n} J^k(t,z) \qquad k = 1,\cdots,K \tag{10}$$

It turns out that we can rewrite (10) as an ordinary differential equation in an explicit form by following the steps in Krener [23] under the assumption that $J^k(t,z)$ is a smooth invertible solution and $\hat{\xi}_k$ is differentiable. In this case J^k has to satisfy the first order optimality condition

$$0 = J_\xi^k\left(t,\hat{\xi}_k\left(t\right)\right). \tag{11}$$

Since $\hat{\xi}_k$ is assumed differentiable at t, we have

$$\begin{aligned}\tfrac{d}{dt}J^k\left(t,\hat{\xi}_k\left(t\right)\right) &= J_t^k\left(t,\hat{\xi}_k\left(t\right)\right) + J_\xi^k\left(t,\hat{\xi}_k\left(t\right)\right)\dot{\hat{\xi}}_k\left(t\right)\\ &= J_t^k\left(t,\hat{\xi}_k\left(t\right)\right).\end{aligned}$$

Differentiating (11) with respect to t yields

$$J_{\xi\xi}^k(t,\hat{\xi}_k)\dot{\hat{\xi}}_k(t) + J_{\xi t}^k(t,\hat{\xi}_k) = 0. \tag{12}$$

Now, take the derivative of (8) w.r.t. ξ, along $\hat{\xi}_k(t)$ and use (11) to obtain

$$J_{\hat{\xi}t}^k\left(t,\hat{\xi}_k\right) + \left(\frac{\epsilon^2}{2}\Delta J^k\left(t,\hat{\xi}_k\right) - \log\left(\frac{p_k^\xi\left(t,\hat{\xi}_k\right)}{p_k^x\left(t,\hat{\xi}_k\right)}\right)\right)_\xi = 0. \tag{13}$$

Replacing (12) in (13), multiplying by $\left(J_{\xi\xi}^k(t,\hat{\xi}_k)\right)^{-1}$, and rearranging the terms we get

$$\dot{\hat{\xi}}_k(t) = \left(J_{\xi\xi}^k(t,\hat{\xi}_k)\right)^{-1}\left(\frac{\epsilon^2}{2}\Delta J^k\left(t,\hat{\xi}_k\right) - \log\left(\frac{p_k^\xi\left(t,\hat{\xi}_k\right)}{p_k^x\left(t,\hat{\xi}_k\right)}\right)\right)_\xi. \tag{14}$$

The system of equations composed by (14) and (8)-(9) provides the optimal population dynamics. In the next result we show that the maximum likelihood (4) is increasing.

Theorem 1. *Let $\mathcal{X} = \left\{x_{(1)}, ..., x_{(N)}\right\}$ be a data set generated by the mixture model* (3)*. Consider a mixture*

$$p^\xi(t,\xi) = \sum_{k=1}^{K} \alpha_k^\xi(t) p_k^\xi(t,\xi) \tag{15}$$

where $p_k^\xi(t,\xi)$ is greater than zero almost everywhere for all t and evolves accordingly to (5)-(6)*, and α_k is given by* (7)*. Then $p^\xi(t,\xi)$ continuously increases the likelihood* (4) *with respect to the distribution of $\mathcal{X}$ given by*

$$p^x(t,\xi) = \sum_{k=1}^{K} \alpha_k^x(t) p_k^x(t,\xi). \tag{16}$$

Proof: The proof is divided in two parts: first, we show that the likelihood $\mathcal{L}$ increases as the mixture coefficients α_k evolve, and second that the optimal control in (5) increases the likelihood (4) of each density function p_k^ξ w.r.t. p_k^x.

1) By realizing that (15) is a mixture with parameters $\Theta = (\alpha_1, ..., \alpha_K)$ and performing the same steps as in [Section 3, 19], it can be shown that the maximum likelihood of the mixture coefficients is given by (7) at each t.
2) Increasing the maximum likelihood of (4) implies that $p_k^\xi \overset{prob}{\to} p_k^x$. Also the maximum likelihood for each p_k^ξ with respect to p_k^x is not dependent on the remaining $p_j^\xi, j \neq k$. The increase of the likelihood follows from the fact that (6) has to satisfy (8), that is derived by dynamic programming optimality principle [22]. In matter of fact, take $T > t$ sufficiently large and $\epsilon > 0$, then (6)

$$\begin{aligned} & J^k\left(t,\xi(t)\right) \\ & = \underbrace{\min_{u_{k:[t,t+\varepsilon]}} \mathbb{E}\left[\int_t^{t+\varepsilon} \left\|u_k\left(\tau,\xi\right)\right\|^2 + \log\left(\frac{p_k^\xi\left(\tau,\xi\right)}{p_k^x\left(\tau,\xi\right)}\right) d\tau\right]}_{(*)} \\ & + \underbrace{\min_{u_{k:[t+\varepsilon,T]}} \mathbb{E}\left[\int_{t+\varepsilon}^{T} \left\|u_k\left(\tau,\xi\right)\right\|^2 + \log\left(\frac{p_k^\xi\left(\tau,\xi\right)}{p_k^x\left(\tau,\xi\right)}\right) d\tau\right]}_{J^k(t+\epsilon,\xi(t+\epsilon))} \end{aligned}$$

where $(*)$ is greater or equal then zero since $\mathbb{E}\left[\int_t^T \|u_k(\tau)\|^2 d\tau \geq 0\right]$ and $\mathbb{E}\left[\int_t^T \log\left(\frac{p_k^\xi(\tau,\xi(\tau))}{p_k^x(\tau,\xi(\tau))}\right) d\tau\right] \geq 0$ (Lemma 2 - proven in Appendix). This way $p_k^\xi \overset{prob}{\to} p_k^x$ due to the fact that $(*) \geq 0$ or because we are imposing $J(T,.) = 0$ as terminal condition.

Since the likelihood (4) is maximized with respect to the mixing coefficients α_k and to the distributions p_k^ξ, it maximizes the likelihood of the mixture presented in (15), concluding the proof. ■

Up to this point, we have shown that we have in the worst case scenario an algorithm that has the same properties as any GEM. In the next section we analyze the case where the data set is generated by a Gaussian mixture model (GMM), and show that the unsupervised learning using MFG for GMM converges exponentially fast to a maximum of the likelihood function.

IV. The Gaussian Mixture Model case

In this section we assume that the mixture model (3) is a convex combination of Gaussians, i.e,

$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mu_k, \Sigma_k), \qquad \sum_{k=1}^{K} \alpha_k = 1$$

where $\mathcal{N}(\mu_k, \Sigma_k)$ is a Gaussian distribution with mean $\mu_k \in \mathbb{R}^n$ and covariance $\Sigma_k \in \mathbb{R}^{n \times n}$. The following lemmas will be useful to characterize the proposed solution.

Lemma 1. *Suppose that we have a mixture* (15)*, where p_k^ξ is driven by* (5)-(6)*, such that:*

i) *$p_k^\xi(0,\xi) \sim \mathcal{N}(\mu_k^\xi(0), \Sigma_k^\xi(0))$ with $\Sigma_k^\xi(0) > 0$ for all k,*
ii) *$p_k^x(t,\xi) \sim \mathcal{N}(\mu_k^x(t), \Sigma_k^x(t))$ with $\Sigma_k^x(t) > 0$ for all k and $t \geq 0$.*

Then $p_k^\xi(t,\xi) \sim \mathcal{N}(\mu_k^\xi(t), \Sigma_k^\xi(t))$ with $\Sigma_k^\xi(t) > 0$ for all k and $t \geq 0$. Moreover the evolution of the mean $\mu^\xi(t)$, $\forall t > 0$ is given by

$$\begin{aligned} & \dot{\hat{\xi}}_k(t) = \left(P^k\right)^{-1} \left(\Sigma_k^x(t)\right)^{-1} \left(\hat{\xi}_k(t) - \mu_k^x(t)\right), \\ & \hat{\xi}(0) = \mu_\xi(0). \end{aligned} \tag{17}$$

where $P^k(t) = J_{\xi\xi}^k(t,\hat{\xi}_k)$ has to satisfy the following Riccati equation

$$\dot{P}^k + 4(P^k)^T P^k + \frac{1}{2}\left(\Sigma_k^\xi\right)^{-1} - \frac{1}{2}\left(\Sigma_k^x\right)^{-1} = 0. \tag{18}$$

□

Proof: The proof of this lemma consists in the following:

1) $p_k^\xi(t,\xi) \sim \mathcal{N}(\mu_k^\xi(t), \Sigma_k^\xi(t))$ with $\Sigma_k^\xi(t) > 0$ for all k and $t \geq 0$ if the initial condition $p_k^\xi(0,\xi)$ for (9) is Gaussian (provided by i)) and u_k in (9) is a linear function.
2) u_k in (9) is a linear function if J^k is quadratic, which holds replacing ii) in the cost function J^k and imposing that p_k^ξ is Gaussian. In other words, the solution of the FP equation is Gaussian if and only if the solution of the HJ equation is quadratic [24].

Replacing i)-ii) in (14) yields (17). The Riccati equation follows from assuming $J_k^\xi = \xi^T P \xi$ and replacing it in (8). ■

The following result adapted from [17] is needed.

Theorem 2 (Theorem 12.1 [17])**.** *Given a system*

$$d\eta = (A\eta + f(\eta,u))dt + \epsilon dw_t, \quad \eta(0) = \eta^0 \tag{19}$$

where $\eta \in \mathbb{R}^n$, $A \in \mathbb{R}^n \times \mathbb{R}^n$, $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ and ϵ, u, w_t as in (5) *and an output function*

$$\zeta = C\eta + h(u) \tag{20}$$

where $C \in \mathbb{R}^n \times \mathbb{R}^n$, $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. If

- (19)-(20) *is uniformly observable [17] for any input u*
- *f and h are Lipschitz and with bounded second derivative with respect to η*
- *$\hat{\eta}$ is the extended Kalman filter solution to the system* (19)-(20)
- *$|\eta(0) - \hat{\eta}(0)|$ is sufficient small,where $\eta(0)$ is the initial condition of* (19) *and $\hat{\eta}(0)$ is the initial condition of the extended Kalman filter.*

Then $|\eta(t) - \hat{\eta}(t)| \to 0$ exponentially as $t \to \infty$. □

We can now state the main result of this paper

Theorem 3. *Under the same assumptions as Theorem 1, for a Gaussian mixture model the proposed solution composed by the system* (5)-(6) *and* (7) *converges exponentially fast to a maximum of the likelihood function.* □

Proof: This proof has two steps: First we show that we can rewrite (5) as (19) and that (17) is the extended Kalman filter for (5), if the conditions of Theorem 2 hold for each k of the mixture (15). Second, we show that $\alpha_k(t)$ depend upon p_k^ξ which is Lipchitz w.r.t. ξ_k, and so for $\epsilon > 0$ we have that $|\alpha_k(t+\epsilon) - \alpha_k(t)| \to 0$ as $t \to \infty$ exponentially fast due to the first step. Concerning the first step, we need to check all the conditions of Theorem 2. We can rewrite (5) as

$$d\xi_k = (\xi_k + f(\xi_k, u_k))dt + \epsilon dw_t, \tag{21}$$

and

$$y_k = \xi_k \tag{22}$$

where $f(\xi_k, u_k) = u_k - \xi_k$. Equation (21) is equivalent to (19) with $A = I_n$ where $I_n \in \mathbb{R}^n \times \mathbb{R}^n$ is the identity matrix. Also equation (22) is equivalent to equation (20) with $C = I_n$ and $h(u_k) = 0$. The system written in this form is uniformly observable, proving the first condition. f is Lipschitz, since $u_k(\xi_k)$ is linear in ξ_k and so is h as it is constant. Let us now focus on the remaining two conditions. Equation (17) can be interpreted as the Kalman filter for system (21) (22) proving the third condition. Finally, $\hat{\xi}_k(0)$ can be made arbitrary small, we define the initial conditions for both systems (21) and (17). From Theorem 2 we can conclude that $\hat{\xi}_k \to \xi_k$ exponentially fast, and from Theorem 1 since the method is a GEM, its properties holds, in particular it converges to a local maximum of the likelihood function. In resume, it converges exponentially fast for a local maximum of the likelihood. We now need to check that $|\alpha(t+\epsilon) - \alpha(t)| \to 0$ as $t \to \infty$ for $\epsilon > 0$. By definition (see (7)) the coefficients α_ks are linear combinations of Gaussian distributions and therefore are Lipschitz with respect to ξ and bounded with respect to t. As a consequence

$$|p_k^\xi(t+\epsilon, \xi(t+\epsilon)) - p_k^\xi(t, \xi(t))| \leq \gamma_{p_k^\xi} \|\xi_k(t+\epsilon) - \xi_k(t)\|,$$

where $\gamma_{p_k^\xi}$ is a Lipschitz constant. From the first part of the proof we know that the right hand side of the inequality converges exponentially fast, thus concluding the proof. ■

V. Simulation results

In this section we illustrate the theoretical results via simulation using artificial data. Since EM is highly sensitive to initialization [4], [8] and it suffers from the *overlap problem* [8], where under certain conditions bi-modal distributions can not be distinguished, we will compare the performance of the MFG and standard EM in both situations.

A. Algorithm

Consider the mixture (15) driven by the system (5)-(6). Since hereafter we are only concerned with GMM, we only need to track the first and second order moments, where the first moment is driven by (17) subject to the initial mean of each of the k Gaussian. The second moment is driven by (18). Explicitly we discretized (17) using Euler method with discretization step of $h = 0.01$. The covariance matrix Σ_k^x is computed as

$$\Sigma_x^k(t) = \frac{\sum_{i=1}^{N} p_k^\xi(t, x_{(i)})(x_{(i)} - \mu_x^k(t))(x_{(i)} - \mu_x^k(t))^T}{\sum_{i=1}^{N} p_k^\xi(t, x_{(i)})}. \tag{23}$$

and μ_k^x as

$$\mu_x^k(t) = \frac{\sum_{i=1}^{N} x_{(i)} p_k^\xi(t, x_{(i)})}{\sum_{i=1}^{N} p_k^\xi(t, x_{(i)})}, \tag{24}$$

as they are ML parameters of the Gaussian [19]. Instead of computing the solution of HJ equation (8) and Fokker-Planck equation (9) as in [15], [16], we consider stationary solutions, following the same strategy as in [9]- [10].

B. Initialization

All the experiments below are initialized in the following way

- let (15) be restricted to the GMM case;
- the number of classes K is given *a priori*;
- initial coefficients $\alpha_i = \frac{1}{N}$;
- means are chosen randomly among the points in the data set $\mathcal{X}$;
- all the distributions share the same variance (we use the same criteria as in [8]), given by $\Sigma_k^\xi(0) = \sigma^2 I_{n\times n}$ and

$$\sigma^2 = \frac{1}{10n} trace\left(\frac{1}{N}\sum_{i=1}^{N}\left(x_{(i)} - m\right)\left(x_{(i)} - m\right)^T\right)$$

where n is the dimension of the data, i.e $x_{(i)} \in \mathbb{R}^n$ and $x_{(i)} \in \mathcal{X},\ i = 1, ..., N$. The trace of a matrix A is given by $trace(A)$ and $m = \frac{1}{N}\sum_{i=1}^{N} x_{(i)}$ is the global data mean.

C. *Experiments*

Two main experiments are carried out. The first one concerns a very simple example which shows how the MFG-based GEM showcases better convergence properties than EM and Variational Bayes EM (VBEM) [5] . The second one is provided in [8] to analyze the overlap problem.

1) Experiment 1: In experiment 1 initial conditions were picked to be the critical case, where the minimum number of iterations required by the methods is maximized. The main drawback of EM [3] is its rate of convergence, which becomes more critical as data set increases. We show that if we start in the worst initial condition, the total number of iterates of EM and VBEM is far greater then the one achieved with MFG scheme. The simulation is based on $N = 2000$ generated data points with the following conditions:

$$\alpha_1 = \alpha_2 = 0.5,$$
$$\mu_1 = [-1\ -1]^T, \mu_2 = [1\ 1]^T,$$
$$\Sigma_1 = \Sigma_2 = I_n.$$

We start with the initial condition as in subsection V-B, except for the expected values

$$\mu_{1_0} = [1\ -1]^T, \mu_{2_0} = [-1\ 1]^T,$$

forced to be the worst case initial conditions (see Bishop [2], page 426).

The algorithm stops when it detects a change in the likelihood (4) lower than 10^{-7}. In the VBEM, we use the program VBEMGMM (by Emtiyaz, CS, UBC) using as priors $\alpha = 0.001$, $\mu = [0\ 0]$, $\beta = 1$, $W = 200 * I$ and $v = 20$ (see [2] for details on these parameters). We generated 100 different datasets, and averaged the number of iterations. MFG outperformed the other methods averaging 45 iterations compared to 240 and 164 of EM and VBEM respectively. For illustrative purposes, Figures 1 and 2 show the result of one run for EM and MFG respectively, where green and red show the result of the classification and the curves show the evolution of the means

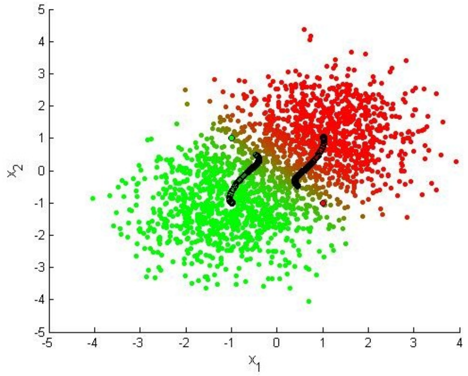

Fig. 1. Execution of EM (standart method)

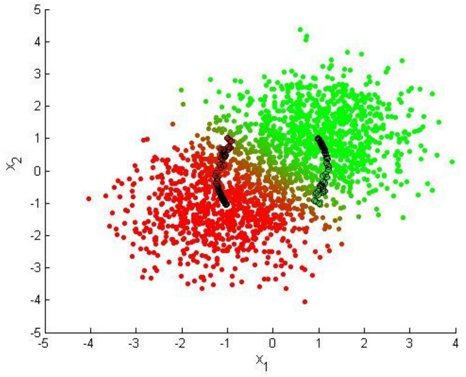

Fig. 2. Execution using the MFG approach (proposed solution)

2) Experiment 2: In this last experiment we compare the two methods when the degree of component overlap varies. For this purpose we used a bivariate Gaussian mixture with two equiprobable components ($\mu_1 = [0\ 0]^T, \mu_2 = [\delta\ 0]^T$ and $\Sigma_1 = \Sigma_2 = I$). In 50 simulations, with data sets of size $N = 800$ all simulations of MFG converged to the correct values for $\delta \geq 0.8$, outperforming EM and VBEM among other methods [8]. For $\delta < 2$ it can be shown that the mixture density is not even bimodal [8]. This is somehow a surprising result. For illustrative purpose, for one database $\mathcal{X}$ with $N = 800$ we can see the final stage (after convergence) of the classification in a green and red class where the circles with bold circunference indicate the evolution of the mean of each class. In Figure 3 we can see one example of the final stage of convergence in the overlap problem with $\delta = 0.8$.

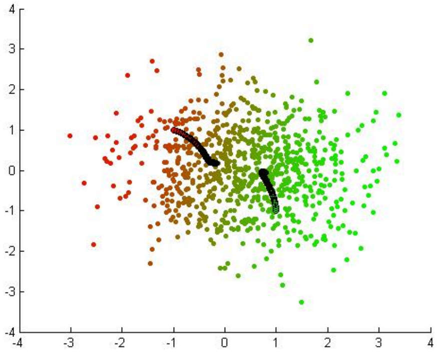

Fig. 3. Execution using the MFG approach for the overlap problem with $\delta = 0.8$

D. *Discussion of Results*

The simulation results corroborate the theoretical findings, showing faster convergence of the MFG classifier with respect EM and VBEM, despite using the stationary solutions of the HJB and FK equations. In addition it is interesting to see that MFG is capable of effectively dealing with the overlap problem, defying the theoretical bounds provided in [8].

Another advantage of the MFG approach is the numerical stability of the solution with respect to initial conditions. As shown by [9], [10] a small change of the initial conditions of (15) does not change the convergence for the same maximum likelihood value. This implies that we can establish regions of convergence to the same maximum likelihood point, a feature not available to existing GEM algorithms.

VI. Conclusions

In this paper we introduced a continuous formulation of a generalized Expectation-Maximization (GEM) algorithm for finite mixture models using Mean Field Games. The proposed solution was proven to share the same properties as GEMs, i.e. convergence to a maximum of the likelihood function. In addition we were able to show that under GMM assumptions the convergence is exponential. Finally we compared our proposed solution to EM and VBEM, showing a significant improvement in the number of iterations needed for convergence. In addition simulation results indicate that the proposed method is effective in dealing with the overlap problem.

Appendix

A. Hamilton-Jacobi-Bellman equation

Let ξ_k^j be an agent j that belongs to population k, i.e, follows a dynamic (5) subject to an initial condition $\xi_k^j(0)$, for $j = 1, ..., M$, where M is the total of players in the game. This agent applies the control $u_k(t, \xi^{(j)})$ that solves the optimization problem (6). By dynamic programming [22] one derives that J is a viscosity solution to (8) [12] and the optimal control is given in feedback form as

$$u_k^j = -J_\xi^k(t, \xi_k^j). \tag{25}$$

B. Fokker-Planck equation

Instead of considering just one agent, take in consideration a very large number M of agents within the same population k distributed throughout space according to a probability density (pdf) $p_k^\xi(t, \xi)$ such that

$$\int\limits_{\mathbb{R}^n} p_k^\xi(t, \eta) d\eta = 1,$$

for each time t.

To simplify, assume that all the agents within a population have identical motivations (in particular, they are all trying to minimize the same cost function (6)), which implies in particular that all agents in a population p_k^ξ at a given point (t, ξ_k^j) in time-space will move in time dt to a slight different location $\xi_k^j(t) + u_k(\xi_k^j)dt + \epsilon dw_t$.

Informally, for an infinitesimal box in space $[\xi_k, \xi_k + d\xi_k]$, the number of agents in that box should be approximately $Mp_k^\xi(t, \xi)\,|d\xi|$. We now suppose that the control u_k is known, as well as the initial density $p_k^\xi(0, \xi)$ of the agents, and ask how the density will evolve as time goes forward. Let us take a distributional viewpoint [14] and test the density $p_k^\xi(t, \xi)$ against various test functions $\varphi(\xi)$ - smooth compactly supported functions of space. The integral

$$\int\limits_{\mathbb{R}^n} p_k^\xi(t, \eta)\varphi(\eta) d\eta,$$

can be viewed as the continuum limit of the sum [9], [10]

$$\frac{1}{M}\sum_{j=1}^{M} \varphi(\xi_k^j(t)).$$

This leads to the heuristic equation, after using the chain rule

$$\int\limits_{\mathbb{R}^n} (p_k^\xi)_t(t, \eta)\varphi(\eta) d\eta \approx \frac{1}{M}\sum_{j=1}^{M} u_k(t, \xi_k^j(t))^T \varphi_\xi(\xi_k^j(t)). \tag{26}$$

Performing a Taylor expansion the right-hand side as before, and passing to the continuum limit, after an integration by parts it takes the form

$$\int\limits_{\mathbb{R}^n} p_k^\xi(t, \eta) \left(\varphi(\eta) + u^T\varphi_\xi(\eta) + \frac{\epsilon^2}{2}\Delta\varphi(\eta)\right) d\eta. \tag{27}$$

When we equal (27) to the left side of (26), we get the Fokker-Planck-Komolgorov equation

$$(p_k^\xi)_t(t, \xi) - \frac{\epsilon^2}{2}\Delta p_k^\xi(t, \xi) + \operatorname{div}\left(p_k^\xi u_k\right)(t, \xi) = 0. \tag{28}$$

where f is a vector field into $\mathbb{R}^n$. Remark, that we assumed that the control was known, and in the matter of fact it is and given by (25) within the MFG framework.

VII. Fact in Theorem 1

In this section we prove the fact used in Theorem 1.

Lemma 2. *Under the same conditions of Theorem 1 we have*

$$\mathbb{E}\left[\int\limits_t^{t+\varepsilon} \log\left(\frac{p_k^\xi(\tau, \xi)}{p_k^x(\tau, \xi)}\right) d\tau\right] \geq 0 \tag{29}$$

Proof: Take the following function

$$\varphi^t(x) = \int\limits_t^{t+\varepsilon} x d\tau$$

that is obviously convex since it is a linear operator. Then by Jensen's inequality [2] we have

$$\varphi(\mathbb{E}[x]) \leq \mathbb{E}[\varphi(x)]$$

and it follows that

$$\begin{aligned}
&\int\limits_t^{t+\varepsilon} \int\limits_{\mathbb{R}} \tfrac{\omega_k(t)}{p_k^\xi(\tau,\xi)} p_k^\xi(\tau, \xi) \log\left(\tfrac{p_k^\xi(\tau,\xi)}{p_k^x(\tau,\xi)}\right) d\xi d\tau \\
&\leq \int\limits_{\mathbb{R}} \omega_k(t) \int\limits_t^{t+\varepsilon} \log\left(\tfrac{p_k^\xi(\tau,\xi)}{p_k^x(\tau,\xi)}\right) d\tau d\xi = \mathbb{E}\left[\int\limits_t^{t+\varepsilon} \log\left(\tfrac{p_k^\xi(\tau,\xi)}{p_k^x(\tau,\xi)}\right) d\tau\right].
\end{aligned}$$

Remark that

$$D_{\mathcal{KL}}(p_k^\xi(\tau, \xi)\,||p_k^x(\tau, \xi)) = p_k^\xi(\tau, \xi) \log\left(\frac{p_k^\xi(\tau, \xi)}{p_k^x(\tau, \xi)}\right) \geq 0$$

where $D_{\mathcal{KL}}(p_k^{\xi}(\tau,\xi)\,||p_k^{x}(\tau,\xi))$ is the Kullback-Leiber divergence [2] and that since $p_k^{\xi}(.,\xi)$ is a probability density function greater than zero almost everywhere and $\omega_k(t)$ is the probability distribution associated with the white gaussian noise that is greater than zero almost everywhere, their ratio is bounded and there exist $\gamma \geq 0$ that does not depend upon t such that

$$\frac{\omega_k(t)}{p_k^{\xi}(\tau,\xi)} \geq \gamma$$

that concludes the proof since the integrand is greater or equal than zero. ■

References

[1] Dempster, A., N. Laird and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society B 39, 138, 1977

[2] Bishop, Christopher M., *Pattern Recognition and Machine Learning* (Information Science and Statistics). Springer, 1st ed. 2006. corr. 2nd printing edition, October 2007.

[3] Xu, Lei and Jordan, Michael I. *On convergence properties of the em algorithm for gaussian mixtures*. Neural Computation, 8:129151, 1995.

[4] McLachlan, Geoffrey J., and Krishnan, T. *The EM algorithm and extensions* / Geoffrey J. McLachlan, Thriyambakam Krishnan Wiley, New York, 1997

[5] Attias, Hagai. *A Variational Bayesian Framework for Graphical Models*. In Advances in Neural Information Processing Systems, volume 12, pp. 209215. MIT Press, 2000.

[6] Lasry, Jean-Michel and Lions, Pierre-Louis. *Jeux a champ moyen. I. Le cas stationnaire*. C. R. Math. Acad. Sci. Paris, 343(9):619625, 2006a.

[7] Lasry, Jean-Michel and Lions, Pierre-Louis. *Jeux a champ moyen. II. Horizon fini et controle optimal.* C. R. Math. Acad. Sci. Paris, 343(10):679684, 2006b.

[8] Figueiredo, Mario A. T. and Jain, Anil K. *Unsupervised learning of finite mixture models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24:381396, 2000.

[9] O. Gueant, *A reference case for mean field games models*, Universit Paris-Dauphine, Open Access publications from Universita Paris-Dauphine urn:hdl:123456789/3983, Sept. 2009.

[10] O. Gueant, *Mean field games and applications to economics*, Tech. Rep.,2009.
URL: http://basepub.dauphine.fr/bitstream/handle/123456789/5789/mean_field_gueant.PDF?sequence=1

[11] J.-M. L. Olivier Guant and P.-L. Lions, *Mean field games and applications*, in Paris-Princeton Lectures in Quantitative Finance, 2009.

[12] L. C. Evans, *Partial differential equations*, 2nd ed., ser. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 2010, vol. 19.

[13] L. C. Evans, *An introduction to mathematical optimal control theory*, 2006.
URL: http://math.berkeley.edu/ẽvans/control.course.pdf

[14] F. G. Friedlander, *Introduction to the theory of distributions*. Cambridge University Press, Cambridge, second edition, 1998.

[15] G. Turinici, J. Salomon, and A. Lachapelle, *A monotonic algorithm for a mean field games model in economics*. 2010.

[16] Y. Achdou and I. Capuzzo-Dolcetta, *Mean field games: Numerical methods*, SIAM Journal on Numerical Analysis, vol. 48, no. 3, pp. 11361162, 2010.

[17] A. J. Krener, *The convergence of the extended Kalman filter*, Tech. Rep. math.OC/0212255, Dec 2002.

[18] Osborne, Martin J. and Rubinstein, Ariel, *A Course in Game Theory*, The MIT Press, 1994

[19] Bilmes, Jeff, *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, 1998
URL: http://crow.ee.washington.edu/people/bulyko/papers/em.pdf

[20] M. Huang, P. E. Caines, and R. P. Malhame. *Large-population costcoupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria*. IEEE Transactions on Automatic Control, vol. 52, no. 9, pp. 1560-1571, 2007.

[21] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, London, U.K., second edition, 1995.

[22] Dimitri P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena Scientific, 1995.

[23] Krener, A. J., *The Convergence of the Minimum Energy Estimator*, in New Trends in Nonlinear Dynamics and Control and Their Applications, W. Kang, M. Xiao and C. Borges, eds. Springer Verlag, Heidelberg, pp. 187-208,2003

[24] M. Bardi, *Explicit solutions of some Linear-Quadratic Mean Field Games*, 2011