A Mean Field Games model for finite mixtures of Bernoulli distributions

Laura Aquilanti, Simone Cacace, Fabio Camilli
and Raul De Maio

April 20, 2020

Abstract

Finite mixture models are an important tool in the statistical analysis of data, for example in data clustering. The optimal parameters of a mixture model are usually computed by maximizing the log-likelihood functional via the Expectation-Maximization algorithm. We propose an alternative approach based on the theory of Mean Field Games, a class of differential games with an infinite number of agents. We show that the solution of a finite state space multi-population Mean Field Games system characterizes the critical points of the log-likelihood functional for a Bernoulli mixture. The approach is then generalized to mixture models of categorical distributions. Hence, the Mean Field Games approach provides a method to compute the parameters of the mixture model, and we show its application to some standard examples in cluster analysis.

AMS-Subject Classification: 62H30, 60J10, 49N70, 91C20.

Keywords: Mixture Models; Bernoulli Distribution; Categorical Distribution; Cluster Analysis; Expectation-Maximization algorithm; Mean Field Games.

1 Introduction

Finite mixture models, given by convex combinations of probability density functions (PDFs in short)

\[ \pi(x) = \sum_{k=1}^{K} \alpha_k \pi_k(x), \quad \text{with } \alpha_k \in [0,1], \sum_{k=1}^{K} \alpha_k = 1, \quad (1.1) \]

are an important mathematical tool in statistical analysis of data. Introduced by the biometrician K. Pearson [14], due to their flexibility, they are employed in a large variety of fields as astronomy, biology, genetic, medicine, marketing and engineering (see [18, Chapter 6-7], [17]). In cluster analysis, a classical problem in unsupervised Machine Learning consisting in the repartition of a data set into subgroups with similar characteristics, finite mixture models can
be used in order to determine the intrinsic structure of clustered data when no information, except for the observed values, are available. For a detailed description of the theory of mixture models and applications, see [19, 13, 17].

Given a data set $X$ representing the measurements of a phenomena, whose different values are related to the membership to unknown categories, a corresponding finite mixture model is built by assuming that the data have been generated by a random variable $X$, whose unknown probability distribution $\pi$ can be described as in (1.1). The parameters of the mixture (1.1) are in general unknown and the aim is to determine them in such a way that they optimally fit the given data set $X$. To this end, different methods can be employed such as the graphical method ([16]), the Bayesian method ([17]) and the likelihood estimator ([4]). The latter one, which is the starting point of our analysis, generates a tough quantity to be maximized, which is usually computed by means of the Expectation Maximization (EM in short) algorithm (see [3]). In our approach, we characterize the optimal parameters of a mixture model through a multi-population Mean Field Games (MFG in short) system, a coupled system of differential or difference equations which characterizes the Nash equilibria in the framework of stochastic games with a very large number of agents. The MFG theory has been introduced simultaneously by Lasry-Lions [12] and Huang-Caines-Malham [11] and it has been successfully applied to different fields, such as economics, biology, environmental policy, etc. (for a plain introduction, see [9]). Recently, its scope has been broadened to Machine Learning applications (see for example [2, 5, 6, 7, 15]).

In [1], we developed a MFG approach to finite mixture models defined over a continuous random variable $X$. Data points are interpreted as agents and the aim is subdivide to the whole population, described by a PDF $\pi(x)$, in $K$ sub-populations, described by PDFs $\pi_k$ and mixing coefficients $\alpha_k$, on the basis of some similar characteristics appropriately encoded in the cost functional of the control problems of the different populations. As a result, we end up with a stationary multi-population Mean Field Games in $\mathbb{R}^d$ which, in the particular case of a quadratic cost, characterizes the critical points of the log-likelihood functional for a mixture of Gaussian distributions.

In this paper, we focus on a mixture model for a discrete random variable $X$ described by a Bernoulli distribution or, more generally, by a categorical one. Assume that the sub-populations can be discriminated on the basis of a single specific characteristic taking only one of $S$ different values (the case of several different characteristics will be discussed further on in the paper). Given a data set $X = \{x_n\}$, in order to characterize the components of the mixture model, i.e. the PDFs $\pi_k$ and the weights $\alpha_k$, we introduce the $K$-populations finite
state MFG system

\[
\begin{aligned}
V_k(i) &= \min_{P_i} \left\{ \sum_{j=1}^{S} P_{ij} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta_k) + V_k(j) \right) \right\} - \lambda_k, \\
\pi_k(i) &= \sum_{j=1}^{S} P_{ji} \pi_k(j), \\
\pi_k(i) &\geq 0, \sum_{i=1}^{S} \pi_k(i) = 1, \sum_{i=1}^{S} V_k(i) = 0, \\
\alpha_k &= \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n),
\end{aligned}
\]

(1.2)

for \( i \in \{1, \ldots, S\} \). The term \( c(P_{ij}) \) is a transition cost between the states \( i \) and \( j \) and \( \varepsilon \log(P_{ij}) \) is an entropy penalization term which forces the agents to diversify their transition choice. Each of the \( K \) sub-populations is characterized by a quadruple \((V_k, \lambda_k, \pi_k, \alpha_k)\), with the couple \((V_k, \lambda_k)\) solving an ergodic Hamilton-Jacobi-Bellman equation and with the probability distribution \( \pi_k \) solving a Fokker-Planck equation where the transition matrix \( P \) is composed of the rows \( P_i = \{P_{ij}\}_{j=1}^{S} \) realizing the minimum in the first equation. The vector \( \theta_k \in \mathbb{R}^{S} \) represents the average probability that the points of the data set have been generated by the \( k^{th} \) distribution. Interaction among the sub-populations is encoded in the weights \( \alpha_k \) and in the coupling cost \( F(i, \theta_k) \), which depend on the responsibilities

\[
\gamma_k(x_n) = \frac{\alpha_k \pi_k(x_n)}{\pi(x_n)} \quad k = 1, \ldots, K, \; x_n \in \mathcal{X}.
\]

These quantities play a crucial role and, in cluster analysis, can be used to assign a point to the class with the highest \( \gamma_k \).

Relying on the theory for finite states MFG system developed in [10], we prove that the system (1.2) admits a solution. Moreover we show that, for the case of a space state of dimension 2, (1.2) characterizes the critical points of the maximum log-likelihood functional for a mixture of Bernoulli distributions and therefore gives an alternative way to compute the critical points of the log-likelihood functional. We show an application of this method to the computation of the optimal parameters for mixture models related to some standard examples in cluster analysis, such as digit classification.

We remark that the goal of the paper is to provide a new perspective to a class of problems which, up to now, have been considered in the classical framework of finite dimensional optimization. For a linear cost \( c \) and \( \varepsilon = 0 \), the algorithm described in Section 5 is equivalent to the classical EM algorithm and, in general, it is not computationally competitive with other algorithms for cluster analysis. But our model is feasible to be generalized in several directions and, in a forthcoming paper, we plan to study more general costs, possibly depending on the state variable, in order to exploit in a deeper way the structure of the data set.

The paper is organized as follows. In Section 2, we briefly review the finite mixture model theory and the corresponding EM algorithm. In Section 3, we introduce the MFG model for a Bernoulli random variable and we show the
connection with the maximization of the log-likelihood functional. In Section 4, we generalize the model of the previous section to a mixture of categorical distributions and we prove existence of a solution to the MFG system. In Section 5, we apply the MFG model to some standard problems in cluster analysis. In Appendix A, we prove some results for a finite-states stationary MFG system we use in the previous sections.

2 A short introduction to mixture models

In this section, we shortly review the parametric mixture model and the corresponding EM algorithm for the optimization of the parameters.

Let $X$ be a random variable, univariate or multivariate and consider a sample $X = \{x_1, \ldots, x_N\}$ of size $N$ of $X$, where the sample space $S$ can be discrete or continuous. Let $p(x)$ be the unknown distribution of $X$, defined with respect to an appropriate reference measure on $S$. We assume that $X$ comes from a finite mixture model, i.e. $p(x)$ can be written as a convex combination of PDFs $p_k$ as

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

(2.1)

where $K$, the number of the components of $p$, is supposed to be known a priori and $\alpha_k$, the weights or mixing coefficients, satisfy $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k \geq 0$.

How to determine the number $K$ of PDFs in the mixture model is an unresolved issue and, in general, a combination of criteria and experimental analysis is used to guide the decision (see [19, Chapter 6]).

Usually, it is assumed that the components of the mixture (2.1) belong to the same parametric family of density distributions, i.e. they can be written as $p(x; \theta_k)$, where $\theta_k$ is the parameter which defines the $k$-th PDF. For example, in the Gaussian mixture model, $p(x; \theta_k) = N(x; \mu_k, \Sigma_k)$ are Gaussian distributions of parameters $\theta_k = (\mu_k, \Sigma_k)$ where $\mu_k, \Sigma_k$ are the mean and covariance matrix; in the Bernoulli mixture model, $p(x; \theta_k) = B(x; \mu_k)$ are Bernoulli distributions of parameter $\theta_k = \mu_k$. The aim is to find the parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\theta = (\theta_1, \ldots, \theta_K)$ which give the best representation of the sample $X$. This can be achieved through the maximization of the log-likelihood functional

$$\mathcal{L}(\alpha, \theta; \mathcal{X}) = \sum_{n=1}^{N} \log \left[ \sum_{k=1}^{K} \alpha_k p(x_n; \theta_k) \right] + \lambda \left( \sum_{k=1}^{K} \alpha_k - 1 \right),$$

(2.2)

where $\lambda$ is a Lagrange multiplier which takes into account the constraint for the mixing coefficients. By writing the necessary condition for the extrema of (2.2),
we have
\[
\frac{\partial L}{\partial \alpha_k} = \sum_{n=1}^{N} \frac{p_k(x_n; \theta_k)}{\sum_{j=1}^{K} \alpha_j p_j(x_n; \theta_j)} - \lambda = 0, \quad (2.3)
\]
\[
\frac{\partial L}{\partial \theta_k} = \sum_{n=1}^{N} \alpha_k \frac{\partial p_k(x_n; \theta_k)}{\partial \theta_k} = 0. \quad (2.4)
\]

The Expectation-Maximization algorithm is an iterative procedure for the computation of a solution of the previous system. We shortly describe its derivation. Firstly, it is introduced a latent, or hidden, $K$-dimensional random variable $Y = (Y_1, \ldots, Y_K)$, with $Y_k \in \{0, 1\}$ and $\sum_{k=1}^{K} Y_k = 1$, saying which component of the mixture (2.1) has generated a given sample point $x_n$. More specifically, we assume that, for each observed point $x_n \in \mathbb{R}^D$, there exists a corresponding unobserved one $y_n \in \mathbb{R}^K$ such that, if the point $x_n$ has been generated by the $k$-component of the mixture, then $y_{n,k} = 1$ and $y_{n,j} = 0$ for $j \neq k$. Hence $Y$ is a multinomial random variable and we assume that $P(Y_k = 1) = \alpha_k$. A simple application of Bayes’ Theorem allows to compute the responsibility, i.e. the probability of $Y$ given $X$,
\[
\gamma_k(x_n) := P(Y_k = 1 | X = x_n) = \frac{\alpha_k p_k(x_n; \theta_k)}{\sum_{j=1}^{K} \alpha_j p_j(x_n; \theta_j)}. \quad (2.5)
\]

Multiplying (2.3) by $\alpha_k$ and summing over $k$ we get that $\lambda = N$, hence replacing (2.5) in (2.3) we get a first condition for an extremum of (2.2)
\[
\alpha_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n). \quad (2.6)
\]

Using (2.5) in (2.4), we have
\[
\sum_{n=1}^{N} \gamma_k(x_n) \frac{\partial \log(p_k(x_n; \theta_k))}{\partial \theta_k} = 0. \quad (2.7)
\]

Since the coefficients $\alpha_k$ in (2.6) depend on $\theta_k$ via the responsibilities $\gamma_k$, the equations (2.6)-(2.7) do not provide a closed form solution for the parameters. However, they suggest the following iterative scheme which alternates two steps. Starting from some arbitrary initialization $(\alpha^{(0)}, \theta^{(0)})$, at $h$-iteration we perform the following steps: in the Expectation step (E-step in short), using the current values $(\alpha^{(h)}, \theta^{(h)})$ of the parameters, we compute the responsibilities by means of formula (2.5) with $\theta^{(h)}$ in place $\theta_k$. In the Maximization step (M-step in short), given the responsibilities as in the E-step, we compute the new parameters $(\alpha^{(h+1)}, \theta^{(h+1)})$ by means of (2.6)-(2.7). It can be proved that at each iteration (E-step, M-step), the log-likelihood function increases its value. For more details see [3], [4, Chapter 9].

For specific parametrized families of distributions, such as Gaussian and Bernoulli
distributions, it is possible to get explicit formulas for the parameters \( \theta_k \). Computed the responsibility \( \gamma_k^{(h+1)}(x) \) by means of (2.5) in the E-step, in the M-step we have the following explicit formulas

\[
\alpha_k^{(h+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n)
\]

and, for a Gaussian mixture,

\[
\mu_k^{(h+1)} = \frac{\sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n) x_n}{\sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n)}
\]

\[
\Sigma_k^{(h+1)} = \frac{\sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n)(x_n - \mu_k^{(h+1)})(x_n - \mu_k^{(h+1)})^t}{\sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n)}
\]

for a Bernoulli mixture,

\[
\mu_k^{(h+1)} = \frac{\sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n) x_n}{\sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n)}
\]

3 A Mean Field Games approach to Bernoulli mixture models

In this section, we describe a Mean Field Games approach for a mixture of multivariate Bernoulli distributions. This method is used to cluster high-dimensional binary data. In order to explain the technique, we start describing an application of the previous model to the classification of handwritten digits (see [4, Cap. 9]). Consider a database \( X = \{x_1, \ldots, x_N\} \) of images representing handwritten digits of \( K \) numbers between 0 and 9. Each image, which is originally given by a square of \( d \times d \) pixels in grey scale, is turned in a binary vector \( x = (x_1, \ldots, x_D) \), of size \( D = d^2 \), by setting the elements whose value exceeds 1/2 to 1 and the remaining to 0, with 1 corresponding to a white pixel, 0 to a black one. Assuming to know \( K \), the aim is to subdivide the data set \( X \) in \( K \) clusters, where each cluster is represented by the component \( \pi_k \) of a mixture of Bernoulli distributions and an image is attributed to the cluster which maximizes the responsibility. In a second phase, these clusters can be used to identify the digit corresponding to a new image, but we will not consider this problem here. In Figure 1, we see some samples of digits taken from database MNIST [20]. We suppose that the digits are the i.i.d. observations of a random variable \( X \) in \( \mathbb{R}^D \). Assuming that the colour of each pixel is independent of all the other ones, this implies that the \( D \) components of the random vector \( X = (X^1, \ldots, X^D) \), with \( X^d \in \{0, 1\} \), are independent from each other. Hence we can describe the distribution of \( X \) as a mixture of \( K \) multivariate Bernoulli distributions

\[
\pi(x) = \sum_{k=1}^{K} \alpha_k \pi_k(x),
\]
Figure 1: Samples of hand-written digits from the MNIST database

where each measure $\pi_k$, defined by

$$
\pi_k(x) = \prod_{d=1}^{D} \pi^d_k(x^d), \quad x = (x^1, \ldots, x^D) \in \{0,1\}^D
$$

represents the distribution of a specific digit in the $k^{th}$ cluster. The aim is to design a $K$-population MFG system to find the unknown parameters $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K$, $\mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^{K \times D}$, where $\mu_k = (\mu^1_k, \ldots, \mu^D_k)$, in such a way that the measure $\pi$ in (3.1) optimally fits the data set $X$. Following the analysis described in Section 2, we introduce a $K$-dimensional latent random variable $Y$ that specifies which element of the mixture (3.1) has generated a given data point $x_n$ and, for $x_n \in X$, we define the responsibilities

$$
\gamma_k(x_n) = \mathbb{P}[Y_k = 1|X = x_n] = \frac{\alpha_k \pi_k(x_n)}{\pi(x_n)} \quad k = 1, \ldots, K.
$$

Note that $\gamma_k(x_n)$ are defined in terms of the unknown measure $\pi$. Given the responsibilities, we define the average value of the data set $\theta_k = (\theta^1_k, \ldots, \theta^D_k) \in \mathbb{R}^D$ with respect to the $k^{th}$ component of the mixture by

$$
\theta^d_k = \mathbb{E}[X^d|Y_k = 1] = \frac{\sum_{n=1}^{N} \gamma_k(x_n)x^d_n}{\sum_{n=1}^{N} \gamma_k(x_n)}, \quad d = 1, \ldots, D.
$$

By definition, $\theta^d_k \in [0,1]$. For each $d = 1, \ldots, D$, we consider the following 2-states $K$-populations MFG system (for a differential game interpretation of
the system we refer to \(10\))

\[
\begin{cases}
V^d_k(0) = \min_{p \in [0,1]} \{ p(-\frac{1-p}{2} + \epsilon \log(p) + V^d_k(0)) \\
+(1-p)(-\frac{q}{2} + \epsilon \log(1-p) + V^d_k(1)) \} - \lambda^d_k + (\theta^d_k)^2 \\
V^d_k(1) = \min_{q \in [0,1]} \{ (1-q)(-\frac{q}{2} + \epsilon \log(1-q) + V^d_k(0)) \\
+q(-\frac{1-q}{2} + \epsilon \log(q) + V^d_k(1)) \} - \lambda^d_k + (1-\theta^d_k)^2 \\
\pi^d_k(0) = p\pi^d_k(0) + (1-q)\pi^d_k(1) \\
\pi^d_k(1) = (1-p)\pi^d_k(0) + q\pi^d_k(1) \\
\pi^d_k \geq 0, \sum_{x\in\{0,1\}} \pi^d_k(x) = 1, \sum_{x\in\{0,1\}} V^d_k(x) = 0.
\end{cases}
\]  

(3.4)

where the values \((p,q)\) in the second couple of equations are given by controls realizing the minimum in the first couple of equations. The family of MFG systems (3.4) is completed with the global coupling condition

\[
\alpha_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n).
\]  

(3.5)

For each \(d = 1, \ldots, D\) fixed, the unknowns in the K-population MFG system (3.4) are the couple \((V^d_k, \lambda^d_k)\), with \(V^d_k \in \mathbb{R}^2\) and \(\lambda^d_k \in \mathbb{R}\), solving the Hamilton-Jacobi-Bellman equation, and the binomial distributions \(\pi^d_k\), solving the Fokker-Planck equation, while the weights \(\alpha_k \in [0,1]\) are global unknowns, independent of the index \(d\). It is important to observe that also the transition matrix

\[
P = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix},
\]

(3.6)

in the Fokker-Planck equation depends on \(k\) and \(d\) and, under appropriate assumptions (see Section 4), is univocally determined by the couple of controls \(p,q\) which minimize the first two Hamilton-Jacobi equations.

For \(d\) fixed, the coupling cost \(((\theta^d_k)^2, (1-\theta^d_k)^2)\) is defined in such a way that a population described by the density function \(\pi^d_k\) distributes in the two states \(\{0,1\}\) so that its average value \(\mu^d_k\) is as close as possible to the average value \(\theta^d_k\) of the data set. Hence this forces the average value of the \(k\)-th component \(\pi_k\) of the mixture to be close to the average of \(X\), given the occurrence of the event \(Y = 1\).

Note that the value \(\theta^d_k\) depends on the responsibilities \(\gamma_k\), see (3.3), and therefore, in turn, on the complete distribution \(\pi\). Hence, the systems (3.4) are coupled also with respect to the index \(d\) by means of the quantities \(\theta^d_k\). Indeed, the covariance matrix of (3.1) is not diagonal (see [4, Equation (9.50)]) and the model can capture correlation among the components of the mixture. The well-posedness of problem (3.4)-(3.5) will be discussed in Section 4 in the more general setting of the categorical distributions.

We now show the connection between the multi-population MFG system (3.4)
and the log-likelihood functional (2.2), which, in the case of a multivariate Bernoulli distribution (3.2), is given by

\[
\mathcal{L}(\alpha, \mu; X) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(x_n) \left( \log(\alpha_k) + \sum_{d=1}^{D} x_n^d \log(\mu_k^d) + (1 - x_n^d) \log(1 - \mu_k^d) \right)
\]

for \( \alpha = (\alpha_1, \ldots, \alpha_k), \mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^{K \times D} \), where \( \mu_k = (\mu_k^1, \ldots, \mu_k^D) \), \( k = 1, \ldots, K \). Note that, since the log-likelihood has to be maximized, we are interested in the critical values of \( \mathcal{L}(\alpha, \mu; X) \) for the values of the parameters \( \alpha_k, \mu_k^d \) inside the interval \((0, 1)\). By (3.7), the necessary conditions (2.6)-(2.7) can be computed explicitly in terms of responsibilities and they are given by

\[
\begin{align*}
\mu_k^d &= \frac{\sum_{n=1}^{N} \gamma_k(x_n)x_n^d}{\sum_{n=1}^{N} \gamma_k(x_n)} \\
\alpha_k &= \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n). 
\end{align*}
\]

In the next proposition, we show that a Bernoulli mixture satisfying (3.8) can be always obtained as a solution of the MFG system (3.4) with \( \varepsilon = 0 \). Moreover, a solution of problem (3.4)-(3.5) with \( \varepsilon = 0 \) gives a mixture satisfying (3.8).

**Proposition 3.1.** We have

(i) Let \( \pi(x) = \sum_{k=1}^{K} \alpha_k \pi_k(x) \), with \( \pi_k \) as in (3.2), be a Bernoulli mixture satisfying (3.8). Then, the family of quadruples \( (V_k, \lambda_k, \pi_k, \alpha_k) \), \( k = 1, \ldots, K \), with

\[
\begin{align*}
V_k &= (V_k^1, \ldots, V_k^D), \quad V_k^d(x) = \frac{(1-2\mu_k^d)^x(2\mu_k^d-1)}{2}1-x, \ x \in \{0, 1\}, \\
\lambda_k &= (\lambda_k^1, \ldots, \lambda_k^D), \quad \lambda_k^d = 0, \\
\pi_k &= (\pi_k^1, \ldots, \pi_k^D), \quad \pi_k^d(x) = (\mu_k^d)^x(1-\mu_k^d)^{1-x}, \ x \in \{0, 1\},
\end{align*}
\]

is a solution of (3.4)-(3.5) with \( \varepsilon = 0 \).

(ii) Let \( \{V_k, \lambda_k, \pi_k, \alpha_k\}_{k=1}^{K} \) be a solution of the MFG system (3.4)-(3.5) with \( \varepsilon = 0 \). Then \( \pi(x) = \sum_{k=1}^{K} \alpha_k \pi_k(x) \) is a Bernoulli mixture verifying (3.8).

**Proof.** (i) Consider a mixture \( \pi \) defined as in (3.1)-(3.2) and satisfying (3.8). Then, each component \( \pi_k^d \) satisfies the 2-states Fokker-Plank equation

\[
\begin{align*}
\pi_k^d(0) &= p\pi_k^d(0) + (1-q)\pi_k^d(1) \\
\pi_k^d(1) &= (1-p)\pi_k^d(0) + q\pi_k^d(1) \\
\pi_k^d &\geq 0, \quad \sum_{x \in \{0, 1\}} \pi_k^d(x) = 1
\end{align*}
\]

(3.10)
where the transition matrix $P$ is defined as in (3.6) with $p = 1 - \mu_k^d$, $q = \mu_k^d$.

By the Hamilton-Jacobi-Bellman equation

$$
\begin{align*}
V_k^d(0) &= \min_{p \in [0,1]} \{ p(-\frac{1-p}{2} + V_k^d(0)) \\
&\quad + (1-p)(-\frac{p}{2} + V_k^d(1)) \} - \lambda_k^d + (\theta_k^d)^2 \tag{3.11}
\end{align*}
$$

$$
\begin{align*}
V_k^d(1) &= \min_{q \in [0,1]} \{ (1-q)(-\frac{q}{2} + V_k^d(0)) \\
&\quad + q(-\frac{1-q}{2} + V_k^d(1)) \} - \lambda_k^d + (1 - \theta_k^d)^2,
\end{align*}
\sum_{x \in \{0,1\}} V_k^d(x) = 0,
$$

we have that the optimal controls are given by $(p, q) = (1 - \theta_k^d, \theta_k^d)$ and the solution is given by the couple $(V_k^d, \lambda_k^d)$ with $V_k^d(x) = (\frac{1-2\theta_k^d}{2})^d (\frac{2\theta_k^d-1}{2})^{1-x}$, $x \in \{0,1\}$, and $\lambda_k^d = 0$. Since, by condition (3.8), the average values $\theta_k^d$, defined in (3.3), coincide with $\mu_k^d$, we conclude that the transition matrix $P$ in (3.10) is given by the optimal controls in (3.11) and therefore (3.9) is a solution of (3.4) with $\varepsilon = 0$.

(ii) Let $\{(V_k, \lambda_k, \pi_k, \alpha_k)\}_{k=1}^K$ be a solution of the MFG system (3.4)-(3.5) with $\varepsilon = 0$. Then, since $(V_k^d, \lambda_k^d)$ is the solution to Hamilton-Jacobi-Bellman equation (3.11), we have $(V_k^d, \lambda_k^d) = \left(\left(\frac{1-2\theta_k^d}{2}\right)^d (\frac{2\theta_k^d-1}{2})^{1-x}, 0\right)$ with the optimal controls $p$ and $q$ given by

$$
p = 1 - \theta_k^d, \quad q = \theta_k^d. \tag{3.12}
$$

It follows that, given the transition matrix $P$ as in (3.6) with $p,q$ as in (3.12), the solution to the Fokker-Planck equation (3.10) is a Bernoulli distribution $\pi_k^d(x)$ of parameter $\mu_k^d = \theta_k^d$. Therefore, by (3.3), $\mu_k^d$ satisfies the first condition in (3.8). Moreover, since the coefficients $\alpha_k$ are given by (3.5), also the second condition in (3.8) is satisfied.

**Remark 3.2.** For $\varepsilon > 0$, the MFG system (3.4) characterizes, in the sense of the previous proposition, the critical points of the functional

$$
L_\varepsilon(\alpha, \mu; \mathcal{X}) = \sum_{n=1}^N \sum_{k=1}^K \gamma_k(x_n) \left( \log(\alpha_k) + \sum_{d=1}^D x_n^d \log(f_\varepsilon(\mu_k^d)) \right) + (1 - x_n^d) \log(f_\varepsilon(1 - \mu_k^d))
$$

where

$$
f_\varepsilon(\mu) = \mu + \frac{\varepsilon}{2} \log \left( \frac{\mu}{1 - \mu} \right).
$$

Note that $f_\varepsilon(\mu) \in (0,1)$ for $\mu \in (\delta_\varepsilon, 1 - \delta_\varepsilon)$, for some appropriate constant $\delta_\varepsilon \in (0,1)$ with $\lim_{\varepsilon \to 0^+} \delta_\varepsilon = 0$. Hence the functional $L_\varepsilon$ is defined for $\mu \in (\delta_\varepsilon, 1 - \delta_\varepsilon)$ and, for $\varepsilon > 0$, the MFG system (3.4) gives a non degenerate multinomial Bernoulli distributions.
4 A Mean Field Games approach to mixture models of categorical distributions

In the classification of handwritten digits previously described, the components of the random variable $X$, which generates the data set $X$, can take only one of two possible values, i.e. 0 and 1 corresponding to a white or a black pixel. In other models, see for example the Fashion-MNIST dataset [21], a discrete random variable can take only one of a certain number $S$ of mutually exclusive states. In this case, $X$ is said a categorical random variable. In this section, we introduce a MFG approach to the optimization of the parameters for a mixture of categorical distributions, i.e. discrete probability measures associated to a categorical random variable.

We denote by $S = \{1, \ldots, S\}$ the state space of a component $X^d$ of $X$. A probability measure on the state space $S$ can be identified with a vector $p \in S$, where $S = \{p = (p(1), \ldots, p(S)) : p(i) \geq 0, \sum_{i=1}^{S} p(i) = 1\}$ is the probability simplex. The set of the $S \times S$ stochastic matrices is identified with $S^S$.

We assume that the data set $X$ is generated by the i.i.d. observations of a random variable $X$ in $\mathbb{R}^D$, whose components $X^d$ are categorical random variables with values in $S$ and independent from each other. In order to find a representation of the distribution of $X$, we consider the mixture

$$\pi(x) = \sum_{k=1}^{K} \alpha_k \pi_k(x), \quad x = (x^1, \ldots, x^D) \in S^D,$$  \hspace{1cm} (4.1)

where each measure $\pi_k$ is given by

$$\pi_k(x) = \prod_{d=1}^{D} \pi_k^d(x^d),$$  \hspace{1cm} (4.2)

with $\pi_k^d \in S$, $d = 1, \ldots, D$, a categorical distribution. Note that (4.2) is consequence of the assumption that the components of $X$ are independent each other.

We denote with $\mathcal{P}$ the space of the multinomial categorical measures, i.e. the space of the measures defined as in (4.2). We also identify a measure $\pi_k$ on $S^D$ with a vector $\pi_k = (\pi_k^1, \ldots, \pi_k^D)$, with $\pi_k^d \in S$.

**Remark 4.1.** The Fashion-MNIST dataset is composed of grey-scale, 28 × 28 pixels images of 10 different types of fashion products. In this case we have $D = 784$, corresponding to the pixels of the images, $S = 256$, corresponding to the grey-scale levels of a single pixel, and $K = 10$, corresponding to the different fashion objects in the dataset.

Given the data set $X = \{x_n\}$, for $k = 1, \ldots, K$ we define the responsibilities by

$$\gamma_k(x_n) = \mathbb{P}[Y_k = 1|X = x_n] = \frac{\alpha_k \pi_k(x_n)}{\pi(x_n)} \quad n = 1, \ldots, N. \hspace{1cm} (4.3)$$
and the vector $\theta_k = (\theta_k^1, \ldots, \theta_k^D) \in \mathbb{S}^D$, $\theta_k^d = (\theta_k^d(1), \ldots, \theta_k^d(S)) \in \mathbb{S}$, by

$$\theta_k^d(i) = \frac{\sum_{n=1}^N \gamma_k(x_n)[x_n^d = i]}{\sum_{n=1}^N \gamma_k(x_n)} \quad i = 1, \ldots, S \quad (4.4)$$

(here the Iverson brackets $[x_n^d = i]$ evaluates 1 if the random variable $X^d$ assumes state $i$, 0 otherwise). The vector $\theta_k$ represents the weighted probabilities of the random variable $X^d$ with respect to the $k^{th}$ component of the mixture computed on the data set $X$.

For $k = 1, \ldots, K$, $d = 1, \ldots, D$ and $i = 1, \ldots, S$, we consider the multi-population MFG system

$$V_k^d(i) = \min_{\lambda_k} \left\{ \sum_{j=1}^S P_{ij} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta_k^d) + V_k^d(j) \right) \right\} - \lambda_k^d,$n\lambda_k^d,$n\pi_k^d(i) = \sum_{j=1}^S P_{ij} \pi_k^d(j),$n\pi_k^d(i) \geq 0,$n\sum_{i=1}^S \pi_k^d(i) = 1,$n\sum_{i=1}^S V_k^d(i) = 0,$n\alpha_k = \frac{1}{N} \sum_{n=1}^N \gamma_k(x_n),$$

where the transition matrix $P$ in the Fokker-Planck equation is composed of rows $P_i = \{P_{ij}\}_{j=1}^S$ which realize the minimum in the Hamilton-Jacobi-Bellman equation, $c(P_{ij})$ is a transition cost from the state $i$ to the state $j$ and the entropy penalization term $\varepsilon \log(P_{ij})$ enforces the agent to diversify their choices, so that $P_{ij} > 0$ for any $i, j = 1, \ldots, S$. We assume that $c \in C^1([0, 1])$ with $pc(p)$ convex for $p \in [0, 1]$ and $F : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}$ is such that $F(i, \cdot)$ is bounded and continuous for all $i \in \mathbb{S}$.

A solution of (4.5) is a family of quadruples $(V_k, \lambda_k, \pi_k, \alpha_k), k = 1, \ldots, K$, where

$$V_k = (V_k^1, \ldots, V_k^D), \quad \lambda_k = (\lambda_k^1, \ldots, \lambda_k^D), \quad \pi_k = (\pi_k^1, \ldots, \pi_k^D), \quad \alpha_k \in [0, 1].$$

As observed in the previous section, for fixed $d \in \{1, \ldots, D\}$, (4.5) gives a $K$-populations MFG system on a $S$-states space. The system is globally coupled by means of the vectors $\theta_k^d$, which depend on the full measure $\pi$ via the responsibilities $\gamma_k$, and by means of the weights $\alpha_k$. Note that, a priori, the responsibilities are not well defined since $\pi$ could vanish for some $x_n \in X$. However, we will prove in Theorem 4.2 that, because of the entropy penalization term, there exists a solution of (4.5) for which $\pi$ cannot vanish on the data set $X$ and therefore $\gamma_k$ and $\theta_k^d$ are well defined.

**Theorem 4.2.** For any $\varepsilon > 0$, there exists a solution $(V_k, \lambda_k, \pi_k, \alpha_k), k = 1, \ldots, K$, of (4.5).
Proof. We define the

\[ D = \{ (\alpha, \pi) = (\alpha_1, \ldots, \alpha_K, \pi_1, \ldots, \pi_K) : \alpha_k \in [0, 1], \sum_{k=1}^{K} \alpha_k = 1, \] \]

\[ \pi_k = \prod_{d=1}^{D} \pi_k^d \in \mathcal{P}, \min_{i \in S} \pi_k^d(i) \geq \delta > 0, \quad d = 1, \ldots, D \}, \tag{4.6} \]

where \( \delta \) is a constant to be fixed later. It is easy to see that \( D \) is a convex and compact set with respect to the topology of \( \mathbb{R}^K \times (\mathcal{S}^D)^K \). We define a map \( \Psi \) on \( D \) in the following way:

Given \((\alpha, \pi) \in D\), we define for \( k = 1, \ldots, K \)

\[ \gamma_k(x_n) = \frac{\alpha_k \pi_k(x_n)}{\sum_{k=1}^{K} \alpha_k \pi_k(x_n)}, \quad n = 1, \ldots, N \tag{4.7} \]

\[ \theta_d^k(i) = \frac{\sum_{n=1}^{N} \gamma_k(x_n)[x_n^d = i]}{\sum_{k=1}^{K} \gamma_k(x_n)} \quad d = 1, \ldots, D \quad i = 1, \ldots, S \tag{4.8} \]

and \( \theta_d^k = (\theta_d^k(1), \ldots, \theta_d^k(S)) \).

Note that, since \( \min_{i \in S} \pi_k^d(i) \geq \delta > 0 \), then \( \pi_k(x) \geq \delta^D \) for any \( x \in S^D \), and therefore

\[ \sum_{k=1}^{K} \alpha_k \pi_k(x_n) \geq \delta^D, \quad n = 1, \ldots, N. \tag{4.9} \]

Hence \( \gamma_k \) and \( \theta_d^k \) in (4.7)-(4.8) are well defined. For each \( d = 1, \ldots, D \), we consider the \( K \)-population \( S \)-states space MFG system

\[
\begin{cases}
V_k^d(i) = \min_{P_{ij}} \left\{ \sum_{j=1}^{S} P_{ij} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta_k^d(j)) + V_k^d(j) \right) \right\} - \lambda_k^d \\
\rho_k^d(i) = \sum_{j=1}^{S} P_{ji} \rho_k^d(j) \\
\rho_k^d(i) \geq 0, \quad \sum_{i=1}^{S} \rho_k^d(i) = 1, \quad \sum_{i=1}^{S} V_k^d(i) = 0
\end{cases}
\tag{4.10}
\]

for \( i = 1, \ldots, S \). Since the coefficient \( \theta_k^d(i) \) are given (see (4.8)), the previous systems are not coupled with respect to the index \( d \). By Theorem A.5, for any \( d = 1, \ldots, D \), there exists a unique solution \((V_k^d, \lambda_k, \rho_k^d)\) to (4.10) with

\[ \min_{i \in S} \rho_k^d(i) \geq C(\varepsilon) > 0, \quad k = 1, \ldots, K. \tag{4.11} \]

Set \( \rho = (\rho_1, \ldots, \rho_K) \), where for any \( k = 1, \ldots, K \) the vector \( \rho_k = (\rho_k^1, \ldots, \rho_k^D) \in \mathcal{P} \) has components given by the solutions \( \rho_k^d \) of (4.10), and

\[ \beta_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n). \]

Then, the map \( \Psi \) is defined by \( \Psi(\alpha, \pi) = (\rho, \beta) \). Fixed the constant \( \delta \) in (4.6) smaller than the constant \( C(\varepsilon) \) in (4.11), then it follows that \( \Psi \) maps the set
We prove that \( \Psi \) is a continuous map on \( \mathcal{D} \) with respect to the topology of \( \mathbb{R}^K \times (\mathbb{S}^D)^K \). Consider a sequence \( (\alpha^{(h)}, \pi^{(h)}) \in \mathcal{D}, h \in \mathbb{N} \), converging to \( (\alpha, \pi) \in \mathcal{D} \) and denote with \( (\rho^{(h)}, \beta^{(h)}), (\rho, \beta) \) the corresponding images by means of \( \Psi \). Given the vector \( (\theta_{k}^{d})^{(h)} \) whose components are defined by

\[
(\theta_{k}^{d})^{(h)}(i) = \frac{\sum_{n=1}^{N} \gamma_{k}^{(h)}(x_{n})[x_{n}^{d} = i]}{\sum_{n=1}^{N} \gamma_{k}^{(h)}(x_{n})}, \quad k = 1, \ldots, K, d = 1, \ldots, D,
\]

and recalling (4.9), it is immediate that \( (\theta_{k}^{d})^{(h)}(i) \to \theta_{k}^{d}(i) \) for \( h \to \infty \), and hence \( (\theta_{k}^{d})^{(h)} \to \theta_{k}^{d} \) for \( h \to \infty \) for any \( k = 1, \ldots, K, d = 1, \ldots, D \). By continuity, \( F(i, (\theta_{k}^{d})^{(h)}) \) converges uniformly to \( F(i, \theta_{k}^{d}) \). By Lemma A.3 and A.4, it follows that the vectors \( P_{i}^{(h)} \), which attain the minimum in the first equation of (4.10) for \( ((V_{k}^{d})^{(h)}, (\lambda_{k}^{d})^{(h)}) \), converge to the corresponding vector \( P_{i} \) which attains the minimum in the first equation of (4.10) for \( (V_{k}^{d}, \lambda_{k}^{d}) \). Since the transition matrices \( P_{i}^{(h)} \), composed of the rows \( P_{i}^{(h)} \), \( i = 1, \ldots, S \), converge to the transition matrix \( P_{i} \), composed of the rows \( P_{i}, i = 1, \ldots, S \), it follows that the corresponding invariant distribution \( (\rho_{k}^{d})^{(h)} \) converge to \( \rho_{k}^{d}, d = 1, \ldots, D \) and therefore \( (\rho_{k})^{(h)} = \prod_{d=1}^{D} (\rho_{k}^{d})^{(h)} \) converges to \( \rho_{k} = \prod_{d=1}^{D} \rho_{k}^{d} \). Moreover, recalling (4.11), it follows that \( \beta_{k}^{(h)} \) converges to \( \beta_{k} \). Hence the map \( \Psi \) is continuous.

By the Brouwer’s fixed point Theorem, there exists \( (\alpha, \pi) \in \mathcal{D} \) such that \( \Psi(\alpha, \pi) = (\alpha, \pi) \) and therefore a solution to (4.5).

Note that, in this framework, it is, in general, not reasonable to obtain uniqueness for the system (4.5), since the log-likelihood functional can be not concave. This also corresponds to the fact that, in cluster analysis, there may be several admissible partitions of a given data set. As for the case \( S = 2 \), see Proposition 3.1, it is possible to show that, identifying the state \( i \in S \) with the vertex of coordinate

\[
T_{i} = (0, \ldots, 1, \ldots, 0)
\]

of the simplex in \( \mathbb{R}^{S} \), taking \( c(p) = -(1 - p)/2 \) and \( F(i, \theta) \) equal to the square of distance of \( \theta \) from \( T_{i} \), then the solutions of the MFG system (4.5) for \( \varepsilon = 0 \) are associated with the critical points of the the log-likelihood functional for a mixture of categorical distributions.

**Remark 4.3.** Theorem 4.2 provides an existence result to the MFG system (4.5) only for \( \varepsilon > 0 \). For \( \varepsilon = 0 \), estimate (4.9) no longer applies and therefore it is not guaranteed that the responsibilities \( \gamma_{k} \) in (4.3) are well defined. On the other hand, for \( \varepsilon \to 0 \), it is possible to prove that, up to a subsequence, the vector \( \theta_{k}^{d} \), defined in (4.4) converges to a vector \( \bar{\theta}_{k}^{d} \in \mathbb{S} \) and, moreover, the solution of (4.5) converges to a solution of (4.5) with \( \varepsilon = 0 \) and cost \( F(i, \bar{\theta}_{k}^{d}) \). The point is that we cannot characterize \( \bar{\theta}_{k}^{d} \) by a formula equivalent to (4.4) since again we cannot exclude that the responsibilities \( \gamma_{k} \), corresponding to the solution of the limit system, vanish for some value \( x_{n} \in \mathcal{X} \).
Remark 4.4. For the simplicity of the notation, we assumed that all the components $X^d$ of the random variable $X$, which generates the data set, are $S$-dimensional categorical random variables. Nevertheless, it is possible to assume that they have state spaces of different dimension $S_d$, for each $d = 1, \ldots, D$. In this case, also the MFG systems (4.5) are defined on states spaces of dimension $S_d$ and the results of this section can be easily reformulated in this more general framework.

5 Numerical approximation and examples

In this section we present the main idea for the numerical solution of the MFG system (4.5), then we apply the resulting algorithm to some classical tests in cluster analysis. For the reader’s convenience, we recall the MFG system here: for $k = 1, \ldots, K$, $d = 1, \ldots, D$ and $i = 1, \ldots, S$,

$$
\begin{cases}
V^d_k(i) = \min_{P_{ij}} \left\{ \sum_{j=1}^{S} P_{ij} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta^d_k) + V^d_k(j) \right) \right\} - \lambda^d_k, \\
\pi^d_k(i) = \sum_{j=1}^{S} P_{ij} \pi^d_k(j), \\
\pi^d_k(i) \geq 0, \sum_{i=1}^{S} \pi^d_k(i) = 1, \sum_{i=1}^{S} V^d_k(i) = 0, \\
\alpha_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n),
\end{cases}
$$

As discussed in Section 4, this system depends on the dimension $S$ of the state space, the number $K$ of populations, and also on the number $D$ of components of the categorical random variable associated to the mixture. This yields a possibly huge non linear problem of size $S \times K \times D$, which is fully coupled, since the parameters $\theta_k$, $k = 1, \ldots, K$, appearing in the cost $F$ depend, through the responsibilities $\gamma_k$, on the whole mixture $\pi$. Moreover, each Hamilton-Jacobi-Bellman equation in the system includes an optimization problem for the optimal controls in the transition matrix $P$.

Following the strategy of the classical EM algorithm, we can mitigate the computational efforts by reducing the above system to $K \times D$ independent subsystems of size $S$. More precisely, we devise the following iterative procedure, in which each iteration is split into two steps: starting from an arbitrary guess $\alpha_k^{(0)}, \pi_k^{(0)}$ for the components of the mixture (4.1), iterate on $h \geq 0$:

E-step: for $k = 1, \ldots, K$ compute the new responsibilities and weights

$$
\gamma_k^{(h+1)}(x_n) = \frac{\alpha_k^{(h)} \pi_k^{(h)}(x_n)}{\sum_{k=1}^{K} \alpha_k^{(h)} \pi_k^{(h)}(x_n)}, \quad \alpha_k^{(h+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n),
$$

and the parameters $\theta_k$ of the components of the mixture

$$
(\theta_k^d)^{(h+1)}(i) = \frac{\sum_{n=1}^{N} \gamma_k^{(h+1)}(x_n)[x_n^d = i]}{\sum_{k=1}^{K} \gamma_k^{(h+1)}(x_n)}, \quad d = 1, \ldots, D, \quad i = 1, \ldots, S.
$$
suitable norm the condition to obtain the new mixture components $k,d,h$

\[ \begin{aligned}
V_k^d(i) &= \min_{P_{ij}} \left\{ \sum_{j=1}^{S} P_{ij} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta_k^d(i+1)) + V_k^d(j) \right) \right\} - \lambda_k^d, \\
\pi_k^d(i) &= \sum_{j=1}^{S} P_{ij} \pi_k^d(j), \\
\pi_k^d(i) &\geq 0, \sum_{i=1}^{S} \pi_k^d(i) = 1, \sum_{i=1}^{S} V_k^d(i) = 0,
\end{aligned} \]

to obtain the new mixture components $\pi_k^{(h+1)}$.

For a fixed tolerance $\tau > 0$, convergence can be checked by evaluating in a suitable norm the condition $\|\theta^{(h+1)} - \theta^{(h)}\| < \tau$. For instance one can reinterpret $\hat{\theta}^{(h)}$ as a vector in $\mathbb{R}^{S \times K \times D}$ and simply take the Euclidean norm.

Note that in this EM-like formulation, the coupling is all embedded in the E-step, whereas the M-step can be completely parallelized. Moreover, each MFG sub-system indexed by $(k,d)$ is also decoupled, since the dependency of $(\theta_k^d)^{(h+1)}$ on $\pi$ is frozen at the previous iteration. Hence, the building block of the algorithm is just to solve the Hamilton-Jacobi-Bellman and the Fokker-Planck equations separately (we remove the indices $k,d,h$ to simplify the notation):

\[ \begin{aligned}
V(i) &= \min_{P_{ij}} \left\{ \sum_{j=1}^{S} P_{ij} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta(i)) + V(j) \right) \right\} - \lambda, \\
\sum_{i=1}^{S} V(i) &= 0, \\
\pi(i) &= \sum_{j=1}^{S} P_{ij} \pi(j), \\
\pi(i) &\geq 0, \sum_{i=1}^{S} \pi(i) = 1.
\end{aligned} \]

For the Hamilton-Jacobi-Bellman equation, we employ a standard policy iteration algorithm. More precisely, starting from a guess for the optimal transition matrix $P^{(0)}$, we introduce an inner iteration $m \geq 0$ by taking $P^{(m)}$ as minimizer:

\[ \begin{aligned}
V^{(m)}(i) &= \sum_{j=1}^{S} P_{ij}^{(m)} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta(i)) + V^{(m)}(j) \right) - \lambda^{(m)}, \\
\sum_{i=1}^{S} V^{(m)}(i) &= 0.
\end{aligned} \]

This results in a very simple linear system of size $S + 1$ for the unknowns $(V^{(m)}(1), \ldots, V^{(m)}(S), \lambda^{(m)})$, whose solution is then plugged back in the optimization problem for $P$ to get $P^{(m+1)}$:

\[ P^{(m+1)} = \arg \min_{P_{ij}} \left\{ \sum_{j=1}^{S} P_{ij} \left( c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta(i)) + V^{(m)}(j) \right) \right\}. \]

Iterations on $m$ are performed up to convergence $\|P^{(m+1)} - P^{(m)}\| < \tau$.

Note that, under the assumptions made on the transition cost, each optimization problem for $P$ is convex with linear constraints $\sum_{i=1}^{S} P_{ij} = 1$, and it can be readily solved with classical algorithms.
Finally, the optimal transition matrix $P$ is plugged in the Fokker-Plank equation, yielding again a simple linear system with linear constraints.

It is important to remark that the solution of the full $K$-populations MFG system is in general not unique, hence the numerical solution obtained with the proposed algorithm depends on the choice of the initial guess $\alpha^{(0)}_k, \pi^{(0)}_k$, for $k = 1, \ldots, K$. In the following experiments we always choose $\alpha^{(0)}_k = 1/K$, while $\pi^{(0)}_k$ is built using random numbers in $(0,1)$. Moreover, we choose the costs $c(p) = -(1-p)/2$ and $F(i, \theta) = |\theta - T_i|^2$, identifying the state $i \in S$ with the vertex of coordinate $T_i = (0, \ldots, 1, \ldots, 0)$ of the simplex in $\mathbb{R}^S$. We also set $\varepsilon = 0.05$ in order to justify the computational efforts of our algorithm, namely ensuring that our solution does not coincide with the explicit one produced by the classical EM algorithm for $\varepsilon = 0$. The case of more general costs, and an experimental analysis on how these choices affect the resulting clusterization, will be addressed in a more computational oriented work.

We now present the numerical results for some classical examples. We first consider the case of Bernoulli mixtures, i.e. $S = 2$, taking as dataset the MNIST database of handwritten digits [20], see Figure 1. We recall that the database contains 60000 images of the digits $\{0, \ldots, 9\}$, each composed by $28 \times 28$ pixels in 256 grey levels, that we turn (via hard-threshold) in monochrome images and represent by binary vectors of size $D = 784$. Moreover, we remark that each sample in the database is already labelled by the number of the corresponding digit, that we use to check the correctness of the clusterization. To this end, given $1 \leq K \leq 10$, we select $K$ digits $d_1, \ldots, d_K$ in $\{0, \ldots, 9\}$, and we run our algorithm to compute the Bernoulli parameters $\mu_k$ of the mixture components, for $k = 1, \ldots, K$. Then we introduce a matrix $H$ of size $K \times K$, whose entries are obtained as follows. For each sample $x$ of type $d_k$, we compute the corresponding responsibilities $\gamma_1(x), \ldots, \gamma_K(x)$, and we accumulate these values in the $k$-th row of $H$, normalizing their sum with respect to the number of samples in $d_k$. More precisely, for $k,j = 1, \ldots, K$

$$H_{kj} = \frac{1}{|d_k|} \sum_{x \in d_k} \gamma_j(x)$$

provides the averaged probability, for a sample of type $d_k$, of belonging to the cluster $j$. Up to a permutation of the rows, namely a reordering of the mixture components, we can always assume that the maximal values of $H$ correspond to the diagonal entries, so that digits of type $d_k$ will belong, with the highest probability, to the cluster $k$. In a perfect clusterization, $H$ is clearly the identity matrix, but we recall that each Bernoulli distribution $\pi_k$ is built as a joint probability of all the observed pixel values (3.2). Since the samples of type $d_k$ can be very different from each other, and also share some similarity with samples of other types (see Figure 2), we can never expect such a sharp partition.

For visualization purposes we represent each row of the matrix $H$ as a histogram, reporting on the $x$-axis the type $d_k$, for $k = 1, \ldots, K$, and assigning $K$ different colors to the values $H_{kj}$ for $j = 1, \ldots, K$, corresponding to the $K$ clusters $C_1, \ldots, C_K$. Moreover, since the state space dimension is $S = 2$, we
can conveniently represent the parameters $\mu_k \in [0, 1]^D$ of the corresponding Bernoulli distributions in the mixture, for $k = 1, \ldots, K$, as grey scale images.

![Figure 2: Different samples of hand-written digits from the MNIST database](image)

Let us start with the case $K = 2$, choosing the digits 1, 3. In Figure 3, we show the clusterization histogram and the corresponding Bernoulli parameters, observing that the samples of the two digits are very well separated.

![Figure 3: Clusterization histogram for digits 1, 3 and the corresponding Bernoulli parameters.](image)

We now choose the digits 3, 5, and we show the results in Figure 4. In this case, we observe that the clusterization is slightly ambiguous, since, in average, the samples of the two types are more similar to each other (see the corresponding Bernoulli parameters). In the histogram we clearly see a repartition of about 60% and 40% (and vice versa) between the two clusters.

We finally consider the case $K = 5$ with even digits 0, 2, 4, 6, 8. In Figure 5, we observe that the chosen digits are, in average, different from each other,
Figure 4: Clusterization histogram for digits 3, 5 and the corresponding Bernoulli parameters.

so that they are quite well clusterized.

Figure 5: Clusterization histogram for even digits and the corresponding Bernoulli parameters.

It is worth noting that the greatest error in the cluster assignment correspond to digit 6, which indeed shares about 15% of its samples with the cluster of the digit 2. Looking at the corresponding Bernoulli parameters, we readily see that the two images have very similar vertical alignments, and also a quite large overlapping bottom region. In particular, the Bernoulli parameter for the digit 2 is visibly more diffused, and this reflects the inhomogeneity of the corresponding
samples. Similar considerations also apply to the pairs of digits $2, 4$ and $4, 8$, with errors in the cluster assignment below $10\%$.

We now consider the case of categorial distributions, i.e. $S > 2$, taking as dataset the Fashion-MNIST database [21], see Figure 6.

![Figure 6: Samples of fashion products from the Fashion-MNIST database](image)

The database contains 60000 images of ten categories of fashion products, namely \{T-shirt, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Boot\}, each composed by $28 \times 28$ pixels in 256 grey levels. We turn these images (via hard-thresholding) in images with $S$ grey levels, and represent them by vectors in $S^D$, where $S = \{1, \ldots, S\}$ and $D = 784$. As in the previous tests, we use the label associated to each sample in the database to build the clusterization histogram. Moreover, we visualize the parameters $\theta_k(i) \in [0, 1]^D$ of the corresponding categorial distributions in the mixture, for $i = 1, \ldots, S$ and $k = 1, \ldots, K$, as grey scale images, i.e. averaging the values with respect to $i$ as $\bar{\theta}_k = \frac{1}{S} \sum_{i=1}^{S} \theta_k(i)$.

In this example, clusterization is very challenging. To give an idea of the issues, we set $S = 32$ and compute separately ten “ideal” categorial distributions associated to the dataset, by simply averaging the pixel values of all the samples of a same type. The result is shown in Figure 7.

![Figure 7: Averaged categorical distributions for the Fashion-MNIST database](image)
We clearly see that types **Pullover**, **Coat** and **Shirt** are almost indistinguishable. Moreover, they all have a very large overlap region with the type **T-shirt**, but also with the types **Trouser** and **Dress**. Finally, we observe that the samples of type **Sandal** are so different from each other that the corresponding distribution is completely smoothed out. As one can expect, such drawbacks dramatically affect the quality of the clusterization, but they also suggest how the present model could be improved, for instance including in the cost functions some geometric correlation between the image pixels. This direction of research is currently under development.

We conclude this section with the following tests, which confirm the above considerations. We set $K = 2$ and choose the types **T-shirt** and **Trouser**. In Figure 8, we show the resulting clusterization histogram and the corresponding categorical parameters.

![Clusterization histogram for types T-shirt and Trouser](image)

**Figure 8**: Clusterization histogram for types **T-shirt**, **Trouser** and the corresponding categorical parameters.

Despite the clusterization is still "acceptable" (in average more than 60% of samples correctly assigned to the corresponding clusters), we clearly observe a weird mixing of the two types. On the other hand, choosing types which are substantially different from each other, we end up with a good clusterization. This is the case for the example shown in Figure 9, where we set $K = 4$ and choose the types **Dress**, **Sneaker**, **Bag** and **Boot**.

![Clusterization histogram for types Dress, Sneaker, Bag and Boot](image)

**Figure 9**: Clusterization histogram for types **Dress**, **Sneaker**, **Bag** and **Boot**.
A two-states stationary MFG system

In this section, we study some properties of a particular MFG system we need for the analysis performed in Section 4. In this case, the cost in the Hamilton-Jacobi-Bellman equation does not depend on the distribution of the population, hence the coupling term is only in the Fokker-Planck equation which characterizes the stationary distribution. Here we mainly follow the notations in [10].

Given a matrix $P \in \mathbb{S}^S$, we denote with $P_i$ the $i$-row of $P$. For $\varepsilon > 0$, we consider MFG system

\[
\begin{align*}
V(i) &= \min_{P_i} \left\{ \sum_{j=1}^S P_{ij} (c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta) + V(j)) \right\} - \lambda \\
\pi(i) &= \sum_{j=1}^S P_{ij} \pi(j) \\
\pi(i) &\geq 0, \sum_{i=1}^S \pi(i) = 1, \sum_{i=1}^S V(i) = 0
\end{align*}
\]

for $i = 1, \ldots, S$, where $\varepsilon > 0$ and the components of the vector $\theta = (\theta(1), \ldots, \theta(S)) \in \mathbb{S}$ are fixed parameters and the transition matrix $P$ in the Fokker-Planck equation is composed of the rows

\[
P_i = \{P_{ij}\}_{j=1}^S
\]

which realize the minimum in the Hamilton-Jacobi-Bellman equation. The transition cost from the state $i$ to the state $j$ is given by

\[
C^c(P_{ij}, \theta) = c(P_{ij}) + \varepsilon \log(P_{ij}) + F(i, \theta), \quad i, j = 1, \ldots, S.
\]

We assume that the cost function $c \in C^1([0, 1])$ with $pc(p)$ convex for $p \in [0, 1]$ and $F : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ is such that $F(i, \cdot)$ is bounded and continuous for all $i \in \mathbb{S}$. 

---

**Figure 9:** Clusterization histogram for types **Dress, Sneaker, Bag, Boot** and the corresponding categorical parameters.
The average cost in the state $i$ for a given choice of the transition matrix $P$ is defined by

$$e_i(P,V) = \sum_{j=1}^{S} (C^{\ast}(P_{ij},\theta) + V(j)) P_{ij} \quad i = 1,\ldots,S,$$

and we denote with $e(P,V)$ the corresponding $S$-dimensional vector. Note that $e_i(P,V)$ depends only on the $i$-th row of the matrix $P$.

We introduce the definition of Nash minimizer for $e(P,V)$ (see [10, Definition 1]). Given the stochastic matrix $P \in \mathcal{S}^S$ and a probability vector $q \in \mathcal{S}$, we denote with $\mathcal{R}(P,q,i)$ the stochastic matrix obtained by replacing the $i$-th row of $P$ with the vector $q$.

**Definition A.1.** Given a cost vector $V \in \mathbb{R}^S$, a stochastic matrix $P$ is said to be a Nash minimizer for $e(P,V)$ if for each $i = 1,\ldots,S$, $q \in \mathcal{S}$, it holds

$$e_i(P,V) \leq e_i(\mathcal{R}(P,q,i),V).$$

**Proposition A.2.** For each vector $V \in \mathbb{R}^S$, there exists a unique Nash minimizer $P$ for $e(P,V)$.

**Proof.** As proved in [10, Theorems 1], existence of a Nash minimizer $P$ follows by the continuity and convexity of the cost $e_i(P,V)$ with respect to the vector $P_i$, $i = 1,\ldots,S$. Moreover, a straightforward computation gives that the function $g : \mathcal{S} \rightarrow \mathcal{S}$, defined by $g_{ij}(P) = \partial e_i(P,V) / \partial P_{ij}$, is diagonally convex, i.e. for all $P^1, P^2 \in \mathcal{S}^S$, $P^1 \neq P^2$, it holds

$$\sum_{i,j}(P^1_{ij} - P^2_{ij})(g_{ij}(P^1) - g_{ij}(P^2)) > 0.$$ 

Indeed, for any $i, j \in \{1,\ldots,S\}$ and $P \in \mathcal{S}^S$, we have

$$\frac{\partial e_i(P,V)}{\partial P_{ij}} = P_{ij}(c'(P_{ij}) + \varepsilon \frac{1}{P_{ij}}) + c_{ij}(P_{ij}) + \varepsilon \log(P_{ij}) + V(j) + F(i,\theta).$$

Hence,

$$(P^1_{ij} - P^2_{ij})(g_{ij}(P^1) - g_{ij}(P^2)) = \varepsilon(P^1_{ij} - P^2_{ij}) \log \left( \frac{P^1_{ij}}{P^2_{ij}} \right)$$

$$+ P^1_{ij}c(P^1_{ij}) - P^2_{ij}c(P^2_{ij}) - (P^2_{ij}c'(P^2_{ij}) + c(P^2_{ij}))(P^1_{ij} - P^2_{ij})$$

$$+ P^2_{ij}c(P^2_{ij}) - P^1_{ij}c(P^1_{ij}) - (P^1_{ij}c'(P^1_{ij}) + c(P^1_{ij}))(P^2_{ij} - P^1_{ij})) > 0,$$

because of the monotonicity of $\ln(p)$ and the convexity of $pc(p)$ for $p \in [0,1]$.

By [10, Theorems 2], we get the uniqueness of the Nash minimizer.

We now prove that $\pi$ has positive mass for any state $i = 1,\ldots,S$. This is a crucial result for the existence of a solution to the multi-population MFG system studied in Section 4. We need some preliminary results.
Lemma A.3. For any vector $\theta \in \mathbb{S}$, there exists a unique solution to $(\lambda_\theta, V_\theta)$ to the Hamilton-Jacobi-Bellman equation

\[
\begin{cases}
V(i) = \min_{P_i} \left\{ \sum_{j=1}^{S} \left( C^\varepsilon((P_{ij}, \theta)) + V(j) \right) P_{ij} \right\} - \lambda \\
\sum_{i=1}^{S} V(i) = 0.
\end{cases}
\tag{A.2}
\]

Moreover, if $\theta \to \bar{\theta}$, then $(\lambda_\theta, V_\theta)$ converges to $(\lambda_{\bar{\theta}}, V_{\bar{\theta}})$.

Proof. Existence and uniqueness of a solution to (A.2) follows from [10, Proposition 8 and Theorem 4]. Indeed, since the cost in the Hamilton-Jacobi-Bellman equation is independent of the distribution $\pi$, the assumptions of Proposition 8 and Theorem 4 in [10] are trivially satisfied. The convergence of $(\lambda_\theta, V_\theta)$ to $(\lambda_{\bar{\theta}}, V_{\bar{\theta}})$ follows by the continuity of the costs $e_i(P, V)$ with respect to $\theta$ and the uniqueness of the solution to (A.2).

Proposition A.2 and the continuity of the average cost $e(P, V)$ implies the continuity of the Nash minimizer (see [10, Proposition 1]).

Lemma A.4. The Nash minimizer $P(V)$ is a continuous function of $V$.

Theorem A.5. For any vector $\theta \in \mathbb{S}$, there exists a unique solution $(V_\theta, \lambda_\theta, \pi_\theta) \in \mathbb{R}^S \times \mathbb{R} \times \mathbb{S}$ to (A.1). Moreover, there exists a positive constant $c(\varepsilon) \in (0, 1)$ such that, for each $\theta \in \mathbb{S}$ and for each $i = 1, \ldots, S$

\[
\pi_\theta(i) \geq c(\varepsilon) > 0.
\]

(A.3)

Proof. Note that, because of the particular structure of the system (A.1), Propositions A.2 and A.3 immediately ensure existence and uniqueness of a solution to (A.1).

To show (A.3), we argue by contradiction assuming that there exists a sequence $\theta_n \in \mathbb{S}$ such that $\pi_{\theta_n}(i_n) \to 0$ for $n \to \infty$. Since the state space is finite we assume that, up to a subsequence, $i_n = i$ for any $n$. Let $\theta \in \mathbb{S}$ be such that, up to a subsequence, $\theta_n \to \theta$. By Lemma A.3 and A.4, it follows that $P(V_{\theta_n}) \to P(V_{\theta})$, where $P(V_{\theta_n})$ and $P(V_{\theta})$ are the Nash minimizers corresponding to the solution of (A.2) with $\theta_n$ and, respectively, with $\theta$. By the equation

\[
\pi_{\theta_n} = \pi_{\theta_n} P(V_{\theta_n}),
\]

satisfied by the invariant distribution associated to $P(V_{\theta_n})$, we get that, up to a subsequence, $\pi_{\theta_n}$ converges to $\bar{\pi} \in \mathbb{S}$ satisfying

\[
\bar{\pi} = \bar{\pi} P(V_{\theta}),
\]

with $\bar{\pi}(i) = 0$. But, because of entropy regularization term, all the entries $P_{ij}$ of the transition matrix $P(V_{\theta})$ are strictly positive, hence $P$ is a positive matrix and the associated invariant distribution satisfies $\bar{\pi}(i) > 0$ for each $i = 1, \ldots, S$. By the contradiction, we obtain (A.3).
References

[1] Aquilanti, L.; Cacace, S.; Camilli, F.; De Maio, R. A mean field games approach to cluster analysis. Applied Math. Optim., DOI: 10.1007/s00245-019-09646-2.

[2] Bertucci, C.; Vassilaras, S.; Lasry, J.-M.; Paschos, G.; Debbah, M.; Lions, P.-L. Transmit strategies for massive machine-type communications based on Mean Field Games. *Proceedings of the International Symposium on Wireless Communication System (ISWCS)*, 2018.

[3] Bilmes, J.A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov model. Technical Report ICSI-TR-97-021, University of Berkeley, 2000.

[4] Bishop, C.M. *Pattern recognition and Machine Learning*. Information Science and Statistics, Springer, New York, 2006.

[5] Carmona, R; Lauriere, M. Convergence analysis of Machine Learning algorithms for the numerical solution of mean field control and games: I – the ergodic case. arXiv:1907.05980, 2019.

[6] Coron, J.L. *Quelques exemples de jeux à champ moyen*. Ph.D. thesis, Université Paris-Dauphine, 2018.

[7] E, W., Han, J.; Li, Q. A mean-field optimal control formulation of deep learning. Res. Math. Sci. 6 (2019), 41p.

[8] Everitt, B. S.; Hand, D.J. *Finite mixture distributions*. Monographs on Applied Probability and Statistics. Chapman & Hall, London-New York, 1981.

[9] Gomes, D.A.; Saúde, J. Mean field games models—a brief survey. Dyn. Games Appl. 4 (2014), no. 2, 110–154.

[10] Gomes, D.A.; Mohr, J.; Souza, R.R. Discrete time, finite state space mean field games. J. Math. Pures Appl. 93 (2010), no. 3, 308-328.

[11] Huang, M. Y.; Malham, R. P.; Caines, P. E. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. Commun. Inf. Syst. 6 (2006), no. 3, 221-252.

[12] Lasry, J.-M.; Lions, P.-L. Mean Field Games. Jpn. J. Math., 2 (2007), 229-260.

[13] McLachlan, G.; Peel, D. *Finite mixture models*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York, 2000.

[14] Pearson, K. Contributions to the mathematical theory of evolution. Philosophical Transaction of the Royal Society, A 185 (1894), 71-110.
[15] Pequito, S.; Aguiar, A.P.; Sinopoli, B.; Gomes, D.A. Unsupervised learning of finite mixture models using Mean Field Games, in Annual Allerton Conference on Communication, Control and Computing, 2011, 321-328.

[16] Tarter, M.E.; Lock, M.D. Model-free curve estimation. Monographs on Statistics and Applied Probability, 56. Chapman & Hall, New York, 1993.

[17] Titterington, D.M.; Smith A.F.M.; Makov, U.E. Statistical analysis of finite mixture distributions. Wiley Series Probability and Mathematical Statistic, 1985.

[18] Wedel, M., and Kamakura, W.A. Market segmentation: conceptual and methodological foundations. Kluwer Academic Press, Dordrecht, 1998.

[19] Everitt, B.S; Landau, S.; Leese, M.; Stahl, D. Cluster Analysis Wiley Series in Probability and Statistics, 2011.

[20] http://yann.lecun.com/exdb/mnist/

[21] https://github.com/zalandoresearch/fashion-mnist

laura.aquilanti@sbai.uniroma1.it
fabio.camilli@sbai.uniroma1.it
SBAI, Sapienza Università di Roma
via A.Scarpa 14, 00161 Roma (Italy)

cace@mat.uniroma3.it
Dipartimento di Matematica e Fisica
Università degli Studi Roma Tre
Largo S. L. Murialdo 1, 00146 Roma (Italy)

r.demaio@iconsulting.biz
IConsulting
Via della Conciliazione 10, 00193 Roma (Italy)