Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Reinforcement Learning Algorithm for Mixed Mean Field Control Games

Andrea Angiuli [*] [1], Nils Detering [†] [2], Jean-Pierre Fouque [‡] [2], Mathieu Laurière [§] [3], and Jimin Lin [¶] [2]

[1]Prime Machine Learning Team, Amazon. 320 Westlake Ave N, SEA83, Seattle, WA, 98109, USA. The work presented here does not relate to this author's position at Amazon.
[2]Department of Statistics and Applied Probability, South Hall, University of California, Santa Barbara, CA 93106, USA.
[3]NYU-ECNU Institute of Mathematical Sciences, and Center for Data Science and Artificial Intelligence, New York University Shanghai, 200122, China.

Abstract. We present a new combined mean field control game (MFCG) problem which can be interpreted as a competitive game between collaborating groups and its solution as a Nash equilibrium between groups. Players coordinate their strategies within each group. An example is a modification of the classical trader's problem. Groups of traders maximize their wealth. They face cost for their transactions, for their own terminal positions, and for the average holding within their group. The asset price is impacted by the trades of all agents. We propose a three-timescale reinforcement learning algorithm to approximate the solution of such MFCG problems. We test the algorithm on benchmark linear-quadratic specifications for which we provide analytic solutions.

Keywords:
Mean Field Control Games,
Reinforcement Learning,
Q-Learning,
Optimal Liquidation.

Article Info.:
Volume: 2
Number: 2
Pages: 108 - 137
Date: June/2023
doi.org/10.4208/jml.220915

Article History:
Received: 15/09/2022
Accepted: 05/04/2023

Communicated by:
Jiequn Han

1 Introduction

Mean field approaches are based on the idea that the main properties of large coupled systems of entities (e.g. agents, players, or particles) can be described by the distribution of one representative entity. To answer many questions related to the system, it is not required to know the individual states of all entities but only the distribution of their representative. This reduces significantly the complexity of large systems.

Mean field approaches were first introduced in the context of statistical physics where propagation of chaos among particles was studied. Under mild assumptions, in a system of particles described by a large system of diffusion processes, the location of one particle becomes independent of the others as the size of the system grows [29]. In the following we think of the entities of the system being agents or players and we have mainly financial

[*]aangiuli@amazon.com
[†]Corresponding author. detering@pstat.ucsb.edu
[‡]fouque@pstat.ucsb.edu
[§]mathieu.lauriere@nyu.edu
[¶]jiminlin@pstat.ucsb.edu

applications in mind. Mean field ideas have later been adapted to differential games with large number of agents in the cooperative setting (mean field control, MFC), and in the competitive framework (mean field games, MFG) [8, 12, 25]. MFC and MFG problems arise in a number of applications ranging from engineering to economics. Mean field type games (MFTG) [18] are games with a finite number of players who are of mean field type, i.e., their dynamics and cost functions may depend on their own distribution.

Recently numerical solution of MFC and MFG problems has received greater attention [1, 7, 11, 21, 22, 26, 30]; see e.g. [2] for a survey. Classical methods of optimization theory have been complemented by deep neural networks [14,15,20,23,24] and by Reinforcement Learning (RL) approaches which aim at calculating optimal strategies without the precise knowledge of the underlying model [16, 17, 19, 27, 28, 31].

In [4], a unified reinforcement Q-learning algorithm is proposed to solve MFG and MFC problems based on the ratio of two learning rates, one for the decision Q-matrix and the other for the distribution of the population. In the present paper, we argue that this algorithm can be adapted for solving a new class of mean field control game (MFCG) problems arising naturally in the context of many large groups where agents are cooperating within each group but in competition with all agents in other groups. In this type of games, a MFC problem is defined at each group level motivating the dependency on the groups' distribution of the agents. At the full system level, a MFG problem is defined between groups explaining the freezing of the full system distribution and the following fixed point problem typical of this framework. Our algorithm naturally involves three learning rates: a fast one for the distribution of the group, a medium one for the agent's Q-matrix, and a slow one for the distribution of the overall population. We illustrate its performance on linear-quadratic examples for which we derive explicit solutions for the optimal strategy.

In [5], the unified reinforcement Q-learning algorithm proposed in [4] is generalized to finite horizon extended MFC and MFG problems. It is applied to the problem of a trader who wants to minimize transaction and inventory costs when trading an asset impacted by all agents' trades. We show in this paper that the algorithm can be naturally adapted for solving MFCG when both the distributions of states and controls (for the group and for the overall population) are involved.

In Section 2.1, we motivate the introduction of the new MFCG problem in the classical context of discrete time, finite horizon, differential games in discrete state and action spaces. Agents control their drifts and minimize an expected cost which may depend on the distributions of their own group and of the entire population. We give an intuitive justification of the fact that the solution of the MFCG provides an approximate Nash equilibrium between groups.

In Section 2.2 we introduce the discrete time and space infinite horizon MFCG considered in this paper. We focus on the asymptotic formulation of the problem introduced in [4] where comparisons with time-dependent and stationary formulations were discussed. In Section 2.3, we generalize the results of [5] regarding finite time extended MFC and MFG problems to the MFCG framework.

The Q-learning approach to solve these problems is described in Section 3 where we state the Bellman equation for the optimal action-value function (Q-matrix), and we intro-

duce its three timescales stochastic approximation based on well-separated three learning rates: one for the states' distribution of the group, one for the action-value Q-matrix, and one for the states' distribution of the overall population.

Algorithm and learning rates are presented in Section 4. Its performance on a linear-quadratic benchmark are shown in Section 5.1. Section 5.2 illustrates the results on the trader's problem, an example where the states' distribution of the group and the controls' distribution of the overall population appear in the objective function of the agent. We compare the strategies learned by our algorithm with the theoretical solutions provided in the Appendix B.

2 Mean field control games

2.1 Motivation

In order to introduce our notion of MFCGs, we consider the familiar context of discrete time finite horizon differential games between agents evolving in a finite state space $\mathcal{X}$ by taking actions in a finite action space $\mathcal{A}$. The population is made of M groups, each of size N. An agent will be indexed by a pair (m,n) where the first index $m = 1,\dots,M$ indicates her group and index $n = 1,\dots,N$ being her identifier in the group. Agents are collaborating within their groups and competing with all agents of other groups. In other words, all N agents of group m will try to collectively minimize the total cost of group m. Between groups, agents play a Nash equilibrium. So every single agent (m,n) interacts, possibly in different ways, on both the distribution within its group and the distribution within the whole population.

We now present the general model that we consider, starting with the dynamics. At time $t = 0,1,\dots,T-1$, agent (m,n) uses the control $\alpha_t^{m,n} \in \mathcal{A}$. The evolution of her state is given by: $X_0^{m,n} \sim \mu_0$ and for $t = 0,1,\dots,T-1$,

$$\mathbb{P}\left(X_{t+1}^{m,n} = x' | X_t^{m,n} = x, \alpha_t^{m,n} = a, \mu_t = \mu\right) = p(x'|x,a,\mu),$$

where x and $x' \in \mathcal{X}$ represent respectively current and next state, $a \in \mathcal{A}$ is the action taken, and $\mu \in \Delta^{|\mathcal{X}|}$ represents the empirical distribution of the whole population. Here, $p : \mathcal{X} \times \mathcal{A} \times \Delta^{|\mathcal{X}|} \to \Delta^{|\mathcal{X}|}$ is a transition kernel interpreted also as a function

$$p : \mathcal{X} \times \mathcal{X} \times \mathcal{A} \times \Delta^{|\mathcal{X}|} \to [0,1], \quad (x,x',a,\mu) \mapsto p(x'|x,a,\mu),$$

which provides the probability to jump to state x' from state x if action a is taken. We assume that this transition kernel depends on the global distribution μ but not on the local distribution μ^m (that is the distribution of the agent's group defined below). In this finite-horizon setting, we allow for time-dependent feedback Markovian controls $\alpha : \{0,1,\dots,T-1\} \times \mathcal{X} \to \mathcal{A}$ that depend only on time and the state. So if agent (m,n) uses control α, then $\alpha_t^{m,n} = \alpha(t, X_t^{m,n})$.

Considering the behavior of other groups as fixed, the goal for agents of group m is to minimize the expected cost of the group

$$J^m(\boldsymbol{\alpha}) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\sum_{t=0}^{T} f\left(t, X_t^{m,n}, \alpha_t^{m,n}, \mu_t, \mu_t^m\right) + g\left(X_T^{m,n}\right)\right],$$

where f is a running cost which may depend on the empirical distribution of the full population $\mu_t = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\delta_{X_t^{m,n}}$ referred to as the global population, and on the empirical distribution $\mu_t^m = \frac{1}{N}\sum_{n=1}^{N}\delta_{X_t^{m,n}}$ of the group m referred to as the local population. The terminal cost g could as well depend on these empirical distributions at terminal time. Note that in this setting agents of group m interact with agents of other groups through the distribution of the global population appearing in the cost.

Additional assumptions on f and g are needed, but we may keep in mind the simple quadratic cost case with, for example,

$$f(t,x,\alpha,\mu,\tilde{\mu}) = \frac{1}{2}\alpha^2 + \frac{c_1}{2}(x-\bar{\mu})^2 + \frac{c_2}{2}\bar{\tilde{\mu}}^2,$$

where $\bar{\mu}$ and $\bar{\tilde{\mu}}$ denote respectively the means of the global population μ and the local population $\tilde{\mu}$. The first term is the classical quadratic cost for controlling the drift, the second term is an incentive to stay close to the global mean and the third term is a group incentive to keep the local mean close to zero. For simplicity we assume a zero terminal cost ($g = 0$) in this example. We will revisit a linear-quadratic (LQ) continuous-time variant of this setting in Section 5.1. The key point is that the interaction through the global mean is of mean field game (MFG) competitive nature, while the interaction through the local mean is of mean field control (MFC) collaborative nature, motivating the name mean field control game (MFCG). In other words, this problem is a competition between M coalitions of N players, all the players being identical in the sense that they have similar dynamics and cost functions. The explicit solution for a continuous time and space version of this finite-player MFCG is given in Appendix A.

Passing to the mean field limit $M \to \infty, N \to \infty$ in a sense made precise in Appendix A.3, a representative agent faces the following problem. Given a sequence of probability distributions $\boldsymbol{\mu} = (\mu_t)_{0\le t\le T}$, the goal is to solve the McKean-Vlasov (MKV) control problem of finding a minimizer $\hat{\alpha}$ for

$$J(\alpha) = \mathbb{E}\left[\sum_{t=0}^{T} f\left(t, X_t^{\alpha,\boldsymbol{\mu}}, \alpha_t, \mu_t, \mathcal{L}\left(X_t^{\alpha,\boldsymbol{\mu}}\right)\right) + g\left(X_T^{\alpha,\boldsymbol{\mu}}\right)\right],$$

subject to

$$X_{t+1}^{\alpha,\boldsymbol{\mu}} \sim p\left(X_t^{\alpha,\boldsymbol{\mu}}, \alpha_t, \mu_t\right), \quad t = 0, 1, \ldots, T-1, \quad X_0^{\alpha,\boldsymbol{\mu}} \sim \mu_0.$$

Allowing for time-dependent feedback Markovian controls means that α_t is given in the form $\alpha_t = \alpha(t, X_t^{\alpha,\boldsymbol{\mu}})$ for some control function α. Then, to find the Nash equilibrium, we need to solve the fixed point compatibility condition

$$\mu_t = \mathcal{L}\left(X_t^{\hat{\alpha},\boldsymbol{\mu}}\right), \quad \forall t \in \{0, 1, \ldots, T\},$$

where $\mathcal{L}(X)$ denotes the law of the random variable X. This problem can be viewed as an MFG in which each player is of McKean-Vlasov type, in the sense that her dynamics

and her cost function depend on her own distribution. As such, this can correspond to the limit of a MFTG [18] when the number of players goes to infinity. Solving this MFCG is justified by showing that the control $\hat{\alpha}$ enables the agents in the mixed finite-player game to achieve an ϵ-Nash equilibrium. This argument is developed in the Appendix A for the LQ example.

The proof of this result in a general setting will be presented in the companion paper [6] in preparation. In particular, the analysis covers the case where the global distribution is involved in the dynamics. However, proving convergence when the local distribution appears in the dynamics is more challenging, which is why we do not include it in the dynamics studied in the present work. The algorithm presented in this paper is in the context of finite state and action spaces. A version for continuous spaces based on a Deep Learning Actor Critic algorithm is a work in progress [3].

2.2 Asymptotic formulation

In this section we present the discrete time infinite horizon setting and we consider the asymptotic formulation of the game introduced in [4]. Our model involves the distribution of states within the collaborative agent's group (also called local distribution), and the distribution of states of the overall competitive population (also called global distribution).

We allow for time homogeneous controls $\alpha : \mathcal{X} \to \mathcal{A}$ that depend only on the state. We denote by μ^α the asymptotic (long time) distribution of the controlled process following the strategy α which we assume to exist and to be unique (the state space being finite, aperiodicity and irreducibility of the discrete time process ensure these properties).

We go from finite horizon to infinite horizon so that the problem will be simpler to tackle with RL and we will look for stationary policy, see Section 2.3. Given a cost function f defined on $\mathcal{X} \times \mathcal{A} \times \Delta^{|\mathcal{X}|} \times \Delta^{|\mathcal{X}|}$ and a discount rate $\gamma < 1$, we now consider the following infinite horizon asymptotic MFCG problem:

Find a strategy $\hat{\alpha}$ and a distribution $\hat{\mu}$ such that:

1. (best response) $\hat{\alpha}$ is the minimizer of

$$J(\alpha; \hat{\mu}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t f\left(X_t^{\alpha,\hat{\mu}}, \alpha\big(X_t^{\alpha,\hat{\mu}}\big), \hat{\mu}, \mu^{\alpha,\hat{\mu}}\right)\right],$$

   where $X_0^{\alpha,\hat{\mu}} \sim \mu_0$ and for $t = 0, 1, \ldots,$

$$\mathbb{P}\left(X_{t+1}^{\alpha,\hat{\mu}} = x' | X_t^{\alpha,\hat{\mu}} = x, \alpha\big(X_t^{\alpha,\hat{\mu}}\big) = a, \mu = \hat{\mu}\right) = p(x'|x, a, \hat{\mu})$$

   and $\mu^{\alpha,\hat{\mu}} = \lim_{t\to\infty} \mathcal{L}(X_t^{\alpha,\hat{\mu}})$.

2. (fixed-point) $\hat{\mu} = \lim_{t\to\infty} \mathcal{L}(X_t^{\hat{\alpha},\hat{\mu}}) = \mu^{\hat{\alpha},\hat{\mu}}$.

In order to make sense of the above problem statement we have to restrict to actions $\alpha : \mathcal{X} \to \mathcal{A}$ which are such that the controlled process $X_t^{\alpha,\mu}$ has a limiting distribution,

i.e., $\lim_{t\to\infty} \mathcal{L}(X_t^{\alpha,\mu})$ exists. For a finite state Markov chain this is the case if $(X_t^{\alpha,\mu})_{t\in\mathbb{N}}$ is irreducible and aperiodic. We therefore assume that the strategy $\hat{\alpha}$ is the minimizer over all strategies such that $(X_t^{\alpha,\hat{\mu}})_{t\in\mathbb{N}}$ is irreducible and aperiodic.

Remark 2.1. We could have also considered a classical formulation of MFCG where $\hat{\mu}$ and $\mu^{\alpha,\hat{\mu}}$ are flows of distributions $(\hat{\mu}_t)_{t\in\mathbb{N}}$ and $(\mu_t^{\alpha,\hat{\mu}})_{t\in\mathbb{N}}$ in which case the fixed-point requirement is $\hat{\mu}_t = \mathcal{L}(X_t^{\alpha,\hat{\mu}})$ for every $t \in \mathbb{N}$, and where the strategy is time-dependent. As well, we could have considered a stationary formulation of MFCG where μ^{α} is the stationary distribution of the controlled process, equal to $\hat{\mu}$ in the fixed-point step. As in [4], it can be shown that the optimal strategies for the asymptotic and stationary problems coincide, and they coincide with the limiting optimal strategy (as $t \to \infty$) of the classical formulation.

2.3 Finite horizon extended formulation

Following [5], we generalize the MFCG problem and its reinforcement learning algorithm to the case with a discrete time finite horizon T, on a finite state space, and mean field of state and control. The state-action space is as described in Section 2. The state follows a random evolution in which X_{t+1} is determined as a function of the current state X_t, the action α_t, and some noise. We introduce the transition probability function $p(x'|x,a,\nu)$, $(x,x',a,\nu) \in \mathcal{X}\times\mathcal{X}\times\mathcal{A}\times\Delta^{|\mathcal{X}\times\mathcal{A}|}$, which provides the probability to jump to state x' given its current state x, the action taken a and the global population distributed as ν. We assume no dependence on the state-action group distribution $\tilde{\nu}$ in order to apply the MKV Bellman equation introduced in [4]. For simplicity, we consider the homogeneous case where this function does not depend on time. Restoring this time-dependence if needed is a straightforward procedure.

We now consider the MFCG cost function given by: for $\nu = (\nu_t)_{t=0,1,\dots,T}$

$$J(\alpha;\nu) = \mathbb{E}\left[\sum_{t=0}^{T-1} f\big(X_t^{\alpha,\nu},\alpha_t,\nu_t,\nu_t^{\alpha,\nu}\big) + g\big(X_T^{\alpha,\nu},\mu_T,\mu_T^{\alpha,\nu}\big)\right],$$

where μ_T (resp. $\mu_T^{\alpha,\nu}$) is the first marginal of ν_T (resp. $\nu_T^{\alpha,\nu}$). Again, for simplicity, we assume that f does not depend on time. The process $X^{\alpha,\nu}$ has a given initial distribution $\mu_0 \in \Delta^{|\mathcal{X}|}$ and follows the dynamics:

$$\mathbb{P}\left(X_{t+1}^{\alpha,\nu} = x' | X_t^{\alpha,\nu} = x, \alpha_t = a, \nu_t = \nu\right) = p(x'|x,a,\nu).$$

3 Q-learning

3.1 Action-value function

Our algorithm to solve the MFCG is based on the concept of Q-learning which is a well known procedure to solve Markov decision problems. However, following [4] we combine the idea of Q-learning with the model agnostic view of reinforcement learning. We first

adapt the Q-learning concepts to our problem at hand. Since the local distribution is not fixed and depends on the control itself, we have to adapt the classical Q-learning in the spirit of [4]. For an admissible control $\alpha : \mathcal{X} \to \mathcal{A}$ and a pair $(x, a) \in \mathcal{X} \times \mathcal{A}$, we define the new control $\alpha_{x,a}$ by

$$\alpha_{x,a}(x') = \begin{cases} a, & \text{if} \quad x' = x, \\ \alpha(x'), & \text{otherwise.} \end{cases} \tag{3.1}$$

Given a global measure μ and a strategy α, the Q-function for our problem is given by

$$Q_\mu^\alpha(x,a) = f(x,a,\mu,\mu^{\alpha_{x,a},\nu}) + \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t f\left(X_t, \alpha\big(X_t^{\alpha,\mu}\big), \mu, \mu^{\alpha,\nu}\right) | X_0^{\alpha,\mu} = x, A_0 = a\right],$$

where $\mu^{\alpha_{x,a},\nu}$ is the local distribution relative to the strategy $\alpha_{x,a}$. The optimal function in the sense of minimizing cost is given by

$$Q_\mu^*(x,a) := \min_\alpha Q_\mu^\alpha(x,a).$$

From the function Q_μ^* one obtains the control

$$\alpha^*(x) = \arg\min_a Q_\mu^*(x,a)$$

(in fact in the algorithm presented in Section 4, we use a randomized policy, which is not taken into account here). Note that the minimizing strategy may depend on the global measure μ. For fixed μ, the function Q_μ^* follows the Bellman equation given by

$$Q_\mu^*(x,a) = f\big(x,a,\mu,\mu_{x,a}^{*,\mu}\big) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q_\mu^*(x',a'). \tag{3.2}$$

Note that using this modified (McKean–Vlasov type) Bellman equation established in [4] allows us to consider the Q-function as a function of state and action only. The measure $\mu_{x,a}^{*,\mu} = \lim_{t\to\infty} \mathcal{L}(X_t^{\alpha^*_{x,a},\mu})$ corresponds to the strategy $\alpha^*_{x,a}$ which is derived from α^* by changing the action in state x to a, see (3.1). The above Bellman equation follows from the results in [5] as the measure μ is fixed and does not depend on α.

3.2 Time-dependent Q-function

The definition of the time-dependent optimal Q-function in the extended framework is given for a fixed flow of state-action global distributions $\nu = (\nu_t)_{t=0,1,\dots,T}$ by

$$\begin{cases} Q_{T,\nu}^*(x,a) = g\big(x,\mu_T,\mu_T^{\alpha,\nu}\big), \quad (x,a) \in \mathcal{X}\times\mathcal{A}, \\ Q_{t,\nu}^*(x,a) \\ = \min_\alpha \mathbb{E}\left[\sum_{t'=t}^{T-1} f\left(X_{t'}^{\alpha,\nu}, \alpha_{t'}(X_{t'}^{\alpha,\nu}), \nu_{t'}, \nu_{t'}^{\alpha,\nu}\right) + g\big(X_T^{\alpha,\nu}, \mu_T, \mu_T^{\alpha,\nu}\big) \,\middle|\, X_t^{\alpha,\nu} = x, A_t = a\right], \\ \qquad t = 0,1,\dots,T-1, \quad (x,a) \in \mathcal{X}\times\mathcal{A}, \end{cases}$$

where μ_T (resp. $\mu_T^{\alpha,\nu}$) is the first marginal of ν_T (resp. $\nu_T^{\alpha,\nu}$), and $\alpha_{t'}(\cdot) = \alpha(t',\cdot)$. Using dynamic programming, it can be shown that $(Q_{t,\nu}^*)_t$ is the solution of the Bellman equation

$$\begin{cases} Q_T^*(x,a) = g\big(x,\mu_T,\mu_T^{\alpha,\nu}\big), \quad (x,a)\in\mathcal{X}\times\mathcal{A}, \\ Q_{t,\nu}^*(x,a) = f\big(x,a,\nu_t,\nu_t^{\tilde{\alpha},\nu}\big) + \sum\limits_{x'\in\mathcal{X}} p(x'|x,a,\nu)\min\limits_{a'} Q_{t+1,\nu}^*(x',a'), \\ \qquad t=0,1,\dots,T-1, \quad (x,a)\in\mathcal{X}\times\mathcal{A}, \end{cases}$$

where $\nu_t^{\tilde{\alpha},\nu}$ takes into account the modification of α due to the decision a at state x. The corresponding optimal value function $(V_{t,\nu}^*)$ is given by

$$V_{t,\nu}^*(x) = \min_a Q_{t,\nu}^*(x,a), \quad t=0,1,\dots,T, \quad x\in\mathcal{X}.$$

One of the main advantages of computing the action-value function instead of the value function is that from the former one obtains the optimal control at time t by computing $\arg\min_{a\in\mathcal{A}} Q_t^*(x,a)$. This is particularly important in order to design model-free methods as we will see in the next section.

The next step consists in describing the updates of the Q_t's tables, the flows of measures ν_t's and $\nu_t^{\tilde{\alpha},\nu}$'s. As for the infinite horizon case discussed in the next section, a three-timescale approach is implemented by introducing three learning rates $\rho_k^{\nu} < \rho_k^{Q} < \rho_k^{\nu^{\alpha,\nu}}$. We skip the details for the finite horizon extended framework as they are similar to [5] where it is presented for the two timescale case. This approach justifies the algorithm presented in Section 4.

3.3 Stochastic approximation

In this section we propose a learning procedure that under reasonable assumptions on the functions p and f approximates the solution of the discrete time MFCG. The algorithm is based on the idea that the local distribution, the Q-function describing the optimal strategy, and the global distribution should be updated at different rates. For the sake of a lighter notation, we will use the notation μ, Q and μ^{α} omitting the mutual dependencies that are fully discussed in the previous sections.

For a pure MFC and a pure MFG problem the authors of [4] use results in [9,10] for classical Q-learning to show that a two-timescale approach involving the system distribution and the optimal response can converge to either the MFC solution or the MFG solution depending on how the learning rates are chosen. For a MFC problem, the system distribution resulting from a chosen strategy has to be updated more frequently than the strategy itself. In contrast, the MFG case requires the strategy to be updated more frequently than the distribution.

To gain some intuition for the three-timescale approach used to approximate our MFCG, we start with the function $Q : \mathcal{X}\times\mathcal{A}\to\mathbb{R}$ inducing a strategy α' such that at each time the system is at state x, the action $\arg\min_a Q(x,a)$ is chosen. Say the global distribution μ is frozen (as it is in part 1 of our MFCG problem) and the local distribution is given by μ^{α}, then the local population will be driven at the next step towards the new distribution $\sum_{x\in\mathcal{X}} \mu^{\alpha}(x)p(x'|x,\arg\min_a Q(x,a),\mu)$ if all players follow the strategy encoded in Q. This continues until a fixed-point $\mu^{\alpha'}$ is reached. When a fixed-point is (approximately)

reached, the strategy has to be updated, taking this new limiting distribution into account. This leads to a new optimal strategy with action values given by

$$f(x,a,\mu,\mu^{\alpha'}) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q(x',a').$$

This procedure continues until an optimal pair of strategy α and resulting limiting measure μ^{α} is reached which depends on the frozen global measure μ. In an outer global optimization the fixed-point for the global measure is now obtained by updating the global measure via $\sum_{x\in\mathcal{X}} \mu(x)p(x'|x, \arg\min_a Q(x,a), \mu)$. The three timescales therefore arise naturally by the different layers of optimization involved in the problem. It is intuitive that in each layer one has to perform sufficiently many iterations to ensure that the optimization in the next layer is based on sufficiently accurate results. This idea leads to a learning rate that decreases from the outer to the inner layer. In addition the ratios of the increasing learning rates (from inner to outer layer) have to be sufficiently large.

These considerations lead to the following updating rules: (μ, Q, μ^{α}) are updated with rates $\rho_k^{\mu} < \rho_k^{Q} < \rho_k^{\mu^{\alpha}}$ by

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^{\mu} \mathcal{P}(\mu_k, Q_k, \mu_k), \\ Q_{k+1} = Q_k + \rho_k^{Q} \mathcal{T}(\mu_k, Q_k, \mu_k^{\alpha}), \\ \mu_{k+1}^{\alpha} = \mu_k^{\alpha} + \rho_k^{\mu^{\alpha}} \mathcal{P}(\mu_k, Q_k, \mu_k^{\alpha}), \end{cases} \tag{3.3}$$

where k denotes the learning episode (see the algorithm below), and

$$\begin{cases} \mathcal{P}(\mu,Q,\nu)(x) = (\nu P^{\mu,Q})(x) - \nu(x), \\ \mathcal{T}(\mu,Q,\mu^{\alpha})(x,a) = f(x,a,\mu,\mu^{\alpha}) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q(x',a') - Q(x,a), \\ P^{\mu,Q}(x,x') = p\big(x'|x, \arg\min_a Q(x,a), \mu\big), \\ \big(\nu P^{\mu,Q}\big)(x) = \sum_{x_0} \nu(x_0) P^{\mu,Q}(x_0,x). \end{cases}$$

To see that the above system does in fact converge to a solution of our MFCG, we assume that $\rho_k^{\mu} \ll \rho_k^{Q}$ so that ρ_k^{μ}/ρ_k^{Q} is of order $\epsilon \ll 1$, and $\rho_k^{Q} \ll \rho_k^{\mu^{\alpha}}$ so that $\rho_k^{Q}/\rho_k^{\mu^{\alpha}}$ is of order $\tilde{\epsilon} \ll 1$.

Now, following [9], we denote by τ a continuous time variable, and we consider the following ODE system:

$$\begin{cases} \dot{\mu}_\tau = \mathcal{P}(\mu_\tau, Q_\tau, \mu_\tau), \\ \dot{Q}_\tau = \dfrac{1}{\epsilon} \mathcal{T}(\mu_\tau, Q_\tau, \mu_\tau^{\alpha}), \\ \dot{\mu}_\tau^{\alpha} = \dfrac{1}{\epsilon \cdot \tilde{\epsilon}} \mathcal{P}(\mu_\tau, Q_\tau, \mu_\tau^{\alpha}), \end{cases}$$

which tracks the system (3.3). Furthermore, we assume that the functions f and p are such that the system fulfills a Lipschitz condition. As shown in [4], this can be ensured by Lipschitz continuity of f and p and by smoothing the minimum in the definition of P.

We refer to [4] where these considerations are treated in more detail for a two-timescale approach.

We start with the fastest timescale. For a fixed global distribution μ and a fixed action table Q, we assume that the ODE

$$\dot{\mu}_\tau^\alpha = \frac{1}{\epsilon \cdot \tilde{\epsilon}} \mathcal{P}(\mu, Q, \mu_\tau^\alpha)$$

has a unique asymptotically ($\tilde{\epsilon} \to 0$) stable equilibrium $\mu^{Q,\mu}$ such that $\mathcal{P}(\mu, Q, \mu^{Q,\mu}) = 0$. Now, we plug this equilibrium $\mu^{Q,\mu}$ into the second equation and obtain the ODE

$$\dot{Q}_\tau = \frac{1}{\epsilon} \mathcal{T}(\mu, Q_\tau, \mu^{Q_\tau,\mu}).$$

Again, we assume that the above ODE has a stable equilibrium ($\epsilon \to 0$), which we call Q^μ and which satisfies that $\mathcal{T}(\mu, Q^\mu, \mu^{Q^\mu,\mu}) = 0$. Now, going to the slowest timescale, the first equation has an asymptotic ($\tau \to \infty$) equilibrium, say μ_∞ that solves $\mathcal{P}(\mu_\infty, Q^{\mu_\infty}, \mu_\infty) = 0$. By uniqueness of this equilibrium, we get that $\mu_\infty = \mu^{Q^{\mu_\infty},\mu_\infty}$, which in turns implies that μ_∞ and the action given by minimizing Q^{μ_∞} solves our MFCG.

4 Reinforcement learning algorithm

4.1 Asymptotic version

The three-timescale mean field Q-learning algorithm (U3-MF-QL) that we propose leverages the two-timescale version (U2-MF-QL) introduced by [4]. It not only encompasses learning the pure MFG and pure MFC problems, but, more importantly, it facilitates learning the generalized MFCG problems. Despite of its advantage and flexibility, it inherits the very simple intuition that by manipulating the relative value of learning rates we can induce the algorithm to updating distributions in either MFG's or MFC's manner, as described in the last Section 3.3. Depending on whether the problem has infinite horizon or finite horizon, the U3-MF-QL algorithm will be specified accordingly. Here we first introduce the infinite horizon version (U3-MF-QL-IH) in Algorithm 1. The intuition underlying the algorithm is based on the asymptotic formulation but, as explained in Remark 2.1, this is equivalent to solving the stationary problem. Furthermore, for the sake of simplicity we present the algorithm for an MFCG involving only the state distribution but it can be adapted to solve an MFCG involving the state-action distribution, i.e., an extended MFCG. In Section 4.2, Algorithm 2 is presented for the finite horizon extended MFCG. Both algorithms can be adapted to solve continuous states and actions problems by applying the necessary truncation and discretization techniques as originally discussed in [4].

Algorithm 1 Three-Timescale Mean Field Q-Learning – Discrete Time Infinite Horizon

Require:

1: Time steps $t = 0, 1, \ldots, T$ with $T \gg 0$,
2: Finite state space: $\mathcal{X} = \{x_0, \ldots, x_{|\mathcal{X}|-1}\}$,

Finite action space: $\mathcal{A} = \{a_0, \dots, a_{|\mathcal{A}|-1}\}$,
Initial distribution of the representative player: μ_0,
Factor of the ε-greedy policy: ε,
Break rule tolerances: tol_Q, tol_μ, $tol_{\tilde{\mu}}$.
Initialization:
$Q^0(x,a) = 0$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,
$\mu_t^0 = \frac{1}{|\mathcal{X}|} J_{|\mathcal{X}|}$ and $\tilde{\mu}_t^0 = \frac{1}{|\mathcal{X}|} J_{|\mathcal{X}|}$ for $t \le T$,
where J_d is a d-dimensional vector.
for each episode $k = 1, 2, \dots$ **do**
  Observe initial state: $X_0^k \sim \mu_T^{k-1}$ and set $Q^k \equiv Q^{k-1}$.
  for $t = 0, \dots, T$ **do**
    Choose action:
      choose A_t^k using the ε-greedy policy derived from $Q^k(X_t^k, \cdot)$.
    Update distributions:
      $\mu_t^k = \mu_t^{k-1} + \rho_k^{\mu}\big(\delta(X_t^k) - \mu_t^{k-1}\big)$,
      $\tilde{\mu}_t^k = \tilde{\mu}_t^{k-1} + \rho_k^{\tilde{\mu}}\big(\delta(X_t^k) - \tilde{\mu}_t^{k-1}\big)$,
      where $\delta(X_t^k) = (\mathbf{1}_x(X_t^k))_{x \in \mathcal{X}}$.
    Observe next state:
      observe X_{t+1}^k from the environment.
    Observe cost:
      observe $f_t = f(X_t^k, A_t^k, \mu_t^k, \tilde{\mu}_t^k)$.
    Update Q table:
      $Q^k(x,a) = Q^k(x,a) + \mathbf{1}_{x,a}\big(X_t^k, A_t^k\big)\rho^Q_{x,a,t,k}$
        $\times \big(f_t + \gamma \min_{a' \in \mathcal{A}} Q^k(X_{t+1}^k, a') - Q^k(x,a)\big)$,
      where γ is the discount parameter.
  end for
  if $\|Q^k - Q^{k-1}\| \le tol_Q$, $\|\mu^k - \mu^{k-1}\| \le tol_\mu$, and $\|\tilde{\mu}^k - \tilde{\mu}^{k-1}\| \le tol_{\tilde{\mu}}$ **then**
    break
  end if
end for

4.1.1 Learning rates

By choosing $\rho_k^{\mu} < \rho_k^Q$, we induce the global distribution μ to converge in the fashion of MFG. On the other hand, by letting $\rho_k^Q < \rho_k^{\tilde{\mu}}$, we allow the local distribution $\tilde{\mu}$ to renew towards the MFC style. Combining both such that $\rho_k^{\mu} < \rho_k^Q < \rho_k^{\tilde{\mu}}$, the algorithm is expected to learn both the global and local distributions simultaneously. In addition, to ensure that the learned Q-table and distributions can stabilize at the end of the episode iteration, all the three learning rates shall also decay as the number of episodes, k, increases. Adapting the learning rate discussed in [4], we design the triplet of learning rates as follows:

$$\rho^Q_{x,a,t,k} := \frac{1}{(1 + \#|(x,a,k,t)|)^{\omega^Q}}, \quad \rho_k^{\mu} := \frac{1}{(1+k)^{\omega^{\mu}}}, \quad \rho_k^{\tilde{\mu}} := \frac{1}{(1+k)^{\omega^{\tilde{\mu}}}}, \tag{4.1}$$

where $\#|(x,a,k,t)|$ counts the visits of the pair (x,a) up to the episode k and time t. The triplet $(\omega^Q, \omega^\mu, \omega^{\tilde{\mu}})$ should be chosen such that $\omega^\mu > \omega^Q > \omega^{\tilde{\mu}}$, so that $\rho_k^\mu < \rho_k^Q < \rho_k^{\tilde{\mu}}$, and it should satisfy $\omega^Q \in (0.5, 1)$.

4.2 Time-dependent version

The three-timescale mean field Q-learning approach specified for the finite horizon (U3-MF-QL-FH) is shown in Algorithm 2. Although its overall structure is similar to that of Algorithm 1, we shall highlight several important differences. First, in the finite horizon problem, the algorithm must learn the optimal control and state-action distribution for each time point. So, the number of Q tables to be learned is $T-1$, each corresponding to a time step, except for the terminal time which is excluded because no action is taken at time T. In contrast, in the infinite horizon problem we had just a single Q table to learn. Second, in each episode, the initial state X_0 is always drawn from the initial distribution μ_0. This is in contrast to the infinite horizon case where the initial state X_0 is drawn from the terminal empirical distribution learned up to the last episode μ_T^{k-1}. Third, within each episode, the algorithm only iterates through the time steps from 0 to $T-1$. It skips the terminal time T, because at time $T-1$, once the action A_{T-1} is chosen, the final state X_T can be generated and henceforth the terminal cost $g(X_T)$ is observed, which already completes the episode. In the infinite horizon case, whether one iterates up to $T-1$ or T does not make a big difference. Fourth, when updating the Q_t table, the table Q_{t+1} for the next time step $t < T$ needs to be taken into account, in contrast to the infinite horizon case. Lastly, the learning rate for the Q-tables in the finite horizon case are

$$\rho_{x,a,k}^{Q_t} := \frac{1}{(1 + T\#|(x,a,k,t)|)^{\omega^Q}}, \tag{4.2}$$

where $\#|(x,a,k,t)|$ counts separately for each time step t the visits of tuples (x,a) up to episode k.

The approximation of this time-dependent version of the algorithm to the MFCG solution can be shown similarly as we did for the asymptotic problem in Section 3.3. We refer the reader to [4] where this is done for the pure mean field control and the pure mean field game problem.

Algorithm 2 Three-Timescale Mean Field Q-Learning – Discrete Time Finite Horizon

Require:

Time steps: $t = 0, 1, \ldots, T$,
Finite state space: $\mathcal{X} = \{x_0, \ldots, x_{|\mathcal{X}|-1}\}$,
Finite action space: $\mathcal{A} = \{a_0, \ldots, a_{|\mathcal{A}|-1}\}$,
Initial distribution of the representative player: μ_0,
Factor of the ε-greedy policy: ε,
Break rule tolerances: tol_Q, tol_v, $tol_{\tilde{v}}$.

Initialization:

$Q_t^0(x,a) = 0$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, for all $t \in \{0, 1, \ldots, T\}$,

$\nu_t^0 = \frac{1}{|\mathcal{X}\times\mathcal{A}|} J_{|\mathcal{X}|\times|\mathcal{A}|}$, $\tilde{\nu}_t^0 = \frac{1}{|\mathcal{X}\times\mathcal{A}|} J_{|\mathcal{X}|\times|\mathcal{A}|}$ for all $t \in \{0, 1, \ldots, T\}$,
where $J_{n\times m}$ is an $n \times m$ unit matrix.
for each episode $k = 1, 2, \ldots$ **do**
  Observe initial state: $X_0 \sim \mu_0$.
  for $t = 0,1, \ldots, T-1$ **do**
    Choose action:
      choose A_t using the ϵ-greedy policy derived from $Q_t^{k-1}(X_t, \cdot)$.
    Update empirical distributions:
      $\nu_t^k = \nu_t^{k-1} + \rho_k^{\nu}\big(\delta(X_t, A_t) - \nu_t^{k-1}\big)$,
      $\tilde{\nu}_t^k = \tilde{\nu}_t^{k-1} + \rho_k^{\tilde{\nu}}\big(\delta(X_t, A_t) - \tilde{\nu}_t^{k-1}\big)$,
      where $\delta(X_t, A_t) = (\mathbf{1}_{x,a}(X_t, A_t))_{x\in\mathcal{X}, a\in\mathcal{A}}$.
    Observe next state:
      observe X_{t+1} from the environment
    Observe cost:
      running cost $f_t = f(X_t, A_t, \nu_t^k, \tilde{\nu}_t^k)$,
      terminal cost $g_T = g(X_T)$ when reach $t + 1 = T$.
    Update Q_t:
      $Q_t^k(x, a) = Q_t^{k-1}(x, a) + \mathbf{1}_{x,a}(X_t, A_t)\rho_{x,a,k,t}^{Q}\big(f_t + B - Q_t^{k-1}(x, a)\big)$,
      where $B = \mathbf{1}_{\{t+1=T\}} g_T + \mathbf{1}_{\{t+1<T\}} \min_{a\in\mathcal{A}} Q_{t+1}^{k-1}(X_{t+1}, a)$.
  end for
  if $\|Q^k - Q^{k-1}\| \le tol_Q$, $\|\nu^k - \nu^{k-1}\| \le tol_\nu$, and $\|\tilde{\nu}^k - \tilde{\nu}^{k-1}\| \le tol_{\tilde{\nu}}$ **then**
    break
  end if
end for

5 Numerical experiments

We illustrate the performance of our algorithms on benchmark models for which we have explicit solutions: in the infinite horizon case (Algorithm 1) in Section 5.1, and in a finite horizon extended game setting (Algorithm 2) in Section 5.2.

5.1 Asymptotic problem

For MFG or MFC problems in finite spaces, explicit solutions are usually not available and we would have to rely on approximate solutions obtained by other numerical methods to compare with the solutions obtained with our RL algorithm. Instead, here we choose to work with a linear-quadratic model in continuous time and space for which we can easily derive explicit solutions (see Appendix B.1). We then apply our algorithm to a discretization in time and space of this model described in Section 5.1.1. We do not address here the quality of this discretization which has been widely studied. We simply compare the results of our algorithm with the explicit solutions of the continuous model.

Specifically, we consider a continuous-time and space benchmark linear-quadratic MFCG problem with a running cost given by

$$f(x,\alpha,\mu,\mu^{\alpha,\mu}) = \frac{1}{2}\alpha^2 + c_1(x - c_2 m)^2 + c_3(x - c_4)^2 + \tilde{c}_1(x - \tilde{c}_2 m^{\alpha,\mu})^2 + \tilde{c}_5(m^{\alpha,\mu})^2, \quad (5.1)$$

where

$$m = \int x \mathrm{d}\mu(x), \quad m^{\alpha,\mu} = \int x \mathrm{d}\mu^{\alpha,\mu}(x),$$

and c_1, $\tilde{c}_1$, and $\tilde{c}_5$ are positive constants. Here, μ and $\mu^{\alpha,\mu}$ are understood as global environment and local environment. The constant c_1 determines the magnitude of the global effect and the constants $\tilde{c}_1$, $\tilde{c}_5$ specify local effects.

The asymptotic formulation of this MFCG problem is given by

$$\begin{aligned}\inf_{\alpha} J(\alpha;\mu) &= \inf_{\alpha} \mathbb{E}\left[\int_0^\infty \mathrm{e}^{-\beta t} f(X_t^{\alpha,\mu}, \alpha_t, \mu, \mu^{\alpha,\mu})\mathrm{d}t\right] \\ &= \inf_{\alpha} \mathbb{E}\left[\int_0^\infty \mathrm{e}^{-\beta t}\left(\frac{1}{2}\alpha_t^2 + c_1\big(X_t^{\alpha,\mu} - c_2 m\big)^2 + c_3\big(X_t^{\alpha,\mu} - c_4\big)^2 \right.\right. \\ &\qquad \left.\left. + \tilde{c}_1\big(X_t^{\alpha,\mu} - \tilde{c}_2 m^{\alpha,\mu}\big)^2 + \tilde{c}_5(m^{\alpha,\mu})^2\right)\mathrm{d}t\right]\end{aligned}$$

subject to $\mathrm{d}X_t^{\alpha,\mu} = \alpha_t \mathrm{d}t + \sigma \mathrm{d}W_t$, $X_0^{\alpha,\mu} \sim \mu_0$, and the fixed point condition

$$m = \lim_{t\to\infty} \mathbb{E}\big(X_t^{\hat{\alpha},\mu}\big) = m^{\hat{\alpha},\mu},$$

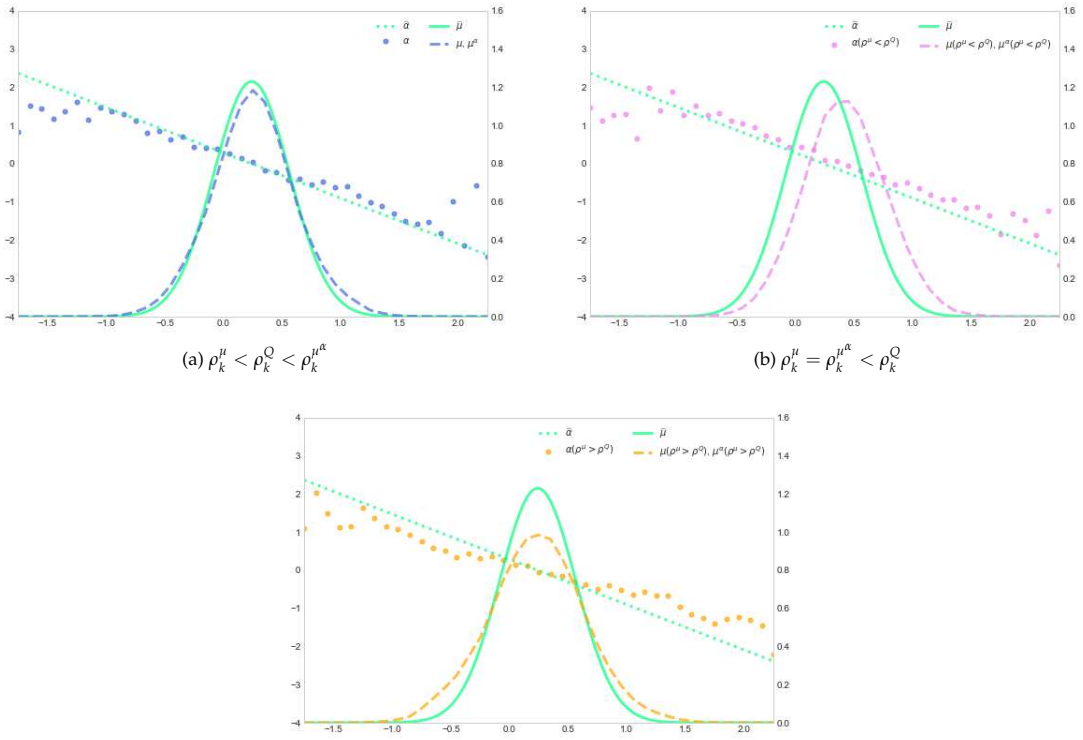
where $\hat{\alpha}$ is the optimal action.

5.1.1 Results

We consider the asymptotic MFCG with the following choice of parameters: $c_1 = 0.5$, $c_2 = 1.5$, $c_3 = 0.5$, $c_4 = 0.25$ $\tilde{c}_1 = 0.3$, $\tilde{c}_2 = 1.25$, $\tilde{c}_5 = 0.25$, discount rate $\beta = 1$, and volatility of the state dynamics $\sigma = 0.5$. We truncate the infinite time horizon at $T = 20$, and discretize the interval $[0, T]$ with time steps of size $\Delta t = 10^{-2}$. The discount factor in the discrete time setting is then given by $\gamma := e^{-\beta \Delta t}$. The state space is

$$\mathcal{X} = \{x_0 = -2 + x_c, \ldots, x_{|\mathcal{X}|-1} = 2 + x_c\}$$

centered at $x_c = 0.25$, and the action space is $\mathcal{A} = \{a_0 = -3, \ldots, a_{|\mathcal{A}|} = 3\}$, where the step sizes are $\Delta x = \Delta a = \sqrt{\Delta t} = 10^{-1}$. The ϵ-greedy policy parameter is 0.01. We remind the reader on the choice of the learning rates. In contrast to the pure MFG and MFC problems which can be learned by the two-timescale parameterization proposed in [4], the MFCG problem requires the three-timescale as explained in Section 3.3. We choose the three learning rates to be $(\omega^\mu, \omega^Q, \omega^{\mu^\alpha}) = (0.85, 0.55, 0.15)$, which satisfy $\rho_k^\mu < \rho_k^Q < \rho_k^{\mu^\alpha}$. In addition, we also demonstrate that if we miss-specify the learning rates $(\omega^\mu, \omega^Q, \omega^{\mu^\alpha})$ by either $(0.85, 0.55, 0.85)$ in which case $\rho_k^\mu = \rho_k^{\mu^\alpha} < \rho_k^Q$ or by $(0.15, 0.55, 0.15)$ in which case $\rho_k^\mu = \rho_k^{\mu^\alpha} > \rho_k^Q$, then the algorithm fails to learn the correct result.



(a) $\rho_k^{\mu} < \rho_k^{Q} < \rho_k^{\mu^\alpha}$ (b) $\rho_k^{\mu} = \rho_k^{\mu^\alpha} < \rho_k^{Q}$

(c) $\rho_k^{\mu} = \rho_k^{\mu^\alpha} > \rho_k^{Q}$

Figure 5.1: Control and distributions for the benchmark asymptotic MFCG learned by Algorithm 1. The x-axis shows the state variable x, the left y-axis refers to the value of the control $\alpha(x)$, and the right y-axis marks the probability mass of $\mu(x)$ and $\mu^\alpha(x)$. The green dotted line (labeled by $\hat{\alpha}$) is the theoretical control function and green curve (labeled by $\hat{\mu}$) shows the theoretical distribution of state, where the global distribution equals to the local distribution. The dots (labeled by α) are the learned controls and the overlapping dashed curves (labeled by μ and μ^α) refer to the overlapping empirical global and local distributions learned by the algorithm, colored in blue, violet, and orange according to the selection of learning rates.

Fig. 5.1 is generated with 5 runs of $K = 100{,}000$ episodes with the above setting. We report the average of the learned control and distribution in the last 10,000 episodes over the 5 runs. In Fig. 5.1(a) the control and distribution learned with the correct three-timescale rates by the algorithm are plotted against the theoretical optimal control obtained in Eq. (B.4) and with the theoretical distribution. The control learned by the algorithm (blue dots) lies well along the theoretical control function (dotted green line) except for states never visited by the algorithm. That is, within the support of the distribution the algorithm learns the optimal control well. Also, we observe that the learned global distribution overlaps the local distribution (both dashed blue curve), and that both match the theoretical distribution (solid green curve). Therefore, the algorithm successfully learns the correct local and global distributions as well. Figs. 5.1(b) and 5.1(c) illustrate the failure of the algorithm in cases where the learning rates are misspecified as described in the last paragraph.

5.2 Trader's problem

As we did for the infinite horizon case, we consider a finite horizon extended game in continuous time and spaces. In particular, we reassess the renowned trader's execution problem presented in [13] starting with the finite number of players case. Instead of a game among traders, we consider a game between groups of traders as follows. Suppose there are M homogeneous trading groups, each with N traders trading on a single stock. Let the index tuple (m,n) denotes the n-th trader in the m-th group.

Trader (m,n) is controlling the drift term in the dynamic of her stock inventory,

$$\mathrm{d}X_t^{m,n} = \alpha_t^m\left(X_t^{m,n}\right)\mathrm{d}t + \sigma_t^{m,n}\mathrm{d}W_t^{m,n},$$

by the trading rate $\alpha_t^m(X_t^{m,n})$ with which all the traders in the m-th group comply. Her cash position $K_t^{m,n}$ evolves as

$$\mathrm{d}K_t^{m,n} = -\left[\alpha_t^m\left(X_t^{m,n}\right)S_t + c_\alpha\left(\alpha_t^m(X_t^{m,n})\right)\right]\mathrm{d}t$$

with $\alpha \mapsto c_\alpha(\alpha)$ a non-negative convex function representing the cost of trading at rate α. The stock price is impacted by a function $h(\cdot)$ of the transactions and follows the dynamic

$$\mathrm{d}S_t = \frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N} h\left(\alpha_t^i(X_t^{i,j})\right)\mathrm{d}t + \sigma_t^0\mathrm{d}W_t^0.$$

The total wealth $V_t^{m,n}$ of her self-financing portfolio consists of her cash position and her stock value,

$$V_t^{m,n} = K_t^{m,n} + X_t^{m,n}S_t$$

with dynamic

$$\begin{aligned}\mathrm{d}V_t^{m,n} &= \mathrm{d}K_t^{m,n} + S_t\mathrm{d}X_t^{m,n} + X_t^{m,n}\mathrm{d}S_t\\ &= \left[-c_\alpha\left(\alpha_t^m(X_t^{m,n})\right) + X_t^{m,n}\frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N} h\left(\alpha_t^i(X_t^{i,j})\right)\right]\mathrm{d}t\\ &\quad + S_t\sigma_t^{m,n}\mathrm{d}W_t^{m,n} + X_t^{m,n}\sigma_t^0\mathrm{d}W_t^0.\end{aligned}$$

We assume that the individual trader is subject to a running liquidation constraint modeled by a function c_X of the average shares held by her own group. In this model, the individual trader's objective function is given by

$$\begin{aligned}&J^{m,n}\left(\alpha^1,\ldots,\alpha^M\right)\\ &= \mathbb{E}\left\{\int_0^T c_X\left(\frac{1}{N}\sum_{j=1}^{N}X_t^{m,j}\right)\mathrm{d}t + g\left(X_T^{m,n}\right) - V_T^{m,n}\right\}\\ &= \mathbb{E}\left\{\int_0^T\left[c_X\left(\frac{1}{N}\sum_{j=1}^{N}X_t^{m,j}\right) + c_\alpha\left(\alpha_t^m(X_t^{m,n})\right) - X_t^{m,n}\frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N} h\left(\alpha_t^i(X_t^{i,j})\right)\right]\mathrm{d}t\right.\\ &\qquad \left. + g\left(X_T^{m,n}\right)\right\}.\end{aligned}$$

In the limit of a large number of large groups without precising the relation between M and N (see Appendix A.3 for more details about this type of limit for a finite horizon linear-quadratic model), and assuming $\sigma_t^{m,n} = \sigma$, this problem leads to the following mixture of MFCG problems: Minimize

$$J(\alpha;\theta) = \mathbb{E}\left\{\int_0^T \left[c_X\left(m_t^{\alpha,\theta}\right) + c_\alpha\left(\alpha_t\right) - X_t^{\alpha,\theta}\int h(a)d\theta_t(a)\right]\mathrm{d}t + g\left(X_T^{\alpha,\theta}\right)\right\},$$

where θ_t is the law of the control α_t, $m_t^{\alpha,\theta} = \mathbb{E}(X_t^{\alpha,\theta})$ and

$$dX_t^{\alpha,\theta} = \alpha_t dt + \sigma dW_t, \quad t \leq T, \quad X_0^{\alpha,\theta} = x.$$

Note that the problem is of MFG style in control through θ_t and MFC style in state through $m_t^{\alpha,\theta}$.

In what follows we focus on the Linear-Quadratic case where

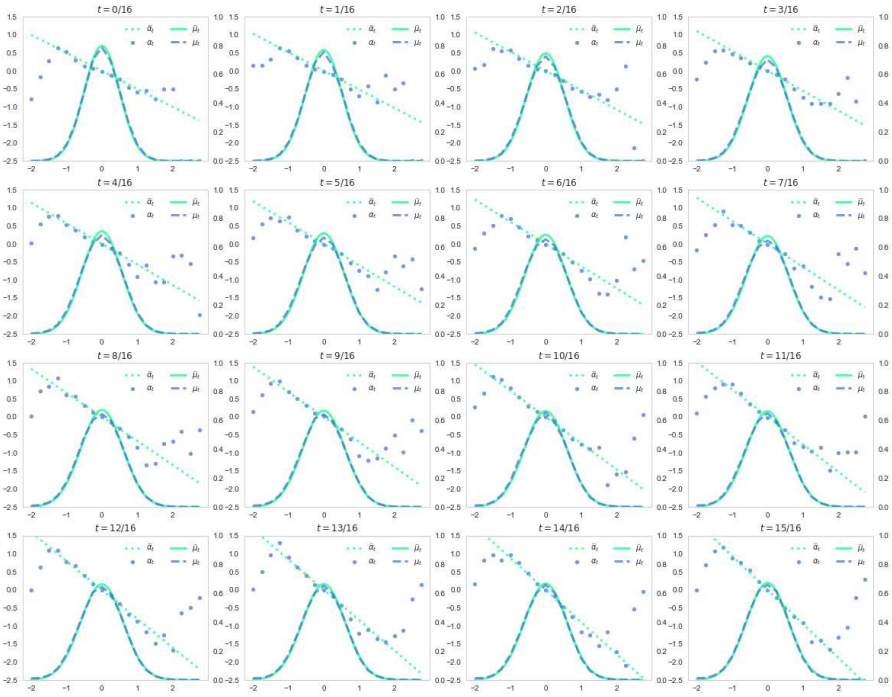
$$c_x(m) = \frac{c_X}{2}m^2, \quad c_\alpha(\alpha) = \frac{c_\alpha}{2}\alpha^2, \quad h(a) = c_h a, \quad g(x) = \frac{c_g}{2}x^2,$$

so that

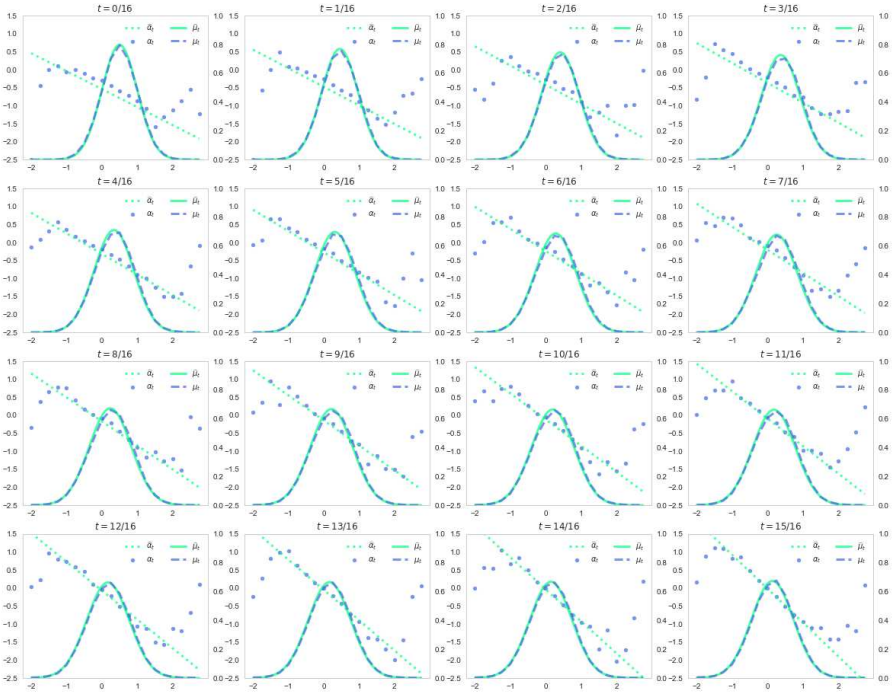
$$J(\alpha;\theta) = \mathbb{E}\left\{\int_0^T \left[\frac{c_X}{2}\left(m_t^{\alpha,\theta}\right)^2 + \frac{c_\alpha}{2}\alpha_t^2 - c_h X_t^{\alpha,\theta}\int a d\theta_t(a)\right]\mathrm{d}t + \frac{c_g}{2}\left(X_T^{\alpha,\theta}\right)^2\right\}.$$

5.2.1 Results

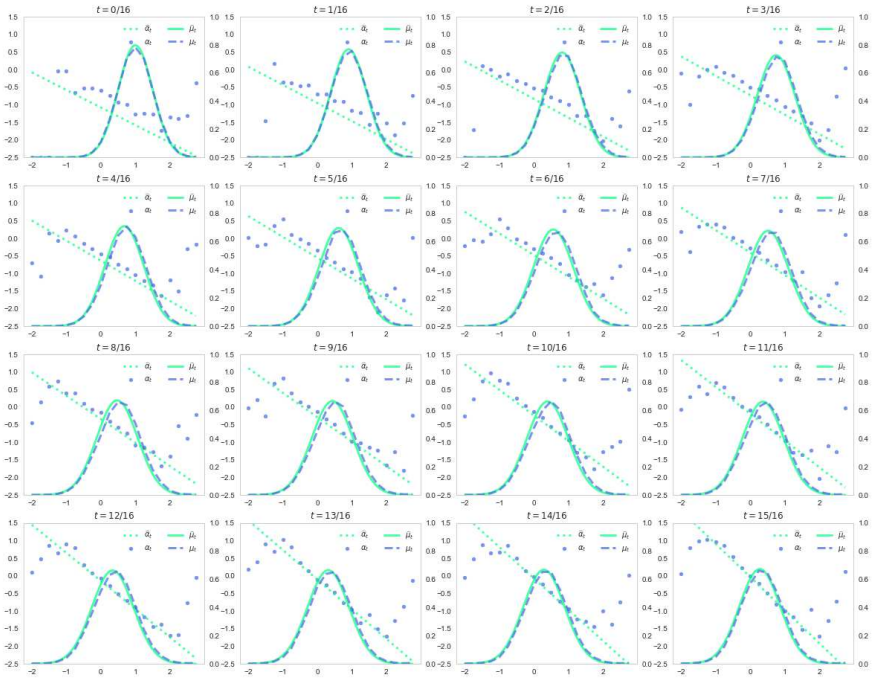
We consider the trader's problem with the choice of parameters: $c_\alpha = 1$, $c_X = 0.75$, $c_h = 1.25$, $c_g = 1$, and with a volatility for the state dynamic $\sigma = 0.75$. We test three distributions for the initial inventory X_0: Gaussian with mean $x_0 = 0$, 0.5, and 1 and the same standard deviation $\sigma = 0.5$. The terminal time is $T = 1$, and we choose a time grid $\tau = \{0, \Delta t, \ldots, T\}$ with time step $\Delta t = 1/16$. We discretize the state space into $\mathcal{X} = \{x_0 = -2, \ldots, x_{|\mathcal{X}|-1} = 2.5\}$, and the action space into $\mathcal{A} = \{a_0 = -2, \ldots, a_{|\mathcal{A}|-1} = 1.5\}$, where the step sizes are $\Delta x = \Delta a = \sqrt{\Delta t} = 1/4$. The triplet of the learning rates is chosen as $(\omega^\theta, \omega^Q, \omega^\mu) = (0.85, 0.55, 0.15)$. For the ϵ-greedy policy we choose $\epsilon = 0.05$. We run the experiment 10 times, each with $K = 200{,}000$ episodes. We average the control and state distributions learned by Algorithm 2 over the last 10,000 episodes and over 10 runs. We report the results in Fig. 5.2. We present the results for every time step in τ, except for the last time step T. The subplots are ordered by time, from left to right and top to bottom. Note that the theoretical optimal control $\hat{\alpha}_t$ (dotted green line) changes over time. As time increases, the slope and intercept of $\hat{\alpha}_t$ increase. Also, the theoretical local state distribution μ_t (green curve) under the optimal control changes over time. As time increases from 0 to T, the center of μ_t moves towards zero and the standard deviation increases. To evaluate the effectiveness of Algorithm 2, we compare the learned action (blue dots) and the learned local state distribution (dashed blue curve) with their theoretical counterparts. Again we observe that except for the tails of the distribution, the control learned by the algorithm is very close to the theoretical value. This means that the algorithm successfully learns the optimal control for states that are frequently sampled. Also, we see that


t = 0/16
t = 1/16
t = 2/16
t = 3/16
t = 4/16
t = 5/16
t = 6/16
t = 7/16
t = 8/16
t = 9/16
t = 10/16
t = 11/16
t = 12/16
t = 13/16
t = 14/16
t = 15/16

(a) $x_0 = 0$


t = 0/16
t = 1/16
t = 2/16
t = 3/16
t = 4/16
t = 5/16
t = 6/16
t = 7/16
t = 8/16
t = 9/16
t = 10/16
t = 11/16
t = 12/16
t = 13/16
t = 14/16
t = 15/16

(b) $x_0 = 0.5$



(c) $x_0 = 1$

Figure 5.2: Control and distributions for the trader's MFCG learned by Algorithm 2. The x-axis shows the state variable x, the left y-axis refers to the value of the control $\alpha(x)$, and the right y-axis marks the probability mass of state, $\mu(x)$. The dotted green lines (labeled by $\hat{\alpha}_t$) are the theoretical control function and the blue dots (labeled by α_t) are the learned control. The green curves (labeled by $\hat{\mu}_t$) show the theoretical distributions of state and dashed blue curves (labeled by μ_t) refer to the empirical distribution of state learned by the algorithm.

the dashed blue curve perfectly overlaps with the solid green curve, hence the algorithm succeeds in capturing the evolution of the state distribution under the correctly learned control.

6 Conclusion

We have introduced a type of mean field control game (MFCG) that models a competitive game between a large number of large collaborative groups. It turns out that the two-timescale reinforcement learning algorithm (U2-MF-QL) that was proposed in [4] for infinite horizon problems and in [5] for finite horizon extended problems, for learning either MFG or MFC problems, is naturally adapted for learning MFCG problems by managing three learning rates in the three-timescale reinforcement learning algorithm (U3-MF-QL) proposed in this paper. We illustrate the results with linear quadratic problems for which we derive explicit formulas. In particular, a new type of trader problem is presented. The theory associated for MFCGs is a work in progress [6], as well as an actor-critic version of the U3-MF-QL algorithm in the context of continuous spaces [3].

Appendix A. Linear-quadratic example

In this appendix we provide additional details about the LQ example presented in Section 5.1, here in a finite horizon setting. In particular, we explain the relation between the finite-player game and its corresponding MFCG limiting problem.

A.1 The finite-player model

There are M competitive groups each made of N collaborative players ($m = 1, \dots, M$ is the group number and $n = 1, \dots, N$ is the player number within the group). The dynamics of the state of player (m, n) is

$$dX_t^{m,n} = \alpha_t^{m,n} dt + \sigma dW_t^{m,n}, \quad X_0^{m,n} \sim \mu_0,$$

where we aim at an open-loop equilibrium (the α's are adapted to the W's). The objective of the collaborative group m is to minimize

$$J^m(\boldsymbol{\alpha}) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E} \int_0^T \left\{ \frac{1}{2} \left(\alpha_t^{m,n}\right)^2 + \frac{c_1}{2} \left(\bar{\mu}_t - X_t^{m,n}\right)^2 + \frac{c_2}{2} \left(\bar{\mu}_t^m\right)^2 \right\} dt,$$

where

$$\bar{\mu}_t^m = \frac{1}{N} \sum_{n=1}^{N} X_t^{m,n}$$

is the empirical mean of group m, and

$$\bar{\mu}_t = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} X_t^{m,n}$$

is the empirical mean of the total population (c_1 and c_2 are positive constants and we assume a zero terminal condition for simplicity).

Accordingly, we introduce the Hamiltonian H^m of group m

$$H^{m,n} = \sum_{m'=1}^{M} \sum_{n'=1}^{N} \alpha^{m',n'} y^{m,m',n,n'} + \sum_{n=1}^{N} \frac{1}{2} (\alpha^{m,n})^2 + \frac{c_1}{2} \sum_{n=1}^{N} (\bar{\mu} - x^{m,n})^2 + \frac{c_2}{2} N (\bar{\mu}^m)^2,$$

where $y^{m,m',n,n'}$ are the adjoint variables.

Minimizing the Hamiltonian $H^{m,n}$ with respect to $\alpha^{m,n}$, we get

$$\frac{\partial H^{m,n}}{\partial \alpha^{m,n}} = y^{m,m,n,n} + \alpha^{m,n} = 0 \quad \Longrightarrow \quad \hat{\alpha}^{m,n} = -y^{m,m,n,n}.$$

The backward stochastic differential equation (BSDE) that the adjoint process $Y_t^{m,m',n,n'}$ must satisfy is

$$dY_t^{m,m',n,n'} = -\partial_{x^{m',n'}} H^{m,n} dt + \sum_{m''=1}^{M} \sum_{n''=1}^{N} Z_t^{m,m',n,n',m'',n''} dW_t^{m'',n''}$$

$$= -\left[c_1 \sum_{k=1}^{N} \left(\bar{\mu}_t - X_t^{m,k}\right) \left(\frac{1}{MN} - \delta_{\{m'=m,k=n'=n\}}\right) + c_2 N \bar{\mu}_t^m \left(\frac{1}{N} \delta_{\{m'=m,n'=n\}}\right)\right] dt$$
$$+ \sum_{m''=1}^{M} \sum_{n''=1}^{N} Z_t^{m,m',n,n',m'',n''} dW_t^{m'',n''}$$

with a zero terminal condition $Y_T^{m,m',n,n'} = 0$. The Z-processes are part of the solution and must be adapted.

The diagonal adjoint process $Y_t^{m,m,n,n} \equiv Y_t^{m,n}$ satisfies

$$\begin{aligned} dY_t^{m,n} &= -\left[c_1 \sum_{k=1}^{N} \left(\bar{\mu}_t - X_t^{m,k}\right) \left(\frac{1}{MN} - \delta_{\{k=n\}}\right) + c_2 \bar{\mu}_t^m\right] dt \\ &\quad + \sum_{m''=1}^{M} \sum_{n''=1}^{N} Z_t^{m,m,n,n,m'',n''} dW_t^{m'',n''} \\ &= \left[c_1 \left(\left(1 - \frac{1}{M}\right) \bar{\mu}_t + \frac{1}{M} \bar{\mu}_t^m - X_t^{m,n}\right) - c_2 \bar{\mu}_t^m\right] dt \\ &\quad + \sum_{m''=1}^{M} \sum_{n''=1}^{N} Z_t^{m,m,n,n,m'',n''} dW_t^{m'',n''}. \end{aligned} \tag{A.1}$$

We omit the non-diagonal adjoint processes which can be treated analogously, and we formulate the ansatz

$$Y_t^{m,n} = \eta_t X_t^{m,n} + \phi_t \bar{\mu}_t + \xi_t \bar{\mu}_t^m,$$

where η_t, ϕ_t and ξ_t are deterministic functions to be determined. We have

$$d\bar{\mu}_t^m = \frac{1}{N} \sum_{n=1}^{N} dX_t^{m,n} = -\bar{Y}_t^m dt + \frac{\sigma}{N} \sum_{n=1}^{N} dW_t^{m,n},$$

where

$$\bar{Y}_t^m = \frac{1}{N} \sum_{n=1}^{N} Y_t^{m,n} = \phi_t \bar{\mu}_t + (\eta_t + \xi_t) \bar{\mu}_t^m,$$

and

$$d\bar{\mu}_t = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} dX_t^{m,n} = -\bar{Y}_t dt + \frac{\sigma}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} dW_t^{m,n},$$

where

$$\bar{Y}_t = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} Y_t^{m,n} = (\eta_t + \phi_t + \xi_t)\, \bar{\mu}_t.$$

Differentiating the ansatz gives

$$\begin{aligned} dY_t^{m,n} &= \eta_t' X_t^{m,n} dt + \eta_t dX_t^{m.n} + \phi_t' \bar{\mu}_t dt + \phi_t d\bar{\mu}_t + \xi_t' \bar{\mu}_t^m dt + \xi_t d\bar{\mu}_t^m \\ &= \eta_t' X_t^{m,n} dt - \eta_t \left(\eta_t X_t^{m,n} + \phi_t \bar{\mu}_t + \xi_t \bar{\mu}_t^m\right) dt + \phi_t' \bar{\mu}_t dt \end{aligned}$$

$$
\begin{aligned}
&- \phi_t \left(\eta_t + \phi_t + \xi_t\right) \bar{\mu}_t dt + \xi_t' \bar{\mu}_t^m dt - \xi_t(\phi_t \bar{\mu}_t + (\eta_t + \xi_t)\bar{\mu}_t^m) dt + d\text{Mart} \\
&= \left[\eta_t' - \eta_t^2\right] X_t^{m,n} dt + \left[\phi_t' - \eta_t\phi_t - \phi_t(\eta_t + \phi_t + \xi_t) - \xi_t\phi_t\right] \bar{\mu}_t dt \\
&\quad + \left[\xi_t' - \eta_t\xi_t - \xi_t(\eta_t + \xi_t)\right] \bar{\mu}_t^m dt + d\text{Mart} \\
&= \left[\eta_t' - \eta_t^2\right] X_t^{m,n} dt + \left[\phi_t' - \phi_t^2 - 2\eta_t\phi_t - 2\xi_t\phi_t\right] \bar{\mu}_t dt \\
&\quad + \left[\xi_t' - \xi_t^2 - 2\eta_t\xi_t\right] \bar{\mu}_t^m dt + d\text{Mart}.
\end{aligned}
$$

Comparing the drift terms with the previous expression (A.1) for $dY_t^{m,n}$, we get the following system of Riccati equations for which explicit solutions can be obtained (omitted here):

$$
\begin{aligned}
&\eta_t' - \eta_t^2 = -c_1, && \eta_T = 0, \\
&\phi_t' - \phi_t^2 - 2\phi_t(\eta_t + \phi_t) = -c_2 + \frac{1}{M}c_1, && \phi_T = 0, \\
&\xi_t' - \xi_t^2 - 2\eta_t\xi_t = c_1\left(1 - \frac{1}{M}\right), && \xi_T = 0.
\end{aligned}
$$

As usual the Z's processes are deterministic (hence adapted) and identified by matching the martingale terms.

One can define $\zeta_t = \eta_t + \xi_t$, so that the system of ODEs becomes

$$
\begin{aligned}
&\eta_t' - \eta_t^2 = -c_1, && \eta_T = 0, \\
&\phi_t' - \phi_t^2 - 2\phi_t\zeta_t = -c_2 + \frac{1}{M}c_1, && \phi_T = 0, \\
&\zeta_t' - \zeta_t^2 = -\frac{1}{M}c_1, && \xi_T = 0,
\end{aligned}
\tag{A.2}
$$

which highlights the limit $M \to \infty$ where ζ_t vanishes.

A.2 The corresponding limiting MFCG

To the previous finite-player model, we propose to associate the following MFCG problem: For a fixed flow of distributions $\mu = (\mu_t)$, one agent controls her state given by

$$
dX_t^{\alpha,\mu} = \alpha_t dt + \sigma dW_t, \quad X_0^{\alpha,\mu} \sim \mu_0.
$$

The agent solves the MKV control problem which consists in minimizing

$$
J(\alpha;\mu) = \mathbb{E}\int_0^T \left\{\frac{1}{2}\alpha_t^2 + \frac{c_1}{2}\left(\bar{\mu}_t - X_t^{\alpha,\mu}\right)^2 + \frac{c_2}{2}\left(\mathbb{E}(X_t^{\alpha,\mu})\right)^2\right\} dt,
$$

where $\bar{\mu}_t = \int x\mu(dx)$. One then solves the fixed point condition

$$
\mathbb{E}\left(X_t^{\hat{\alpha},\mu}\right) = \bar{\mu}_t, \quad \forall t \le T.
$$

Introducing the adjoint process Y_t and using the lighter notation X_t for the state process, the optimal strategy $\hat{\alpha}_t$ is given by $-Y_t$ which satisfies the BSDE

$$dY_t = \big[c_1(\mathbb{E}(X_t) - X_t) - c_2\mathbb{E}(X_t)\big]dt + Z_t dW_t, \quad Y_T = 0.$$

The term $c_2\mathbb{E}(X_t)$ comes form the differentiation of $\frac{c_2}{2}(\mathbb{E}(X_t^{\alpha,\mu}))^2$ with respect to the measure.

One verifies easily that the solution is

$$Y_t = -\eta_t(\bar{\mu}_t - X_t) + \phi_t\bar{\mu}_t$$

with

$$\eta_t' - \eta_t^2 = -c_1, \quad \eta_T = 0, \tag{A.3}$$
$$\phi_t' - \phi_t^2 = -c_2, \quad \phi_T = 0, \tag{A.4}$$

and

$$d\bar{\mu}_t = -\phi_t\bar{\mu}_t, \quad \bar{\mu}_0 = x_0,$$

that is

$$\bar{\mu}_t = x_0 e^{-\int_0^t \phi_s ds}.$$

Note that the functions η and ϕ are given explicitly by

$$\eta_t = \sqrt{c_1}\frac{e^{2\sqrt{c_1}(T-t)} - 1}{e^{2\sqrt{c_1}(T-t)} + 1}, \quad \phi_t = \sqrt{c_2}\frac{e^{2\sqrt{c_2}(T-t)} - 1}{e^{2\sqrt{c_2}(T-t)} + 1}.$$

A.3 From finite-player to MFCG

The limit $N \to \infty$ ensures that $\bar{\mu}_t^m = \bar{\mu}_t$ for every m, and the limit $M \to \infty$ ensures that the coefficient functions given by (A.2) converge to those given by (A.3).

Our goal is to show that the strategy obtained from the limiting problem in Section A.2 provides an ϵ-Nash equilibrium for the finite-player game described in Section A.1.

Here we use the notation $(\eta_t^\infty, \phi_t^\infty, \bar{\mu}_t^\infty)$ for the quantities obtained in the system of equations (A.3) (not to be confused with the corresponding quantities obtained in (A.2)).

We denote by α^∞ the optimal strategy obtained in Section A.2, that is

$$\alpha_t^\infty = -Y_t = \eta_t^\infty\,(\bar{\mu}_t^\infty - X_t) - \phi_t^\infty\bar{\mu}_t^\infty,$$

which we apply to all the players in the finite-player game. The value function for the m-th group is given by

$$J^m(\alpha^\infty) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\int_0^T\left\{\frac{1}{2}\big(\alpha_t^{m,n}\big)^2 + \frac{c_1}{2}\big(\bar{\mu}_t - X_t^{m,n}\big)^2 + \frac{c_2}{2}\big(\bar{\mu}_t^m\big)^2\right\}dt,$$

where

$$\alpha_t^{m,n} = \eta_t^\infty\,(\bar{\mu}_t^\infty - X_t^{m,n}) - \phi_t^\infty\bar{\mu}_t^\infty,$$

and

$$dX_t^{m,n} = \alpha_t^{m,n} dt + \sigma dW_t^{m,n}, \quad X_0^{m,n} \sim \mu_0.$$

Note that $\bar{\mu}_t^m = \frac{1}{N}\sum_{n=1}^N X_t^{m,n}$ is given by

$$d\bar{\mu}_t^m = [\eta_t^\infty (\bar{\mu}_t - \bar{\mu}_t^m) - \phi_t^\infty \bar{\mu}_t]dt + \frac{1}{N}\sum_{n=1}^N dW_t^{m,n}, \quad \bar{\mu}_0 = \frac{1}{N}\sum_{n=1}^N X_0^{m,n},$$

and

$$\bar{\mu}_t = \frac{1}{MN}\sum_{m=1}^M \sum_{n=1}^N X_t^{m,n}$$

is given by

$$d\bar{\mu}_t = [-\phi_t^\infty \bar{\mu}_t]\, dt + \frac{1}{MN}\sum_{m=1}^M \sum_{n=1}^N dW_t^{m,n}, \quad \bar{\mu}_0 = \frac{1}{MN}\sum_{m=1}^M \sum_{n=1}^N X_0^{m,n}.$$

Now we consider a strategy $(\alpha^{\infty,-m}, \beta^m)$ where the players from group m use $\beta_t^{m,n}$ instead of $\alpha_t^{m,n}$, and the players from the other groups m' continue using $\alpha_t^{m',n}$. We denote by $\tilde{X}_t^{m,n}$ the state of player (m,n) which satisfies

$$d\tilde{X}_t^{m,n} = \beta_t^{m,n} dt + \sigma dW_t^{m,n}, \quad X_0^{m,n} \sim \mu_0.$$

We denote the corresponding group empirical mean by $\tilde{\mu}_t^m$, and the population empirical mean by $\tilde{\mu}_t$. We also denote

$$f(x,\alpha,\bar{\mu},\bar{\mu}^m) = \frac{1}{2}\alpha^2 + F(x,\bar{\mu},\bar{\mu}^m), \quad F(x,\bar{\mu},\bar{\mu}^m) = \frac{1}{2}\left\{c_1(\bar{\mu}-x)^2 + c_2(\bar{\mu}^m)^2\right\}.$$

Principle of the proof. First, show that for $\epsilon > 0$ there exists M_0 and N_0 such that for $M \geq M_0$ and $N \geq N_0$, we have

$$\left| \frac{1}{N}\sum_{n=1}^N \mathbb{E}\int_0^T \left(F(\tilde{X}_t^{m,n}, \tilde{\mu}_t, \tilde{\mu}_t^m) - F(X_t^{m,n}, \bar{\mu}_t, \bar{\mu}_t^m)\right) dt \right| < \frac{\epsilon}{2}, \tag{A.5}$$

where in the first term the strategy $(\alpha^{\infty,-m}, \beta^m)$ is used, while in the second term the strategy $(\hat{\alpha}^{-m}, \beta^m)$ is used with $\hat{\alpha}$ the optimal strategy obtained in Section A.1 for the finite-player game. Adding $\frac{1}{2}(\beta_t^{m,n})^2$ to both terms, we obtain

$$J^m(\alpha^{\infty,-m}, \beta^m) > J^m(\hat{\alpha}^{-m}, \beta^m) - \frac{\epsilon}{2}.$$

Using the fact that $\hat{\alpha}$ is a Nash equilibrium for the finite-player game, we get

$$J^m(\alpha^{\infty,-m}, \beta^m) > J^m(\hat{\alpha}) - \frac{\epsilon}{2}.$$

As in the first step we can derive

$$|J^m(\alpha^\infty) - J^m(\hat{\alpha})| < \frac{\epsilon}{2} \tag{A.6}$$

and, therefore

$$J^m(\alpha^{\infty,-m}, \beta^m) > J^m(\alpha^\infty) - \epsilon$$

that is α^∞ is an ϵ-Nash equilibrium for the finite-player game.

Of course, (A.5) and (A.6) require some technical work which will be given in a general setting in [6].

Finally we observe that the limits $N \to \infty$ and $M \to \infty$ can be taken sequentially.

If $N \to \infty$ for M fixed, we obtain a game between competitive MKV agents called mean field type game in [18]. Then, our limit describes the MFG limit between these MKV agents. If N is fixed and $M \to \infty$, on can consider each group as one player in the higher dimension N and this is a classical MFG. Our MFCG describes the subsequent limit $N \to \infty$.

Appendix B. Analytic solutions

In this appendix we provide details about the solutions of the problems discussed in Sections 5.

B.1 Solution of the asymptotic problem

The corresponding HJB equation is given by

$$\beta V(x) - H(x, \alpha, \mu, \mu^{\alpha,\mu}) - \int_{\mathbb{R}} \frac{\partial H}{\partial \mu^{\alpha,\mu}} H\left(h, \alpha, \mu, \mu^{\alpha,\mu}\right)(x) \mathrm{d}\mu^{\alpha,\mu}(h) = 0$$

with the Hamiltonian

$$\begin{aligned} H(x, \alpha, \mu, \mu^{\alpha,\mu}) &= \inf_\alpha \left\{ \mathcal{A}^X V(x) + f(x, \alpha, \mu, \mu^{\alpha,\mu}) \right\} \\ &= \inf_\alpha \Big\{ \alpha \dot{V}(x) + \frac{1}{2}\sigma^2 \ddot{V}(x) + \frac{1}{2}\alpha^2 + c_1(x - c_2 m)^2 \\ &\qquad + c_3(x - c_4)^2 + \tilde{c}_1(x - \tilde{c}_2 m^{\alpha,\mu})^2 + \tilde{c}_5 \left(m^{\alpha,\mu}\right)^2 \Big\} \\ &= -\frac{1}{2}\dot{V}(x)^2 + \frac{1}{2}\sigma^2 \ddot{V}(x) + c_1(x - c_2 m)^2 + c_3(x - c_4)^2 \\ &\quad + \tilde{c}_1(x - \tilde{c}_2 m^{\alpha,\mu})^2 + \tilde{c}_5 \left(m^{\alpha,\mu}\right)^2, \end{aligned}$$

and the derivative with respect to $\mu^{\alpha,\mu}$ due to the MFC part, calculated at the optimal $\hat{\alpha}(x) = -\dot{V}(x)$ as follows:

$$\frac{\partial H}{\partial \mu^{\alpha,\mu}}\left(h, -\dot{V}(h), \mu, \mu^{\alpha,\mu}\right) = \frac{\partial}{\partial \mu^{\alpha,\mu}} \left(\tilde{c}_1(h - \tilde{c}_2 m^{\alpha,\mu})^2 + \tilde{c}_5 \left(m^{\alpha,\mu}\right)^2 \right)(x)$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \mu^{\alpha,\mu}} \left(\tilde{c}_1 \left(h - \tilde{c}_2 \int_{\mathbb{R}} y \mathrm{d}\mu^{\alpha,\mu}(y) \right)^2 + \tilde{c}_5 \left(\int_{\mathbb{R}} y \mathrm{d}\mu^{\alpha,\mu}(y) \right)^2 \right)(x) \\
&= -2\tilde{c}_1\tilde{c}_2 x \left(h - \tilde{c}_2 \int_{\mathbb{R}} y \mathrm{d}\mu^{\alpha,\mu}(y) \right) + 2\tilde{c}_5 x \int_{\mathbb{R}} y \mathrm{d}\mu^{\alpha,\mu}(y) \\
&= -2\tilde{c}_1\tilde{c}_2 x \left(h - \tilde{c}_2 m^{\alpha,\mu} \right) + 2\tilde{c}_5 x m^{\alpha,\mu},
\end{aligned}
$$

and

$$
\int_{\mathbb{R}} \frac{\partial H}{\partial \mu^{\alpha,\mu}} \left(h, -\dot{V}(h), \mu, \mu^{\alpha,\mu} \right)(x) \mathrm{d}\mu^{\alpha,\mu}(h) = -2\tilde{c}_1\tilde{c}_2(1-\tilde{c}_2) x m^{\alpha,\mu} + 2\tilde{c}_5 x m^{\alpha,\mu}.
$$

Finally, the HJB equation reduces to

$$
\begin{aligned}
&\beta V(x) + \frac{1}{2}\dot{V}(x)^2 - \frac{1}{2}\sigma^2 \ddot{V}(x) - c_1(x - c_2 m)^2 - c_3(x - c_4)^2 - \tilde{c}_1(x - \tilde{c}_2 m^{\alpha,\mu})^2 \\
&\quad - \tilde{c}_5 \left(m^{\alpha,\mu} \right)^2 + 2\tilde{c}_1\tilde{c}_2(1-\tilde{c}_2) x m^{\alpha,\mu} - 2\tilde{c}_5 x m^{\alpha,\mu} = 0.
\end{aligned} \tag{B.1}
$$

Using the following ansatz for the value function and its derivatives

$$
\begin{aligned}
V(x) &= \Gamma_2 x^2 + \Gamma_1 x + \Gamma_0, \\
\dot{V}(x) &= 2\Gamma_2 x + \Gamma_1, \\
\ddot{V}(x) &= 2\Gamma_2,
\end{aligned} \tag{B.2}
$$

we obtain the optimal control

$$
\hat{\alpha}(x) = -\dot{V}(x) = -2\Gamma_2 x - \Gamma_1. \tag{B.3}
$$

Plugging the ansatz (B.2) into the HJB (B.1) we have

$$
\begin{aligned}
&\left(\beta\Gamma_2 + 2\Gamma_2^2 - (c_1 + c_3 + \tilde{c}_1) \right) x^2 \\
&\quad + \left(\beta\Gamma_1 + 2\Gamma_2\Gamma_1 + 2c_1c_2 m + 2\tilde{c}_1\tilde{c}_2 m^{\alpha,\mu} + 2c_3c_4 + 2\tilde{c}_1\tilde{c}_2(1-\tilde{c}_2) m^{\alpha,\mu} - 2\tilde{c}_5 m^{\alpha,\mu} \right) x \\
&\quad + \beta\Gamma_0 + \frac{1}{2}\Gamma_1^2 - \sigma^2\Gamma_2 - c_1c_2^2 m^2 - \left(\tilde{c}_1\tilde{c}_2^2 + \tilde{c}_5 \right) (m^{\alpha,\mu})^2 - c_3c_4^2 = 0.
\end{aligned}
$$

The solution is given by

$$
\begin{aligned}
\Gamma_2 &= \frac{-\beta + \sqrt{\beta^2 + 8\left(c_1 + c_3 + \tilde{c}_1\right)}}{4}, \\
\Gamma_1 &= \frac{2\tilde{c}_5 m^{\alpha,\mu} - 2\tilde{c}_1\tilde{c}_2(2-\tilde{c}_2) m^{\alpha,\mu} - 2c_1c_2 m - 2c_3c_4}{\beta + 2\Gamma_2}, \\
\Gamma_0 &= \frac{c_1c_2^2 m^2 + \left(\tilde{c}_1\tilde{c}_2^2 + \tilde{c}_5 \right) (m^{\alpha,\mu})^2 + \sigma^2\Gamma_2 - \Gamma_1^2/2 + c_3c_4^2}{\beta}.
\end{aligned}
$$

Taking the expectation of the dynamics of $X_t^{\alpha,\mu}$ with the control $\hat{\alpha}(x)$, we obtain the following ODE for $m^{\hat{\alpha},\mu}$:

$$\dot{m}_t^{\hat{\alpha},\mu} = -2\Gamma_2 m_t^{\hat{\alpha},\mu} - \Gamma_1,$$

which is solved by

$$\begin{aligned} m^{\hat{\alpha},\mu} &= \lim_{t\to\infty} m_t^{\hat{\alpha},\mu} = \lim_{t\to\infty} \left(-\frac{\Gamma_1}{2\Gamma_2} + \left(m_0^{\hat{\alpha},\mu} + \frac{\Gamma_1}{\Gamma_2} \right) \mathrm{e}^{-2\Gamma_2 t} \right) \\ &= -\frac{\Gamma_1}{2\Gamma_2} = -\frac{2\tilde{c}_5 m^{\hat{\alpha},\mu} - 2\tilde{c}_1\tilde{c}_2(2-\tilde{c}_2) m^{\hat{\alpha},\mu} - 2c_1c_2m - 2c_3c_4}{2\Gamma_2(\beta + 2\Gamma_2)}. \end{aligned}$$

From the fixed point condition $m = m^{\hat{\alpha},\mu}$, we deduce

$$\hat{m} = m^{\hat{\alpha},\hat{\mu}} = \frac{c_3c_4}{c_1(1-c_2) + \tilde{c}_1(1-\tilde{c}_2)^2 + c_3 + \tilde{c}_5},$$

and the explicit form of the optimal control (B.3)

$$\hat{\alpha}(x) = -2\Gamma_2 \left(x - \hat{m} \right). \tag{B.4}$$

Note that $\mu^{\hat{\alpha},\hat{\mu}} = \mathcal{N}(\hat{m}, \sigma^2/(4\Gamma_2))$ is the limiting distribution of the OU process $(X_t^{\hat{\alpha},\hat{\mu}})$.

B.2 Solution of the Traders' problem

In order to solve this problem, one first freezes the flow (θ_t) as in the MFG problem, and then solves the control problem which is of MKV type due to the term $m_t = \mathbb{E}(X_t)$ of MFC style. Differentiating the corresponding Hamiltonian with respect to α, one gets

$$\hat{\alpha}_t = -\frac{1}{c_\alpha} Y_t.$$

On the other hand,

$$dY_t = -\big(- c_h \mathbb{E}[\hat{\alpha}_t] + c_X \mathbb{E}[X_t] \big) dt + Z_t dW_t,$$

which leads to the following FBSDE:

$$\begin{cases} dX_t = -\dfrac{1}{c_\alpha} Y_t dt + \sigma dW_t, & X_0 \sim \mu_0, \\ dY_t = -\left(\dfrac{c_h}{c_\alpha} \mathbb{E}[Y_t] + c_X \mathbb{E}[X_t] \right) dt + Z_t dW_t, & Y_T = c_g X_T. \end{cases}$$

Note that this is a different system than the one studied in [5, Section 6.2.]. Taking expectation in this system one obtains

$$\begin{cases} d\mathbb{E}[X_t] = -\dfrac{1}{c_\alpha} \mathbb{E}[Y_t] dt, & \mathbb{E}[X_0] = x_0, \\ d\mathbb{E}[Y_t] = -\left(\dfrac{c_h}{c_\alpha} \mathbb{E}[Y_t] + c_X \mathbb{E}[X_t] \right) dt, & \mathbb{E}[Y_T] = c_g \mathbb{E}[X_T]. \end{cases}$$

Solving this system leads to

$$\mathbb{E}[Y_t] = \bar{\eta}(t)\mathbb{E}[X_t],$$

where

$$\bar{\eta}_t = \frac{-C(e^{(\delta^+-\delta^-)(T-t)}-1) - c_g(\delta^+ e^{(\delta^+-\delta^-)(T-t)} - \delta^-)}{(\delta^- e^{(\delta^+-\delta^-)(T-t)} - \delta^+) - c_g B(e^{(\delta^+-\delta^-)(T-t)}-1)}$$

for $t \in [0,T]$, $B = 1/c_\alpha$, $C = c_X$, $\delta^\pm = -D \pm \sqrt{R}$, with $D = -c_h/(2c_\alpha)$ and $R = D^2 + BC$. Subsequently,

$$\mathbb{E}[X_t] = x_0 e^{-\int_0^t \frac{\bar{\eta}(s)}{c_\alpha} ds}.$$

From the FBSDE system for (X_t, Y_t, Z_t) and centering X_t and Y_t, one gets

$$\begin{aligned}
Y_t &= \eta(t)X_t + \psi(t),\\
\eta(t) &= \frac{c_\alpha c_g}{c_\alpha + c_g(T-t)},\\
Z_t &= \sigma\eta(t),\\
\psi(t) &= \big(\bar{\eta}(t) - \eta(t)\big)\mathbb{E}[X_t].
\end{aligned}$$

Finally, we recall that the optimal control is given by

$$\hat{\alpha}_t = -\frac{1}{c_\alpha} Y_t = -\frac{1}{c_\alpha}\big(\eta(t)X_t + \psi(t)\big).$$

Assuming that X_0 is $\mathcal{N}(x_0, \sigma_0^2)$-distributed and independent of W, X_t is normally-distributed with mean given above by

$$\mathbb{E}[X_t] = x_0 e^{-\int_0^t \frac{\bar{\eta}(s)}{c_\alpha} ds}$$

and variance easily computed from

$$dX_t = -\frac{1}{c_\alpha}\big(\eta(t)X_t + \psi(t)\big)dt + \sigma dW_t$$

to obtain

$$Var(X_t) = \sigma_0^2 e^{-\frac{2}{c_\alpha}\int_0^t \eta(s)ds} + \sigma^2 \int_0^t e^{-\frac{2}{c_\alpha}\int_s^t \eta(s')ds'} ds.$$

Acknowledgments

Work of J.-P. Fouque was supported by NSF grants DMS-1814091 and DMS-1953035. Use was made of computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 1720256) at UC Santa Barbara.

References

[1] Y. Achdou and I. Capuzzo-Dolcetta, Mean field games: Numerical methods, *SIAM J. Numer. Anal.*, **48**(3), 2010.

[2] Y. Achdou and M. Laurière, Mean field games and applications: Numerical aspects. In: *Mean Field Games. Lecture Notes in Mathematics*, P. Cardaliaguet and A. Porretta (Eds.), Springer, Vol. 2281, 249–307, 2020.

[3] A. Angiuli, J.-P. Fouque, R. Hu, and A. Raydan, Reinforcement learning for mean field control games in continuous spaces. In preparation.

[4] A. Angiuli, J.-P. Fouque, and M. Laurière, Unified reinforcement q-learning for mean field game and control problems, *Math. Control Signals Systems*, **34**:217–271, 2022.

[5] A. Angiuli, J.-P. Fouque, and M. Laurière, Reinforcement Learning for Mean Field Games, with Applications to Economics. In: *Machine Learning in Financial Markets: A Guide to Contemporary Practises*, A. Capponi and C.-A. Lehalle (Eds.), Cambridge University Press, 2023.

[6] A. Angiuli, J.-P. Fouque, M. Laurière, and M. Zhang, Convergence of a multiscale reinforcement q-learning algorithm for mixed mean field control games. In preparation.

[7] A. Angiuli, C. V. Graves, H. Li, J.-F. Chassagneux, F. Delarue, and R. Carmona, Cemracs 2017: Numerical probabilistic approach to mfg, *ESAIM Proc. Surveys*, **65**:84–113, 2019.

[8] A. Bensoussan, J. Frehse, and S. C. P. Yam, Mean Field Games and Mean Field Type Control Theory, *Springer Briefs in Mathematics*, Springer, 2013.

[9] V. S. Borkar, Stochastic approximation with two time scales, *Syst. Control. Lett.*, **29**(5):291–294, 1997.

[10] V. S. Borkar, Stochastic Approximation. A Dynamical Systems Viewpoint. In: *Texts and Readings in Mathematics*, Hindustan Book Agency Gurgaon, 2008.

[11] P. Cardaliaguet and S. Hadikhanloo, Learning in mean field games: The fictitious play, *ESAIM Control Optim. Calc. Var.*, **23**(2), 2017.

[12] R. Carmona, F. Delarue, Probabilistic Theory of Mean Field Games with Applications I-II, Springer, 2018.

[13] R. Carmona and D. Lacker, A probabilistic weak formulation of mean field games and applications, *Ann. Appl. Probab.*, **25**(3):1189–1231, 2015.

[14] R. Carmona and M. Laurière, Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: I–The ergodic case, *arXiv:1907.05980*, 2019.

[15] R. Carmona and M. Laurière, Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: II–The finite horizon case, *arXiv:1908.01613*, 2019.

[16] R. Carmona, M. Laurière, and Z. Tan, Linear-quadratic mean-field reinforcement learning: Convergence of policy gradient methods, Preprint, 2019.

[17] R. Carmona, M. Laurière, and Z. Tan, Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning, Preprint, 2019.

[18] B. Djehiche, A. Tcheukam and H. Tembine, Mean-field-type games in engineering, *arxiv.org/abs/1605.03281*, 2017.

[19] R. Elie, J. Perolat, M. Laurière, M. Geist, and O. Pietquin, On the convergence of model free learning in mean field games. In: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[20] J.-P. Fouque and Z. Zhang, Deep learning methods for mean field control problems with delay, *Front. Appl. Math. Stat.*, **6**(11), 2020.

[21] X. Guo, A. Hu, R. Xu, and J. Zhang, Learning mean-field games. In: *Advances in Neural Information Processing Systems*, 4966–4976, 2019.

[22] J. Han and R. Hu, Deep fictitious play for finding Markovian nash equilibrium in multi-agent games, *arxiv.org/abs/1912.01809*, 2020.

[23] M. Huang, P.E. Caines, and R.P. Malhamé, Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria, *IEEE Trans. Automat. Control*, **52**(9):1560–1571, 2007.

[24] M. Huang, R.P. Malhamé, and P.E. Caines, Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle, *Commun. Inf. Syst.*, **6**(3):221–251, 2006.

[25] J.-M. Lasry and P.-L. Lions, Mean field games, *Jpn. J. Math.*, **2**(1):229–260, 2007.
[26] M. Laurière and O. Pironneau, Dynamic programming for mean-field type control, *C. R. Math. Acad. Sci. Paris*, **352**(9):707–713, 2014.
[27] D. Mguni, J. Jennings, and E. M. de Cote, Decentralised learning in systems with many, many strategic agents. In: *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[28] J. Subramanian and A. Mahajan, Reinforcement learning in stationary mean-field games. In: *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, 2019.
[29] A.-S. Sznitman, Topics in propagation of chaos. In: *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, P.-L. Hennequin, (Ed.), Springer, 165–251 1991.
[30] J. Yang, X. Ye, R. Trivedi, H. Xu, and H. Zha, Deep mean field games for learning optimal behavior policy of large populations. In: *International Conference on Learning Representations*, 2018.
[31] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, Mean field multi-agent reinforcement learning. In: *International Conference on Machine Learning*, 5567–5576, 2018.