Mathematics of Control, Signals, and Systems (2022) 34:217–271
https://doi.org/10.1007/s00498-021-00310-1

ORIGINAL ARTICLE

Unified reinforcement Q-learning for mean field game and control problems

Andrea Angiuli[1] · **Jean-Pierre Fouque**[1] · **Mathieu Laurière**[2]

Received: 15 October 2020 / Accepted: 8 October 2021 / Published online: 15 January 2022

Abstract

We present a Reinforcement Learning (RL) algorithm to solve infinite horizon asymptotic Mean Field Game (MFG) and Mean Field Control (MFC) problems. Our approach can be described as a unified two-timescale Mean Field Q-learning: The *same* algorithm can learn either the MFG or the MFC solution by simply tuning the ratio of two learning parameters. The algorithm is in discrete time and space where the agent not only provides an action to the environment but also a distribution of the state in order to take into account the mean field feature of the problem. Importantly, we assume that the agent cannot observe the population's distribution and needs to estimate it in a model-free manner. The asymptotic MFG and MFC problems are also presented in continuous time and space, and compared with classical (non-asymptotic or stationary) MFG and MFC problems. They lead to explicit solutions in the linear-quadratic (LQ) case that are used as benchmarks for the results of our algorithm.

Keywords Q-learning · Mean field game · Mean field control · Timescales · Linear-quadratic control

Mathematics Subject Classification 93E35 · 60J20 · 90C40

Published in the topical collection *Machine Learning for Control Systems and Optimal Control*.

J. Fouque: Work supported by NSF Grant DMS-1814091 M. Laurière: Work supported by NSF Grant DMS-1716673 and ARO Grant W911NF-17-1-0578.

✉ Mathieu Laurière
lauriere@princeton.edu

Andrea Angiuli
angiuli@pstat.ucsb.edu

Jean-Pierre Fouque
fouque@pstat.ucsb.edu

[1] Department of Statistics and Applied Probability, University of California, South Hall 5504, Santa Barbara, CA 93106, USA

[2] Department of Operations Research and Financial Engineering, Princeton University, Princeton, USA

1 Introduction

Reinforcement learning (RL) is a branch of machine learning (ML) which studies the interactions of an agent within an environment in order to maximize a reward signal. RL algorithms solve Markov Decision Processes (MDP) based on trials and errors. At each discrete time n, the agent observes the state of the environment X_n and chooses an action A_n. Due to the agent's action, the environment evolves to a state X_{n+1} and assigns a reward r_{n+1}. The goal of the agent is to find the optimal strategy π which assigns to each state of the environment the optimal action in order to maximize the aggregate discounted rewards. A complete overview on the evolution of this field is given in [28]. The Q-learning method was introduced by [29] to solve a discrete time MDP with finite state and action spaces. It is based on the evaluation of the optimal action-value table, $Q(x, a)$, which represents the expected aggregate discounted rewards when starting in state x and choosing the first action a, i.e.,

$$Q^*(x, a) = \max_{\pi} \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n r_{n+1} \,\middle|\, X_0 = x, A_0 = a\right], \tag{1}$$

where $r_{n+1} = r(X_n, \pi(X_n))$ is the instantaneous reward, $\gamma \in (0, 1)$ is a discounting factor, and $X_{n+1} = b(X_n, \pi(X_n))$. The maximum is taken over strategies (or policies) π, which are functions of the state taking values in some action space. Since the state's dynamics b (and sometimes the reward function r) are unknown to the agent, the algorithm is characterized by the trade-off between exploration of the environment and exploitation of the current available information. This is typically accomplished by the implementation of an ϵ-greedy policy. The greedy action which maximizes the immediate reward is chosen with probability $1 - \epsilon$ and a random action otherwise, i.e.,

$$\pi^{\epsilon}(x) = \begin{cases} a \in Unif(A), & \text{with probability } \epsilon, \\ a^* = \arg\max_{a \in A} Q(x, a), & \text{with probability } 1 - \epsilon. \end{cases} \tag{2}$$

Note that this is the randomized policy which will be used in the algorithm presented in Sect. 4, but as the optimal strategies will turn out to be deterministic (as ϵ goes to zero over learning episodes), in the following, we present the problems and the Q-learning approach only using deterministic policies called controls and denoted by α instead of π (see [25] for additional details on randomized policies).

On the other hand, and to summarize, mean field games are the result of the application of mean field techniques from physics into game theory. The mean field interaction is introduced to describe the behavior of a large number N of indistinguishable players with symmetric interactions. The complexity of the system would be intractable if we were to describe all the pairwise interactions. A solution to this problem is given by describing the interactions of each player i with the empirical distribution of the other players. As the number of players increases, the impact of each of them on the empirical distribution decreases. By the principle of propagation of chaos (law of large numbers), each player becomes asymptotically independent from the others and its

interaction is with its own distribution making the statistical structure of the system simpler. Two types of mean field problems can be distinguished between a mean field game and a mean field control depending on the goal the agents try to achieve. The aim of a mean field game is to find an equivalent of a Nash equilibrium in a non-cooperative N-player game when the number of players becomes large. On the other hand, a mean field control problem analyzes the social optimum in a cooperative game within a large population. Since the seminal works [23], and [21,22], the research in mean field game theory attracted a huge interest. We refer to the extensive works [8], and [4] for further details. Connections between machine learning and mean field theory have been proposed in the recent literature. Some model-based methods have first been introduced in [9,10,15] by combining neural network approximation tools and stochastic gradient descent. Furthermore, model-free methods and links with reinforcement learning have also attracted a surge of interest. [32] analyzes the benefits that a mean field (local) interaction brings in a multi-agent reinforcement learning (MARL) algorithm when the number of player is finite. [31] uses inverse reinforcement learning to learn the dynamics of a mean field game on a graph. [19] defines a simulator-based Q-learning algorithm to solve a mean field game with finite state and action spaces. [27] designs a gradient-based algorithm to solve cooperative games (MFC) and a two-timescale approach to solve non-cooperative games (MFG) with finite state and action spaces, analogously to [24]. Convergence of actor-critic method for linear-quadratic MFG [16] and convergence regularized Q-learning for MFG with finite state and action spaces [1] have also been proved. To learn MFC optima, model-free policy gradient methods have been proved to converge for LQ problems in [11], whereas Q-learning for a "lifted" MDP on the space of distributions has been introduced in [12]. To learn MFG equilibria, the fictitious play scheme has been introduced in [7], assuming the best response can be computed exactly. [13] analyses the propagation of error when the best response is computed approximately in a model-free setting, while [26] extends the analysis of the fictitious play scheme in continuous time of learning. Similarly to our approach, [30] studies a single-loop fictitious play algorithm in which the state and the policy are updated at each iteration. Fictitious play combined with deep neural networks has also been used to compute Nash equilibria in multi-agent games [20].

In this paper, we propose a mean field Q-learning algorithm which is able to solve the mean field game or mean field control problem depending on the tuning of the parameters and the rate of update of the distribution. Differently from the approach developed by [19], the algorithm does not require a simulator of the population simplifying its application to real-world problems. It exploits the mean field limit transposing the interaction of the player with the population to the interaction of the player with herself.

In Sect. 2, we formulate in discrete time and space the type of infinite horizon Asymptotic MFG and MFC problems that our algorithm will address. Comparison with classical (non-asymptotic) and stationary problems is also made. In Sect. 3, we recast them as a two-timescale problem of Borkar's type [5,6] which provides convergence results. The algorithm itself is presented in Sect. 4. In Sect. 5, we show numerical results with comparison to the benchmark case of discrete time and space approximations for continuous time and space linear-quadratic problems for which we have explicit formulas derived in Appendix A.

2 Mean field game and mean field control problems

We start by presenting three formulations of MFG and MFC problems: non-asymptotic, asymptotic, and stationary. All these problems are on an infinite horizon and for the sake of consistency with the RL literature, we present them in a discrete time and space framework. We will, however, resort to continuous time and space models in Sect. 5 in order to obtain simple benchmarks. Note that, as customary in the MFG literature, without loss of generality, we minimize a cost instead of maximizing a reward.

Let $\mathcal{X}$ and $\mathcal{A}$ be finite sets corresponding to states and actions. We denote by $\Delta^{|\mathcal{X}|}$ the simplex in dimension $|\mathcal{X}|$, which we identify with the space of probability measures on $\mathcal{X}$. Let $p : \mathcal{X} \times \mathcal{A} \times \Delta^{|\mathcal{X}|} \to \Delta^{|\mathcal{X}|}$ be a transition kernel. We will sometimes view it as a function:

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

which will be interpreted as the probability, at any given time step, to jump to state x' when starting from state x and using action a and when the population distribution is μ.

Let $f : \mathcal{X} \times \mathcal{A} \times \Delta^{|\mathcal{X}|} \to \mathbb{R}$ be a running cost function. We interpret $f(x, a, \mu)$ as the one-step cost, at any given time step, incurred to a representative agent who is at state x and uses action a while the population distribution is μ. For a random variable X, we denote its law by $\mathcal{L}(X)$. We will focus on feedback controls, i.e., functions of the state of the agent and possibly of time.

2.1 Non-asymptotic formulations

In the usual formulation for time-dependent MFG and MFC, the interactions between the players are through the distribution of states at the current time. More precisely, in a MFG, one typically looks for $(\hat{\alpha}, \hat{\boldsymbol{\mu}})$ where $\hat{\alpha} : \mathbb{N} \times \mathcal{X} \to \mathcal{A}$ and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_n)_{n\geq 0} \in (\Delta^{|\mathcal{X}|})^{\mathbb{N}}$ is a flow of probability distributions on $\mathcal{X}$, such that the following two conditions hold:

1. Optimality of the best response map: $\hat{\alpha}$ is the minimizer of

$$\alpha \mapsto J^{MFG}(\alpha; \hat{\boldsymbol{\mu}}) = \mathbb{E}\left[\sum_{n=0}^{+\infty} \gamma^n f(X_n^{\alpha,\hat{\boldsymbol{\mu}}}, \alpha_n(X_n^{\alpha,\hat{\boldsymbol{\mu}}}), \hat{\mu}_n)\right],$$

   where $\alpha_n(\cdot) := \alpha(n, \cdot)$ and the process $X^{\alpha,\hat{\boldsymbol{\mu}}}$ follows the dynamics given by:

$$X_{n+1}^{\alpha,\hat{\boldsymbol{\mu}}} \sim p\left(\cdot|X_n^{\alpha,\hat{\boldsymbol{\mu}}}, \alpha_n(X_n^{\alpha,\hat{\boldsymbol{\mu}}}), \hat{\mu}_n\right)$$

   with initial distribution $X_0^{\alpha,\hat{\boldsymbol{\mu}}} \sim \mu_0$;
2. Fixed point condition: $\hat{\mu}_n = \mathcal{L}(X_n^{\hat{\alpha},\hat{\boldsymbol{\mu}}})$ for every $n \geq 0$.

In a MFC problem, the goal is to find α^* such that the following condition holds: α^* is the minimizer of

$$\alpha \mapsto J^{MFC}(\alpha) = \mathbb{E}\left[\sum_{n=0}^{+\infty} \gamma^n f(X_n^\alpha, \alpha_n(X_n^\alpha), \mathcal{L}(X_n^\alpha))\right],$$

where the process X^α follows the dynamics:

$$X_{n+1}^\alpha \sim p\left(\cdot | X_n^\alpha, \alpha_n(X_n^\alpha), \mathcal{L}(X_n^\alpha)\right)$$

with initial distribution $X_0^\alpha \sim \mu_0$. Note that p is the same transition probability function as for the MFG above, but we plug the law $\mathcal{L}(X_n^\alpha)$ of X_n^α instead of a given distribution $\hat{\mu}_n$. In other words, the MFC problem is of McKean–Vlasov (MKV) type.

We will sometimes use the notation $\boldsymbol{\mu}^* = \boldsymbol{\mu}^{\alpha^*}$ for the optimal distribution in the MFC. Note that the objective function in the MFC setting can be written in terms of the objective function in the MFG as:

$$J^{MFC}(\alpha) = J^{MFG}(\alpha; \boldsymbol{\mu}^\alpha),$$

where $\mu_n^\alpha = \mathcal{L}(X_n^\alpha)$ for all $n \geq 0$. However, in general,

$$J^{MFC}(\alpha^*) = J^{MFG}(\alpha^*; \boldsymbol{\mu}^*) \neq J^{MFG}(\hat{\alpha}; \hat{\boldsymbol{\mu}}).$$

In these two problems, the equilibrium control $\hat{\alpha}$ or the optimal control α^* usually depends on time due to the dependence of p and f on the mean field flow, which evolves with time.

Although these are the usual formulations of MFG and MFC problems, in order to draw connections with reinforcement learning more directly, we turn our attention to formulations in which the control is independent of time. That is naturally the case in some applications, and, roughly speaking, it is also in the spirit of an individual player trying to optimally join a crowd of players already in the long-time asymptotic equilibrium. This will be made more precise in the following section.

2.2 Asymptotic formulations

We consider the following MFG problem: Find $(\hat{\alpha}, \hat{\mu})$ where $\hat{\alpha} : \mathcal{X} \to \mathcal{A}$ and $\hat{\mu} \in \Delta^{|\mathcal{X}|}$, such that the following two conditions hold:

1. $\hat{\alpha}$ is the minimizer of

$$\alpha \mapsto J^{AMFG}(\alpha; \hat{\mu}) = \mathbb{E}\left[\sum_{n=0}^{+\infty} \gamma^n f(X_n^{\alpha,\hat{\mu}}, \alpha(X_n^{\alpha,\hat{\mu}}), \hat{\mu})\right],$$

where the process $X^{\alpha,\hat{\mu}}$ follows the transitions:

$$X_{n+1}^{\alpha,\hat{\mu}} \sim p\left(\cdot|X_n^{\alpha,\hat{\mu}}, \alpha(X_n^{\alpha,\hat{\mu}}), \hat{\mu}\right)$$

with initial distribution $X_0^{\alpha,\hat{\mu}} \sim \mu_0$;

2. $\hat{\mu} = \lim_{n\to+\infty} \mathcal{L}(X_n^{\hat{\alpha},\hat{\mu}})$.

We stress that in this problem the control is a function of the state only and does not depend on time, as b and f depend only on the limiting distribution but not on time. Intuitively, this problem corresponds to the situation in which an infinitesimal player wants to join a crowd of players who are already in the asymptotic regime (as time goes to infinity). This stationary distribution is a Nash equilibrium if the new player joining the crowd has no interest in deviating from this asymptotic behavior.

We also consider the following MFC problem: Find α^* such that the following condition holds: α^* is the minimizer of

$$\alpha \mapsto J^{AMFC}(\alpha) = \mathbb{E}\left[\sum_{n=0}^{+\infty} \gamma^n f(X_n^{\alpha}, \alpha(X_n^{\alpha}), \mu^{\alpha})\right],$$

where the process X^{α} follows the transitions

$$X_{n+1}^{\alpha} \sim p\left(\cdot|X_n^{\alpha}, \alpha(X_n^{\alpha}), \mu^{\alpha}\right)$$

with initial distribution $X_0^{\alpha} \sim \mu_0$, and with the notation $\mu^{\alpha} = \lim_{n\to+\infty} \mathcal{L}(X_n^{\alpha})$.

We will sometimes use the shorthand notation $\mu^* = \mu^{\alpha^*}$ for the optimal distribution in the MFC setting. Here too, the control is independent of time, and p and f depend only on the limiting distribution. Intuitively, this problem can be viewed as the one posed to a central planner who wants to find the optimal stationary distribution such that the cost for the society is minimal when a new agent joins the crowd.

Note that in this formulation again, the objective function in the MFC setting can be written in terms of the objective function in the MFG as:

$$J^{AMFC}(\alpha) = J^{AMFG}(\alpha; \mu^{\alpha}),$$

with the notation $\mu^{\alpha} = \lim_{n\to+\infty} \mathcal{L}(X_n^{\alpha})$.

Remark 1 Although the AMFG and AMFC problems in this section are defined using an initial distribution μ_0 for the state process, one expects that under suitable conditions, *ergodicity* in particular, the optimal controls $\hat{\alpha}$ and α^* are independent of this initial distribution.

2.3 Stationary formulations

Another formulation with controls independent of time consists in looking at the situation in which the new agent joining the crowd starts with a position drawn according

to the ergodic distribution of the equilibrium control or the optimal control. This type of problems has been considered, e.g., in [19,27] and can be described as follows.

The stationary MFG problem is to find $(\hat{\alpha}, \hat{\mu})$ where $\hat{\alpha} : \mathcal{X} \to \mathcal{A}$ and $\hat{\mu} \in \Delta^{|\mathcal{X}|}$, such that the following two conditions hold:

1. $\hat{\alpha}$ is the minimizer of

$$\alpha \mapsto J^{SMFG}(\alpha; \hat{\mu}) = \mathbb{E}\left[\sum_{n=0}^{+\infty} \gamma^n f(X_n^{\alpha,\hat{\mu}}, \alpha(X_n^{\alpha,\hat{\mu}}), \hat{\mu})\right],$$

where the process $X^{\alpha,\hat{\mu}}$ follows the SDE

$$X_{n+1}^{\alpha,\hat{\mu}} \sim p\left(\cdot | X_n^{\alpha,\hat{\mu}}, \alpha(X_n^{\alpha,\hat{\mu}}), \hat{\mu}\right),$$

and starts with distribution $X_0^{\alpha,\hat{\mu}} \sim \hat{\mu}$;

2. The process $X^{\hat{\alpha},\hat{\mu}}$ admits $\hat{\mu}$ as invariant distribution (so $\hat{\mu} = \mathcal{L}(X_n^{\hat{\alpha},\hat{\mu}})$ for all $n \geq 0$).

The key difference with the Asymptotic MFG formulation is that here the process starts with the invariant distribution $\hat{\mu}$. The control is a function of the state only and does not depend of time, and p and f depend only on this stationary distribution.

The stationary MFC problem is defined as follows: Find α^* such that the following condition holds: α^* is the minimizer of

$$\alpha \mapsto J^{SMFC}(\alpha) = \mathbb{E}\left[\sum_{n=0}^{+\infty} \gamma^n f(X_n^{\alpha}, \alpha(X_n^{\alpha}), \mu^{\alpha})\right],$$

where the process X^{α} follows the MKV dynamics

$$X_{n+1}^{\alpha} \sim p\left(\cdot | X_n^{\alpha}, \alpha(X_n^{\alpha}), \mu^{\alpha}\right),$$

with initial distribution $X_0^{\alpha} \sim \mu^{\alpha}$, and such that μ^{α} is the invariant distribution of X^{α} (assuming it exists).

To conclude, let us mention that there is yet another formulation, in which the solution is stationary but depends on the initial distribution, see [4, Chapter 7].

2.4 Connecting the three formulations

Denoting by $\hat{\alpha}^{MFG}$, $\hat{\alpha}^{AMFG}$, and $\hat{\alpha}^{SMFG}$, the MFG equilibrium strategies, respectively, in the non-asymptotic, asymptotic, and stationary formulations, we expect

$$\begin{cases} \hat{\alpha}_n^{MFG}(x) \to \hat{\alpha}^{AMFG}(x), & \forall x, \quad \text{as} \quad n \to +\infty, \\ \hat{\alpha}^{AMFG}(x) = \hat{\alpha}^{SMFG}(x), & \forall x. \end{cases} \tag{3}$$

Similarly denoting by $\alpha^{*MFC}, \alpha^{*AMFC}$, and α^{*SMFC}, the MFC optimal controls, respectively, in the non-asymptotic, asymptotic, and stationary formulations, we expect

$$\begin{cases} \alpha_n^{*MFC}(x) \to \alpha^{*AMFC}(x), & \forall x, \quad \text{as} \quad n \to +\infty, \\ \alpha^{*AMFC}(x) = \alpha^{*SMFC}(x), & \forall x. \end{cases} \tag{4}$$

In fact, we have the following result.

Theorem 1 *Consider the set of admissible controls to be defined as the set of controls α such that the process $(X_n^\alpha)_{n\geq 0}$ is an irreducible and aperiodic Markov process on the finite space X. If a solution for the asymptotic MFG (resp. MFC) exists, then it is equal to the solution of the corresponding stationary MFG (resp. MFC) and vice versa.*

Proof Let us consider the pair $(\hat{\alpha}^{AMFG}, \hat{\mu}^{AMFG})$ solution of an asymptotic MFG. The optimal control $\hat{\alpha}^{AMFG}$ is an optimizer over the set of admissible controls such that the process $(X_n^\alpha)_{n\geq 0}$ is an irreducible Markov process and admits a limiting distribution which is then the unique invariant distribution using the control $\hat{\alpha}^{AMFG}$. Note that the control $\hat{\alpha}^{AMFG}$ does not depend on the initial distribution μ_0 and consequently $\hat{\mu}^{AMFG}$ does not either. Therefore, $(\hat{\alpha}^{AMFG}, \hat{\mu}^{AMFG})$ is the solution of the AMFG starting from $\hat{\mu}^{AMFG}$, which is the corresponding stationary MFG problem. Thus, we deduce the desired relation $\hat{\alpha}^{AMFG} = \hat{\alpha}^{SMFG}$. A similar argument for MFC problems applies and shows that $\alpha^{*AMFC} = \alpha^{*SMFC}$. □

Remark 2 In terms of practical applications, the asymptotic formulation (AMFG and AMFC) seems to be the most appropriate, and if one is interested in the optimal controls, Theorem 1 shows that solving the asymptotic games also gives the solutions to the corresponding stationary games. Additionally, (3) and (4) indicate that it also gives the long-time solutions to the corresponding time-dependent games. Developing Q-learning algorithms for solving time-dependent finite horizon games is addressed in our forthcoming paper [2].

In Appendix A, we provide explicit solutions for MFG, AMFG, SMFG, MFC, AMFC, and SMFC, in the case of continuous time, continuous space Linear-Quadratic stochastic differential games. We verify that (3) and (4), and therefore, Theorem 1, are satisfied in that case as well. In Sect. 5, discrete approximations of these games will also serve as benchmarks for our algorithm described in Sect. 4.

3 A unified view of learning for MFG and MFC

In this section, we draw a connection between MFG, MFC, Q-learning, and Borkar's two timescale approach [5,6]. The definitions of MFG and MFC reveal that the two formulations are very similar and both involve an optimization and a distribution. This leads to the idea of designing an iterative procedure which would update the value function and the distribution. However, in the MFG, the distribution is frozen during

the optimization and then, a fixed point condition is enforced, whereas in the MFC problem the distribution is directly linked to the control, which implies that it should change instantaneously when the control function is modified. Hence, to compute the solutions using an iterative algorithm, the updates should be done differently for each problem: intuitively, in a MFG, the value function should be updated in an inner loop and the distribution in an outer loop, whereas it should be the converse for MFC. More generally, we can update both functions in turn but at different rates. Then, to compute the MFG solution, the distribution should be updated at a lower rate than the value function. For MFC, it should be the converse. In the rest of this subsection, we formalize these ideas.

3.1 Action-value function in the classical Q-learning setup

One of the most popular methods in RL is the so-called Q-learning [29]. Instead of looking at the value function V as in a PDE approach for optimal control, this method is based on the action-value function, also called Q-function, which takes as inputs not only a state x but also an action a. Intuitively, in a standard (non mean-field) MDP, this function quantifies the optimal cost-to-go of an agent starting at x, using action a for the first step and then acting optimally afterward. In other words, the value of (x, a) is the cost of using a when in state x, plus the minimal cost possible after that, i.e., the cost induced by using the optimal control; see, e.g., [28, Chapter 3] for more details. The definition of the optimal Q-function, denoted by Q^*, is similar to (1), up to a change of sign since we consider a cost f and a minimization problem instead of a reward r and a maximization problem, namely,

$$Q^*(x,a) = \min_{\alpha} \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n f(X_n, \alpha(X_n)) \,\middle|\, X_0 = x,\, A_0 = a\right].$$

Using dynamic programming, it can be shown that Q^* is the solution of the Bellman equation:

$$Q^*(x,a) = f(x,a) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a) \min_{a'} Q^*(x',a'), \qquad (x,a) \in \mathcal{X} \times \mathcal{A}.$$

The corresponding value function V^* is given by:

$$V^*(x) = \min_{a} Q^*(x,a), \qquad x \in \mathcal{X}.$$

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

3.2 Action-value function for asymptotic MFG

In the context of Asymptotic MFG introduced in Sect. 2.2, we can view the problem faced by an infinitesimal agent among the crowd as an MDP *parameterized* by the population distribution. Hence, given a population distribution μ, standard RL techniques can be applied to compute the Q-function of an infinitesimal agent against this given μ.

Then, the optimal Q-function is defined, for a given μ, by

$$Q^*_\mu(x,a) = \min_\alpha \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu) \,\middle|\, X_0 = x,\, A_0 = a\right], \tag{5}$$

where the cost function $f(x,a,\mu)$ depends on the fixed μ as well as the transition probabilities $p(x'|x,a,\mu)$. Since μ is fixed, as in the classical case, one obtains the Bellman equation:

$$Q^*_\mu(x,a) = f(x,a,\mu) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\mu) \min_{a'} Q^*_\mu(x',a'), \qquad (x,a) \in \mathcal{X} \times \mathcal{A}. \tag{6}$$

This function characterizes the optimal cost-to-go for an agent starting at state x, using action a for the first step, and then acting optimally for the rest of the time steps, while the population distribution is given by μ (for every time step). Note that $\min_a Q^*_\mu(x,a) = \min_\alpha J^{AMFG}(\alpha;\mu)$ in the notation of Sect. 2.2.

3.3 Action-value function for asymptotic MFC

For MFC, it is not obvious how to use the same Q-function because, as noticed earlier, the distribution appearing in the definition of MFC is directly linked to the control and not fixed a priori. One possibility is to look at MFC as an MDP on the space of distributions and then to introduce a Q-function which takes a distribution as an input [12,17,18,25].

We take a different route and consider a modified Q- function as follows. For an admissible control $\alpha(x)$, we define the MKV- dynamics $p(x'|x,a,\mu^\alpha)$ so that μ^α is the limiting distribution of the associated process (X^α_n). We define the control $\tilde{\alpha}$ by

$$\tilde{\alpha}(x') = \begin{cases} a & \text{if} \quad x' = x, \\ \alpha(x) & \text{for } x' \neq x. \end{cases} \tag{7}$$

Note that $\tilde{\alpha}$ depends on x and a. Our modified Q-function is given by

$$Q^\alpha(x,a) = f(x,a,\mu^{\tilde{\alpha}}) + \mathbb{E}\left[\sum_{n=1}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^\alpha) \,\middle|\, X_0 = x,\, A_0 = a\right].$$

We then obtain that the optimal $Q^*(x,a) = \min_\alpha Q^\alpha(x,a)$ satisfies the Bellman equation

$$Q^*(x,a) = f(x,a,\tilde{\mu}^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\tilde{\mu}^*) \min_{a'} Q^*(x',a'), \qquad (x,a) \in \mathcal{X} \times \mathcal{A}, \tag{8}$$

where the optimal control α^* is given by $\alpha^*(x) = \arg\min_a Q^*(x,a)$, the control $\tilde{\alpha}^*$ is defined by (7) for x and a, and $\tilde{\mu}^* := \mu^{\tilde{\alpha}^*}$. The optimal value function is $V^*(x) = \min_a Q^*(x,a)$ ($= J^{AMFC}(\alpha^*)$ in the notation of Sect. 2.2). The details of the derivation of these equations are given in Appendix C.

Note that, compared with the Q_μ-function used for MFG, our MFC modified Q-function involves the differences $\Delta_\mu f := f(x,a,\tilde{\mu}) - f(x,a,\mu)$ and $\Delta_\mu p := p(\cdot|x,a,\tilde{\mu}) - p(\cdot|x,a,\mu)$ which play the role of derivatives with respect to the probability distribution in the classical continuous time and space Mean Field Control problems.

3.4 Unification through a two timescale approach

The goal is now to design a learning procedure which can approximate, for either MFG or MFC, not only Q but also the corresponding μ. For MFG, the usual fixed point iterations are on the distribution and at each iteration, the best response against this distribution (which can be deduced from the corresponding Q table) is computed. For MFC, the iterations are on the control (here again, it can be deduced from the Q table) and the distribution corresponding to this control is computed at each iteration. Instead of completely freezing the distribution (resp. the control) in the first case (resp. the second case), we can imagine that letting it evolve at a slow rate would still lead to the same limit. In other words, the definitions of MFG and MFC seem to lie at the two opposite sides of a spectrum.

Based on this viewpoint, we consider the following iterative procedure, where both variables (Q and μ) are updated at each iteration but with different rates. Starting from an initial guess $(Q_0, \mu_0) \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \times \Delta^{|\mathcal{X}|}$, define iteratively for $k = 0, 1, \ldots$:

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^\mu \mathcal{P}(Q_k, \mu_k), & \text{(9a)} \\ Q_{k+1} = Q_k + \rho_k^Q \mathcal{T}(Q_k, \mu_k), & \text{(9b)} \end{cases}$$

where

$$\begin{cases} \mathcal{P}(Q,\mu)(x) = (\mu P^{Q,\mu})(x) - \mu(x), \qquad x \in \mathcal{X}, \\ \mathcal{T}(Q,\mu)(x,a) = f(x,a,\mu) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q(x',a') \\ -Q(x,a), \qquad (x,a) \in \mathcal{X} \times \mathcal{A}, \end{cases}$$

and

$$P^{Q,\mu}(x,x') = p(x'|x, \arg\min_a Q(x,a), \mu),$$

$$(\mu P^{Q,\mu})(x) = \sum_{x_0} \mu(x_0) P^{Q,\mu}(x_0, x),$$

$P^{Q,\mu}$ is the transition matrix when the population distribution is μ and the agent uses the optimal control according to Q. The learning rates ρ_k^μ and ρ_k^Q are assumed to satisfy usual Robbins–Monro type conditions, namely: $\sum_k \rho_k^\mu = \sum_k \rho_k^Q = +\infty$ and $\sum_k |\rho_k^\mu|^2 = \sum_k |\rho_k^Q|^2 < +\infty$.

If $\rho_k^\mu < \rho_k^Q$, the approximate Q-function evolves faster, while it is the converse if $\rho_k^\mu > \rho_k^Q$. This suggests that these two regimes should converge to different limit points. These ideas have been studied by Borkar [5,6] in connection with reinforcement learning methods under the name of two timescales approach. More precisely, from Borkar [6, Chapter 6, Theorem 2], we expect to have the following two situations. We assume that the operators $\mathcal{T}$ and $\mathcal{P}$ are Lipschitz continuous, which, as explained in Appendix B, can be obtained from the Lipschitz continuity of f and p in the model, as well as a slight modification of $\mathcal{P}$ to regularize the minimizer. Furthermore, for every Q, the function $\mu \mapsto \mu + \rho^\mu \mathcal{P}(Q, \mu)$ is a strict contraction for ρ^μ small enough (depending on p), which ensures existence and uniqueness of a fixed point to this function. Similarly, for every μ the function $Q \mapsto Q + \rho^Q \mathcal{T}(Q, \mu)$ is a strict contraction for ρ^Q small enough (depending on f, p and γ), which ensures existence and uniqueness of a fixed point to this function.

– **Two timescale approach for MFG**.
If $\rho_k^\mu / \rho_k^Q \to 0$ as $k \to +\infty$, the system (9a)–(9b) tracks the ODE system

$$\begin{cases} \dot{\mu}_t = \mathcal{P}(Q_t, \mu_t), \\ \dot{Q}_t = \dfrac{1}{\epsilon} \mathcal{T}(Q_t, \mu_t), \end{cases}$$

where ρ_k^μ / ρ_k^Q is thought of being of order $\epsilon \ll 1$. We consider, for any fixed μ, the ODE

$$\dot{Q}_t = \frac{1}{\epsilon} \mathcal{T}(Q_t, \mu),$$

and we assume it has a unique globally asymptotically stable equilibrium Q_μ. In particular, $\mathcal{T}(Q_\mu, \mu) = 0$, meaning by (6) that Q_μ is the value function of an infinitesimal agent facing the crowd distribution μ. We further assume that Q_μ is Lipschitz continuous with respect to μ. Convergence to Q_μ can be obtained following standard arguments for Q-learning (see, *e.g.*, [6, Section 10.3]) and the Lipschitz continuity of Q_μ can be guaranteed through Lipschitz continuity of f, p and the minimizer in (5). Then, the first ODE becomes

$$\dot{\mu}_t = \mathcal{P}(Q_{\mu_t}, \mu_t).$$

Assuming it has a unique globally asymptotically stable equilibrium μ_∞, this distribution satisfies

$$\mathcal{P}(Q_{\mu_\infty}, \mu_\infty) = 0.$$

This condition implies that μ_∞ and the associated control given by $\hat{\alpha}(x) = \arg\min_a Q_{\mu_\infty}(x, a)$ form a Nash equilibrium. From [6, Chapter 6, Theorem 2], the system (9a)–(9b) with discrete time updates also converges to this Nash equilibrium when $\rho_k^\mu / \rho_k^Q \to 0$ as $k \to +\infty$.

– **Two timescale approach for MFC**.
If $\rho_k^Q / \rho_k^\mu \to 0$ as $k \to +\infty$, the system (9a)–(9b) tracks the ODE system

$$\begin{cases} \dot{\mu}_t = \dfrac{1}{\epsilon}\mathcal{P}(Q_t, \mu_t), \\ \dot{Q}_t = \mathcal{T}(Q_t, \mu_t), \end{cases}$$

where ρ_k^Q / ρ_k^μ is thought of being of order $\epsilon \ll 1$. We consider, for any fixed Q, the ODE

$$\dot{\mu}_t = \frac{1}{\epsilon}\mathcal{P}(Q, \mu_t),$$

and we assume it has a unique globally asymptotically stable equilibrium μ_Q. In particular, $\mathcal{P}(Q, \mu_Q) = 0$, meaning that μ_Q is the asymptotic distribution of a population in which every agent uses the control $\alpha(x) = \arg\min_a Q(x, a)$. We further assume that μ_Q is Lipschitz continuous with respect to Q. Then, the second ODE becomes

$$\dot{Q}_t(x, a) = \mathcal{T}(Q_t(x, a), \widetilde{\mu}_{Q_t}),$$

where $\widetilde{\mu}_{Q_t}$ is defined by (7) at (x, a) for $\alpha(\cdot) = \arg\min_{a'} Q_t(\cdot, a')$. This is consistent with the update of Q and what the algorithm proposed in Sect. 4 does. Assuming this ODE has a unique globally asymptotically stable equilibrium Q_∞, this Q-table satisfies

$$\mathcal{T}(Q_\infty, \widetilde{\mu}_{Q_\infty}) = 0.$$

This last condition means that $Q_\infty = Q^*$ satisfies the MFC Bellman equation (8), and that the control $\alpha^*(x) = \arg\min_a Q_\infty(x, a)$ is an MFC optimum for the asymptotic formulation and the induced optimal distribution is μ_{Q_∞}. From [6, Chapter 6, Theorem 2], the system (9a)–(9b) with discrete time updates also converges to this social optimum when $\rho_k^Q / \rho_k^\mu \to 0$ as $k \to +\infty$.

3.5 Stochastic approximation

The above (deterministic) algorithm relies on the operators $\mathcal{P}$, $\mathcal{T}$ which, in many practical situations are not known, for instance because the agent does not know for sure the dynamics or the reward function. In such situations, the agent can only rely on random samples (more details are provided in the next section). The algorithm can be modified to account for such stochastic approximations. Indeed, let us assume that, for any Q, μ, x, a, the agent can know the value $f(x, a, \mu)$ and can sample a realization of the random variable

$$X'_{x,a,\mu} \sim p(\cdot|x, a, \mu).$$

Then, she can compute the realization of the following random variables $\check{\mathcal{T}}_{Q,\mu,x,a}$ and $\check{\mathcal{P}}_{Q,\mu,x,a}$ taking values, respectively, in $\mathbb{R}$ and $\Delta^{|\mathcal{X}|}$:

$$\check{\mathcal{T}}_{Q,\mu,x,a} = f(x, a, \mu) + \gamma \min_{a'} Q(X'_{x,a,\mu}, a') - Q(x, a),$$

and

$$\check{\mathcal{P}}_{Q,\mu,x,a}(x'') = \mathbf{1}_{\{X'_{x,a,\mu}=x''\}} - \mu(x''), \qquad \forall x'' \in \mathcal{X}.$$

Observe that

$$\begin{aligned}\mathbb{E}[\check{\mathcal{T}}_{Q,\mu,x,a}] &= \sum_{x'} p(x'|x, a, \mu) \left[f(x, a, \mu) + \gamma \min_{a'} Q(x', a') - Q(x, a) \right] \\ &= \mathcal{T}(Q, \mu)(x, a), \end{aligned} \tag{12}$$

and

$$\mathbb{E}[\check{\mathcal{P}}_{Q,\mu,x,a}(x'')] = \sum_{x'} p(x'|x, a, \mu) \left(\mathbf{1}_{\{x'=x''\}} - \mu(x'') \right) = p(x''|x, a, \mu) - \mu(x'').$$

If the starting point x comes from a random variable $X \sim \mu$ and if a is chosen to be an optimal action at X according to a given table Q, i.e., $a \in \arg\min_{\mathcal{A}} Q(X, \cdot)$, then we obtain

$$\begin{aligned}&\mathbb{E}[\check{\mathcal{P}}_{Q,\mu,X,\arg\min_a Q(X,a)}(x'')] \\ &= \sum_x \mu(x) \sum_{x'} p(x'|x, \arg\min_a Q(x, a), \mu) \left(\mathbf{1}_{\{x'=x''\}} - \mu(x'') \right) \\ &= \sum_x \mu(x) \left(p(x''|x, \arg\min_a Q(x, a), \mu) - \mu(x'') \right) \\ &= (\mu P^{Q,\mu})(x'') - \mu(x'') \\ &= \mathcal{P}(Q, \mu)(x''). \end{aligned} \tag{13}$$

We can thus replace the deterministic updates (9a)–(9b) by the following stochastic ones, starting from some initial Q_0, μ_0: for $k = 0, 1, \ldots,$

$$\begin{cases} \mu_{k+1}(x) = \mu_k(x) + \rho_k^{\mu} \breve{\mathcal{P}}_{Q_k,\mu_k,X_k,\arg\min_a Q(X_k,a)}(x) \\ \qquad\qquad = \mu_k(x) + \rho_k^{\mu} \mathcal{P}(Q_k,\mu_k)(x) + \mathbf{P}^k(x), \qquad \forall x \in \mathcal{X} \qquad (14a) \\ Q_{k+1}(x,a) = Q_k(x,a) + \rho_k^{Q} \breve{\mathcal{T}}_{Q_k,\mu_k,x,a} \\ \qquad\qquad = Q_k + \rho_k^{Q} \mathcal{T}(Q_k,\mu_k)(x,a) + \mathbf{T}^k(x,a), \qquad \forall (x,a) \in \mathcal{X} \times \mathcal{A}, \\ \qquad X_k \sim \mu_k, \qquad (14b) \end{cases}$$

where we introduced the notation:

$$\mathbf{P}^k(x) = \rho_k^{\mu} \Big(\breve{\mathcal{P}}_{Q_k,\mu_k,X_k,\arg\min_a Q_k(X_k,a)}(x) - \mathcal{P}(Q_k,\mu_k)(x) \Big), \qquad \forall x,$$

and

$$\mathbf{T}^k(x,a) = \rho_k^{Q} \Big(\breve{\mathcal{T}}_{Q_k,\mu_k,x,a} - \mathcal{T}(Q_k,\mu_k)(x,a) \Big), \qquad \forall (x,a),$$

with X_k sampled from μ_k. Note that $\mathbf{T}^k$ and $\mathbf{P}^k$ are martingales by the above remarks, see (12)–(13). Hence, under suitable conditions, we expect convergence to hold by classical stochastic approximation results [6].

However, the procedure (14a)–(14b) is synchronous (it updates all the coefficients of the Q-table and the distribution at each iteration k) and it requires having access to a generative model, i.e., to a simulator which can provide samples of transitions drawn according to $p(\cdot|x, a, \mu_k)$ for arbitrary state x. In the next section, we propose a procedure which works even with a more restricted setting, which uses episodes: In each episode, the learner is constrained to follow the trajectory sampled by the environment without choosing arbitrarily its state.

4 Reinforcement learning algorithm

As recalled in the Introduction, RL studies the algorithms to solve a Markov decision process (MDP) based on trials and errors. An MDP can be described through the interactions of an agent with an environment. At each time n, the agent observes its current state $X_n \in \mathcal{X}$ and chooses an action $A_n \in \mathcal{A}$. Due to the agent's action, the environment provides the new state of the agent X_{n+1} and incurs a cost f_{n+1}. The goal of the agent is to find an optimal strategy (or policy) π^* which assigns to each state an action in order to minimize the aggregated discounted costs. The idea is then to design methods which allow the agent to learn (an approximation of) π^* by making repeated use of the environment's outputs but without knowing how the environment produces the new state and the associated cost. A detailed overview of this field can be found in [28] (although RL methods are often presented with reward maximization objectives, we consider cost minimization problems for the sake of consistency with the MFG literature).

As presented in Sect. 3.1, the optimal strategy can be derived from the optimal action-value function. However, Q^* is a priori unknown. In order to learn Q^* by

trials and errors, an approximate version Q of the table Q^* is constructed through an iterative procedure. At each step, an action is taken, which leads to a cost and to a new state. On the one hand, it is interesting to act efficiently in order to avoid high costs, and on the other hand, it is important to improve the quality of the table Q by trying actions and states which have not been visited many times so far. This is the so-called exploitation–exploration trade-off. The trade-off between exploration of the unknown environment and exploitation of the currently available information can be taken care of by an ϵ-greedy policy based on Q. The algorithm chooses the action that minimizes the immediate cost with probability $1-\epsilon$, and a random action otherwise, as in (2) with an arg min.

4.1 U2-MF-QL: unified two timescales mean field Q-learning

In order to apply the RL paradigm to mean field problems, the first step consists in defining the connection between these two frameworks. In a MFG (resp. a MFC), the goal of a typical agent is to find the pair $(\hat{\alpha}, \hat{\mu})$ (resp. (α^*, μ^*)) where $\hat{\alpha} : \mathcal{X} \mapsto \mathcal{A}$ (resp. $\alpha^* : \mathcal{X} \mapsto \mathcal{A}$) represents the equilibrium (resp. optimal) strategy which assigns at each state the equilibrium (resp. optimal) action in order to minimize the aggregated discounted costs and $\hat{\mu}$ (resp. μ^*) is the ergodic distribution of the population at equilibrium (resp. optimum). The traditional definition of an MDP based on the agent–environment pair is augmented with the distribution of the population. In this new framework, the agent corresponds to the representative player of the mean field problem.

We now define the type of environment to which the agent is assumed to have access. A key difference with prior works on RL for mean field problems is that we do not assume that agent can witness the evolution of the population's distribution. Instead, the environment estimates the distribution of the population by exploiting the symmetry property of the problem. Indeed, when the system is at equilibrium, the law of the representative player matches the distribution of the population. As showed in the diagram of Fig. 1, at each time n, the agent observes its current state $X_n \in \mathcal{X}$ and then chooses an action $A_n \in \mathcal{A}$. An approximation of the distribution μ_n is computed by the environment based on the observed states of the representative player. Provided with the choice of the action and the estimate of the distribution, the environment generates the new state of the agent X_{n+1} and assigns a cost f_{n+1}.

The algorithm is designed to solve infinite horizon problems through an online approach, i.e., interacting with the environment. The learning procedure is based on splitting the infinite horizon in successive episodes in order to promote the exploration of the environment. The first episode is initialized based on the initial distribution of the representative player. Within a given episode, the agent updates her strategy at each learning step aiming to optimize the expected aggregated cost based on the current estimate of the distribution of the population μ_n. Changes in the representative player's strategy have an effect on the population requiring to update μ_n accordingly. After an assigned number of steps T, the episode is terminated. A new episode is initialized based on the current version of the environment represented by the estimate of the population obtained at the last time point of the previous episode. One may think at the

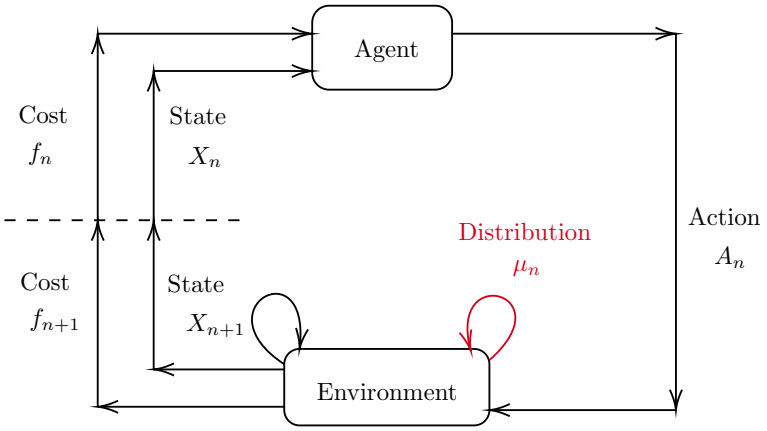

Fig. 1 MDP with mean field interactions: interaction of the representative agent with the environment. When the current state of the representative agent is X_n, given an action A_n, the environment produces an estimate of the distribution μ_n, the new state X_{n+1} and incurs a cost f_{n+1} calculated by starting from the current state of the environment X_n and using the transition controlled by A_n and parameterized by μ_n

initialization step as a change in the choice of the representative player who provides the data flow. As the number of episodes increases, one expects the distribution of the representative player to converge to the limiting distribution. Within a given learning step, the environment computes an estimate of μ_n based on the current state of the agent X_n, provides the next state X_{n+1} and assigns the cost f_{n+1} given the triple (X_n, A_n, μ_n). In other words, the environment consists of the dynamics of the agent and the cost structure. The case of our interest corresponds to the one in which the dynamics of the agent and the cost structure are unknown. In this way, introducing the RL paradigm is equivalent to define a data-driven approach to solve mean field models which may scale their applicability to real-world problems.

In contrast with standard Q-learning, since in the mean field framework the cost function also depends on the distribution of the population, the goal here consists in learning the optimal strategy along with the corresponding ergodic distribution of the population, i.e., $(\hat{\alpha}, \hat{\mu})$ in the MFG setting and (α^*, μ^*) in the MFC setting. Based on the intuition provided in Sect. 3 related to the two timescale approach, we propose Algorithm 1. At each step, we update the Q-table at the observed state-action pair $Q(X_n, A_n)$. With a different learning rate, the estimate of the distribution is updated based on the operator $\boldsymbol{\delta} : \mathcal{X} \mapsto \Delta^{|\mathcal{X}|}$ which maps the next observed state $X_{n+1} \in \mathcal{X}$ to the corresponding one-hot vector measure. To be specific, we identify the simplex $\Delta^{|\mathcal{X}|}$ with the subset $\left\{[\mu(x_i)]_{i=0,\dots,|\mathcal{X}|-1} \,:\, \mu(x_i) \in [0,1] \text{ and } \sum_i \mu(x_i) = 1\right\}$ of $\mathbb{R}^{|\mathcal{X}|}$. Then, $\boldsymbol{\delta}$ is the function which associates to each element of $\mathcal{X} = \{x_0, \dots, x_{|\mathcal{X}|-1}\}$ the corresponding element of the canonical basis $(e_0, \dots, e_{|\mathcal{X}|-1})$ of $\mathbb{R}^{|\mathcal{X}|}$, i.e., for each $i = 0, \dots, |\mathcal{X}| - 1$, $\boldsymbol{\delta}(x_i) = e_i$, which is an element of $\Delta^{|\mathcal{X}|}$ by the above identification. In order to learn the limiting distribution of the population through successive learning episodes, an estimate μ_{n_i} is computed for each step n_i based on the sample $X^k_{n_i}$ collected from episodes $k = 1, 2, \dots$. This approach attempts to minimize the

correlation of the sampled states. The update rule presented in algorithm 1 allocates more weight on the most recent samples allowing to forget progressively the initial sample that were obtained by a distribution far from the limiting one. At convergence, one may expect each μ_{n_i} to be an estimate of the limiting distribution.

The algorithm returns both an approximation μ_T^k of the distribution and an approximation Q^k of the Q-function, from which an approximation of the optimal control can be recovered as $x \mapsto \arg\min_{a\in\mathcal{A}} Q^k(x,a)$.

Algorithm 1 Unified Two Timescales Mean Field Q-learning - Tabular version

Require: T : number of time steps in a learning episode,
$\mathcal{X} = \{x_0, \dots, x_{|\mathcal{X}|-1}\}$: finite state space,
$\mathcal{A} = \{a_0, \dots, a_{|\mathcal{A}|-1}\}$: finite action space,
μ_0 : initial distribution of the representative player,
ϵ : parameter related to the ϵ-greedy policy,
tol_μ, tol_Q : break rule tolerances.

Initialization: $Q^0(x,a) = 0$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, $\mu_n^0 = \left[\frac{1}{|\mathcal{X}|}, \dots, \frac{1}{|\mathcal{X}|}\right]$ for $n = 0, \dots, T$
for each episode $k = 1, 2, \dots$ **do**
  Initialization: Sample $X_0^k \sim \mu_T^{k-1}$ and set $Q^k \equiv Q^{k-1}$
  for $n \leftarrow 0$ to $T-1$ **do**
    Update μ:
    $\mu_n^k = \mu_n^{k-1} + \rho_k^\mu(\boldsymbol{\delta}(X_n^k) - \mu_n^{k-1})$ where $\boldsymbol{\delta}(X_n^k) = \left[\mathbf{1}_{x_0}(X_n^k), \dots, \mathbf{1}_{x_{|\tilde{\mathcal{X}}|-1}}(X_n^k)\right]$
    Choose action A_n^k using the ϵ-greedy policy derived from $Q^k(X_n^k, \cdot)$
    Observe cost $f_{n+1} = f(X_n^k, A_n^k, \mu_n^k)$ and state X_{n+1}^k provided by the environment
    Update Q:
    $Q^k(X_n^k, A_n^k) = Q^k(X_n^k, A_n^k) + \rho^Q_{k,n,X_n^k,A_n^k}[f_{n+1} + \gamma \min_{a'\in\mathcal{A}} Q^k(X_{n+1}^k, a') - Q^k(X_n^k, A_n^k)]$
  end for
  if $\delta(\mu_T^{k-1}, \mu_T^k) \le tol_\mu$ and $\|Q^k - Q^{k-1}\|_{1,1} < tol_Q$ **then**
    break
  end if
end for
return (μ^k, Q^k)

The Unified Two Timescales Mean Field Q-learning (U2-MF-QL) algorithm represents a unified approach to solve mean field problems. On the one hand, by choosing the learning rate for the distribution of the population slower than the one for the Q-table, we obtain the solution to the MFG problem. Similarly to the scheme presented in Sect. 3, the iterations in Q perceive the quantity μ as quasi-static mimicking the freezing of the flow of measures characteristic in the solving scheme of a MFG. On the other hand, by choosing the learning rate for the mean-field term faster than the one for the Q-table, we obtain the solution to the MFC problem. Indeed, this choice of the parameters guarantees that the distribution changes instantaneously for each variation of the control function (Q-table) replicating the structure of the MFC problem.

4.2 Application to continuous problems

Although it is presented in a setting with finite state and action spaces, the application of the algorithm U2-MF-QL can be extended to continuous problems. Such adaptation

requires truncation and discretization procedures to time, state, and action spaces which should be calibrated based on the specific problem.

In practice, the learning episode will correspond to a uniform discretization $\tau = \{t_n\}_{n\in\{0,\dots,|\tau|-1\}}$ of a time interval $[0, T]$ with T large enough. The environment will provide the new state and reward at these discrete times. We assume that T is large enough to reach the ergodic regime. The continuous state space will be represented as the disjoint union of equally sized neighbors. Each of them will be identified by its centroid, and it will correspond to a row of the Q table. Likewise, actions will be provided to the environment in a finite set $\mathcal{A} = \{a_0, \dots, a_{|\mathcal{A}|-1}\} \subset \mathbb{R}^k$, and the distribution μ will be estimated on the set of centroids $\mathcal{X} = \{x_0, \dots, x_{|\mathcal{X}|-1}\} \subset \mathbb{R}^k$ identifying $\mu(x_i)$ as the probability of the neighbor centered in x_i. Then, Algorithm 1 is ran on those spaces.

We will use the benchmark linear-quadratic models given in continuous time and space for which we have explicit formulas given in Appendix A. In that case, we use an Euler discretization. We do not address here the error of approximation since the purpose of this comparison with a benchmark is mainly for illustration.

5 Numerical experiments

In this section, we illustrate our algorithm on a benchmark problem which admits an analytical solution.

5.1 Benchmark problem

We illustrate our algorithm on the following model, in which the mean-field interactions are through the first moment. We take $d = k = 1$,

$$f(x, \alpha, \mu) = \frac{1}{2}\alpha^2 + c_1\,(x - c_2 m)^2 + c_3\,(x - c_4)^2 + c_5 m^2, \quad b(x, \alpha, \mu) = \alpha, \qquad (15)$$

where $m = \int_{\mathbb{R}} x\mu(x)dx$. Here, the parameters $c_2, c_4 \in \mathbb{R}$ and $c_1, c_3, c_5 \in \mathbb{R}_+$ are constant such that $c_1 + c_3 - c_1c_2 \neq 0$. In this model, the drift is simply the control, while the running cost can be understood as follows: the first term is a quadratic cost for controlling the diffusion, which penalizes high velocity, the second term incorporates mean field interactions and encourages the agents to be close to $c_2 m$ (if $c_2 = 1$, this has a mean-reverting effect), the third term creates an incentive for each agent to be close to the target position c_4, and the fourth term penalizes the population when its mean m is far away from zero. We thus obtain a complex combination of various effects, which can be balanced depending on the choice of parameters.

We consider both the corresponding MFG and MFC problems in the asymptotic formulation. The details on the solutions of these problems and their connection to the non-asymptotic formulation are given in the appendix.

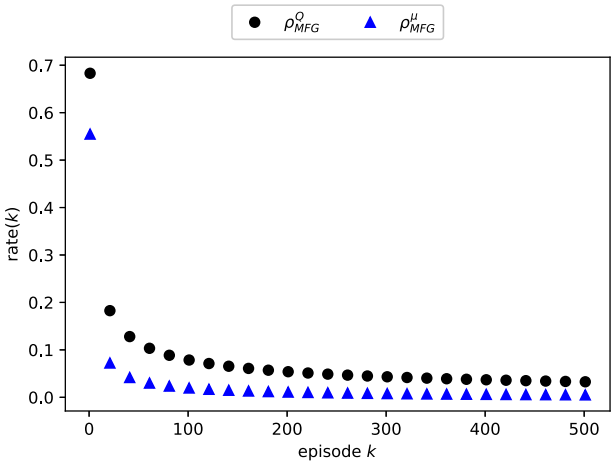

Fig. 2 MFG: learning rates over the first 500 episodes

5.2 Numerical results

We present the results obtained by applying the U2-MF-QL algorithm to the mean field problems based on the running cost and drift specified in (15). These results show how the algorithm successfully learns the MFG solution or the MFC solution based on simply tuning the learning rates. Moreover, this shows that the algorithm manages to solve problems defined on continuous time and continuous state, action spaces even though it is conceived for discrete problems. Such applications require to apply truncation and discretization procedures to time, state, and actions which should be calibrated on a problem base.

We consider the problem defined by the choice of parameters: $c_1 = 0.25$, $c_2 = 1.5$, $c_3 = 0.50$, $c_4 = 0.6$, $c_5 = 5$, discount parameter $\beta = 1$ and volatility $\sigma = 0.3$. The infinite time horizon is truncated at time $T = 20$. The continuous time is discretized using step $\Delta t = 10^{-2}$. Recall that γ in the discrete time setting corresponds to $e^{-\beta \Delta t}$ in the continuous time setting. The action space is given by $\mathcal{A} = \{a_0 = -1, \ldots, a_{N_{\mathcal{A}}} = 1\}$ and the state space by $\mathcal{X} = \{x_0 = -2 + x_c, \ldots, x_{N_{\mathcal{X}}} = 2 + x_c\}$, where x_c is the center of the state space. The step size for the discretization of the spaces $\mathcal{X}$ and $\mathcal{A}$ is given by $\Delta_{\cdot} = \sqrt{\Delta t} = 10^{-1}$. The state space $\mathcal{X}$ and the action space $\mathcal{A}$ have been chosen large enough to make sure that the state is within the boundary most of the time. In practice, this would have to be calibrated in a model-free way through experiments. In this example, for the numerical experiments, we used the knowledge of the model. In particular, we choose $x_c = 0.5$ for both examples. Note that if the problem under consideration is posed on finite spaces, this issue does not occur since the domain is fixed. The exploitation–exploration trade off is tackled on each episode using an ϵ-greedy policy, see (2). In particular, the value of ϵ is fixed to 0.15.
We present the following results for both the MFG and MFC benchmark examples:

1. learning rates analyses;
2. learning of the controls and the ergodic distribution;

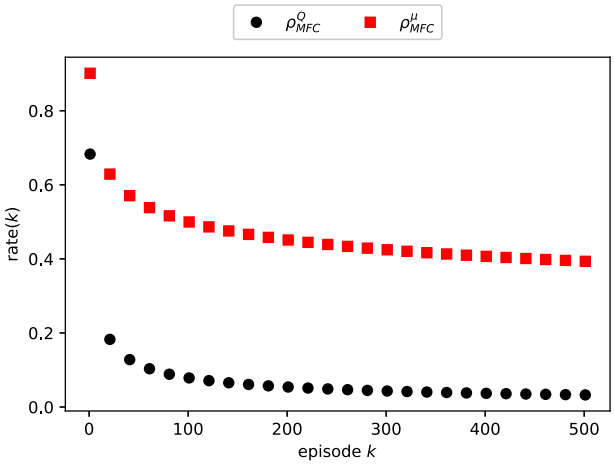

Fig. 3 MFC: learning rates over the first 500 episodes

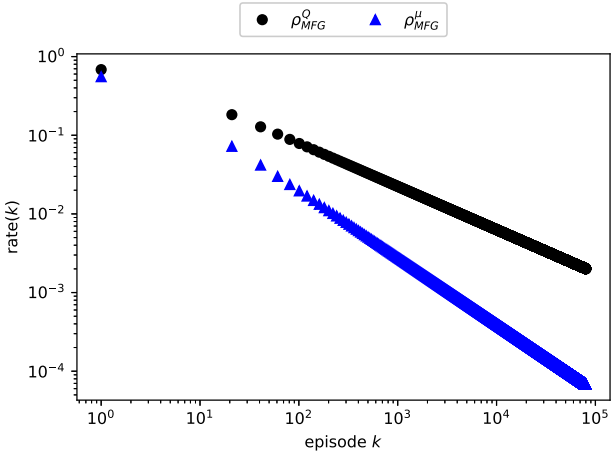

Fig. 4 MFG: learning rates over 80×10^3 episodes

3. empirical error analyses;
4. empirical analyses of the stopping criteria.

5.2.1 Learning rates analyses

It is important to observe that even if in the MFC case, the choice of ρ_k^μ below does not satisfy the classical Robbins–Monro summability condition recalled in Sect. 3.4, the numerical convergence of the algorithm is obtained, suggesting that these requirements may be relaxed in this framework. Failing in satisfying these conditions generates a noisy approximation of the distribution μ in the MFC problem. However, averaging over the last 10k episodes allows to minimize such noise as showed in the Figures below. Based on the theoretical results given in [14], we define the learning rates

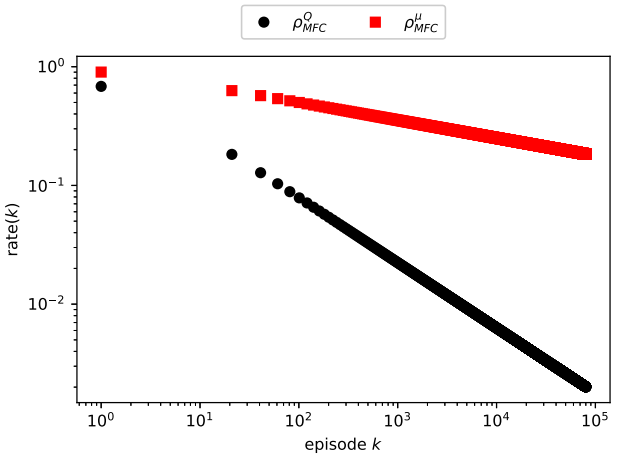

Fig. 5 MFC: learning rates over 80×10^3 episodes

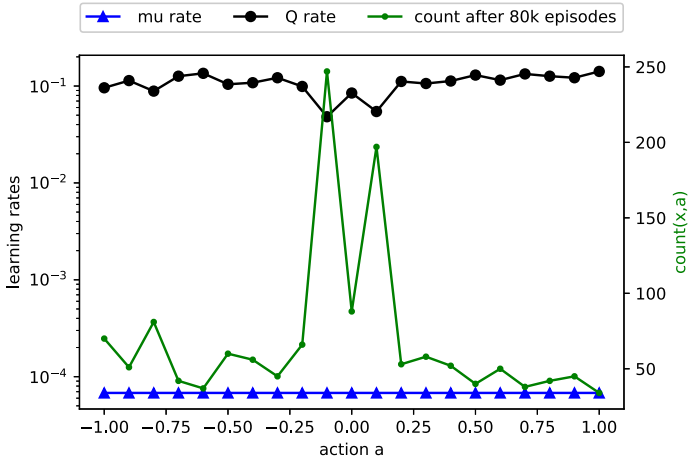

Fig. 6 MFG: comparison learning rates for state $x = -1.50$

appearing in Algorithm 1 as follows:

$$\rho^{Q}_{k,n,x,a} = \frac{1}{(1 + \#|(x,a,k,n)|)^{\omega^Q}}, \qquad \rho^{\mu}_{k} = \frac{1}{(1+k)^{\omega^\mu}}, \tag{16}$$

where $\#|(x,a,k,n)|$ is the number of times that the algorithm visited state x and performed action a until episode k and time t_n. The exponent ω^Q can take values in $(\frac{1}{2}, 1)$. The value of ω^μ is chosen depending on the value of ω^Q and the cooperative or non-cooperative nature of the problem we want to solve. The algorithm is run over 80×10^3 episodes over the interval $[0, T]$.

Figures 2, 3, 4, 5: **comparison of the learning rates.** The solution of the MFG benchmark is reached based on the choice $(\omega^Q, \omega^\mu) = (0.55, 0.85)$, such that $\rho^\mu <$

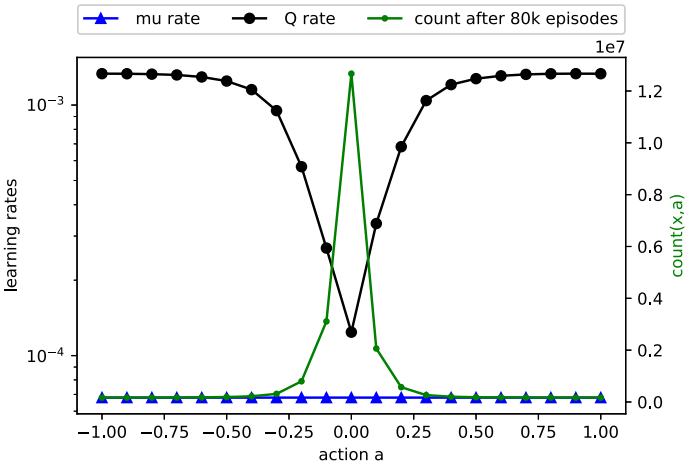

Fig. 7 MFG: comparison learning rates for state $x = 0.80$

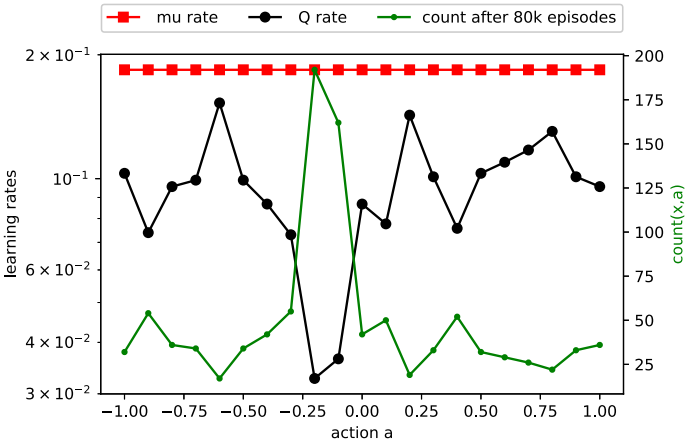

Fig. 8 MFC: comparison learning rates for state $x = -1.50$

ρ^Q. As pointed out in Sect. 3.4, by satisfying this relation the Q-function evolves faster than the estimation of the distribution mimicking the solving scheme of a MFG. On the other hand, the solution of the MFC benchmark can be obtained by opting for the pair of parameters $(\omega^Q, \omega^\mu) = (0.65, 0.15)$ such that $\rho^\mu > \rho^Q$. In Figs. 2, 3, 4, 5, we suppose that $\#|(x, a, k, 1)| = k$. The x-axis refers to the episode. The y-axis represents the rate evaluated at episode k.

Figures 6, 7, 8, 9: **Empirical check of the two timescale conditions.** The U2-MF-QL algorithm is based on an asynchronous QL approach which makes use of different learning rates for each $Q(x, a)$ based on the number of visits to the relative state-action pair. An empirical check of the two timescale conditions presented in Sect. 3.4 is presented in the following plots. The number of visits to each state depends on their proximity to the mean of the ergodic distribution. As a proof of concept, the

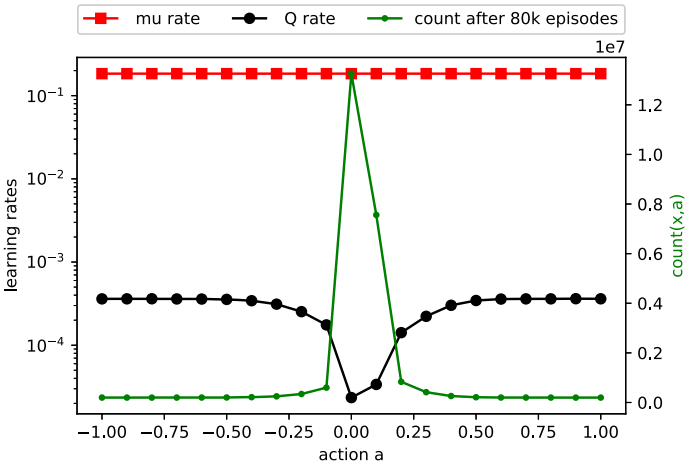

Fig. 9 MFC: comparison learning rates for state $x = 0.10$

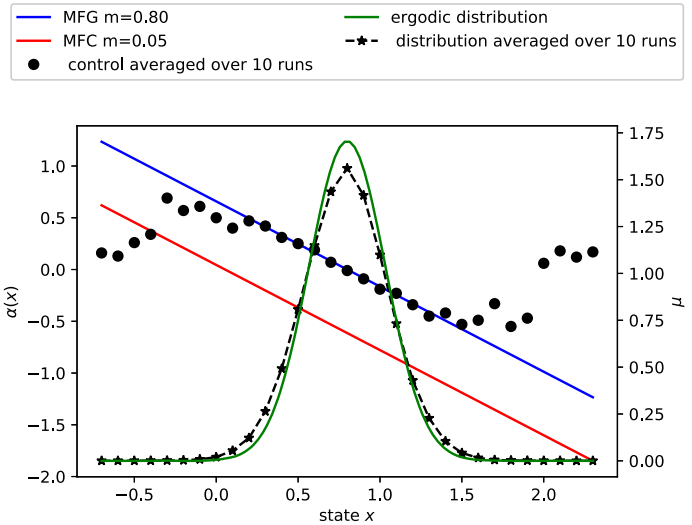

Fig. 10 MFG: results averaged over 10 runs

learning rates for two different states in the MFG and MFC frameworks are analyzed after 80×10^3 learning epochs. The plots on the left are relative to the state on the left bound of $\mathcal{X}$, while the plots on the right are relative to the closest state to the theoretical mean. Each plot shows the value of the learning rates ρ_k^{μ} and $\rho_{k,n,x,a}^{Q}$ together with the counter of visits to each pair (x, a). The two timescale conditions are satisfied in each plot. The number of visits changes from order 10^2 for the state on the border of $\mathcal{X}$ to order 10^7 for the closest state to the ergodic mean. The x-axis refers to the action. The left y-axis represents the learning rate. The right y-axis represents the counter of visits for each state-action pair.

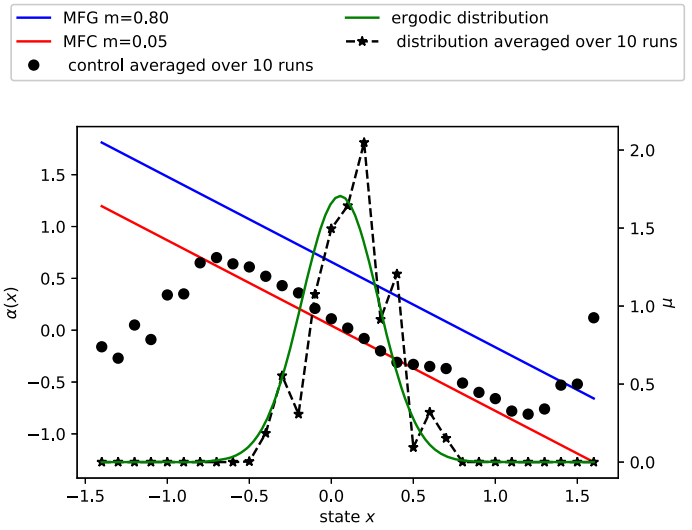

Fig. 11 MFC: results averaged over 10 runs

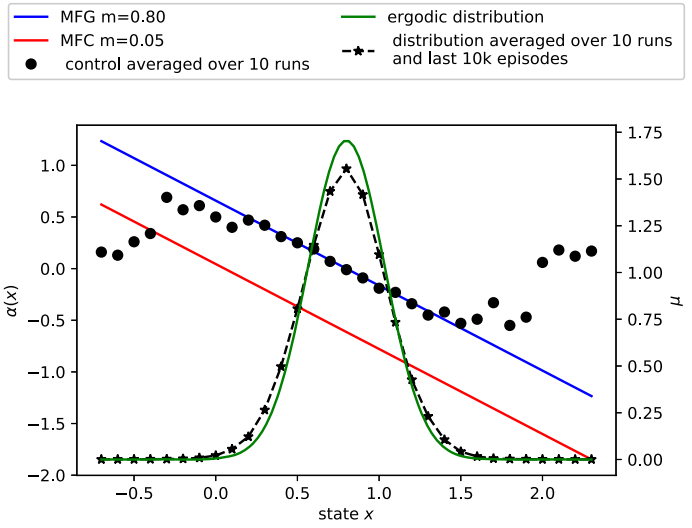

Fig. 12 MFG: results averaged over 10 runs and last $10k$ episodes

5.2.2 Learning of the controls and the ergodic distribution

Figures 10, 11, 12, 13, 14, 15: **controls, distributions and value functions learned by the algorithm.** The controls and the distribution learned by the algorithm are compared with the theoretical solution obtained in the appendix A. As presented in Sect. 3, the control $\alpha(x)$ is obtained as the $\arg\min_a Q(x,a)$. Similarly, the value function $V(x)$ can be recovered as $\min_a Q(x,a)$. The x-axis represents the state variable x. In Figs. 10, 11, 12, 13, 14, 15, the left y-axis relates to the action $\alpha(x)$. The right y-axis refers

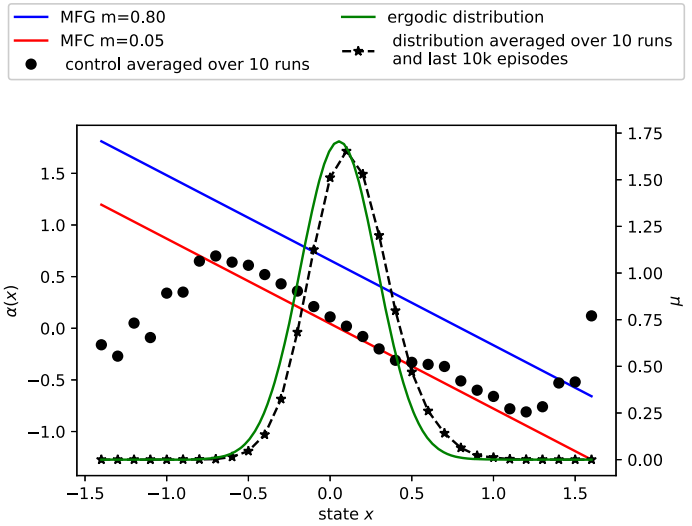

Fig. 13 MFC: results averaged over 10 runs and last $10k$ episodes

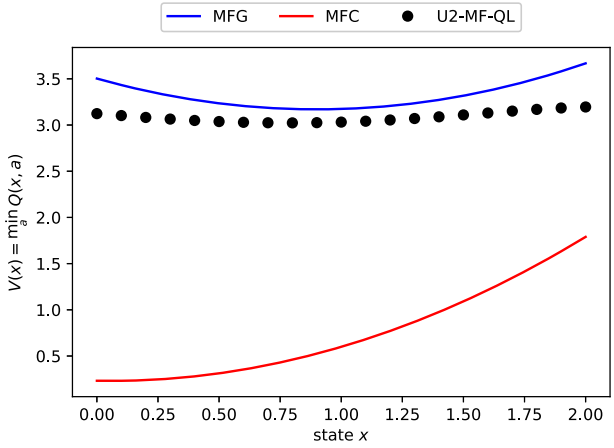

Fig. 14 MFG: value function

to the probability mass $\mu(x)$. The red (resp. blue) line shows the theoretical control function for the MFG (resp. MFC) problem. The black dots are the controls learned by the algorithm. Note that the peak of the distribution μ is not located at the same point x for MFG and MFC. Note that the peak of the distribution μ is not located at the same point x for MFG and MFC. In Figs. 10, 12, the y-axis corresponds to the value function $V(x)$. The continuous lines refer to the theoretical solution. The black dots are the numerical approximation recovered by the Q-function. We observe that the algorithm converges to different solutions based on the choice of the pair (ω^Q, ω^μ). On the left, the choice $(\omega^Q, \omega^\mu) = (0.55, 0.85)$ produces the approximation of the solution of the MFG. On the right, the set of parameters $(\omega^Q, \omega^\mu) = (0.65, 0.15)$

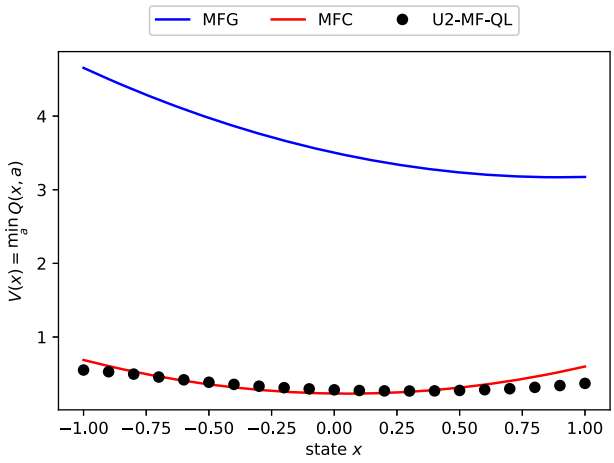

Fig. 15 MFC: value function

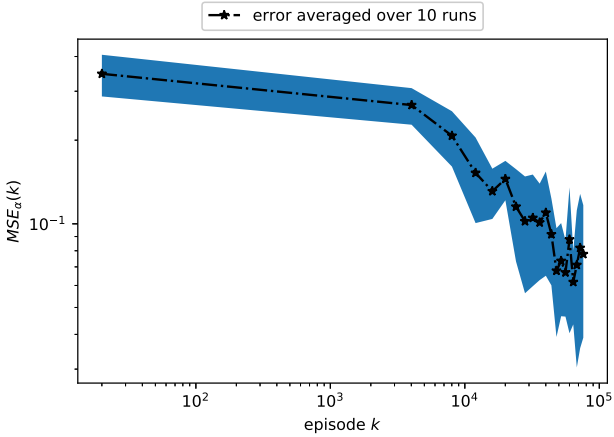

Fig. 16 MFG: squared root of $\mathrm{MSE}_\alpha(k)$

lets the algorithm learn the solution of the MFC problem. In Figs. 10, 11 the learned controls and the learned ergodic distribution are averaged over 10 runs. In Figs. 12, 13, the learned controls and the learned distribution μ_T are averaged over 10 runs and the last 10^4 episodes.

5.2.3 Empirical error analyses

Figures 16, 17: **MSE error on the control.** A metric used to evaluate the numerical results consists in the mean squared error (MSE) of the controls learned by episode k with respect to the theoretical solution presented in Appendix A. In particular, this metric considers the states $x \in \mathcal{X}$ where the ergodic distribution $\hat{\mu}$ is mostly concentrated. Let $\mathcal{C}_{MFG} \subset \mathcal{X}$ be centered in $\hat{m}$ s.t. $\hat{\mu}(\mathcal{C}_{MFG}) = 0.99$, then the mean

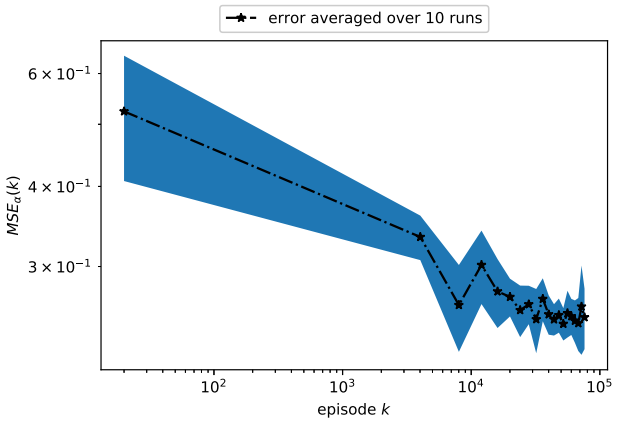

Fig. 17 MFC: squared root of $\mathrm{MSE}_\alpha(k)$

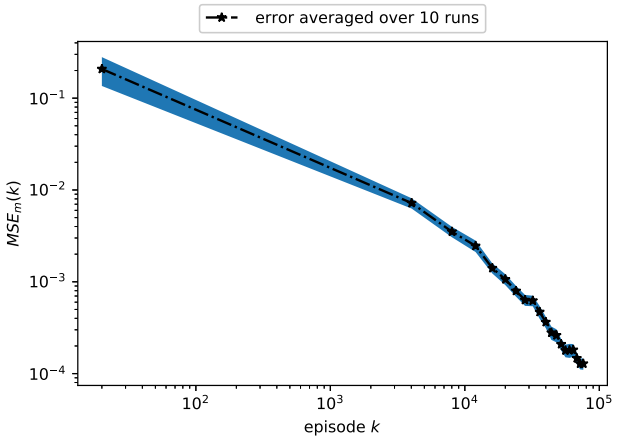

Fig. 18 MFG: mean squared error on $\hat{m}$

squared error by episode k for run i, and its average over all runs are defined as

$$\mathrm{MSE}_\alpha(i,k) = \frac{1}{|\mathcal{C}_{MFG}|} \sum_{j=0}^{|\mathcal{C}_{MFG}|-1} (\alpha^{i,k}(x_j) - \hat{\alpha}(x_j))^2,$$

$$\mathrm{MSE}_\alpha(k) = \frac{1}{\#runs} \sum_{i=0}^{\#runs} \mathrm{MSE}_\alpha(i,k).$$

The x-axis represents the number of episodes used for learning. The y-axis represents the mean squared error averaged over 10 runs (solid line) and its standard deviation (shaded region).

Figures 18, 19: **MSE on the ergodic mean.** A metric used to evaluate the numerical results consists in the squared error of the ergodic mean learned by episode k compared

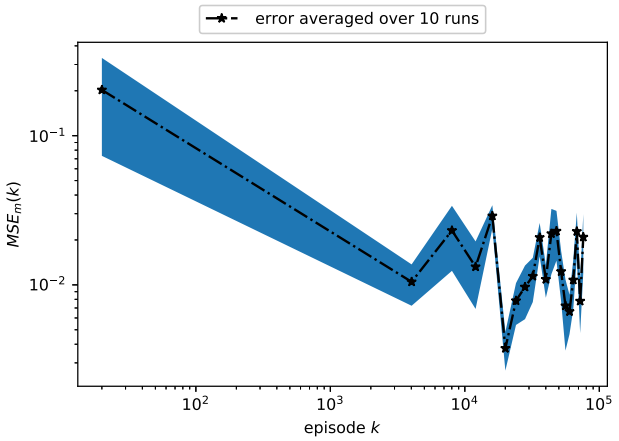

Fig. 19 MFC: mean squared error on $\hat{m}$

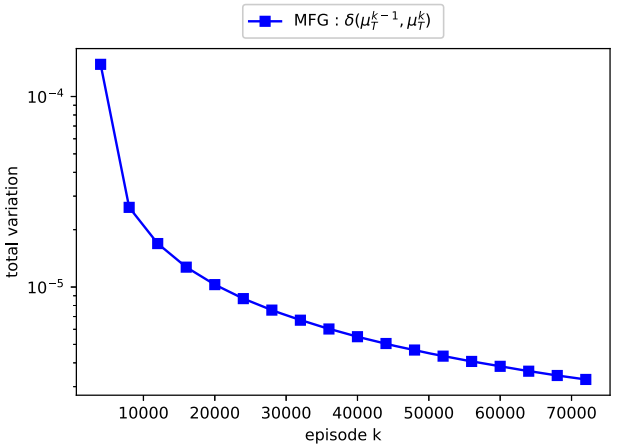

Fig. 20 MFG: total variation on μ

with its theoretical value obtained in Appendix A averaged over the total numbers of runs, i.e.,

$$\mathrm{MSE}_m(k) = \frac{1}{\#runs} \sum_{i=0}^{\#runs} (m_T^{i,k} - \hat{m})^2.$$

The x-axis represents the number of episodes used for learning. The y-axis represents the error averaged over 10 runs (solid line) and its standard deviation (shaded region). For the MFG, the error in the approximation of the ergodic mean reduces both in mean and standard deviation by increasing the number of episodes. For the MFC case, an oscillating behavior is observed. The choice of $\omega_\mu = 0.15$ in the learning rates defined in 16 allows to quicker adjustment of the mean by allocating more weights

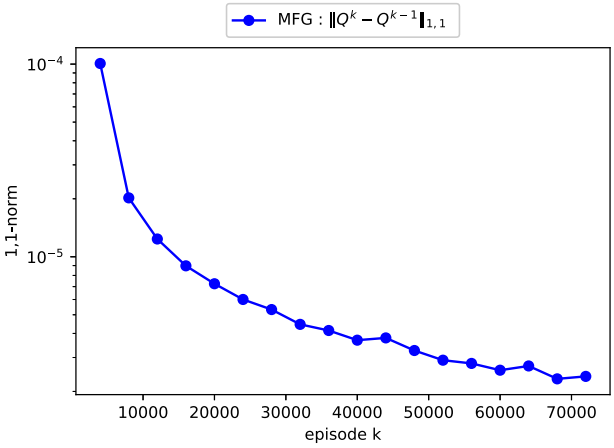

Fig. 21 MFG: total variation on Q

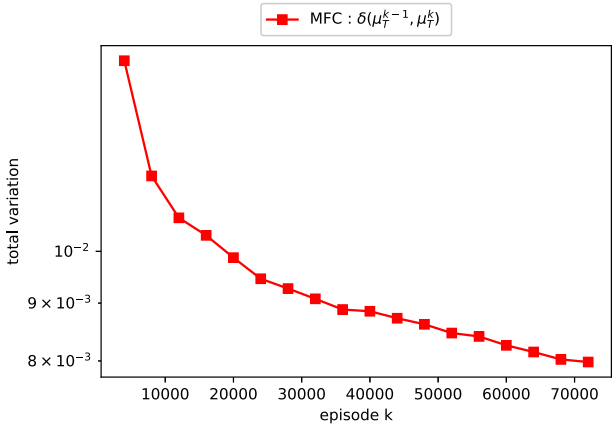

Fig. 22 MFC: total variation on μ

on the most recent sample. In this way, the algorithm mimics the nature of the MFC problem at the expense of a slower and more oscillating convergence.

5.2.4 Empirical analyses of the stopping criteria

Figures 20, 22, 21, 23: **stopping criteria.** The goal of the U2-MF-QL is to obtain a good approximation of the optimal controls and the ergodic distribution. As presented in algorithm 1, the stopping criteria are based on the analyses of the progresses in learning the optimal Q function and the ergodic distribution. The total variation and the 1, 1-norm between the start and the end of each episode is evaluated for the distribution

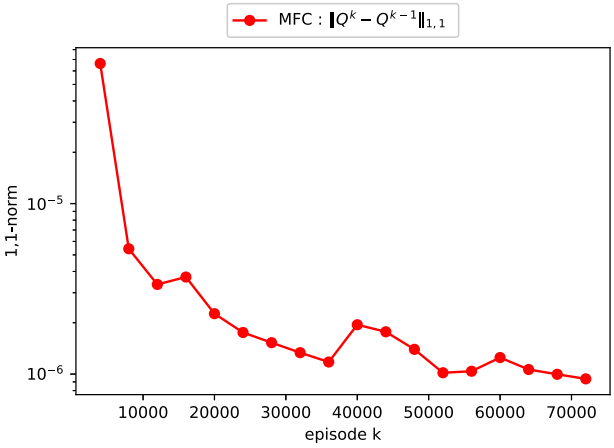

Fig. 23 MFC: total variation on Q

and the Q-table, respectively, as follows

$$\delta(\mu_T^{k-1}, \mu_T^k) = \sum_{x_i \in \mathcal{X}} \left| \mu_T^k(x_i) - \mu_T^{k-1}(x_i) \right|,$$
$$\|Q^k - Q^{k-1}\|_{1,1} = \sum_{i,j} \left| Q_{i,j}^k - Q_{i,j}^{k-1} \right|.$$

The algorithm stops when the increments are not significant anymore based on a threshold given as input. The value of the threshold depends on the user's needs, and it may be calibrated by a trial and error approach. The remaining plots show how these quantities decrease as the number of episodes increase. The x-axis represents the number of episodes used for learning. The y-axis represents the value of the total variation.

A Theoretical solutions for the benchmark examples

In this appendix, the solutions of the following benchmark problems are presented for the linear-quadratic models given by (15).

A.1 Non-asymptotic Mean Field Game,
A.2 Asymptotic Mean Field Game,
A.3 Stationary Mean Field Game,
A.4 Non-asymptotic Mean Field Control,
A.5 Asymptotic Mean Field Control.
A.6 Stationary Mean Field Control.

In particular, we check that the relations (3) and (4) are satisfied. The explicit formulas for the optimal controls (AMFG and AMFC) are used as benchmarks for our algorithm.

A.1 Solution for non-asymptotic MFG

We present the solution for the following MFG problem

1. Fix $\boldsymbol{m} = (m_t)_{t\geq 0} \subset \mathbb{R}$ and solve the stochastic control problem:

$$\begin{aligned}\min_{\boldsymbol{\alpha}} J^{\boldsymbol{m}}(\boldsymbol{\alpha}) &= \min_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^\infty e^{-\beta t} f(X_t^{\boldsymbol{\alpha}}, \alpha_t, m_t) dt\right] \\ &= \min_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^{+\infty} e^{-\beta t}\left(\frac{1}{2}\alpha_t^2 + c_1\left(X_t^{\boldsymbol{\alpha}} - c_2 m_t\right)^2 + c_3\left(X_t^{\boldsymbol{\alpha}} - c_4\right)^2 + c_5 m_t^2\right) \mathrm{d}t\right],\end{aligned}$$

subject to

$$\begin{aligned} dX_t^{\boldsymbol{\alpha}} &= \alpha_t \mathrm{d}t + \sigma dW_t, \\ X_0^{\boldsymbol{\alpha}} &\sim \mu_0. \end{aligned}$$

2. Find the fixed point, $\hat{\boldsymbol{m}} = (\hat{m}_t)_{t\geq 0}$, such that $\mathbb{E}\left[X_t^{\hat{\boldsymbol{\alpha}}}\right] = \hat{m}_t$ for all $t \geq 0$.

This problem can be solved by two equivalent approaches: PDE and FBSDEs. Both approaches start by solving the problem defined by a finite horizon T. Then, the solution to the infinite horizon problem is obtained by taking the limit T goes to infinity. Let $V^{\boldsymbol{m}^T,T}(t,x)$ be the optimal value function for the finite horizon problem conditioned on $X_0 = x$, i.e.,

$$\begin{aligned} V^{\boldsymbol{m}^T,T}(t,x) &= \inf_{\boldsymbol{\alpha}} J^{\boldsymbol{m},x}(\boldsymbol{\alpha}) \\ &= \inf_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_t^T e^{-\beta s} f(X_s^{\boldsymbol{\alpha}}, \alpha_s, m_s^T) ds \,\middle|\, X_0^{\boldsymbol{\alpha}} = x\right], \quad V^{\boldsymbol{m}^T,T}(T,x) = 0. \end{aligned}$$

where $\boldsymbol{m}^T = \{m_t^T\}_{0\leq t\leq T} \subset \mathbb{R}$. Let us consider the following ansatz with its derivatives

$$\begin{aligned} V^{\boldsymbol{m}^T,T}(t,x) &= \Gamma_2^T(t)x^2 + \Gamma_1^T(t)x + \Gamma_0^T(t), \\ \partial_t V^{\boldsymbol{m}^T,T}(t,x) &= \dot{\Gamma}_2^T(t)x^2 + \dot{\Gamma}_1^T(t)x + \dot{\Gamma}_0^T(t), \\ \partial_x V^{\boldsymbol{m}^T,T}(t,x) &= 2\Gamma_2^T(t)x + \Gamma_1^T(t), \\ \partial_{xx} V^{\boldsymbol{m}^T,T}(t,x) &= 2\Gamma_2^T(t), \end{aligned} \tag{17}$$

Then, the HJB equation for the value function reads:

$$\begin{aligned} &\partial_t V^{\boldsymbol{m}^T,T} - \beta V^{\boldsymbol{m}^T,T} + \inf_{\alpha}\{\mathcal{A}^X V^{\boldsymbol{m}^T,T} + f(x,\alpha,m^T)\} \\ &\quad = \partial_t V^{\boldsymbol{m}^T,T} - \beta V^{\boldsymbol{m}^T,T} + \inf_{\alpha}\left\{\alpha \partial_x V^{\boldsymbol{m}^T,T}\right. \\ &\qquad \left. + \frac{1}{2}\sigma^2 \partial_{xx} V^{\boldsymbol{m}^T,T} + \frac{1}{2}\alpha^2 + c_1(x - c_2 m^T)^2 + c_3(x - c_4)^2 + c_5(m^T)^2\right\} \\ &\quad = \partial_t V^{\boldsymbol{m}^T,T} - \beta V^{\boldsymbol{m}^T,T} + \left\{-\partial_x V^{\boldsymbol{m}^T,T^2} + \frac{1}{2}\sigma^2 \partial_{xx} V^{\boldsymbol{m}^T,T}\right. \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\partial_x V^{\boldsymbol{m}^T,T^2} + c_1(x - c_2 m^T)^2 + c_3(x - c_4)^2 + c_5(m^T)^2\Big\} \\
&= \partial_t V^{\boldsymbol{m}^T,T} - \beta V^{\boldsymbol{m}^T,T} - \frac{1}{2}\partial_x V^{\boldsymbol{m}^T,T^2} \\
&\quad + \frac{1}{2}\sigma^2 \partial_{xx} V^{\boldsymbol{m}^T,T} + c_1(x - c_2 m^T)^2 + c_3(x - c_4)^2 + c_5(m^T)^2 = 0,
\end{aligned}
$$

where in the third line we evaluated the infimum at $\hat{\alpha}^T = -V_x^{\boldsymbol{m}^T,T}$. The following ODEs system is obtained by replacing the ansatz and its derivatives in the HJB equation:

$$
\begin{cases}
\dot{\Gamma}_2^T - 2(\Gamma_2^T)^2 - \beta\Gamma_2^T + c_1 + c_3 = 0, & \Gamma_2^T(T) = 0, \\
\dot{\Gamma}_1^T = (2\Gamma_2^T + \beta)\Gamma_1^T + 2c_1c_2m^T + 2c_3c_4, & \Gamma_1^T(T) = 0, \\
\dot{\Gamma}_0^T = \beta\Gamma_0^T + \frac{1}{2}(\Gamma_1^T)^2 & \\
-\sigma^2\Gamma_2^T - c_3{c_4}^2 - (c_1{c_2}^2 + c_5)(m^T)^2, & \Gamma_0^T(T) = 0, \\
\dot{m}^T = -2\Gamma_2^T m^T - \Gamma_1^T, & m^T(0) = \mathbb{E}[\mu_0] = m_0,
\end{cases} \tag{18}
$$

where the last equation is obtained by considering the expectation of $X_t^{\boldsymbol{\alpha}}$ after replacing $\hat{\alpha}^T = -\partial_x V^{\boldsymbol{m}^T,T} = -(\Gamma_2^T x + \Gamma_1^T)$. The first equation is a Riccati equation. In particular, the solution Γ_2^T converges to $\hat{\Gamma}_2 = \frac{-\beta+\sqrt{\beta^2+8(c_1+c_3)}}{4}$ as T goes to infinity. The second and fourth ODEs are coupled, and they can be written in matrix notation as

$$
\dot{\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}} = \begin{bmatrix} -2\Gamma_2^T & -1 \\ 2c_1c_2 & 2\Gamma_2^T + \beta \end{bmatrix} \begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix} + \begin{pmatrix} 0 \\ 2c_3c_4 \end{pmatrix}, \quad \begin{pmatrix} m^T(0) \\ \Gamma_1^T(T) \end{pmatrix} = \begin{pmatrix} m_0 \\ 0 \end{pmatrix}. \tag{19}
$$

We start by solving the homogeneous equation, i.e.,

$$
\dot{\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}} = K_t^T \begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix} := \begin{bmatrix} -2\Gamma_2^T & -1 \\ 2c_1c_2 & 2\Gamma_2^T + \beta \end{bmatrix} \begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}, \quad \begin{pmatrix} m^T(0) \\ \Gamma_1^T(T) \end{pmatrix} = \begin{pmatrix} m_0 \\ 0 \end{pmatrix}. \tag{20}
$$

We introduce the propagator P^T, i.e.,

$$
\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix} = P_t^T \begin{pmatrix} m^T(0) \\ \Gamma_1^T(0) \end{pmatrix}. \tag{21}
$$

By deriving $\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}$ and expressing the initial conditions in terms of the inverse of P^T and $\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}$, we obtain

$$\dot{\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}} = \dot{P_t^T} \begin{pmatrix} m^T(0) \\ \Gamma_1^T(0) \end{pmatrix} = \dot{P_t^T} (P_t^T)^{-1} \begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}. \tag{22}$$

By comparing the last system with (20), we obtain

$$\begin{cases} \dot{P_t^T} &= K_t^T P_t^T \\ P_0^T &= \mathbb{I}_2 \end{cases} \tag{23}$$

where $\mathbb{I}_2$ is the identity matrix in dimension 2. The solution is given by $P_t^T = e^{\int_0^t K_s^T ds} := e^{L_t^T}$. In particular, the exponent is equal to

$$L_t^T = \int_0^t K_s^T ds = \begin{bmatrix} -2\int_0^t \Gamma_2^T(s)ds & -t \\ 2c_1c_2t & 2\int_0^t \Gamma_2^T(s)ds + \beta t \end{bmatrix} = \begin{bmatrix} g_t^T & d_t \\ b_t & a_t^T \end{bmatrix}. \tag{24}$$

We evaluate the exponential $P^T(t) = e^{L_t^T}$ by using the Taylor's expansion and diagonalizing the matrix L_t^T. The eigenvalues/eigenvectors of L_t^T are given by

$$\begin{aligned} \lambda_{1\backslash 2,t}^T &:= \frac{a_t^T + g_t^T \pm \sqrt{(a_t^T - g_t^T)^2 + 4b_t d_t}}{2}, \\ v_{1,t}^T &:= \begin{pmatrix} d_t \\ \lambda_{1,t}^T - g_t^T \end{pmatrix}, \quad v_{2,t}^T := \begin{pmatrix} d_t \\ \lambda_{2,t}^T - g_t^T \end{pmatrix}. \end{aligned} \tag{25}$$

P_t is obtained by

$$\begin{aligned} P_t^T &= \begin{pmatrix} p_t^T(1,1) & p_t^T(1,2) \\ p_t^T(2,1) & p_t^T(2,2) \end{pmatrix} \\ &= e^{L_t^T} = \sum_{k=0}^{\infty} \begin{bmatrix} v_{1,t}^T & v_{2,t}^T \end{bmatrix} \frac{\begin{pmatrix} \lambda_{1,t}^T & 0 \\ 0 & \lambda_{2,t}^T \end{pmatrix}^k}{k!} \begin{bmatrix} v_{1,t}^T & v_{2,t}^T \end{bmatrix}^{-1} \\ &:= S_t^T \sum_{k=0}^{\infty} \frac{{D_t^T}^k}{k!} (S_t^T)^{-1} \\ &= S_t^T \begin{pmatrix} e^{\lambda_{1,t}^T} & 0 \\ 0 & e^{\lambda_{2,t}^T} \end{pmatrix} (S_t^T)^{-1} \\ &= \frac{1}{d_t(\lambda_{2,t}^T - \lambda_{1,t}^T)} \\ &\quad \begin{pmatrix} d_t e^{\lambda_{1,t}^T}(\lambda_{2,t}^T - g_t^T) + d_t e^{\lambda_{2,t}^T}(g_t^T - \lambda_{1,t}^T) & d_t^2(e^{\lambda_{2,t}^T} - e^{\lambda_{1,t}^T}) \\ (\lambda_{1,t}^T - g_t^T)(\lambda_{2,t}^T - g_t^T)(e^{\lambda_{1,t}^T} - e^{\lambda_{2,t}^T}) & d_t e^{\lambda_{2,t}^T}(\lambda_{2,t}^T - g_t^T) + d_t e^{\lambda_{1,t}^T}(g_t^T - \lambda_{1,t}^T) \end{pmatrix}. \end{aligned} \tag{26}$$

In order to solve the non-homogeneous case, we introduce an extra term $\begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix}$, i.e.,

$$\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix} = P_t^T \begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix}. \tag{27}$$

By deriving $\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}$, we obtain

$$\begin{aligned} \dot{\widehat{\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}}} &= \dot{P}_t^T \begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix} + P_t^T \dot{\widehat{\begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix}}} = K_t^T P_t^T \begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix} + P_t^T \dot{\widehat{\begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix}}} \\ &= K_t^T \begin{pmatrix} m_t^T \\ \Gamma_1^T \end{pmatrix} + P_t^T \dot{\widehat{\begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix}}}. \end{aligned} \tag{28}$$

By comparing (19) with (28), we obtain

$$\dot{\widehat{\begin{pmatrix} h_1^T \\ h_2^T \end{pmatrix}}} = (P_t^T)^{-1} \begin{pmatrix} 0 \\ 2c_4c_4 \end{pmatrix} = \frac{1}{|P_t^T|} \begin{pmatrix} p_t^T(2,2) & -p_t^T(1,2) \\ -p_t^T(2,1) & p_t^T(1,1) \end{pmatrix} \begin{pmatrix} 0 \\ 2c_3c_4 \end{pmatrix}. \tag{29}$$

By integration, we obtain

$$\begin{aligned} h_1^T(t) &= h_1^T(0) - 2c_3c_4 \int_0^t \frac{p_s^T(1,2)}{|P_s^T|} ds, \\ h_2^T(t) &= h_2^T(0) + 2c_3c_4 \int_0^t \frac{p_s^T(1,1)}{|P_s^T|} ds, \end{aligned} \tag{30}$$

where $h_1^T(0) = m_0$ and $h_2^T(0) = \Gamma_1^T(0)$.

We use the terminal condition $\Gamma_1^T(T) = 0$ to obtain an evaluation of $h_2^T(0) = \Gamma_1^T(0)$ in terms of P_T^T and m_0, i.e.,

$$\begin{aligned} \Gamma_1^T(T) &= p_T^T(2,1) h_1^T(T) + p_T^T(2,2) h_2^T(T) = 0, \\ \Gamma_1^T(T) &= p_T^T(2,1) \\ &\quad \left(m_0 - 2c_3c_4 \int_0^T \frac{p_s^T(1,2)}{|P_s^T|} ds \right) \\ &\quad + p_T^T(2,2) \left(\Gamma_1^T(0) + 2c_3c_4 \int_0^T \frac{p_s^T(1,1)}{|P_s^T|} ds \right) = 0, \\ \Gamma_1^T(0) &= -\frac{p_T^T(2,1)}{p_T^T(2,2)} \left(m_0 - 2c_3c_4 \int_0^T \frac{p_s^T(1,2)}{|P_s^T|} ds \right) - 2c_3c_4 \int_0^T \frac{p_s^T(1,1)}{|P_s^T|} ds. \end{aligned} \tag{31}$$

In order to evaluate the limit of $\Gamma_1^T(0)$ as T goes to infinity, we analyze the different terms separately. First, we evaluate the following limit:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \Gamma_2^T(s)ds = \lim_{T\to\infty}\Gamma_2^T(s_1) = \hat{\Gamma}_2, \quad s_1\in[0,T], \tag{32}$$

where we applied the mean value integral theorem and $\hat{\Gamma}_2 = \frac{-\beta+\sqrt{\beta^2+8(c_1+c_3)}}{4}$ is the limit of the solution of the Riccati equation obtained previously, i.e., $\hat{\Gamma}_2 = \lim_{T\to\infty}\Gamma_2^T(s)$. We recall that

$$\lambda_{2,T}^T-\lambda_{1,T}^T = \sqrt{(a_T^T-g_T^T)^2+4b_T^Td_T} = T\sqrt{\left(\frac{4}{T}\int_0^T\Gamma_2^T(s)ds+\beta\right)^2-8c_1c_2}>0$$

which goes to infinity as T goes to ∞ when the term under square root is well defined. We observe that

$$\begin{aligned}
\hat{g}_t &:= \lim_{T\to\infty} g_t^T = \lim_{T\to\infty} -2\int_0^t\Gamma_2^T(s)ds = -2\hat{\Gamma}_2 t := gt,\\
b_t &= 2c_1c_2t,\\
\hat{a}_t &:= \lim_{T\to\infty} a_t^T = \lim_{T\to\infty} 2\int_0^t\Gamma_2^T(s)ds+\beta t = 2\hat{\Gamma}_2t+\beta t,\\
d_t &= -t,\\
\hat{\lambda}_{1\backslash 2,t} &:= \lim_{T\to\infty}\lambda_{1\backslash 2,t}^T = \frac{\hat{a}_t+\hat{g}_t\pm\sqrt{(\hat{a}_t-\hat{g}_t)^2+4b_td_t}}{2}\\
&= t\frac{\beta\pm\sqrt{(4\hat{\Gamma}_2+\beta)^2-8c_1c_2}}{2} := t\lambda_{1\backslash 2},\\
\hat{P}_t &:= \lim_{T\to\infty} P_t^T\\
&= \frac{1}{d_t(\hat{\lambda}_{2,t}-\hat{\lambda}_{1,t})}\\
&\quad\begin{pmatrix} d_te^{\hat{\lambda}_{1,t}}(\hat{\lambda}_{2,t}-\hat{g}_t)+d_te^{\hat{\lambda}_{2,t}}(\hat{g}_t-\hat{\lambda}_{1,t}) & d_t^2(e^{\hat{\lambda}_{2,t}}-e^{\hat{\lambda}_{1,t}})\\ (\hat{\lambda}_{1,t}-\hat{g}_t)(\hat{\lambda}_{2,t}-\hat{g}_t)(e^{\hat{\lambda}_{1,t}}-e^{\hat{\lambda}_{2,t}}) & d_te^{\hat{\lambda}_{2,t}}(\hat{\lambda}_{2,t}-\hat{g}_t)+d_te^{\hat{\lambda}_{1,t}}(\hat{g}_t-\hat{\lambda}_{1,t})\end{pmatrix}.
\end{aligned} \tag{33}$$

To evaluate $\hat{\Gamma}_1(0) = \lim_{T\to\infty}\Gamma_1^T(0)$, we study the limit of the remaining terms:

$$\begin{aligned}
\lim_{T\mapsto\infty} -\frac{p_T^T(2,1)}{p_T^T(2,2)} &= \lim_{T\mapsto\infty}\frac{(\lambda_{1,T}^T-g_T^T)(\lambda_{2,T}^T-g_T^T)(e^{\lambda_{2,T}^T}-e^{\lambda_{1,T}^T})}{d_Te^{\lambda_{2,T}^T}(\lambda_{2,T}^T-g_T^T)+d_Te^{\lambda_{1,T}^T}(g_T^T-\lambda_{1,T}^T)}\\
&= \lim_{T\mapsto\infty}\frac{1}{\frac{d_T}{(\lambda_{1,T}^T-g_T^T)(1-e^{\lambda_{1,T}^T-\lambda_{2,T}^T})}+\frac{d_T}{(\lambda_{2,T}^T-g_T^T)(1-e^{\lambda_{2,T}^T-\lambda_{1,T}^T})}}\\
&= -(\lambda_1-g)\\
&= -(\lambda_1+2\hat{\Gamma}_2),\\
\lim_{T\mapsto\infty}\int_0^T\frac{p_s^T(1,2)}{|P_s^T|}ds &= \lim_{T\mapsto\infty}\int_0^T\frac{d_s(e^{\lambda_{2,s}^T}-e^{\lambda_{1,s}^T})}{(\lambda_{2,s}^T-\lambda_{1,s}^T)(e^{\lambda_{1,s}^T+\lambda_{2,s}^T})}ds
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda_2-\lambda_1}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\\
\lim_{T\mapsto\infty}\int_0^T \frac{p_s^T(1,1)}{|P_s^T|}ds &= \lim_{T\mapsto\infty}\int_0^T \frac{1}{e^{\lambda_{1,s}^T+\lambda_{2,s}^T}}\left(e^{\lambda_{1,s}^T}\frac{\lambda_{2,s}^T-g_s^T}{\lambda_{2,s}^T-\lambda_{1,s}^T}+e^{\lambda_{2,s}^T}\frac{g_s^T-\lambda_{1,s}^T}{\lambda_{2,s}^T-\lambda_{1,s}^T}\right)ds\\
&= \frac{\lambda_2-g}{\lambda_2(\lambda_2-\lambda_1)}+\frac{g-\lambda_1}{\lambda_1(\lambda_2-\lambda_1)}.
\end{aligned}
\tag{34}
$$

Finally, the value of $\hat{\Gamma}_1(0)$ is given by

$$
\hat{\Gamma}_1(0) = -(\lambda_1-g)m_0 - 2\frac{c_3c_4}{\lambda_2}. \tag{35}
$$

Given $\hat{\Gamma}_1(0)$, we evaluate the limit as T goes to ∞ of (30), i.e.,

$$
\begin{aligned}
h_1(t) := \lim_{T\mapsto\infty} h_1^T(t) &= m_0 - 2c_3c_4\lim_{T\mapsto\infty}\int_0^t \frac{p_s^T(1,2)}{|P_s^T|}ds\\
&= m_0 + 2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{1}{\lambda_2}e^{-t\lambda_2}-\frac{1}{\lambda_1}e^{-t\lambda_1}+\frac{1}{\lambda_1}-\frac{1}{\lambda_2}\right),\\
h_2(t) := \lim_{T\mapsto\infty} h_2^T(t) &= \lim_{T\mapsto\infty}\left(\Gamma_1^T(0)+2c_3c_4\int_0^t \frac{p_s^T(1,1)}{|P_s^T|}\mathrm{d}s\right)\\
&= \hat{\Gamma}_1(0) + 2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{\lambda_2-g}{\lambda_2}(1-e^{-t\lambda_2})+\frac{g-\lambda_1}{\lambda_1}(1-e^{-t\lambda_1})\right).
\end{aligned}
\tag{36}
$$

We can conclude that

$$
\begin{aligned}
\hat{m}_t &= \lim_{T\to\infty} m_t^T\\
&= \hat{p}_t(1,1)h_1(t)+\hat{p}_t(1,2)h_2(t)\\
&= \left(m_0+2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_2}\right)\right)e^{t\lambda_1}+2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right),\\
\hat{\Gamma}_1(t) &= \lim_{T\to\infty}\Gamma_1^T(t)\\
&= \hat{p}_t(2,1)h_1(t)+\hat{p}_t(2,2)h_2(t)\\
&= m_0(g-\lambda_1)e^{t\lambda_1}+2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{\lambda_2-g}{\lambda_2}-\frac{\lambda_1-g}{\lambda_1}\right).
\end{aligned}
\tag{37}
$$

Finally, the third ODE in (18) can be solved by plugging in the solution of the previous ones and integrating. Since our interest is into the evolution of the mean and the control function, we omit these calculations, but we recall that:

$$
\hat{\alpha}_t = -(\hat{\Gamma}_2 x+\hat{\Gamma}_1(t)),\quad \hat{\Gamma}_2 = \frac{-\beta+\sqrt{\beta^2+8(c_1+c_3)}}{4}, \tag{38}
$$

and we observe that

$$
\lim_{t\to\infty}\hat{\alpha}_t = -(\hat{\Gamma}_2 x+\hat{\Gamma}_1),\quad \hat{\Gamma}_1 = -\frac{4c_1c_2\hat{\Gamma}_2}{\lambda_2} = \frac{c_3c_4\hat{\Gamma}_2}{2(c_1+c_3-c_1c_2)}. \tag{39}
$$

A.2 Solution for asymptotic MFG

The asymptotic version of the problem presented above is given by:

1. Fix $m \in \mathbb{R}$ and solve the stochastic control problem:

$$\begin{aligned}\min_{\boldsymbol{\alpha}} J^m(\boldsymbol{\alpha}) &= \min_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^\infty e^{-\beta t} f(X_t^{\boldsymbol{\alpha}}, \alpha_t, m)\mathrm{d}t\right] \\ &= \min_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^\infty e^{-\beta t}\left(\frac{1}{2}\alpha_t^2 + c_1\left(X_t^{\boldsymbol{\alpha}} - c_2 m\right)^2 + c_3\left(X_t^{\boldsymbol{\alpha}} - c_4\right)^2 + c_5 m^2\right)\mathrm{d}t\right], \\ \text{subject to:}\quad & dX_t^{\boldsymbol{\alpha}} = \alpha_t \mathrm{d}t + \sigma dW_t, \quad X_0^{\boldsymbol{\alpha}} \sim \mu_0.\end{aligned}$$

2. Find the fixed point, $\hat{m}$, such that $\hat{m} = \lim_{t\to+\infty} \mathbb{E}\left[X_t^{\hat{\alpha},\hat{m}}\right]$.

Let $V^m(x)$ be the optimal value function given $m \in \mathbb{R}$ and conditioned on $X_0 = x$, i.e.,

$$\begin{aligned}V^m(x) &= \inf_{\boldsymbol{\alpha}} J^{m,x}(\boldsymbol{\alpha}) \\ &= \inf_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^{+\infty} e^{-\beta t}\left(\frac{1}{2}\alpha_t^2 + c_1\left(X_t^{\boldsymbol{\alpha}} - c_2 m\right)^2 + c_3\left(X_t^{\boldsymbol{\alpha}} - c_4\right)^2 + c_5 m^2\right)\Big| X_0^{\boldsymbol{\alpha}} = x\right].\end{aligned}$$

We consider the following ansatz with its derivatives with respect to x:

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

Let us consider the HJB equation

$$\begin{aligned}&\beta V^m(x) - \inf_{\alpha}\{\mathcal{A}^X V^m(x) + f(x,\alpha,m)\} \\ &\quad= \beta V^m(x) - \inf_{\alpha}\left\{\alpha\dot{V}(x) + \frac{1}{2}\sigma^2\ddot{V}^m(x) + \frac{1}{2}\alpha^2 + c_1(x - c_2 m)^2\right. \\ &\qquad\left. + c_3(x - c_4)^2 + c_5 m^2\right\} \\ &\quad= \beta V^m(x) - \left\{-(\dot{V}^m)^2(x) + \frac{1}{2}\sigma^2\ddot{V}^m(x) + \frac{1}{2}(\dot{V}^m)^2(x)\right. \\ &\left. + c_1(x - c_2 m)^2 + c_3(x - c_4)^2 + c_5 m^2\right\} \\ &\quad= \beta V^m(x) + \frac{1}{2}(\dot{V}^m)^2(x) \\ &\quad- \frac{1}{2}\sigma^2\ddot{V}^m(x) - c_1(x - c_2 m)^2 - c_3(x - c_4)^2 - c_5 m^2 = 0,\end{aligned}$$

where in the third line we evaluated the infimum at $\hat{\alpha}(x) = -\dot{V}^m(x)$. Replacing the ansatz and its derivatives in the HJB equation, it follows that

$$\left(\beta\Gamma_2 + 2\Gamma_2^2 - c_1 - c_3\right)x^2 + (\beta\Gamma_1 + 2\Gamma_2\Gamma_1 + 2c_1c_2m + 2c_3c_4)x + \beta\Gamma_0 + \frac{1}{2}\Gamma_1^2 - \sigma^2\Gamma_2 - (c_1{c_2}^2 + c_5)m^2 - c_3{c_4}^2 = 0.$$

An easy computation gives the values

$$\begin{aligned}\Gamma_2 &= \frac{-\beta + \sqrt{\beta^2 + 8(c_1 + c_3)}}{4},\\ \Gamma_1 &= -\frac{2c_1c_2m + 2c_3c_4}{\beta + 2\Gamma_2},\\ \Gamma_0 &= \frac{c_5m^2 + c_3{c_4}^2 + c_1{c_2}^2m^2 + \sigma^2\Gamma_2 - \frac{1}{2}\Gamma_1^2}{\beta}.\end{aligned}$$

By plugging the control $\hat{\alpha}(x) = -(2\Gamma_2 x + \Gamma_1)$ into the dynamics of X_t and taking the expected value, we obtain an ODE for m_t

$$\dot{m}_t = -(2\Gamma_2 m_t + \Gamma_1). \tag{40}$$

The solution of (40) is used to derive m as follows

$$\begin{aligned}m &= \lim_{t\mapsto\infty} m_t = \lim_{t\mapsto\infty} -\frac{\Gamma_1}{2\Gamma_2} + \left(m_0 + \frac{\Gamma_1}{\Gamma_2}\right)e^{-2\Gamma_2 t} = -\frac{\Gamma_1}{2\Gamma_2} = \frac{2c_1c_2m + 2c_3c_4}{2\Gamma_2(\beta + 2\Gamma_2)},\\ m &= \frac{c_3c_4}{\Gamma_2(\beta + 2\Gamma_2) - c_1c_2}\end{aligned} \tag{41}$$

To summarize, we derived that $\hat{\alpha}(x) = -(2\Gamma_2 x + \Gamma_1)$ with $\Gamma_2 = \hat{\Gamma}_2$ and $\Gamma_1 = \hat{\Gamma}_1$ obtained in (39). In other words, we have checked that

$$\lim_{t\to\infty} \hat{\alpha}_t^{MFG}(x) = \hat{\alpha}^{AMFG}(x), \quad \forall x,$$

that is the first part of (3) for this LQ MFG.

A.3 Solution for stationary MFG

The only difference with the derivation above in the case of asymptotic MFG is that m_t should be a constant which, from (40), should satisfy $2\Gamma_2 m + \Gamma_1 = 0$. Therefore, m takes the same value as in (41), and we deduce

$$\hat{\alpha}^{SMFG}(x) = \hat{\alpha}^{AMFG}(x), \quad \forall x,$$

that is the second part of (3) for this LQ MFG.

A.4 Solution for non-asymptotic MFC

We present the solution for the following non-asymptotic MFC problem

$$\begin{aligned}\min_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) &= \min_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^{\infty} e^{-\beta t} f(X_t^{\boldsymbol{\alpha}}, \alpha_t, \mathbb{E}\left[X_t^{\boldsymbol{\alpha}}\right]) \mathrm{d}t\right] \\ &= \min_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^{+\infty} e^{-\beta t}\left(\frac{1}{2}\alpha_t^2 + c_1\left(X_t^{\boldsymbol{\alpha}} - c_2\mathbb{E}\left[X_t^{\boldsymbol{\alpha}}\right]\right)^2 + c_3\left(X_t^{\boldsymbol{\alpha}} - c_4\right)^2 + c_5\mathbb{E}\left[X_t^{\boldsymbol{\alpha}}\right]^2\right)\mathrm{d}t\right], \\ \text{subject to:}\quad & dX_t^{\boldsymbol{\alpha}} = \alpha_t \mathrm{d}t + \sigma dW_t, \quad X_0^{\boldsymbol{\alpha}} \sim \mu_0.\end{aligned}$$

Note that here the mean $\mathbb{E}\left[X_t^{\boldsymbol{\alpha}}\right]$ of the population changes instantaneously when $\boldsymbol{\alpha}$ changes.

This problem can be solved by two equivalent approaches: PDE and FBSDEs. Both approaches start by solving the problem defined by a finite horizon T. Then, the solution to the infinite horizon problem is obtained by taking the limit for T goes to infinity. Let $V^T(t,x)$ be the optimal value function for the finite horizon problem conditioned on $X_0 = x$, i.e.,

$$\begin{aligned}V^T(t,x) &= \inf_{\boldsymbol{\alpha}} J^{\boldsymbol{m}^{\boldsymbol{\alpha}},x}(\boldsymbol{\alpha}) \\ &= \inf_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_t^T e^{-\beta s} f(X_s^{\boldsymbol{\alpha}}, \alpha_s, m_s^{\boldsymbol{\alpha}}) ds \,\middle|\, X_0^{\boldsymbol{\alpha}} = x\right], \quad V^T(T,x) = 0.\end{aligned}$$

Let us consider the following ansatz with its derivatives

$$\begin{aligned}V^T(t,x) &= \Gamma_2^T(t)x^2 + \Gamma_1^T(t)x + \Gamma_0^T(t), \quad V^T(T,x) = 0, \\ \partial_t V^T(t,x) &= \dot{\Gamma}_2^T(t)x^2 + \dot{\Gamma}_1^T(t)x + \dot{\Gamma}_0^T(t), \\ \partial_x V^T(t,x) &= 2\Gamma_2^T(t)x + \Gamma_1^T(t), \\ \partial_{xx} V^T(t,x) &= 2\Gamma_2^T(t),\end{aligned} \tag{42}$$

Starting by the MFC-HJB equation (4.12) given in [4], we extended it to the asymptotic case as follows

$$\beta V^T - V_t^T - H\left(t, x, \boldsymbol{\mu}, \alpha\right) - \int_{\mathbb{R}} \frac{\delta H}{\delta \mu}\left(t, h, \boldsymbol{\mu}, -\partial_x V^T\right)(x)\mu_t(h)dh = 0,$$

where $m_t = \int_{\mathbb{R}} y\mu_t(dy)$ and $\alpha^* = -\partial_x V^T$. We have:

$$\begin{aligned}&H\left(t, x, \boldsymbol{\mu}, \alpha\right) \\ &:= \inf_{\alpha}\left\{\mathcal{A}^X V^T + f\left(t, x, \alpha, \boldsymbol{\mu}\right)\right\} \\ &= \inf_{\alpha}\left\{\alpha\partial_x V^T + \frac{1}{2}\sigma^2\partial_{xx}V^T + \frac{1}{2}\alpha^2 + c_1(x - c_2 m_t)^2 + c_3(x - c_4)^2 + c_5 {m_t}^2\right\} \\ &= -\frac{1}{2}(\partial_x V^T)^2 + \frac{1}{2}\sigma^2\partial_{xx}V^T + c_1(x - c_2 m_t)^2 + c_3(x - c_4)^2 + c_5 {m_t}^2,\end{aligned}$$

$$
\begin{aligned}
&\frac{\delta H\left(t, h, \boldsymbol{\mu}, \alpha\right)}{\delta \mu}(x) \\
&= \frac{\delta}{\delta \mu}\left(c_1(h - c_2 m_t)^2 + c_5 {m_t}^2\right)(x) \\
&= \frac{\delta}{\delta \mu}\left(c_1\left(h - c_2 \int_{\mathbb{R}} y \mu_t(dy)\right)^2 + c_5\left(\int_{\mathbb{R}} y \mu_t(dy)\right)^2\right)(x) \\
&= -2c_1c_2x\left(h - c_2 \int_{\mathbb{R}} y \mu_t(dy))\right) + 2c_5 x \int_{\mathbb{R}} y \mu_t(dy) \\
&= -2c_1c_2x(h - c_2 m_t) + 2c_5 x m_t,
\end{aligned}
$$

$$
\int_{\mathbb{R}} \frac{\delta H}{\delta \mu}\left(t, h, \boldsymbol{\mu}, -\partial_x V^T\right)(x)\mu_t(h)dh = -2c_1c_2x(m_t - c_2 m_t) + 2c_5 x m_t,
$$

and finally

$$
\begin{aligned}
&\beta V^T - \partial_t V^T + \frac{1}{2}(\partial_x^T)^2 - \frac{1}{2}\sigma^2 \partial_{xx} V^T - c_1(x - c_2 m_t)^2 - c_3(x - c_4)^2 \\
&\quad - c_5 {m_t}^2 + 2c_1c_2x(m_t - c_2 m_t) - 2c_5 x m_t = 0.
\end{aligned}
$$

The following system of ODEs is obtained by replacing the ansatz and its derivatives in the MFC-HJB:

$$
\begin{cases}
\dot{\Gamma}_2^T - 2(\Gamma_2^T)^2 - \beta\Gamma_2^T + c_1 + c_3 = 0, & \Gamma_2^T(T) = 0, \\
\dot{\Gamma}_1^T = (2\Gamma_2^T + \beta)\Gamma_1^T & \\
+(2c_1c_2(2 - c_2) - 2c_5)m_t^T + 2c_3c_4, & \Gamma_1^T(T) = 0, \\
\dot{\Gamma}_0^T = \beta\Gamma_0^T + \frac{1}{2}(\Gamma_1^T)^2 - \sigma^2\Gamma_2^T & \\
-c_3c_4{}^2 - (c_1c_2{}^2 + c_5)(m_t^T)^2, & \Gamma_0^T(T) = 0, \\
\dot{m}_t^T = -2\Gamma_2^T m^T - \Gamma_1^T, & m^T(0) = \mathbb{E}\left[X_0^{\boldsymbol{\alpha}}\right] = m_0,
\end{cases}
\tag{43}
$$

where the last equation is obtained by considering the expectation of $X_t^{\boldsymbol{\alpha}}$ after replacing $\alpha^*(x) = -\partial_x V^T(x) = -(\Gamma_2^T x + \Gamma_1^T)$. The first equation is a Riccati equation. In particular, the solution Γ_2^T converges to $\Gamma_2^* = \frac{-\beta + \sqrt{\beta^2 + 8(c_1 + c_3)}}{4}$ as T goes to infinity. The second and fourth ODEs are coupled, and they can be written in matrix notation as

$$
\begin{aligned}
\dot{\begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix}} &= \begin{bmatrix} -2\Gamma_2^T & -1 \\ (2c_1c_2(2 - c_2) - 2c_5) & 2\Gamma_2^T + \beta \end{bmatrix} \begin{pmatrix} m^T \\ \Gamma_1^T \end{pmatrix} \\
&\quad + \begin{pmatrix} 0 \\ 2c_3c_4 \end{pmatrix}, \quad \begin{pmatrix} m^T(0) \\ \Gamma_1^T(T) \end{pmatrix} = \begin{pmatrix} m_0 \\ 0 \end{pmatrix}.
\end{aligned}
\tag{44}
$$

By similar calculations to the non-asymptotic MFG case, the following solutions can be obtained

$$\begin{aligned} m_t^* &= \lim_{T\to\infty} m_t^T = p_t^*(1,1)h_1(t) + p_t^*(1,2)h_2(t) \\ &= \left(m_0 + 2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_2}\right)\right)e^{t\lambda_1} + 2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right), \\ \Gamma_1^*(t) &= \lim_{T\to\infty} \Gamma_1^T(t) = p_t^*(2,1)h_1(t) + p_t^*(2,2)h_2(t) \\ &= m_0(g-\lambda_1)e^{t\lambda_1} + 2\frac{c_3c_4}{\lambda_2-\lambda_1}\left(\frac{\lambda_2-g}{\lambda_2}-\frac{\lambda_1-g}{\lambda_1}\right), \end{aligned} \tag{45}$$

where

$$\begin{aligned} g &:= -2\Gamma_2^*, \\ b &:= 2(c_1c_2(2-c_2)-c_5), \\ a &:= 2\Gamma_2^* + \beta, \\ d &:= -1, \\ \lambda_{1\backslash 2} &:= \frac{a+g\pm\sqrt{(a-g)^2+4bd}}{2} = t\frac{\beta\pm\sqrt{(4\Gamma_2^*+\beta)^2-8(c_1c_2(2-c_2)-c_5)}}{2}. \end{aligned} \tag{46}$$

As in the MFG case, the third ODE in (43) can be solved by plugging in the solution of the previous ones and integrating. Since our interest is into the evolution of the mean and the control function, we omit the calculation for this ODE.

A.5 Solution for asymptotic MFC

The asymptotic version of the problem presented above is given by:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) &= \inf_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^\infty e^{-\beta t} f(X_t^{\boldsymbol{\alpha}}, \alpha_t, m^{\boldsymbol{\alpha}})\mathrm{d}t\right] \\ &= \inf_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^{+\infty} e^{-\beta t}\left(\frac{1}{2}\alpha_t^2 + c_1\left(X_t^{\boldsymbol{\alpha}} - c_2 m^{\boldsymbol{\alpha}}\right)^2 + c_3\left(X_t^{\boldsymbol{\alpha}} - c_4\right)^2 + c_5(m^{\boldsymbol{\alpha}})^2\right)\mathrm{d}t\right], \\ \text{subject to:}\quad & dX_t^{\boldsymbol{\alpha}} = \alpha_t \mathrm{d}t + \sigma dW_t, \quad X_0^{\boldsymbol{\alpha}} \sim \mu_0, \end{aligned}$$

where $m^{\boldsymbol{\alpha}} = \lim_{t\to+\infty} \mathbb{E}\left[X_t^{\alpha}\right]$.
Let $V(x)$ be the optimal value function conditioned on $X_0 = x$, i.e.,

$$\begin{aligned} V(x) &= \inf_{\boldsymbol{\alpha}} J^x(\boldsymbol{\alpha}) \\ &= \inf_{\boldsymbol{\alpha}} \mathbb{E}\left[\int_0^{+\infty} e^{-\beta t}\left(\frac{1}{2}\alpha_t^2 + c_1\left(X_t^{\boldsymbol{\alpha}} - c_2 m^{\boldsymbol{\alpha}}\right)^2 + c_3\left(X_t^{\boldsymbol{\alpha}} - c_4\right)^2 + c_5(m^{\boldsymbol{\alpha}})^2\right)\mathrm{d}t \,\middle|\, X_0^{\boldsymbol{\alpha}} = x\right]. \end{aligned}$$

We consider the following ansatz with its derivative

$$V(x) = \Gamma_2 x^2 + \Gamma_1 x + \Gamma_0,$$

$$\dot{V}(x) = 2\Gamma_2 x + \Gamma_1,$$
$$\ddot{V}(x) = 2\Gamma_2.$$

Starting by the MFC-HJB equation (4.12) given in [4], we extended it to the asymptotic case as follows

$$\beta V(x) - H\left(x, \mu^{\boldsymbol{\alpha}}, \alpha\right) - \int_{\mathbb{R}} \frac{\delta H}{\delta \mu}\left(h, \mu^{\boldsymbol{\alpha}}, -\dot{V}(h)\right)(x)\mu^{\boldsymbol{\alpha}}(h)dh = 0,$$

where $m^{\boldsymbol{\alpha}} = \int_{\mathbb{R}} y\mu^{\boldsymbol{\alpha}}(dy)$. We have:

$$\begin{aligned}
&H\left(x, \mu^{\boldsymbol{\alpha}}, \alpha\right)\\
&\quad := \inf_{\alpha}\left\{\mathcal{A}^X V(x) + f\left(x, \alpha, \mu^{\boldsymbol{\alpha}}\right)\right\}\\
&\quad = \inf_{\alpha}\left\{\alpha\dot{V}(x) + \frac{1}{2}\sigma^2\ddot{V}(x) + \frac{1}{2}\alpha^2 + c_1(x - c_2 m^{\boldsymbol{\alpha}})^2 + c_3(x - c_4)^2 + c_5(m^{\boldsymbol{\alpha}})^2\right\}\\
&\quad = -\frac{1}{2}\dot{V}(x)^2 + \frac{1}{2}\sigma^2\ddot{V}(x) + c_1(x - c_2 m^{\boldsymbol{\alpha}})^2 + c_3(x - c_4)^2 + c_5(m^{\boldsymbol{\alpha}})^2,\\
&\frac{\delta H\left(h, \mu^{\boldsymbol{\alpha}}, \alpha\right)}{\delta \mu}(x)\\
&\quad = \frac{\delta}{\delta \mu}\left(c_1(h - c_2 m^{\boldsymbol{\alpha}})^2 + c_5(m^{\boldsymbol{\alpha}})^2\right)(x)\\
&\quad = \frac{\delta}{\delta \mu}\left(c_1\left(h - c_2\int_{\mathbb{R}} y\mu^{\boldsymbol{\alpha}}(dy)\right)^2 + c_5\left(\int_{\mathbb{R}} y\mu^{\boldsymbol{\alpha}}(dy)\right)^2\right)(x)\\
&\quad = -2c_1c_2x\left(h - c_2\int_{\mathbb{R}} y\mu^{\boldsymbol{\alpha}}(dy))\right) + 2c_5x\int_{\mathbb{R}} y\mu^{\boldsymbol{\alpha}}(dy)\\
&\quad = -2c_1c_2x(h - c_2 m^{\boldsymbol{\alpha}}) + 2c_5 x m^{\boldsymbol{\alpha}},
\end{aligned}$$

$$\int_{\mathbb{R}} \frac{\delta H}{\delta \mu}\left(h, \mu^{\boldsymbol{\alpha}}, -\dot{V}(h)\right)(x)\mu^{\boldsymbol{\alpha}}(h)dh = -2c_1c_2x(m^{\boldsymbol{\alpha}} - c_2 m^{\boldsymbol{\alpha}}) + 2c_5 x m^{\boldsymbol{\alpha}},$$

and finally, the HJB equation becomes:

$$\begin{aligned}
\beta V(x) + \frac{1}{2}\dot{V}(x)^2 - \frac{1}{2}\sigma^2\ddot{V}(x) - c_1(x - c_2 m^{\boldsymbol{\alpha}})^2 - c_3(x - c_4)^2\\
- c_5(m^{\boldsymbol{\alpha}})^2 + 2c_1c_2x(m^{\boldsymbol{\alpha}} - c_2 m^{\boldsymbol{\alpha}}) - 2c_5 x m^{\boldsymbol{\alpha}} = 0.
\end{aligned}$$

A system of ODEs is obtained by replacing the ansatz and its derivatives in the MFC-HJB and cancelling terms in x^2, and x and constant:

$$\begin{aligned}
\left(\beta\Gamma_2 + 2\Gamma_2^2 - c_1 - c_3\right)x^2 + \left(\beta\Gamma_1 + 2\Gamma_2\Gamma_1 + 2c_1c_2m^{\boldsymbol{\alpha}}(2 - c_2) + 2c_3c_4 - 2c_5 m^{\boldsymbol{\alpha}}\right)x\\
+ \beta\Gamma_0 + \frac{1}{2}\Gamma_1^2 - \sigma^2\Gamma_2 - (c_1c_2{}^2 + c_5)(m^{\boldsymbol{\alpha}})^2 - c_3c_4{}^2 = 0.
\end{aligned}$$

An easy computation gives the values

$$\begin{aligned}\Gamma_2 &= \frac{-\beta + \sqrt{\beta^2 + 8(c_1 + c_3)}}{4},\\ \Gamma_1 &= \frac{2c_5 m^{\boldsymbol{\alpha}} - 2c_1 c_2 m^{\boldsymbol{\alpha}}(2 - c_2) - 2c_3 c_4}{\beta + 2\Gamma_2},\\ \Gamma_0 &= \frac{c_5 (m^{\boldsymbol{\alpha}})^2 + c_3 {c_4}^2 + c_1 {c_2}^2 (m^{\boldsymbol{\alpha}})^2 + \sigma^2 \Gamma_2 - \frac{1}{2}\Gamma_1^2}{\beta}.\end{aligned}$$

By plugging the control $\alpha^*(x) = -(2\Gamma_2 x + \Gamma_1)$ into the dynamics of $X_t^{\boldsymbol{\alpha}}$ and taking the expected value, we obtain an ODE for $m_t^{\boldsymbol{\alpha}}$

$$\dot{m}_t^{\boldsymbol{\alpha}} = -(2\Gamma_2 m_t^{\boldsymbol{\alpha}} + \Gamma_1). \tag{47}$$

The solution of (47) is used to derive m as follows

$$\begin{aligned} m^{\boldsymbol{\alpha}} &= \lim_{t\mapsto\infty} m_t^{\boldsymbol{\alpha}} = \lim_{t\mapsto\infty} \left(-\frac{\Gamma_1}{2\Gamma_2} + \left(m_0 + \frac{\Gamma_1}{\Gamma_2} \right) e^{-2\Gamma_2 t} \right) \\ &= -\frac{\Gamma_1}{2\Gamma_2} = -\frac{2c_5 m^{\boldsymbol{\alpha}} - 2c_1 c_2 m^{\boldsymbol{\alpha}}(2 - c_2) - 2c_3 c_4}{2\Gamma_2(\beta + 2\Gamma_2)} \\ m^{\boldsymbol{\alpha}} &= \frac{c_3 c_4}{\Gamma_2(\beta + 2\Gamma_2) + c_5 - c_1 c_2 (2 - c_2)} \end{aligned} \tag{48}$$

We remark that the values of $m_t^{\boldsymbol{\alpha}}$ and $\Gamma_1(t)$ obtained in the non-asymptotic case converge to m^α and Γ_1, respectively, as t goes to ∞. Therefore, we have obtained that

$$\lim_{t\to\infty} \alpha_t^{*MFC}(x) = \alpha^{*AMFG}(x), \quad \forall x,$$

that is the first part of (4) for this LQ MFC problem.

A.6 Solution for stationary MFC

The only difference with the derivation above in the case of asymptotic MFC is that m_t^α should be a constant which, from (47), should satisfy $2\Gamma_2 m^\alpha + \Gamma_1 = 0$. Therefore, m^α takes the same value as in (48), and we deduce

$$\alpha^{*SMFG}(x) = \alpha^{*AMFG}(x), \quad \forall x,$$

that is the second part of (4) for this LQ MFC problem .

B Lipschitz property of the 2 scale operators

B.1 Generic setting

We modify the original operators using the softmin operator on $\mathbb{R}^{|\mathcal{A}|}$ defined as:

$$\text{soft-min}(z) = \left(\frac{e^{-z_i}}{\sum_j e^{-z_j}} \right)_{i=1,\dots,|\mathcal{A}|} \in \Delta^{|\mathcal{A}|}, \qquad z \in \mathbb{R}^{|\mathcal{A}|}.$$

Intuitively, it gives a probability distribution on the indices $i = 1, \dots, |\mathcal{A}|$ which has higher values on indices whose corresponding values are closer to be a minimum. In particular, the elements of $\min\{i = 1, \dots, |\mathcal{A}| : z_i = \arg\min_j z_j\}$ have equal weight and this weight is the largest among $\left(\frac{e^{-z_i}}{\sum_j e^{-z_j}} \right)_{i=1,\dots,|\mathcal{A}|}$. We recall that the function soft-min is Lipschitz continuous for the 2-norm. Denoting by L_s its Lipschitz constant, it means that

$$\| \text{soft-min}(z) - \text{soft-min}(z') \|_2 \leq L_s \| z - z' \|_2, \qquad z, z' \in \mathbb{R}^{|\mathcal{A}|}.$$

Moreover, since $|\mathcal{A}|$ is finite, all the norms on $\mathbb{R}^{|\mathcal{A}|}$ are equivalent, so there exists a positive constant $c_{2,\infty}$ such that

$$\| \text{soft-min}(z) - \text{soft-min}(z') \|_\infty \leq L_s c_{2,\infty} \| z - z' \|_\infty, \qquad z, z' \in \mathbb{R}^{|\mathcal{A}|}.$$

To alleviate the notation, we will write $Q(x) := (Q(x, a))_{a \in \mathcal{A}}$ for any $Q \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{A}|}$. We also introduce a more general version $\underline{p}$ of the transition kernel p, which can take as an input a probability over actions instead of a single action: for $x, x' \in \mathcal{X}$, $\nu \in \Delta^{|\mathcal{A}|}$, $\mu \in \Delta^{|\mathcal{X}|}$,

$$\underline{p}(x' | x, \nu, \mu) = \sum_a \nu(a) p(x' | x, a, \mu).$$

Intuitively, this is the probability for an agent at x to move to x' when the population distribution is μ and the agent picks a random action following the distribution ν.

We now consider the following iterative procedure, which is a slight modification of (9a)–(9b). Here again, both variables (Q and μ) are updated at each iteration but with different rates. Starting from an initial guess $(Q_0, \mu_0) \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{A}|} \times \Delta^{|\mathcal{X}|}$, define iteratively for $k = 0, 1, \dots$:

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^\mu \underline{\mathcal{P}}(Q_k, \mu_k), & \text{(49a)} \\ Q_{k+1} = Q_k + \rho_k^Q \mathcal{T}(Q_k, \mu_k), & \text{(49b)} \end{cases}$$

where

$$\begin{cases} \mathcal{T}(Q,\mu)(x,a) = f(x,a,\mu) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q(x',a') \\ -Q(x,a), \qquad (x,a) \in \mathcal{X} \times \mathcal{A}, \\ \underline{\mathcal{P}}(Q,\mu)(x) = (\mu \underline{P}^{Q,\mu})(x) - \mu(x), \qquad x \in \mathcal{X}, \end{cases}$$

with

$$\underline{P}^{Q,\mu}(x,x') = \underline{p}(x'|x, \text{soft-min}\, Q(x), \mu),$$
$$\text{and} \qquad (\mu \underline{P}^{Q,\mu})(x) = \sum_{x_0} \mu(x_0) \underline{P}^{Q,\mu}(x_0,x),$$

is the transition matrix when the population distribution is μ and the agent uses an approximately optimal randomized control according to the soft-min of Q.

Lemma 1 *Assume that f is Lipschitz continuous with respect to μ and that $\underline{p}$ is Lipschitz continuous with respect to ν and μ. Then,*

- *the operator $\mathcal{T}$ is Lipschitz continuous w.r.t. μ (with a Lipschitz constant possibly depending on $\|Q\|_\infty$), and Lipschitz continuous in Q (uniformly in μ);*
- *the operator $\underline{\mathcal{P}}$ is Lipschitz continuous in both variables.*

If p is independent of μ, then both $\mathcal{T}$ and $\underline{\mathcal{P}}$ are Lipschitz continuous.

Proof Let us denote by L_p and L_f the Lipschitz constants of p and f, respectively. Let $(Q,\mu), (Q',\mu') \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \times \Delta^{|\mathcal{X}|}$. We first consider $\mathcal{T}$. We have

$$\begin{aligned} &\|\mathcal{T}(Q,\mu) - \mathcal{T}(Q',\mu)\|_\infty \\ &\quad \le \gamma \sum_{x'} \max_{x,a} p(x'|x,a,\mu) \left| \min_{a'} Q(x',a') - \min_{a'} Q'(x',a') \right| + \left\| Q - Q' \right\|_\infty \\ &\quad \le (\gamma+1) \left\| Q - Q' \right\|_\infty . \end{aligned}$$

Moreover,

$$\begin{aligned} \|\mathcal{T}(Q,\mu) - \mathcal{T}(Q,\mu')\|_\infty &\le |f(x,a,\mu) - f(x,a,\mu')| \\ &\quad + \gamma \sum_{x'} |p(x'|x,a,\mu) - p(x'|x,a,\mu')| \, |\min_{a'} Q(x',a')| \\ &\le (L_f + \gamma L_p \|Q\|_\infty) |\mathcal{X}| \|\mu - \mu'\|_\infty, \end{aligned}$$

where L_f and L_p are, respectively, the Lipschitz constants of f and p with respect to μ. If p is independent of μ, we obtain

$$\|\mathcal{T}(Q,\mu) - \mathcal{T}(Q,\mu')\|_\infty \le L_f \|\mu - \mu'\|_\infty.$$

We then show that the operator $\underline{\mathcal{P}}$ is Lipschitz continuous. We have

$$\begin{aligned}
&\|\underline{\mathcal{P}}(Q,\mu)-\underline{\mathcal{P}}(Q,\mu')\|_\infty \\
&\quad\le \|\mu \underline{P}^{Q,\mu}-\mu' \underline{P}^{Q,\mu'}\|_\infty + \|\mu-\mu'\|_\infty \\
&\quad\le \left\| \sum_x \Big(\underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu)\mu(x) - \underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu')\mu'(x) \Big) \right\|_\infty \\
&\qquad + \|\mu-\mu'\|_\infty .
\end{aligned}$$

For the first term, we note that, for every $x \in \mathcal{X}$,

$$\begin{aligned}
&\left\| \Big(\underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu)\mu(x) - \underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu')\mu'(x) \Big) \right\|_\infty \\
&\quad\le \left\| \Big(\underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu) - \underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu') \Big)\mu(x) \right\|_\infty \\
&\qquad + \left\| \underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu')\Big(\mu(x)-\mu'(x)\Big) \right\|_\infty \\
&\quad\le (L_p+1)\left\|\mu-\mu'\right\|_\infty ,
\end{aligned}$$

where we used the fact that discrete probability measures are non-negative and bounded by 1.

Moreover, we have

$$\begin{aligned}
\|\underline{\mathcal{P}}(Q,\mu)-\underline{\mathcal{P}}(Q',\mu)\|_\infty &\le \|\mu(\underline{P}^{Q,\mu}-\underline{P}^{Q',\mu'})\|_\infty \\
&\le \sum_x \|\underline{p}(\cdot|x, \text{soft-min}\, Q(x), \mu) \\
&\quad - \underline{p}(\cdot|x, \text{soft-min}\, Q'(x), \mu)\|_\infty \\
&\le \sum_x L_p \| \text{soft-min}\, Q(x) - \text{soft-min}\, Q'(x)\|_\infty \\
&\le |\mathcal{X}|\, L_p\, L_s\, c_{2,\infty}\, \|Q-Q'\|_\infty ,
\end{aligned}$$

which concludes the proof. □

B.2 Application to a discrete model for the LQ problem

Recall that the continuous linear-quadratic model we consider is defined by (15). Here, we propose a finite space MDP which approximates the dynamics of a typical agent in this continuous LQ model. We consider that the action space is given by $\mathcal{A} = \{a_0 = -1, a_1 = -1 + \Delta_\cdot, \ldots, a_{N_\mathcal{A}} = 1 - \Delta_\cdot, a_{N_\mathcal{A}} = 1\}$ and the state space by $\mathcal{X} = \{x_0 = x_c - 2, x_1 = x_c - 2 - \Delta_\cdot, \ldots, x_{N_\mathcal{X}-1} = x_c + 2 - \Delta_\cdot, x_{N_\mathcal{X}} = x_c + 2\}$, where x_c is the center of the state space. The step size for the discretization of the spaces $\mathcal{X}$ and $\mathcal{A}$ is given by $\Delta_\cdot = \sqrt{\Delta t} = 10^{-1}$.

Consider the transition probability:

$$\begin{aligned} p(x, x', a, \mu) &= \mathbb{P}(Z^{x+a,\Delta t} \in [x' - \Delta_./2, x' + \Delta_./2]) \\ &= \Phi_{x+a,\sigma^2\Delta t}(x' + \Delta_./2) - \Phi_{x+a,\sigma^2\Delta t}(x' - \Delta_./2), \end{aligned}$$

where $Z \sim \mathcal{N}(x + a, \sigma^2\Delta t)$ and $\Phi_{x+a,\sigma^2\Delta t}$ is the cumulative distribution function of the $\mathcal{N}(x + a, \sigma^2\Delta t)$ distribution. Moreover, consider that the one-step cost function is given by $f(x, a, \mu)\Delta t$ with

$$\begin{aligned} f(x, a, \mu) = \frac{1}{2}a^2 + c_1 \left(x - c_2 \sum_{\xi \in S} \mu(\xi) \right)^2 + c_3 \, (x - c_4)^2 \\ + c_5 \left(\sum_{\xi \in S} \mu(\xi) \right)^2, \qquad b(x, a, \mu) = a, \end{aligned}$$

For simplicity, we write $\bar{\mu} = \sum_{\xi \in S} \mu(\xi)$.

Lemma 2 *In this model, f is Lipschitz continuous with respect to μ and $\underline{p}$ is Lipschitz continuous with respect to ν and μ*

Proof We start with f. For the μ component, we have:

$$\begin{aligned} |f(x, a, \mu) - f(x, a, \mu')| &\leq c \left| (x - c_2\bar{\mu})^2 - \left(x - c_2\bar{\mu}'\right)^2 \right| + c \left| (\bar{\mu})^2 - \left(\bar{\mu}'\right)^2 \right| \\ &\leq c \left(\bar{\mu}' - \bar{\mu}\right) \cdot \left(2x + (\bar{\mu}' - \bar{\mu})\right) + c(\bar{\mu} - \bar{\mu}')(\bar{\mu} + \bar{\mu}') \\ &\leq c \max_{x \in S} \|x\|_\infty \left(\bar{\mu}' - \bar{\mu}\right) \\ &\leq c \max_{x \in S} \|x\|_\infty \sum_{x \in S} \left(\mu'(x) - \mu(x)\right) \\ &\leq c \max_{x \in S} \|x\|_\infty \, |S| \, \|\mu' - \mu\|_\infty, \end{aligned}$$

where $c > 0$ is a constant depending only on the parameters of the model and whose value may change from line to line.

Then, we consider $\underline{p}$. It is independent of μ in this model. For the action component, we have:

$$\begin{aligned} &|\underline{p}(x, x', \nu, \mu) - \underline{p}(x, x', \nu', \mu)| \\ &\quad = \left| \sum_a \nu(a) \Big(\Phi_{x+a,\sigma^2\Delta t}(x' + \Delta_./2) - \Phi_{x+a,\sigma^2\Delta t}(x' - \Delta_./2) \Big) \right. \\ &\qquad \left. - \sum_{a'} \nu'(a') \Big(\Phi_{x+a',\sigma^2\Delta t}(x' + \Delta_./2) - \Phi_{x+a',\sigma^2\Delta t}(x' - \Delta_./2) \Big) \right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \sum_a \left(\nu(a) \Phi_{x+a,\sigma^2\Delta t}(x' + \Delta_./2) - \nu'(a) \Phi_{x+a,\sigma^2\Delta t}(x' + \Delta_./2) \right) \right| \\
&\quad + \left| \sum_a \left(\nu(a) \Phi_{x+a,\sigma^2\Delta t}(x' - \Delta_./2) \right) - \nu'(a) \Phi_{x+a,\sigma^2\Delta t}(x' - \Delta_./2) \right) \Bigg| \\
&= \int_{-\infty}^{x'+\Delta_./2} \frac{1}{\sigma\sqrt{2\pi\,\Delta t}} \left| \sum_a (\nu(a) - \nu'(a)) e^{-\frac{(y-(x+a))^2}{2\sigma^2\Delta t}} \right| dy \\
&\quad + \int_{-\infty}^{x'-\Delta_./2} \frac{1}{\sigma\sqrt{2\pi\,\Delta t}} \left| \sum_a (\nu(a) - \nu'(a)) e^{-\frac{(y-(x+a))^2}{2\sigma^2\Delta t}} \right| dy \\
&\leq c \|\nu - \nu'\|_\infty,
\end{aligned}
$$

where c is a constant depending only on the model (and in particular on the state space, the action space and Δt). □

C The Bellman equation for the optimal Q function in the asymptotic MFC framework

In this appendix, we provide the derivation of the Bellman equation (8) for the modified Q-function presented in Sect. 3.3.

Let $\mathcal{X}$ and $\mathcal{A}$ be discrete and finite state and action spaces. Let $V^\alpha : \mathcal{X} \mapsto \mathcal{R}$ and $Q^\alpha : \mathcal{X} \times \mathcal{A} \mapsto \mathcal{R}$ be value function relative to the policy α and the corresponding modified Q-function defined as follows

$$
V^\alpha(x) := \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^\alpha) \,\middle|\, X_0 = x \right], \tag{50}
$$

$$
Q^\alpha(x,a) := f(x,a,\mu^{\tilde{\alpha}}) + \mathbb{E}\left[\sum_{n=1}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^\alpha) \,\middle|\, X_0 = x, A_0 = a \right], \tag{51}
$$

where

$$
\mu^\alpha = \lim_{n\mapsto\infty} \mathcal{L}(X_n^\alpha) \quad \text{and} \quad \tilde{\alpha}(s) = \begin{cases} \alpha(s), & \forall s \neq x, \\ a, & \text{if } s = x. \end{cases}
$$

Theorem 2 *The optimal $Q^*(x,a) = \min_\alpha Q^\alpha(x,a)$ satisfies the Bellman equation*

$$
Q^*(x,a) = f(x,a,\tilde{\mu}^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\tilde{\mu}^*) \min_{a'} Q^*(x',a'), \qquad (x,a) \in \mathcal{X} \times \mathcal{A}, \tag{52}
$$

where the optimal control α^* is given by $\alpha^*(x) = \arg\min_a Q^*(x,a)$, the modification $\tilde{\alpha}^*(x)$ is based on the pair (x,a) and $\tilde{\mu}^* := \mu^{\tilde{\alpha}^*}$.

Remark 3 The population distribution $\tilde{\mu}^*$ based on the modification of α^* given the pair $(x, \alpha^*(x))$ is equal to μ^* . Indeed, $\tilde{\alpha}^*$ is equal to α^* itself, i.e.,

$$\tilde{\alpha}^*(s) = \begin{cases} \alpha^*(s), & \forall s \neq x, \\ \alpha^*(s), & \text{if } s = x. \end{cases}$$

Remark 4 The term $\min_{a'} Q^*(x', a')$ does not depend on $\tilde{\mu}^*$, i.e.,

$$\begin{aligned} \min_{a'} Q^*(x', a') &= Q^*(x', \alpha^*(x')) \\ &= f(x', \alpha^*(x'), \tilde{\mu}^*) + \gamma \sum_{x'' \in \mathcal{X}} p(x''|x', \alpha^*(x'), \tilde{\mu}^*) \min_{a'} Q^*(x'', a') \\ &\overset{\square}{=} f(x', \alpha^*(x'), \mu^*) + \gamma \sum_{x' \in \mathcal{X}} p(x''|x', \alpha^*(x'), \mu^*) \min_{a'} Q^*(x'', a') \end{aligned}$$

where step $\square$ is due to Remark 3. It follows that (52) depends on $\tilde{\mu}^*$ only through the cost due to the first step.

In order to prove Theorem 2, the following results are required.

Theorem 3 *The Bellman equation for Q^α is given by*

$$Q^\alpha(x, a) = f(x, a, \mu^{\tilde{\alpha}}) + \gamma \mathbb{E}\Big[Q^\alpha(X_1, \alpha(X_1)) \,\Big|\, X_0 = x, A_0 = a \Big], \tag{53}$$

Lemma 3 *The value function relative to the policy α is equivalent to the corresponding Q-function evaluated on the pair $(x, \alpha(x))$, i.e.,*

$$V^\alpha(x) = Q^\alpha(x, \alpha(x)). \tag{54}$$

Theorem 4 (Policy improvement) *Let $\tilde{\alpha}$ be a policy derived by α*

$$\tilde{\alpha}(s) = \begin{cases} \alpha(s), & \text{for } s \neq x, \\ a, & \text{for } s = x. \end{cases}$$

such that

$$Q^\alpha(x, \tilde{\alpha}(x)) > V^\alpha(x). \tag{55}$$

Then,

$$V^{\tilde{\alpha}}(x') > V^\alpha(x') \quad \forall x' \in \mathcal{X}. \tag{56}$$

Theorem 5 *Let $V^* : \mathcal{X} \mapsto \mathcal{R}$ be defined as $V^*(x) = \max_\alpha V^\alpha(x)$. Then,*

$$V^*(x) = \max_a \max_\alpha Q^\alpha(x, a), \tag{57}$$

Proof (Theorem 3)

$$
\begin{aligned}
&Q^{\alpha}(x,a)\\
&= f(x,a,\mu^{\tilde{\alpha}})+\\
&\quad+\gamma\mathbb{E}\left[\mathbb{E}\left[\sum_{n=1}^{\infty}\gamma^{n-1}f(X_n,\alpha(X_n),\mu^{\alpha})\,\middle|\,X_0=x,A_0=\alpha(x),X_1\right]\,\middle|\,X_0=x,A_0=a\right]\\
&= f(x,a,\mu^{\tilde{\alpha}})+\gamma\mathbb{E}\left[\mathbb{E}\left[\sum_{n=1}^{\infty}\gamma^{n-1}f(X_n,\alpha(X_n),\mu^{\alpha})\,\middle|\,X_1\right]\,\middle|\,X_0=x,A_0=a\right]\\
&= f(x,a,\mu^{\tilde{\alpha}})+\\
&\quad+\gamma\mathbb{E}\left[f(X_1,\alpha(X_1),\mu^{\alpha})+\gamma\mathbb{E}\left[\sum_{n=2}^{\infty}\gamma^{n-2}f(X_n,\alpha(X_n),\mu^{\alpha})\,\middle|\,X_1\right]\,\middle|\,X_0=x,A_0=a\right]\\
&= f(x,a,\mu^{\tilde{\alpha}})+\gamma\mathbb{E}\Big[Q^{\alpha}(X_1,\alpha(X_1))\,\Big|\,X_0=x,A_0=a\Big],
\end{aligned}
$$

□

Proof (Lemma 3)

$$
\begin{aligned}
V^{\alpha}(x) &= f(x,\alpha(x),\mu^{\alpha})+\mathbb{E}\left[\sum_{n=1}^{\infty}\gamma^{n}f(X_n,\alpha(X_n),\mu^{\alpha})\,\middle|\,X_0=x,A_0=\alpha(x)\right]\\
&= f(x,\alpha(x),\mu^{\tilde{\alpha}})+\mathbb{E}\left[\sum_{n=1}^{\infty}\gamma^{n}f(X_n,\alpha(X_n),\mu^{\alpha})\,\middle|\,X_0=x,A_0=\alpha(x)\right]\\
&\overset{(51)}{=} Q^{\alpha}(x,\alpha(x))
\end{aligned}
$$

where we used that the modification of α given the pair $(x,\alpha(x))$ is equal to α itself and consequently $\mu^{\alpha}=\mu^{\tilde{\alpha}}$. □

Proof (Theorem 4) **Step 1** Show that $V^{\alpha}(x)<V^{\tilde{\alpha}}(x)$.
We observe that

$$
\begin{aligned}
V^{\alpha}(x) &< Q^{\alpha}(x,\tilde{\alpha}(x))\\
&\overset{(53)}{=} f(x,\tilde{\alpha}(x),\mu^{\tilde{\alpha}})+\gamma\mathbb{E}\Big[Q^{\alpha}(X_1,\alpha(X_1))\,\Big|\,X_0=x,A_0=\tilde{\alpha}(x)\Big]\\
&\overset{(54)}{=} f(x,\tilde{\alpha}(x),\mu^{\tilde{\alpha}})+\gamma\mathbb{E}\Big[V^{\alpha}(X_1)\,\Big|\,X_0=x,A_0=\tilde{\alpha}(x)\Big]\leq\\
&\overset{(55)}{\leq} f(x,\tilde{\alpha}(x),\mu^{\tilde{\alpha}})+\gamma\mathbb{E}\Big[Q^{\alpha}(X_1,\tilde{\alpha}(X_1))\,\Big|\,X_0=x,A_0=\tilde{\alpha}(x)\Big]\\
&\overset{(53)}{=} f(x,\tilde{\alpha}(x),\mu^{\tilde{\alpha}})\\
&\quad+\gamma\mathbb{E}\Big[f(X_1,\tilde{\alpha}(X_1),\mu^{\tilde{\alpha}})+\gamma Q^{\alpha}(X_{t_2},\alpha(X_{t_2}))\,\Big|\,X_0=x,A_0=\tilde{\alpha}(x)\Big]\leq\\
&\ \ \vdots\\
&\leq\mathbb{E}\left[\sum_{n=0}^{k}\gamma^{n}f(X_n,\tilde{\alpha}(X_n),\mu^{\tilde{\alpha}})+\gamma^{k+1}V^{\alpha}(X_{k+1})\,\middle|\,X_0=x\right]
\end{aligned}
$$

Considering the limit as $k \to \infty$, it follows that

$$V^{\alpha}(x) < \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n f(X_n, \tilde{\alpha}(X_n), \mu^{\tilde{\alpha}}) \,\middle|\, X_0 = x\right] = V^{\tilde{\alpha}}(x)$$

Step 2 Show that $V^{\alpha}(x') < V^{\tilde{\alpha}}(x') \quad \forall x' \in \mathcal{X} \setminus \{x\}$.
Let define $\tau_x = \min\{n : X_n = x\}$. Then,

$$\begin{aligned}
V^{\alpha}(x') &= \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^{\alpha}) \,\middle|\, X_0 = x'\right] \\
&= \mathbb{E}\left[\sum_{n=0}^{\tau_x - 1} \gamma^n f(X_n, \alpha(X_n), \mu^{\alpha}) + \sum_{n=\tau_x}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^{\alpha}) \,\middle|\, X_0 = x'\right] \\
&= \mathbb{E}\left[\sum_{n=0}^{\tau_x - 1} \gamma^n f(X_n, \alpha(X_n), \mu^{\alpha}) \,\middle|\, X_0 = x'\right] \\
&\quad + \mathbb{E}\left[\sum_{n=\tau_x}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^{\alpha}) \,\middle|\, X_0 = x'\right] = \\
&:= T_1 + T_2
\end{aligned}$$

We start analyzing the first term observing that $X_n \neq x$ and $\alpha(X_n) = \tilde{\alpha}(X_n)$ for all $n <= \tau_x - 1$. Then,

$$T_1 = \mathbb{E}\left[\sum_{n=0}^{\tau_x - 1} \gamma^n f(X_n, \tilde{\alpha}(X_n), \mu^{\tilde{\alpha}}) \,\middle|\, X_0 = x'\right]$$

The analyses of the term T_2 are based on the tower property (TP), the Markov property (MP) and Step 1 (S1). It follows that

$$\begin{aligned}
T_2 &\overset{\text{(TP)}}{=} \mathbb{E}\left[\mathbb{E}\left[\sum_{n=\tau_x}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^{\alpha}) \,\middle|\, X_0 = x', X_1, \ldots, X_{\tau_x}\right] \,\middle|\, X_0 = x'\right] \\
&\overset{\text{(MP)}}{=} \mathbb{E}\left[\gamma^{\tau_x} \mathbb{E}\left[\sum_{n=\tau_x}^{\infty} \gamma^{n-\tau_x} f(X_n, \alpha(X_n), \mu^{\alpha}) \,\middle|\, X_{\tau_x}\right] \,\middle|\, X_0 = x'\right] \\
&= \mathbb{E}\left[\gamma^{\tau_x} V^{\alpha}(X_{\tau_x}) \,\middle|\, X_0 = x'\right] < \\
&\overset{\text{(S1)}}{<} \mathbb{E}\left[\gamma^{\tau_x} V^{\tilde{\alpha}}(X_{\tau_x}) \,\middle|\, X_0 = x'\right]
\end{aligned}$$

Combining the analyses of T_1 and T_2, it follows that

$$\begin{aligned} V^{\alpha}(x') &= T_1 + T_2 \\ &< \mathbb{E}\left[\sum_{n=0}^{\tau_x - 1} \gamma^n f(X_n, \tilde{\alpha}(X_n), \mu^{\tilde{\alpha}}) \,\middle|\, X_0 = x'\right] + \mathbb{E}\left[\gamma^{\tau_x} V^{\tilde{\alpha}}(X_{\tau_x}) \,\middle|\, X_0 = x'\right] \\ &= \mathbb{E}\left[\sum_{n=0}^{\tau_x - 1} \gamma^n f(X_n, \tilde{\alpha}(X_n), \mu^{\tilde{\alpha}}) + \gamma^{\tau_x} \sum_{n=\tau_x}^{\infty} \gamma^{n-\tau_x} f(X_n, \tilde{\alpha}(X_n), \mu^{\tilde{\alpha}}) \,\middle|\, X_0 = x'\right] \\ &= \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n f(X_n, \tilde{\alpha}(X_n), \mu^{\tilde{\alpha}}) \,\middle|\, X_0 = x'\right] \\ &= V^{\tilde{\alpha}}(x') \end{aligned}$$

□

Proof (Theorem 5) Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{A} = \{a_0, \ldots, a_m\}$ be the state and action spaces.

Step 1 Let α^0 be an initial policy and define α^1 as follows

$$\alpha^1(x) = \begin{cases} \arg\max_a Q^{\alpha^0}(x, a), & \text{if } x = x_1, \\ \alpha_0(x), & \text{o.w.} \end{cases}$$

Then,

$$Q^{\alpha^0}(x_1, \alpha^1(x_1)) \geq V^{\alpha^0}(x_1) \overset{(56)}{\Longrightarrow} V^{\alpha^1}(x) \geq V^{\alpha^0}(x), \quad \forall x$$

Step 2 Consider α^2 defined as follows

$$\begin{aligned} \alpha^2(x) &= \begin{cases} \arg\max_a Q^{\alpha^1}(x, a), & \text{if } x = x_2, \\ \alpha_1(x), & \text{o.w.} \end{cases} \\ &= \begin{cases} \arg\max_a Q^{\alpha^1}(x, a), & \text{if } x = x_2, \\ \arg\max_a Q^{\alpha^0}(x, a), & \text{if } x = x_1, \\ \alpha_0(x), & \text{o.w.} \end{cases} \end{aligned}$$

Then,

$$Q^{\alpha^1}(x_2, \alpha^2(x_2)) \geq V^{\alpha^1}(x_1) \overset{(56)}{\Longrightarrow} V^{\alpha^2}(x) \geq V^{\alpha^1}(x) \geq V^{\alpha^0}(x), \quad \forall x$$

Step *n* Consider α^n defined as follows

$$\begin{aligned} \alpha^n(x) &= \begin{cases} \arg\max_a Q^{\alpha^{n-1}}(x, a), & \text{if } x = x_n, \\ \alpha_{n-1}(x), & \text{o.w.} \end{cases} \\ &= \arg\max_a Q^{\alpha^{k-1}}(x, a), \qquad \text{if } x = x_k, \text{ for } k = 1, \ldots, n, \end{aligned}$$

Then,

$$Q^{\alpha^{n-1}}(x_n, \alpha^n(x_n)) \geq V^{\alpha^{n-1}}(x_n) \overset{(56)}{\Longrightarrow} V^{\alpha^n}(x) \geq V^{\alpha^{n-1}}(x) \geq V^{\alpha^0}(x), \quad \forall x$$

Step *N* Since the state and action spaces are finite, the policy can be improved only a finite number of times. In other words, $\exists N > 0$ such that

$$\alpha^N(x) = \arg\max_a Q^{\alpha^N}(x, a), \quad \forall x \in \mathcal{X}$$

and

$$V^{\alpha^N}(x) = Q^{\alpha^N}(x, \alpha^N(x)) = \max_a Q^{\alpha^N}(x, a), \quad \forall x \in \mathcal{X}.$$

Can α^N be still suboptimal? No, by extending Bellman and Dreyfus's Optimality Theorem (1962), [3]. □

Proof (Theorem (2))

$$\begin{aligned}
\text{RHS} &= f(x, a, \mu^{\tilde{\alpha}}) + \gamma \mathbb{E}\left[\max_{a'} Q^*(X_1, a') \,\middle|\, X_0 = x, A_0 = a\right] \\
&\overset{(57)}{=} f(x, a, \mu^{\tilde{\alpha}}) + \gamma \mathbb{E}\left[V^*(X_1) \,\middle|\, X_0 = x, A_0 = a\right] \\
&\overset{(54)}{=} f(x, a, \mu^{\tilde{\alpha}}) + \gamma \mathbb{E}\left[Q^{\alpha^*}(X_1, \alpha^*(X_1)) \,\middle|\, X_0 = x, A_0 = a\right] \\
&\overset{(53)}{=} Q^{\alpha^*}(x, a) = Q^*(x, a),
\end{aligned}$$

where the last step is due to what shown in the proof of equation (57), i.e., the same policy α^* optimizes V^α and Q^α. □

References

1. Anahtarci B, Kariksiz CD, Saldi N (2020) Q-learning in regularized mean-field games. arXiv preprint arXiv:2003.12151
2. Angiuli A, Fouque J-P, Laurière M (2021) Reinforcement learning for mean field games, with applications to economics. To appear in the Handbook on Machine Learning in Financial Markets: A guide to contemporary practises, editors: A. Capponi and C.-A. Lehalle, Cambridge University Press.
3. Bellman RE, Dreyfus SE (2015) Applied dynamic programming, vol 2050. Princeton University Press
4. Bensoussan A, Frehse J, Chi PYS (2013) Mean field games and mean field type control theory. Springer Briefs in Mathematics, Springer, New York
5. Borkar VS (1997) Stochastic approximation with two time scales. Syst Control Lett 29(5):291–294
6. Borkar VS (2008) Stochastic approximation. Cambridge University Press, Cambridge, Hindustan Book Agency, New Delhi. A dynamical systems viewpoint
7. Cardaliaguet P, Hadikhanloo S (2017) Learning in Mean Field Games: the Fictitious Play. COCV 23:569–591.
8. Carmona R, Delarue F (2018) Probabilistic theory of mean field games with applications I–II. Springer
9. Carmona R, Mathieu L (2019) Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: I–the ergodic case. arXiv preprint arXiv:1907.05980
10. Carmona R, Laurière M (2019) Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: II–the finite horizon case. arXiv preprint arXiv:1908.01613

11. Carmona R, Laurière M, Zongjun T (2019) Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. Preprint
12. Carmona R, Laurière M, Zongjun T (2019) Mean-field MDP and mean-field Q-learning: model-free mean-field reinforcement learning. Preprint
13. Elie R, Perolat J, Laurière M, Geist M, Pietquin O (2020) On the convergence of model free learning in mean field games. In: Proceedings of AAAI
14. Even-DE Mansour Y (2003) Learning rates for q-learning. J Mach Learn Res 5(Dec):1–25
15. Fouque JP, Zhang Z (2020) Deep learning methods for mean field control problems with delay. Front Appl Math Stat 6(11)
16. Fu Z, Yang Z, Chen Y, Wang Z (2019) Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. arXiv preprint arXiv:1910.07498
17. Gu H, Guo X, Wei X, Xu R (2019) Dynamic programming principles for learning MFCS. arXiv preprint arXiv:1911.07314
18. Gu H, Guo X, Wei X, Xu R (2020) Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. arXiv preprint arXiv:2002.04131
19. Guo X, Hu A, Xu R, Zhang J (2019) Learning mean-field games. In: Advances in neural information processing systems, pp 4966–4976
20. Han J, Hu R (2020) Deep fictitious play for finding Markovian Nash equilibrium in multi-agent games. arXiv:1912.01809
21. Huang M, Caines PE, Malhamé RP (2007) Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria. IEEE Trans Autom Control 52(9):1560–1571
22. Huang M, Malhamé RP, Caines PE (2006) Large population stochastic dynamic games: closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. Commun Inf Syst 6(3):221–251
23. Lasry J-M, Lions P-L (2007) Mean field games. Jpn J Math 2(1):229–260
24. Mguni D, Jennings J, de Cote EM (2018) Decentralised learning in systems with many, many strategic agents. In: Thirty-second AAAI conference on artificial intelligence
25. Motte M, Pham H (2019) Mean-field Markov decision processes with common noise and open-loop controls. arXiv preprint arXiv:1912.07883
26. Perrin S, Pérolat J, Laurière M, Geist M, Elie R, Olivier P (2020) Fictitious play for mean field games: continuous time analysis and applications. In preparation
27. Subramanian J, Mahajan A (2019) Reinforcement learning in stationary mean-field games. In: Proceedings. 18th international conference on autonomous agents and multiagent systems
28. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT press
29. Watkins CJCH (1989) Learning from delayed rewards. PhD thesis, King's College, Cambridge
30. Xie Q, Yang Z, Wang Z, Minca A (2020) Provable fictitious play for general mean-field games. arXiv preprint arXiv:2010.04211
31. Yang J, Ye X, Trivedi R, Xu H, & Zha H (2018) Deep mean field games for learning optimal behavior policy of large populations. In International Conference on Learning Representations.
32. Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J (2018) Mean field multi-agent reinforcement learning. In: International conference on machine learning, pp 5567–5576