Reinforcement Learning for Intra-and-Inter-Bank Borrowing and Lending Mean Field Control Game

Andrea Angiuli
Prime Machine Learning Team, Amazon
Seattle, USA
aangiuli@amazon.com

Nils Detering
Department of Statistics and Applied Probability, University of California, Santa Barbara
Santa Barbara, USA
detering@pstat.ucsb.edu

Jean-Pierre Fouque
Department of Statistics and Applied Probability, University of California, Santa Barbara
Santa Barbara, USA
fouque@pstat.ucsb.edu

Mathieu Laurière
Mathematics and Data Science, NYU Shanghai
Shanghai, China
mathieu.lauriere@nyu.edu

Jimin Lin
Department of Statistics and Applied Probability, University of California, Santa Barbara
Santa Barbara, USA
jiminlin@pstat.ucsb.edu

ABSTRACT

We propose a *mean field control game* (MFCG) model for the intra-and-inter-bank borrowing and lending problem. This framework allows to study the competitive game arising between groups of collaborative banks. The solution is provided in terms of an asymptotic Nash equilibrium between the groups in the infinite horizon. A three-timescale reinforcement learning algorithm is applied to learn the optimal borrowing and lending strategy in a data driven way when the model is unknown. An empirical numerical analysis shows the importance of the three-timescale, the impact of the exploration strategy when the model is unknown, and the convergence of the algorithm.

CCS CONCEPTS

• **Theory of computation** → **Reinforcement learning**; **Solution concepts in game theory**; • **Mathematics of computing** → **Stochastic control and optimization**.

KEYWORDS

reinforcement learning, mean field control game, systemic risk

ACM Reference Format:
Andrea Angiuli, Nils Detering, Jean-Pierre Fouque, Mathieu Laurière, and Jimin Lin. 2022. Reinforcement Learning for Intra-and-Inter-Bank Borrowing and Lending Mean Field Control Game. In *3rd ACM International Conference on AI in Finance (ICAIF '22), November 2–4, 2022, New York, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3533271.3561743

ICAIF '22, November 2–4, 2022, New York, NY, USA

ACM ISBN 978-1-4503-9376-8/22/11…$15.00
https://doi.org/10.1145/3533271.3561743

1 INTRODUCTION

Many problems in finance involve a large number of strategic agents. A typical example is how traders interact in a common market through the price of some assets. At a larger scale, another example is how banks interact through the money they borrow from or lend to each other or to a central bank. When the agents are competing, one can represent the problem as a game and look for a Nash equilibrium. On the other hand, when the agents are cooperating, one can look for a social optimum. The problems have different solutions, and the non-cooperative equilibrium has a higher average cost per player, which is interpreted as a lack of efficiency. This leads to the notion of price of anarchy [19].

When the number of agents is large, studying every pairwise interactions becomes intractable. To simplify the analysis, a mean field approximation can be used, assuming that the population is homogeneous and the interactions are symmetric. This idea led to the notion of mean field games (MFGs) and mean field control (MFC) problems (also known as McKean Vlasov control problems) depending on whether the agents are competitive or cooperative, [6, 9, 17, 21]. A related notion is the concept of mean field type game (MFTG), in which a finite number of players compete and each player's problem is of MFC type [12]. MFG, MFC and MFTG have found applications in energy production and management [1, 11, 15], crowd trading [7], systemic risk [10], to cite just a few examples. See e.g. [8] for a recent survey of applications to finance and economics.

In the past few years, the question of *learning* solutions to MFG and MFC problems using model-free methods based on reinforcement learning (RL) has gained momentum. Many of these methods rely on updating a value function and a distribution. In particular, stationary MFG solutions have been approximated in [16] using fixed point iterations and Q-learning, and in [13] using fictitious play and deep RL. Two-timescale analysis to learn MFG solutions have been used in [22, 23]. Recently, a two timescale algorithm has been introduced in [5] to solve MFG or MFC depending on the choice of learning rates for the distribution and the value function. This allows to have a unified point of view on these two types

of problems and a common RL method. In [3], the approach has been extended to mean field control games (MFCG) using a three-timescale RL algorithm. It was developed in a finite horizon setting for extended MFCGs arising naturally in the context of the trader's liquidation problem between competitive groups of collaborative traders who share the inventory cost of their group.

In the present paper, our main contributions are threefold. First, in Section 2, we introduce a model of intra-and-inter-bank borrowing and lending, which can be viewed as an extension of the model studied in [10] where there are local coalitions inside each bank. Second, in Section 3, we apply a three-timescale RL algorithm to solve this class of infinite horizon problems. Last, in Section 4, we show numerical results that illustrate the performance of our method on the model of intra-and-inter-bank borrowing and lending.

2 INTRA-AND-INTER-BANK BORROWING AND LENDING PROBLEM

A model of inter-bank borrowing and lending has been introduced in [10] as a linear-quadratic stochastic differential game between banks which control their drifts and minimize a quadratic cost with incentive to stay close to the average capitalization of the system. The model has been studied as a finite-player game in finite horizon. Open-loop and closed-loop Nash equilibria have been computed using Forward-Backward Stochastic Differential Equations (FBSDE) and Hamilton-Jacobi-Bellman (HJB) partial differential equations. In this model the central bank acts as the clearing house. Systemic risk has then been considered as a large deviation event. In addition to the finite player game a mean field game (MFG) limit has been discussed as well. In the present paper we first propose an extension of the aforementioned model where the competitive banks are made of collaborative branches leading to a mean field control game (MFCG) model, and, second, we use a *three-timescales reinforcement learning algorithm* to solve this problem when the structures of the dynamics and of the cost are unknown to the agents. This represents a natural and interesting development of the *two-timescales reinforcement learning algorithm* introduced in [5] to solve MFG or mean field control (MFC) problems. The following model of intra-and-inter-bank borrowing and lending provides a benchmark for our algorithm, which can be applied to a wide range of models.

2.1 System of M Banks with N Branches

In the model considered below, we consider $M \in \mathbb{N}$ bank groups. Each bank has $N \in \mathbb{N}$ local branches and is involved in both intra-and-inter bank borrowing and lending activity. Let tuple (m, n) for $m \in \{1, \dots, M\}$ and $n \in \{1, \dots, N\}$ index the n-th branch of the m-th bank. The one-dimensional diffusion process $(X_t^{m,n})_{t\in[0,\infty)}$ stands for the *log-monetary reserve* of the branch (m, n) over an infinite time horizon, whose dynamics has the following form:

$$dX_t^{m,n} = \left[\kappa\left(\frac{1}{N}\sum_{j=1}^{N} X_t^{m,j} - X_t^{m,n}\right) + \alpha_t^m(X_t^{m,n})\right]dt + \sigma dW_t^{m,n}, \tag{1}$$

with $X_0^{m,n} \sim \mu_0$. The first term in the drift

$$\kappa\left(\frac{1}{N}\sum_{j=1}^{N} X_t^{m,j} - X_t^{m,n}\right) = \frac{\kappa}{N}\sum_{j=1}^{N}\left(X_t^{m,j} - X_t^{m,n}\right), \quad \kappa \geq 0,$$

represents borrowing and lending activity between branch (m, n) with the other branches of the same bank. As can be seen on the right-hand side, branches with more liquidity lend to branches with less liquidity at a rate κ, normalized by the number N of branches. The left-hand side can be interpreted as mean-reversion to the average liquidity reserve of the branches of that bank. To some degree (depending on κ) this mean reversion will be facilitated by the branches at no cost because branches that are well equipped with liquidy have an interest in investing their excess liquidity and branches with too little liquidity have an interest in borrowing. In addition to this mean-reversion behavior, local branch (m, n) has the possibility to borrow and lend from a central bank. This borrowing happens at a rate that depends on the liquidity reserve of (m, n) but needs to comply with the (time-dependent) feedback-form policy of bank group m, which is reflected in the *control* term $\alpha^m : \mathbb{R} \to \mathbb{R}$. The entire system is driven by $M \times N$ independent standard Brownian motions $(W_t^{m,n})_{t,m,n}$. For simplicity, we assume the same constant diffusion rate $\sigma > 0$. Bank group m designs its policy of control α_t^m of the borrowing and lending rate for all of its branches at time t in order to minimize the *group objective function*

$$J(\alpha^m; \alpha^{-m}) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left\{\int_0^\infty e^{-\beta t} f\left(X_t^{m,n}, \alpha_t^m(X_t^{m,n}), \mu_t, \widetilde{\mu}_t^m\right) dt\right\}, \tag{2}$$

where $\beta > 0$ denotes the time discount rate, and the interaction with the other banks is through the *global empirical distribution* $\mu_t = \frac{1}{MN}\sum_{i=}^{M}\sum_{j=1}^{N}\delta_{X_t^{i,j}}$ of reserves of the entire system across all branches and all banks, while the interaction within the branches of bank m is through the *local empirical distribution* $\widetilde{\mu}_t^m = \frac{1}{N}\sum_{j=1}^{N}\delta_{X_t^{m,j}}$. We denote by α^{-m} the control profile for all bank groups except m, i.e., $\alpha^{-m} = (\alpha^1, \dots, \alpha^{m-1}, \alpha^{m+1}, \dots, \alpha^M)$. Here, we consider a quadratic *running cost* function given by

$$f(x, \alpha, \mu, \widetilde{\mu}) = \frac{1}{2}\alpha^2 + c_1(x - c_2\overline{\mu})^2 + \widetilde{c}_1(x - \widetilde{c}_2\overline{\widetilde{\mu}})^2 + \widetilde{c}_3(\overline{\widetilde{\mu}} - \widetilde{c})^2, \tag{3}$$

which depends on the global and local empirical distributions $\mu, \widetilde{\mu} \in \mathcal{P}(\mathbb{R})$ only through their first moments, denoted respectively by $\overline{\mu}, \overline{\widetilde{\mu}}$. So in (2), the cost at time t depends only on the global and local empirical means $\overline{\mu}_t = \frac{1}{MN}\sum_{i=}^{M}\sum_{j=1}^{N} X_t^{i,j}$ and $\overline{\widetilde{\mu}}_t^m = \frac{1}{N}\sum_{j=1}^{N} X_t^{m,j}$. Here, $c_1, c_2, \widetilde{c}_1, \widetilde{c}_2, \widetilde{c}_3, \widetilde{c} \in \mathbb{R}$ are some constants. The running cost is interpreted as follows. The first term represents the quadratic cost of control on borrowing and lending rate. The second and third term shows the bank's intention to keep the reserve of its branch close to both global average reserve $\overline{\mu}$ and local average reserve $\overline{\widetilde{\mu}}$ to some extend quantified by $c_1, c_2, \widetilde{c}_1, \widetilde{c}_2$. Meanwhile, the bank prefers its local average centering around a target level $\widetilde{c}$.

The above system of M banks constitutes a competitive game between the M banks, while it is a collaborative (distributed) game within each bank group. We are looking for a closed-loop Nash

equilibrium between the banks. This kind of mixed competitive-collaborative game is described in [3] in the context of finite horizon extended games applied to the liquidation trader's problem. Mathematically, the problem is defined as follows: find a control profile $(\hat{\alpha}^m)_m$ such that for every $m = 1, \dots, M$, $\hat{\alpha}^m$ minimizes:

$$\alpha^m \mapsto J(\alpha^m; \hat{\alpha}^{-m}).$$

We are interested in the *mean field control game limit* when both M and N go to infinity and its solution using reinforcement learning as presented in Section 3.

2.2 Mean Field Control Game Limit

Associated to the finite-player game introduced above, we associate the MFCG obtained in the asymptotic limit where both M and N go to ∞. The problem is to find a pair $(\hat{\alpha}, \hat{\mu})$ such that the following two conditions are satisfied:

(1) A representative bank confronted with a fixed flow of probability distributions $\hat{\mu} := (\hat{\mu}_t)$ solves the McKean-Vlasov (MKV) control problem of finding a minimizer $\hat{\alpha}$ for

$$\alpha \mapsto J(\alpha; \hat{\mu}) = \mathbb{E}\left\{\int_0^\infty e^{-\beta t} f\left(X_t, \alpha_t(X_t), \hat{\mu}_t, \mathcal{L}(X_t)\right) dt\right\}, \quad (4)$$

subject to

$$dX_t = \left[\kappa(\mathbb{E}(X_t) - X_t) + \alpha_t(X_t)\right] dt + \sigma dW_t, \quad X_0 \sim \mu_0. \quad (5)$$

(2) The law of the state X_t controlled by $\hat{\alpha}$ satisfies the fixed point condition

$$\mathcal{L}(X_t) = \hat{\mu}_t, \quad t \in [0, \infty). \quad (6)$$

The justification of such a limit is treated mathematically in the forthcoming paper [2]. See also the Appendix in [3] for a formal justification in the case of the linear-quadratic trader's liquidation problem.

2.2.1 Value Function and HJB Equation. Since we are looking for an equilibrium among Markovian feedback strategies, we solve the MFCG system (4)-(6) through the Hamilton-Jacobi-Bellman (HJB) equation approach. Following the computation detailed in the Appendix A of [5], we first solve the finite horizon problem with zero terminal condition when the global distribution flow is given by $(\mu_t)_{t\in[0,T]}$:

$$V^T(t,x) = \inf_\alpha \mathbb{E}\left\{\int_t^T e^{-\beta s} f\left(X_s, \alpha_s(X_s), \mu_s, \mathcal{L}(X_s)\right) ds\right\}, \quad (7)$$

subject to:

$$dX_s = \left[\kappa(\mathbb{E}(X_s) - X_s) + \alpha_s(X_s)\right] ds + \sigma dW_s, \quad X_t = x$$

and with the fixed point condition (6) over $[t, T]$. Denoting by $\mathcal{A}$ the infinitesimal generator of X, the Hamiltonian is given by

$$H(t, x, \hat{\alpha}(t,x), \mu_t, \widetilde{\mu}_t) = \inf_\alpha \left\{\mathcal{A}V^T(t,x) + f(x, \alpha, \mu_t, \widetilde{\mu}_t)\right\}, \quad (8)$$

which attains its minimum at $\hat{\alpha}(t,x) = -\partial_x V^T(t,x)$ in our case where f is given by (3), and the dynamics of X by (5). The HJB equation with MKV dynamic reads (see e.g. [6, Section 4.1])

$$\partial_t V^T(t,x) - \beta V^T(t,x) + H(t, x, \hat{\alpha}(t,x), \mu_t, \widetilde{\mu}_t) + \int_{\mathbb{R}} \frac{\partial H}{\partial \widetilde{\mu}_t}(t, \xi, \hat{\alpha}(t,\xi), \mu_t, \widetilde{\mu}_t)(x) d\widetilde{\mu}_t(\xi) = 0, \quad (9)$$

with $V^T(T,x) = 0$. We compute

$$H(t, x, \hat{\alpha}(t,x), \mu_t, \widetilde{\mu}_t) = -\frac{1}{2}(\partial_x V^T(t,x))^2 + \frac{1}{2}\sigma^2 \partial_{xx} V^T(t,x) + \kappa(\overline{\widetilde{\mu}}_t - x) V_x^T(t,x) + c_1(x - c_2\overline{\mu}_t)^2 + \widetilde{c}_1(x - \widetilde{c}_2\overline{\widetilde{\mu}}_t)^2 + \widetilde{c}_3(\overline{\widetilde{\mu}}_t - \widetilde{c})^2,$$

and

$$\int_{\mathbb{R}} \frac{\partial H}{\partial \widetilde{\mu}_t}(t, \xi, \hat{\alpha}(t,\xi), \mu_t, \widetilde{\mu}_t)(x) d\widetilde{\mu}_t(\xi) = -2\widetilde{c}_1\widetilde{c}_2(1 - \widetilde{c}_2)\overline{\widetilde{\mu}}_t x + 2\widetilde{c}_3(\overline{\widetilde{\mu}}_t - \widetilde{c})x.$$

We then formulate the following ansatz for the value function

$$V^T(t,x) = \Gamma_2^T(t)x^2 + \Gamma_1^T(t)x + \Gamma_0^T(t), \quad (10)$$

with the zero terminal conditions $\Gamma_2^T(T) = \Gamma_1^T(T) = \Gamma_0^T(T) = 0$. We have $\hat{\alpha}(t,x) = -2\Gamma_2^T(t)x - \Gamma_1^T(t)$. Plugging the ansatz and its partial derivatives into (9) and identifying the coefficients of powers of x leads to a system of ODEs for $\Gamma_1^T, \Gamma_2^T, \Gamma_0^T$ with zero terminal conditions. This system is complemented with the forward equation

$$d\overline{\mu}_t = \mathbb{E}\left(\hat{\alpha}(t, X_t)\right) dt = -\left[2\Gamma_2^T(t)\overline{\mu}_t + \Gamma_1^T(t)\right] dt, \quad \overline{\mu}_0 = x, \quad (11)$$

obtained by taking expectation in (5) and using the expression of the control $\hat{\alpha}$. The ODE system for $(\Gamma_2^T(t), \Gamma_1^T(t), \Gamma_0^T(t), \overline{\mu}_t)_{t\in[0,T]}$ is a two-point boundary value problem which can be solved explicitly as in in the Appendix A of [5].

2.2.2 Explicit Formulas. The solution to our infinite horizon problem is obtained by taking the limit $T \to \infty$. Furthermore, since we are interested in the *asymptotic solution*, or equivalently the *stationary solution*, we take the limit $t \to \infty$ to obtain that the limiting value function

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

where $\Gamma_0, \Gamma_1, \Gamma_2$ are constants, must satisfy (9) with $\partial_t V^T = 0$, no terminal condition at $T = +\infty$, and $\overline{\widetilde{\mu}}_t = \overline{\widetilde{\mu}}$ being the stationary point of (11) satisfying $2\Gamma_2\overline{\widetilde{\mu}} + \Gamma_1 = 0$. We deduce the formulas:

$$\begin{aligned}
\hat{\alpha}(x) &= -2\Gamma_2 x - \Gamma_1, \\
\Gamma_2 &= \frac{-(\beta + 2\kappa) + \sqrt{(\beta + 2\kappa)^2 + 8(c_1 + \widetilde{c}_1)}}{4}, \\
\Gamma_1 &= \frac{2\widetilde{c}_3(\overline{\mu} - \widetilde{c}) - 2\widetilde{c}_1\widetilde{c}_2(2 - \widetilde{c}_2)\overline{\mu} - 2c_1c_2\overline{\mu}}{\beta + \kappa + 2\Gamma_2}, \\
\Gamma_0 &= \frac{-\kappa\overline{\mu} - \frac{1}{2}\Gamma_1^2 + \sigma^2\Gamma_2 + c_1c_2^2\overline{\mu} + \widetilde{c}_1\widetilde{c}_2^2\overline{\mu} + \widetilde{c}_3(\overline{\mu} - \widetilde{c})^2}{\beta}, \\
\overline{\mu} &= -\frac{\Gamma_1}{2\Gamma_2} = \frac{\widetilde{c}_3\widetilde{c}}{c_1(1 - c_2) + \widetilde{c}_1(1 - \widetilde{c}_2)^2 + \widetilde{c}_3 - \kappa\Gamma_2}.
\end{aligned}$$

Note that at Nash equilibrium and asymptotically when time is large, X_t behaves like an Ornstein–Uhlenbeck process with a rate of mean-reversion $\kappa + 2\Gamma_2$ around the center $\overline{\mu}$ and diffusion σ^2. Consequently, the equilibrium asymptotic distribution is Gaussian with $\mu = \mathcal{N}\left(\overline{\mu}, \frac{\sigma^2}{2\kappa + 4\Gamma_2}\right)$. In other words, each bank is trying to anchor the log-monetary reserve of its branches to the average level $\overline{\mu}$ of the entire bank system. If a branch has sufficient reserve above the average, it will lend money to other branches at a rate which is linear to its current reserve until the average level is reached, and

vice versa. Corresponding to such mean-reversion dynamic, the reserve over all branches is normally distributed around the mean reserve and its diffusion depends on the mean-reversion speed.

3 THREE-TIMESCALE Q-LEARNING ALGORITHM

3.1 Discrete time formulation and Q-learning

We now describe our algorithm to learn the solution to the mixed Control Game problem (4)-(6). Since the algorithm itself is only a minor modification of the algorithm used in [3], we keep this paragraph brief. The algorithm rests on the concept of Q learning, a well established method to solve Markov Decision problems. We first discretize the time interval $[0, \infty]$ into an equally spaced grid $0 = t_0 < t_1 < \dots$ and assume for notational simplicity that $t_i = i$. We then recast the problem (4)-(6) into a discrete time mean field control game problem given by:

(1) Given $\{\mu_n\}_{n\in\mathbb{N}}$, find a minimizer $\hat{\alpha}$ for

$$J(\alpha;\mu) = \mathbb{E}\left[\sum_{n=0}^{\infty} e^{-\beta n} f\left(X_n, \alpha_n(X_n), \mu_n, \mathcal{L}(X_n)\right)\right], \tag{12}$$

subject to

$$\mathbb{P}\big(X_{n+1} = x' | X_n = x, \alpha_n(X_n) = a, \mu_n = \mu, \mathcal{L}(X_n) = \widetilde{\mu}\big) = p(x'|x, a, \mu, \widetilde{\mu}),$$

where the transition kernel $p : \mathcal{X}\times\mathcal{A}\times\Delta^{|\mathcal{X}|}\times\Delta^{|\mathcal{X}|} \to \Delta^{|\mathcal{X}|}$ arises from a discrete counterpart to (5).

(2) The law of the state X_n matches the fixed point condition

$$\mathcal{L}(X_n) = \mu_n, \quad n \in \mathbb{N}. \tag{13}$$

In order to solve this discrete time problem we discretize the state space and action space into $\mathcal{X} = \{x_0, \dots, x_{|\mathcal{X}|-1}\}$, and $\mathcal{A} = \{a_0, \dots, a_{|\mathcal{A}|-1}\}$ respectively.

Our reinforcement learning algorithm to solve the discrete time and discrete state problem (12) and (13) follows [3]. The algorithm is based on well established ideas from Q-learning. The algorithm is model agnostic which means that no information is needed about the model that generates the data. In the control part of our problem (12), the local distribution $\mathcal{L}(X_n)$ depends on the control that is chosen. For this reason it can not simply be treated as an additional parameter but the Q-learning has to be adapted slightly. For an admissible control $\alpha : \mathcal{X} \to \mathcal{A}$, we define the new control $\alpha_{x,a}$ that deviates from α only at the state x where it takes the value a:

$$\alpha_{x,a}(x') = \begin{cases} a & \text{if } x' = x \\ \alpha(x) & \text{otherwise.} \end{cases} \tag{14}$$

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

$$\begin{aligned} Q_\mu^\alpha(x,a) &= f(x,a,\mu,\mu^{\alpha_{x,a}}) \\ &+ \mathbb{E}\left[\sum_{n=1}^{\infty} e^{-\beta n} f(X_n, \alpha(X_n), \mu, \mu^\alpha) | X_0 = x, A_0 = a\right]. \end{aligned}$$

One can then consider the optimal cost function

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

which, conditioned on being in state x and choosing action a at time 0, minimizes the cost over all strategies α chosen in all steps to follow. From the function Q_μ^* one obtains the optimal control $\alpha^*(x) = \arg\min_a Q_\mu^*(x,a)$. In Section 3.3 we will see that actually a randomized counterpart of α^* should be chosen to ensure a wide enough exploration range of the possible actions. We stress that the minimizing strategy usually depends on the global measure μ. For fixed μ, it follows from [4], as the measure μ is fixed and does not depend on α, that the function Q_μ^* follows a Bellman equation given by:

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

The measure $\mu_{x,a}^* = \lim_{n\to\infty} \mathcal{L}(X_n^{\alpha_{x,a}^*,\mu})$ corresponds to the strategy $\alpha_{x,a}^*$ as derived from α^* by changing the action in state x to a, see (14).

3.2 Three-Timescale Updating Rates

Our algorithm to approximate the Q-function, optimal policy and the equilibrium distribution mimics the idea of nested optimization. For a given global distribution, the Q-function that describes the optimal action has to be found, and based on this, the local distribution. This idea of nested simulation leads to a Three-Timescale approach which is sketched in the following. With updating rates ρ_k^μ for the global distribution, ρ_k^Q for the Q table, and $\rho_k^{\mu^\alpha}$ for the local distribution, where we assume $\rho_k^\mu < \rho_k^Q < \rho_k^{\mu^\alpha}$, the updates that can be derived from the Bellman equation are described by

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

with

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

Note that in our model agnostic approach, the transition probabilities p need to be estimated from the data. As samples from the state and the rewards are obtained incrementally, we update these estimates with Robbins–Monro rates. We refer the reader to [3] for more details. Note that in the example we introduced in Section 2.1, a representative agent interacts with both the local distribution and the global distribution. For this reason, the three timescale algorithm is more natural than a two timescale algorithm, since it allows the algorithm to treat differently the two distributions on top of the Q-function.

3.3 Action Exploration

An efficient algorithm is designed to well balance the tendencies between exploring a range of policies and staying in the current

best choice, i.e. exploration and exploitation. An over-exploring algorithm is less likely to converge to the optimal policy while the over-exploiting one will possibly be stuck in a local optimal, which is the well known exploration-exploitation dilemma [18]. As other reinforcement learning algorithms, our three-time scale Q-learning algorithm is confronted with this dilemma. Therefore, we shall develop methods to balance the exploration-exploitation trade-off.

Over the recent decades, various action exploration techniques have been developed to overcome the exploration and exploitation dilemma. Those can roughly be distinguished into two categories: *undirected* and *directed* [24]. Undirected exploration takes actions based on some probability distribution and does not account for the learning progress itself. Widely applied undirected methods include ϵ-greedy, Boltzmann, and Max-Boltzmann [26]. In contrast, directed exploration adapts the action preference by the learning progress, such as the number of times of a state-action pair being visited (counter-based), the environment with large errors from previous exploration (error-based), states not being visited recently (recency-based).

Depending on specific learning tasks, sophisticated directed exploration might require more efforts to calibrate but does not necessarily outperform simple undirected heuristics [20, 25]. Therefore, for the new three-timescale algorithm that has not been comprehensively tested, we shall first focus on the undirected methods, with preference for its generality and simplicity. It can then serve as a benchmark for the application of more complicated directed exploration methods. In particular we consider the following three undirected exploration methods:

(1) *ϵ-greedy.*

$$\pi_t^{\epsilon}(x) = \begin{cases} a \sim \text{Unif}(\mathcal{A}), & \text{w.p. } \epsilon, \\ \arg\max_{a\in\mathcal{A}} Q_t(x,a), & \text{w.p. } 1-\epsilon. \end{cases} \tag{15}$$

Parameter ϵ is the *exploration rate.*

(2) *Boltzmann exploration.*

$$\pi_t^{Boltz}(x,a) \sim Boltz(Q_t(x,a); Q_t(x,\cdot), \tau) \tag{16}$$

with $Boltz(x; X, \tau) := \frac{e^{-x/\tau}}{\sum_{x'\in X} e^{-x/\tau}}$ known as the Boltzmann distribution. Parameter τ is referred as the *temperature.*

(3) *Max-Boltzmann* combines the ϵ-greedy with Boltzmann exploration by replacing Unif($\mathcal{A}$) in (15) by *Boltz* distribution in (16),

where the exploration propensity of the algorithm is controlled by the exploration rate ϵ or constant temperature τ. To search for the appropriate exploration heuristic, for each of the three heuristics, we consider the following three configurations: (1) constant rate; (2) linearly decaying rate w.r.t episode; and (3) exponentially decaying rate w.r.t episode, which will be specified in Section 4.

3.4 Algorithm

The Algorithm 1 applied to learn the asymptotic solution discussed in section 2.2.2 is the three-timescale mean field Q-learning algorithm (U3-MF-QL) presented in [3]. By interacting with the environment in a trial and error fashion, we are able to learn the optimal Q table, together with the local and global distribution at equilibrium.

As discussed in the previous section, the learning rates assume a core role and they are defined as

$$\rho^{Q}_{x,a,n,k} := \frac{1}{(1+\#|(x,a,k,n)|)^{\omega^Q}}, \quad \rho^{\nu}_{k} := \frac{1}{(1+k)^{\omega^\nu}},$$

where ν is replaced by μ and $\widetilde{\mu}$ for the local and global distribution respectively, and $\#|(x,a,k,n)|$ counts the visits of the pair (x,a) up to the episode k and time n. This form of learning rates is inspired by the ones used for Q-learning with provable convergence guarantees [14]. The triplet $(\omega^Q, \omega^\mu, \omega^{\widetilde{\mu}})$ should be chosen such that $\omega^\mu > \omega^Q > \omega^{\widetilde{\mu}}$, so that $\rho_k^\nu < \rho_k^Q < \rho_k^{\widetilde{\nu}}$, and it should satisfy $\omega^Q \in (0.5, 1)$.

Algorithm 1 Three-Timescales Mean Field Q-Learning - Infinite Horizon

Require:
T: number of time steps in a learning episode,
Truncated state space: $\mathcal{X} = \{x_0, \dots, x_{|\mathcal{X}|-1}\}$,
Truncated action space: $\mathcal{A} = \{a_0, \dots, a_{|\mathcal{A}|-1}\}$,
Initial distribution of the representative player: μ_0,
Exploration rule s.t. $\pi^v \in \Delta^{|\mathcal{A}|}$ for any $|\mathcal{A}|-$dim vector v,
Break rule tolerances: tol_Q, tol_μ, $tol_{\widetilde{\mu}}$.
Initialization:
$Q^0(x,a) = 0$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$,
$\mu_n^0 = \frac{1}{|\mathcal{X}|} J_{|\mathcal{X}|}$ and $\widetilde{\mu}_n^0 = \frac{1}{|\mathcal{X}|} J_{|\mathcal{X}|}$ for $n = 0, \dots, T$,
where J_m is an m-dimensional unit vector.
for each episode $k = 1, 2, \dots$ **do**
  Set $Q^k \equiv Q^{k-1}$
  Observe initial state: $X_0^k \sim \mu_T^{k-1}$.
  for $n = 0, \dots, T$ **do**
    Choose action:
    choose A_n^k using the exploration policy $\pi^{Q^k(X_n^k,\cdot)}$.
    Update distributions:
    $\mu_n^k = \mu_n^{k-1} + \rho_k^{\mu}(\boldsymbol{\delta}(X_n^k) - \mu_n^{k-1})$,
    $\widetilde{\mu}_n^k = \widetilde{\mu}_n^{k-1} + \rho_k^{\widetilde{\mu}}(\boldsymbol{\delta}(X_n^k) - \widetilde{\mu}_n^{k-1})$,
    where $\boldsymbol{\delta}(X_n^k) = \left(\mathbf{1}_x(X_n^k)\right)_{x\in\mathcal{X}}$.
    Observe next state:
    observe X_{n+1}^k from the environment.
    Observe cost:
    observe $f_n = f(X_n^k, A_n^k, \mu_n^k, \widetilde{\mu}_n^k)$.
    Update Q table:
    $Q^k(x,a) = Q^k(x,a) + \mathbf{1}_{x,a}(X_n^k, A_n^k)\rho^Q_{x,a,n,k}$
    $(f_n + \beta \min_{a'\in\mathcal{A}} Q^k(X_{n+1}^k, a') - Q^k(x,a))$,
    where β is the discount parameter.
  end for
  if $\begin{cases}\delta(\mu^k, \mu^{k-1}) \le tol_\mu, \\ \delta(\widetilde{\mu}^k, \widetilde{\mu}^{k-1}) \le tol_{\widetilde{\mu}}, \\ \|Q^k - Q^{k-1}\|_{1,1} \le tol_Q, \end{cases}$ **then** break
  end if
end for

4 NUMERICAL RESULTS

For the MFCG problem setting, we choose the following parameters: $(c_1, c_2, \widetilde{c}_1, \widetilde{c}_2, \widetilde{c}_3, \widetilde{c}) = (1.5, 0.75, 2.5, 0.5, 4, 2)$ and discount rate $\beta = 1$ for the running cost f; $(\kappa, \sigma) = (1, 2)$ for the dynamic of state dX. We truncate the infinite time horizon by $[0, T]$ with $T = 20$ and discretize it by steps of size $\delta t = 1/16$. The state and action spaces are trimmed into $\mathcal{X} = \{x_0 = -1.5, x_1 = -1.5 + \delta x, \ldots, x_{|\mathcal{X}|-1} = 4.5\}$ and $\mathcal{A} = \{a_0 = -6, a_1 = -6 + \delta a, \ldots, a_{|\mathcal{A}|-1} = 6\}$ by $\delta x = \delta a = \sqrt{\delta t} = 1/4$. For the reinforcement learning setup, we take $K = 50,000$ episodes and consider the specifications for the action exploration in Table 1. The initial exploration rate is set small for the constant ϵ-greedy action explorer, mildly greater for the linearly decaying rate, and large for the exponentially decaying rate. The initial temperature for Boltzmann explorers are the same. The Max-Boltzmann explorers takes in a constant exploration rate combined with the Boltzmann explorers.

Table 1: Action Exploration Heuristics

ϵ-greedy	$\epsilon(k)$	Boltzmann	$\tau(k)$
ϵ_{Con}	0.01	$Boltz_{Con}$	5
ϵ_{Lin}	$0.05(K-k)/K$	$Boltz_{Lin}$	$5(K-k)/K$
ϵ_{Exp}	0.9995^k	$Boltz_{Exp}$	5×0.9999^k
Max-Boltz	$(\epsilon, \tau(k))$		
MB_{Con}	$(0.05, 5)$		
MB_{Lin}	$(0.05, 5(K-k)/K)$		
MB_{Exp}	$(0.05, 5 \times 0.9999^k)$		

Algorithm 1 learns the solution of the mean field control game based on three different learning rates for the Q-table and local/global distributions. Figure 1 shows the results obtained when the learning rate parameters $(\omega_\mu, \omega_Q, \omega_{\widetilde{\mu}})$ are equal to $(0.75, 0.55, 0.15)$. The x-axis represents the state variable x while the left, right y- axes correspond to the action $\alpha(x)$ and the probability mass $\hat{\mu}(x)$ respectively. The green dot-marked line and continuous curves show the theoretical solutions of the MFCG discussed in section 2.2.2 in terms of the control function and the asymptotic distribution at equilibrium. The blue dots and curve are the corresponding action and distribution learned by the algorithm, averaged over the last 5k episodes. Only the global distribution is plotted because the local distribution perfectly aligns with it. From Figure 1 we clearly observe the linear pattern of optimal control and the normal distribution of state, which were discussed in Section 2.2.2.

Figure 2 shows how different choices of the learning rate parameters let the algorithm converge to different solutions. The green set of line and curve refers to the same theoretical solution to the MFCG problem as in Figure 1. The violet (resp. orange) set shown is obtained when $\omega_\mu = \omega_{\widetilde{\mu}}$ and their values are set to 0.75 (resp. 0.15) such that $\rho^\mu < \rho^Q$ (resp. $\rho^\mu > \rho^Q$). The values of actions and distributions plotted are the average of the last 5k episodes. These choices reduce the algorithm to the two-timescale approach discussed in [5]. The algorithm then converges to the corresponding MFG and MFC versions of our model depending on the choice of the learning rates, where the support of the MFG deviates from the current trimmed state space.

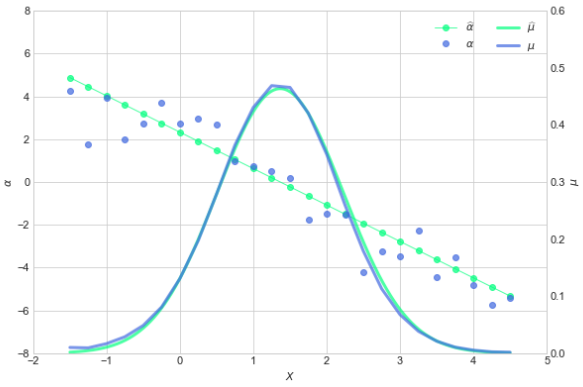

Figure 1: MFCG three-timescale Q-learning result

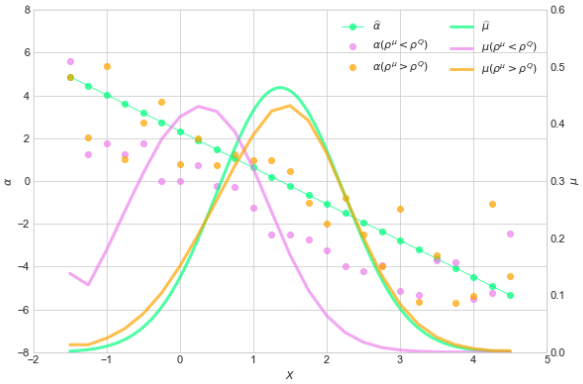

Figure 2: Two-timescale Q-learning result

The convergence of the algorithm 1 is analyzed in terms of the evolution of the estimations of the optimal Q table and the local/global distributions at equilibrium w.r.t. the learning episodes. The changes are evaluated through the total variation and the 1, 1-norm as follows

$$\delta(\nu^k, \nu^k) = \sum_{x_i \in \mathcal{X}} \left| \nu^k(x_i) - \nu^{k-1}(x_i) \right|,$$

$$\|Q^k - Q^{k-1}\|_{1,1} = \sum_{i,j} \left| Q^k_{i,j} - Q^{k-1}_{i,j} \right|,$$

where the episode is tracked by the index k and ν is replaced by μ and $\widetilde{\mu}$. Figure 3 shows how the convergence improves w.r.t. the number of episodes. The x-axis represents the learning episode k. The y-axis represents the value of the 1, 1-norm and the total variation respectively with the averaged values over 10 runs (solid line) and standard deviations (shaded region).

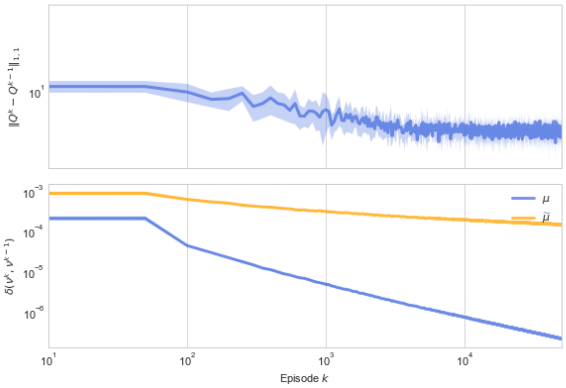

Figure 3: Total variation of Q, μ, and $\widetilde{\mu}$

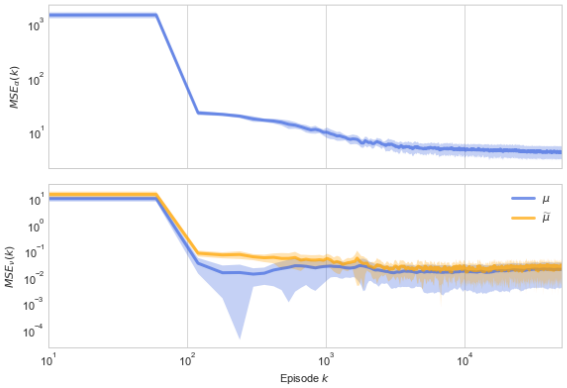

Figure 4: Mean squared error of α, μ, and $\widetilde{\mu}$

The optimal control function learned by the algorithm is evaluated w.r.t. the limiting distribution of the population at the equilibrium. In particular, we analyze the mean square error averaged over multiple runs as follows

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

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

where $\hat{\mu}(x_j) = \int_{x_{j-1}}^{x_j} d\mu(x)$ is obtained by the asymptotic distribution at equilibrium μ using the convention $x_{-1} = -\infty$. Similarly, we evaluate the learning of the first moment of the asymptotic distribution at equilibrium as

$$\mathrm{MSE}_{\overline{\mu}}(k) = \frac{1}{\#runs} \sum_{i=0}^{\#runs} (\overline{\mu}_T^{i,k} - \overline{\hat{\mu}})^2,$$

$$\mathrm{MSE}_{\overline{\widetilde{\mu}}}(k) = \frac{1}{\#runs} \sum_{i=0}^{\#runs} (\overline{\widetilde{\mu}}_T^{i,k} - \overline{\hat{\mu}})^2.$$

Figure 4 shows the decrease of the errors w.r.t. the number of learning episodes. The x-axis corresponds to the learning episode k. The y-axis represents the errors averaged over 10 runs (solid line) and their standard deviations (shaded region). As shown in the above results, the three timescale algorithm learns a solution that approximately matches the analytical solution in terms of control, Q-function and distribution. So if the agents use the policy obtained by RL algorithms, we can expect them to have a behavior close to the Nash equilibrium one.

We conclude by presenting an empirical comparison of the action exploration strategies discussed in section 3.3. Figures 5 and 6 show the results obtained by applying the ϵ-greedy (red set of lines), Boltzman (green set of lines), and Max-Boltzman (purple set of lines) exploration rules when the rate is constant, linear, or exponential decaying w.r.t. the episodes, as in Table 1. The subplot on top of Figure 5 shows the 1, 1-norm of the learned Q table and the two subplots following are total variations of learned μ and $\widetilde{\mu}$ distribution. The x-axis is the log-scaled episode number, and the y-axes correspond to the value of those total variations. The ϵ-greedy with constant exploration rate ϵ_{Con} surprisingly outperforms any other heuristic in converging speed. The worst result is obtained by the Boltzmann exploration group. B_{Con} fails to converge in Q table, while B_{Lin} and B_{Exp} waste almost 10k episodes before the variation is reduced. Obviously, the Boltzmann exploration set is under-tuned with a high initial temperature, and reducing it will hopefully improve its performance. Recall that we aim to control the exploration propensity via the probability distribution, however, the Boltzmann distribution (16) depends on the value of the Q table whose scale is previously unknown. Thus, it requires extra investigation to figure out both the temperature range and the decaying rate. The Max-Boltzmann exploration set performs mediocrely, which is due to its under-tuned Boltzmann component. On the contrary, in Figure 6, we observe that most of the heuristics result in lower mean squared error on α than ϵ_{Con}, except B_{Con}, B_{Exp}, and ϵ_{Exp}. Despite that the ϵ_{Con} still achieves the lowest mean squared error, this result indicates that the Boltzmann and Max-Boltzmann explorations could possibly lead to better results if well-tuned. Therefore, in the linear-quadratic bank borrowing and lending MFCG, the naive ϵ-greedy heuristic handles the learning task well and can serve as a useful benchmark for developing more sophisticated exploration strategies.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the anonymous referees, whose feedback helped to improve the paper.
A. Angiuli's work presented here does not relate to his position at Amazon.
J.-P. Fouque's research is supported by NSF grants DMS-1814091 and DMS-1953035.
J. Lin's work makes use of computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 1720256) at UC Santa Barbara.

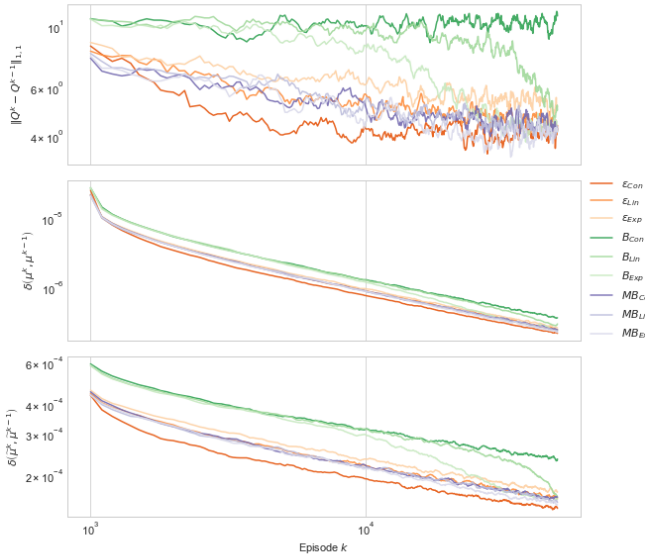

Figure 5: Comparison on total variation of Q, μ, and $\widetilde{\mu}$

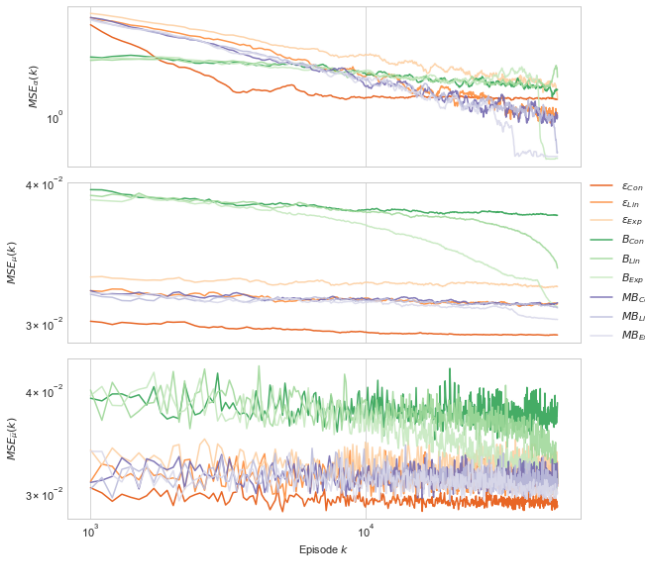

Figure 6: Comparison on mean squared error of α, μ, and $\widetilde{\mu}$

REFERENCES

[1] Clémence Alasseur, Imen Ben Taher, and Anis Matoussi. 2020. An extended mean field game for storage in smart grids. *Journal of Optimization Theory and Applications* 184, 2 (2020), 644–670.

[2] Andrea Angiuli, Nils Detering, Jean-Pierre Fouque, Mathieu Laurière, and Jimin Lin. 2022. Mean Field Control Games. *In preparation* (2022).

[3] Andrea Angiuli, Nils Detering, Jean-Pierre Fouque, and Jimin Lin. 2022. Reinforcement Learning Algorithm for Mixed Mean Field Control Games. *arXiv preprint arXiv:2205.02330* (2022).

[4] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. 2022. Reinforcement Learning for Mean Field Games, with Applications to Economics. *To appear in Machine Learning in Financial Markets: A guide to contemporary practices. arXiv:2106.13755* (2022).

[5] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. 2022. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems* (2022), 1–55.

[6] Alain Bensoussan, Jens Frehse, Phillip Yam, et al. 2013. *Mean field games and mean field type control theory*. Vol. 101. Springer.

[7] Pierre Cardaliaguet and Charles-Albert Lehalle. 2018. Mean field game of controls and an application to trade crowding. *Mathematics and Financial Economics* 12, 3 (2018), 335–363.

[8] René Carmona. 2020. Applications of mean field games in financial engineering and economic theory. *arXiv preprint arXiv:2012.05237* (2020).

[9] René Carmona and François Delarue. 2018. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer.

[10] René Carmona, Jean-Pierre Fouque, and Li-Hsien Sun. 2015. Mean Field Games and Systemic Risk. *Communications in Mathematical Sciences* 13, 4 (2015), 911–933.

[11] Patrick Chan and Ronnie Sircar. 2015. Bertrand and Cournot mean field games. *Applied Mathematics & Optimization* 71, 3 (2015), 533–569.

[12] Boualem Djehiche, Alain Tcheukam, and Hamidou Tembine. 2016. Mean-field-type games in engineering. *arXiv preprint arXiv:1605.03281* (2016).

[13] Romuald Elie, Julien Perolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. 2020. On the Convergence of Model Free Learning in Mean Field Games. In *proc. of AAAI*.

[14] Eyal Even-Dar and Yishay Mansour. 2003. Learning rates for Q-learning. *Journal of machine learning Research* 5, Dec (2003), 1–25.

[15] Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. 2011. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*. Springer, 205–266.

[16] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2019. Learning mean-field games. In *Advances in Neural Information Processing Systems*. 4966–4976.

[17] Minyi Huang, Roland P Malhamé, and Peter E Caines. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems* 6, 3 (2006), 221–252.

[18] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

[19] Elias Koutsoupias and Christos Papadimitriou. 1999. Worst-case equilibria. In *Annual symposium on theoretical aspects of computer science*. Springer, 404–413.

[20] Volodymyr Kuleshov and Doina Precup. 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028* (2014).

[21] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese journal of mathematics* 2, 1 (2007), 229–260.

[22] David Mguni, Joel Jennings, and Enrique Munoz de Cote. 2018. Decentralised learning in systems with many, many strategic agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[23] Jayakumar Subramanian and Aditya Mahajan. 2019. Reinforcement Learning in Stationary Mean-field Games. In *Proceedings. 18th International Conference on Autonomous Agents and Multiagent Systems*.

[24] Sebastian B Thrun. 1992. Efficient exploration in reinforcement learning. (1992).

[25] Marco Wiering and Jürgen Schmidhuber. 1998. Efficient model-based exploration. In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, Vol. 6. MIT Press Cambridge, MA, 223–228.

[26] Marco A Wiering. 1999. *Explorations in efficient reinforcement learning*. Ph.D. Dissertation. University of Amsterdam.