Generalization in Mean Field Games by Learning Master Policies

Sarah Perrin[1, *], **Mathieu Laurière**[2, *], **Julien Pérolat**[3], **Romuald Élie**[3], **Matthieu Geist**[2, †], **Olivier Pietquin**[1, †]

[1] Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL
[2] Google Research, Brain Team
[3] DeepMind
sarah.perrin@inria.fr, lauriere@google.com, perolat@deepmind.com, relie@deepmind.com, mfgeist@google.com, pietquin@google.com

Abstract

Mean Field Games (MFGs) can potentially scale multi-agent systems to extremely large populations of agents. Yet, most of the literature assumes a single initial distribution for the agents, limiting the practical applications of MFGs. Machine Learning has the potential to solve a wider diversity of MFG problems thanks to generalization capacities. We study how to leverage these generalization properties to learn policies enabling a typical agent to behave optimally against any population distribution. In reference to the Master equation in MFGs, we coin the term "Master policies" to describe them and we prove that a single Master policy provides a Nash equilibrium, whatever the initial distribution. We propose a method to learn such Master policies. Our approach relies on three ingredients: adding the current population distribution as part of the observation, approximating Master policies with neural networks, and training via Reinforcement Learning and Fictitious Play. We illustrate not only the efficiency of the learned Master policy but also its generalization capabilities beyond the distributions used for training.

Introduction

Although learning in games has a long history (Shannon 1959; Samuel 1959), most of recent breakthroughs remain limited to a small number of players, *e.g.*, for chess (Campbell, Hoane Jr, and Hsu 2002), Go (Silver et al. 2016, 2017, 2018), poker (Brown and Sandholm 2018; Moravčík et al. 2017) or even video games such as Starcraft (Vinyals et al. 2019) with a large number of agents but only a handful of competing players. The combination of game theory with multi-agent reinforcement learning has proved to be efficient (Lanctot et al. 2017), but learning in games involving a large number of players remains very challenging. Recently, Mean Field Games (MFGs), introduced concurrently by Lasry and Lions (2007) and Huang et al. (2006), have been considered as a promising approach to address this problem. They indeed model games with an infinite number of players. Instead of taking into account interactions between individuals, MFGs model the interaction between a so-called representative agent (sampled from the population distribution) and the full population itself. As in many multi-player games, solving an MFG boils down to finding a Nash equilibrium. Intuitively, it corresponds to a situation where no player can increase their reward (or decrease their cost) by changing their strategy, given that other players keep their current behavior. MFGs are classically described with a forward-backward system of partial differential equations (PDEs) or stochastic differential equations (SDEs) and can only be solved analytically in some specific cases. When an analytical solution is not available, numerical methods such as finite differences can be called to solve the PDE system. However, these techniques do not scale well with the dimensions of the state and action spaces. Another issue with PDE methods is that they are very sensitive to initial conditions. Especially, the policy obtained is only valid for a single initial distribution μ_0 for the population over the state space. This is a strong limitation for practical applications. For example, in an evacuation or traffic-flow scenario, the solution found by a PDE solver could potentially lead to an unforeseen congestion if the agents are not initially distributed as the model expected. This could have dramatic consequences. On the other hand, solving for every possible initial distribution is of course infeasible. Following the traditional trend in the literature, even solutions to MFGs that use most recent Machine Learning methods consider that the initial distribution is fixed and thus compute policies that are agnostic to the current population. A sensible idea to alleviate the sensitivity issue is to incorporate the population as part of the observation for the representative agent, such that it can behave optimally against the population, and not only w.r.t. its current state. Yet, using such a modification of the observation cannot be done seamlessly as the uniqueness of the initial distribution is a core assumption of existing methods, including very recent ones based on Machine Learning.

Here we do a first crucial step in this direction using Deep Reinforcement Learning (Deep RL), which sounds particularly well fitted to overcome the aforementioned difficulty. Our core contribution is to propose the first Deep RL algorithm that calculates an optimal policy independently of the initial population distribution.

Main contributions. First, we extend the basic framework of MFGs by introducing a class of *population-dependent policies* enabling agents to react to any population distribution. Within this class, we identify a *Master policy* and establish its connection with standard population-agnostic policies arising in MFG Nash equilibria (Thm. 1).

Second, we propose an algorithm, based on Fictitious Play and Deep RL, to learn a Master policy. We analyze a continuous time version of Fictitious Play and prove convergence at a linear rate (Thm. 2). Last, we provide empirical evidence that not only this method learns the Master policy on a training set of distributions, but that the learned policy *generalizes* to unseen distributions. Our approach is the first to tackle this question in the literature on MFGs.

Background and Related Works

We consider a finite state space X and finite action space A. The set of probability distributions on X and A are denoted by Δ_X and Δ_A. Let $p : X \times A \times \Delta_X \to \Delta_X$ be a transition probability function and $r : X \times A \times \Delta_X \to \mathbb{R}$ be a reward function. Let $\gamma \in (0, 1)$ be a discount parameter. In this section, we introduce the key concepts needed to explain our main contributions. Although there is no prior work tackling explicitly the question of generalization in MFG, we review along the way several related studies.

Mean Field Games

In the usual MFG setup (Lasry and Lions 2007; Huang et al. 2006), a stationary policy is a function $\pi : X \to \Delta_A$ and a non-stationary policy $\boldsymbol{\pi}$ is an infinite sequence of stationary policies. Let Π and $\mathbf{\Pi} = \Pi^{\mathbb{N}}$ be the sets of stationary and non-stationary policies respectively. Unless otherwise specified, by policy we mean a non-stationary policy. A mean-field (MF) state is a $\mu \in \Delta_X$. It represents the state of the population at one time step. An MF flow $\boldsymbol{\mu}$ is an infinite sequence of MF states. We denote by $M = \Delta_X$ and $\mathbf{M} = M^{\mathbb{N}}$ the sets of MF states and MF flows. For $\mu \in M$, $\pi \in \Pi$, let

$$\phi(\mu, \pi) : x \mapsto \sum_{x' \in X} \sum_{a \in A} p(x|x', a, \mu)\pi(a|x')\mu(x')$$

denote the next MF state. The MF flow starting from μ_0 and controlled by $\boldsymbol{\pi} \in \mathbf{\Pi}$ is denoted by $\Phi(\mu_0, \boldsymbol{\pi}) \in \mathbf{M}$:

$$\Phi(\mu_0, \boldsymbol{\pi})_0 = \mu_0, \ \Phi(\mu_0, \boldsymbol{\pi})_{n+1} = \phi(\Phi(\mu_0, \boldsymbol{\pi})_n, \boldsymbol{\pi}_n), n \geq 0.$$

Facing such a population behavior, an infinitesimal agent seeks to solve the following Markov Decision Process (MDP). Given an initial μ_0 and a flow $\boldsymbol{\mu}$, maximize:

$$\boldsymbol{\pi} \mapsto J(\mu_0, \boldsymbol{\pi}; \boldsymbol{\mu}) = \mathbb{E}\Big[\sum_{n=0}^{+\infty} \gamma^n r(x_n, a_n, \boldsymbol{\mu}_n)\Big],$$

subject to: $x_0 \sim \mu_0$, $x_{n+1} \sim p(.|x_n, a_n, \boldsymbol{\mu}_n)$, $a_n \sim \boldsymbol{\pi}_n(.|x_n)$. Note that, at time n, the reward and transition depend on the current MF state $\boldsymbol{\mu}_n$. So this MDP is non-stationary but since the MF flow $\boldsymbol{\mu}$ is fixed and given, it is an MDP in the classical sense. In an MFG, we look for an equilibrium situation, in which the population follows a policy from which no individual player is interested in deviating.

Definition 1 (MFG Nash equilibrium)**.** Given $\mu_0 \in M$, $(\hat{\boldsymbol{\pi}}^{\mu_0}, \hat{\boldsymbol{\mu}}^{\mu_0}) \in \mathbf{\Pi} \times \mathbf{M}$ is an MFG Nash equilibrium (MFG-NE) consistent with μ_0 if: (1) $\hat{\boldsymbol{\pi}}^{\mu_0}$ maximizes $J(\mu_0, \cdot; \hat{\boldsymbol{\mu}}^{\mu_0})$, and (2) $\hat{\boldsymbol{\mu}}^{\mu_0} = \Phi(\mu_0, \hat{\boldsymbol{\pi}}^{\mu_0})$.

Being an MFG-NE amounts to saying that the *exploitability* $\mathcal{E}(\mu_0, \hat{\boldsymbol{\pi}}^{\mu_0})$ is 0, where the exploitability of a policy $\boldsymbol{\pi} \in \mathbf{\Pi}$ given the initial MF state μ_0 is defined as:

$$\mathcal{E}(\mu_0, \boldsymbol{\pi}) = \max_{\boldsymbol{\pi}'} J(\mu_0, \boldsymbol{\pi}'; \Phi(\mu_0, \boldsymbol{\pi})) - J(\mu_0, \boldsymbol{\pi}; \Phi(\mu_0, \boldsymbol{\pi})).$$

It quantifies how much a representative player can be better off by deciding to play another policy than $\boldsymbol{\pi}$ when the rest of the population uses $\boldsymbol{\pi}$ and the initial distribution is μ_0 for both the player and the population. Similar notions are widely used in computational game theory (Zinkevich et al. 2007; Lanctot et al. 2009).

In general, $\hat{\boldsymbol{\pi}}^{\mu_0}$ is not an MFG-NE policy consistent with $\mu_0' \neq \mu_0$. Imagine for example a game in which the agents need to spread uniformly throughout a one-dimensional domain (see the experimental section). Intuitively, the movement of an agent at the center depends on where the bulk of the population is. If μ_0 is concentrated on the left (resp. right) side, this agent should move towards the right (resp. left). Hence the optimal policy depends on the whole population distribution.

Equilibria in MFG are traditionally characterized by a forward-backward system of equations (Lasry and Lions 2007; Carmona and Delarue 2018). Indeed, the value function of an individual player facing an MF flow $\boldsymbol{\mu}$ is:

$$V_n(x; \boldsymbol{\mu}) = \sup_{\boldsymbol{\pi} \in \mathbf{\Pi}} \mathbb{E}_{x, \boldsymbol{\pi}}\Big[\sum_{n'=n}^{+\infty} \gamma^{n'-n} r(x_{n'}, a_{n'}, \boldsymbol{\mu}_{n'})\Big],$$

where $x_n = x$ and $a_{n'} \sim \boldsymbol{\pi}_{n'}(\cdot|x_{n'})$, $n' \geq n$. Dynamic programming yields:

$$V_n(x; \boldsymbol{\mu}) = \sup_{\pi \in \Pi} \mathbb{E}_{x, \pi}\Big[r(x_n, a_n, \boldsymbol{\mu}_n) + \gamma V_{n+1}(x'; \boldsymbol{\mu})\Big],$$

where $x_n = x$, $a_n \sim \pi(\cdot|x, \boldsymbol{\mu}_n)$ and $x' \sim p(\cdot|x, a, \boldsymbol{\mu}_n)$. Taking the maximizer gives an optimal policy for a player facing $\boldsymbol{\mu}$. To find an equilibrium policy, we replace $\boldsymbol{\mu}$ by the equilibrium MF flow $\hat{\boldsymbol{\mu}}$: $\hat{V}_n(\cdot) = V_n(\cdot; \hat{\boldsymbol{\mu}})$. But $\hat{\boldsymbol{\mu}}$ is found by using the corresponding equilibrium policy. This induces a coupling between the backward equation for the representative player and the forward population dynamics.

The starting point of our Master policy approach is to notice that $V_n(\cdot; \boldsymbol{\mu})$ depends on n and $\boldsymbol{\mu}$ only through $(\boldsymbol{\mu}_{n'})_{n' \geq n}$ hence V_n depends on n only through $(\boldsymbol{\mu}_{n'})_{n' \geq n}$:

$$V_n(x; \boldsymbol{\mu}) = V(x; (\boldsymbol{\mu}_{n'})_{n' \geq n})$$

where, for $\boldsymbol{\mu} \in \mathbf{M}, x \in X$,

$$V(x; \boldsymbol{\mu}) = \sup_{\pi \in \Pi} \mathbb{E}_{x, \pi}\Big[r(x, a, \boldsymbol{\mu}_0) + \gamma V(x'; (\boldsymbol{\mu}_n)_{n \geq 1})\Big], \tag{1}$$

where $a \sim \pi(\cdot|x, \boldsymbol{\mu}_0)$ and $x' \sim p(\cdot|x, a, \boldsymbol{\mu}_0)$.

From here, we will express the equilibrium policy $\hat{\boldsymbol{\pi}}_n$ as a stationary policy (independent of n) which takes $\hat{\boldsymbol{\mu}}_n$ as an extra input. Replacing n by $\hat{\boldsymbol{\mu}}_n$ increases the input size but it opens new possibilities in terms of *generalization in MFGs*.

Learning in Mean Field Games

We focus on methods involving Reinforcement Learning, or Dynamic Programming when the model is known. Learning

in MFGs can also involve methods that approximate directly the forward-backward system of equations with function approximations (such as neural networks), but we will not address them here; see, *e.g.*, (Al-Aradi et al. 2018; Carmona and Laurière 2021).

In the literature, *Learning* in MFGs indistinctly refers to the optimization algorithm (being most of the time the fixed point or variations of Fictitious Play), or to the subroutines involving learning that are used to compute the policy (Reinforcement Learning) or the distribution. We make here a distinction between these notions for the sake of clarity.

Optimization algorithm. From a general point of view, learning algorithms for MFGs approximate two types of objects: (1) a policy for the representative agent, and (2) a distribution of the population, resulting from everyone applying the policy. This directly leads to a simple fixed-point iteration approach, in which we alternatively update the policy and the mean-field term. This approach has been used, e.g., by Guo et al. (2019). However without strong hypothesis of regularity and a strict contraction property, this scheme does not converge to an MFG-NE. To stabilize the learning process and to ensure convergence in more general settings, recent papers have either added regularization (Anahtarci, Kariksiz, and Saldi 2020; Guo, Xu, and Zariphopoulou 2020; Cui and Koeppl 2021) or used Fictitious Play (Cardaliaguet and Hadikhanloo 2017; Cardaliaguet and Lehalle 2018; Mguni, Jennings, and Munoz de Cote 2018; Perrin et al. 2020; Delarue and Vasileiadis 2021), while Hadikhanloo (2017) and Perolat et al. (2021) have introduced and analyzed Online Mirror Descent.

Reinforcement learning subroutine. For a given population distribution, to update the representative player's policy or value function, we can rely on RL techniques. For instance Guo et al. (2019); Anahtarci, Kariksiz, and Saldi (2020) rely on Q-learning to approximate the Q-function in a tabular setting, Fu et al. (2019) study an actor-critic method in a linear-quadratic setting, and Elie et al. (2020); Perrin et al. (2021) solve continuous spaces problems by relying respectively on deep deterministic policy gradient (Lillicrap et al. 2016) or soft actor-critic (Haarnoja et al. 2018). Two time-scales with policy gradient has been studied by Subramanian and Mahajan (2019) for stationary MFGs. Policy iterations together with sequential decomposition has been proposed by Mishra, Vasal, and Vishwanath (2020) while Guo et al. (2020) proposes a method relying on Trust Region Policy Optimization (TRPO, Schulman et al. (2015)).

Distribution embedding. Another layer of complexity in MFGs is to take into consideration population distributions for large spaces or even continuous spaces. To compute MFG solutions through a PDE approach, Al-Aradi et al. (2018); Carmona and Laurière (2021) used deep neural networks to approximate the population density in high dimension. In the context of RL for MFGs, recently, Perrin et al. (2021) have used Normalizing Flows (Rezende and Mohamed 2015) to approximate probability measures over continuous state space in complex environments.

Generalization in MFGs through Master policies

So far, learning approaches for MFGs have considered only two aspects: optimization algorithms (*e.g.*, Fictitious Play or Online Mirror Descent), or model-free learning of a representative player's best response based on samples (*e.g.*, Q-learning or actor-critic methods). Here, we build upon the aforementioned notions and add to this picture another dimension of learning: *generalization* over population distributions. We develop an approach to learn the representative player's best response as a function of any current population distribution and not only the ones corresponding to a fixed MFG-NE. This is tightly connected with the so-called Master equation in MFGs (Lions; Bensoussan, Frehse, and Yam 2015; Cardaliaguet et al. 2019). Introduced in the continuous setting (continuous time, continuous state and action spaces), this equation is a partial differential equation (PDE) which corresponds to the limit of systems of Hamilton-Jacobi-Bellman PDEs characterizing Nash equilibria in symmetric N-player games. In our discrete context, we introduce a notion of Master Bellman equation and associated Master policy, which we then aim to compute with a new learning algorithm based on Fictitious Play.

Master Policies for MFGs

We introduce the notion of Master policy and connect it to standard population-agnostic policies arising in MFG-NE.

Consider an MFG-NE $(\hat{\boldsymbol{\pi}}^{\mu_0}, \hat{\boldsymbol{\mu}}^{\mu_0})$ consistent with some μ_0. Let $\hat{V}(\cdot;\mu_0) = V(\cdot;\hat{\boldsymbol{\mu}}^{\mu_0})$, *i.e.*,

$$\hat{V}(x;\mu_0) = \sup_{\pi\in\Pi} \mathbb{E}_\pi\Big[r(x,a,\mu_0) + \gamma V(x';(\hat{\boldsymbol{\mu}}_n^{\mu_0})_{n\geq 1})\Big],$$

where $a\sim\pi(\cdot|x,\mu_0)$ and $x'\sim p(\cdot|x,a,\mu_0)$. By definition, $\hat{\boldsymbol{\pi}}_0^{\mu_0}$ is a maximizer in the sup above. Moreover, in the right-hand side,

$$V(x';(\hat{\boldsymbol{\mu}}_n^{\mu_0})_{n\geq 1}) = \hat{V}(x';\hat{\boldsymbol{\mu}}_1^{\mu_0}), \qquad \hat{\boldsymbol{\mu}}_1^{\mu_0} = \phi(\mu_0,\hat{\boldsymbol{\pi}}_0^{\mu_0}).$$

By induction, the equilibrium can be characterized as:

$$\begin{cases} \hat{\boldsymbol{\pi}}_n^{\mu_0} \in \operatorname{argmax}_{\pi\in\Pi} \mathbb{E}_\pi\Big[r(x,a,\hat{\boldsymbol{\mu}}_n^{\mu_0}) + \gamma\hat{V}(x';\hat{\boldsymbol{\mu}}_{n+1}^{\mu_0})\Big] \\ \hat{V}(x;\hat{\boldsymbol{\mu}}_n^{\mu_0}) = \mathbb{E}_{\hat{\boldsymbol{\pi}}_n^{\mu_0}}\Big[r(x,a,\hat{\boldsymbol{\mu}}_n^{\mu_0}) + \gamma\hat{V}(x';\hat{\boldsymbol{\mu}}_{n+1}^{\mu_0})\Big] \\ \hat{\boldsymbol{\mu}}_{n+1}^{\mu_0} = \phi(\hat{\boldsymbol{\mu}}_n^{\mu_0},\hat{\boldsymbol{\pi}}_n^{\mu_0}). \end{cases}$$

Note that $\hat{\boldsymbol{\mu}}_{n+1}^{\mu_0}$ and $\hat{\boldsymbol{\pi}}_n^{\mu_0}$ depend on each other (and also on μ_0), which creates a forward-backward structure.

In the sequel, we will refer to this function V as the *Master value function*. Computing the value function $(x,\mu)\mapsto\hat{V}(x;\mu)$ would allow us to know the value of any individual state x facing an MFG-NE starting from any MF state μ. However, it would not allow to easily find the corresponding equilibrium policy, which still depends implicitly on the equilibrium MF flow. For this reason, we introduce the notion of population-dependent policy. The set of population-dependent policies $\widetilde{\pi}: X\times\Delta_X\to\Delta_A$ is denoted by $\widetilde{\Pi}$.

Definition 2. A population-dependent $\widetilde{\pi}^*\in\widetilde{\Pi}$ is a Master policy if for every μ_0, $(\boldsymbol{\pi}^{\mu_0,\widetilde{\pi}^*},\boldsymbol{\mu}^{\mu_0,\widetilde{\pi}^*})$ is an MFG-NE, where: $\boldsymbol{\mu}_0^{\mu_0,\widetilde{\pi}^*}=\mu_0$ and for $n\geq 0$,

$$\begin{cases} \boldsymbol{\pi}_n^{\mu_0,\widetilde{\pi}^*}(x) = \widetilde{\pi}^*(x,\boldsymbol{\mu}_n^{\mu_0,\widetilde{\pi}^*}) \\ \boldsymbol{\mu}_{n+1}^{\mu_0,\widetilde{\pi}^*} = \phi(\boldsymbol{\mu}_n^{\mu_0,\widetilde{\pi}^*},\boldsymbol{\pi}_n^{\mu_0,\widetilde{\pi}^*}). \end{cases} \tag{2}$$

A Master policy allows recovering the MFG-NE starting from any initial MF state. A core question is the existence of such a policy, which we prove in Theorem 1 below. Hence, if there is a unique Nash equilibrium MF flow (*e.g.*, thanks to monotonicity), the MF flow $\boldsymbol{\mu}^{\mu_0,\widetilde{\pi}^*}$ obtained with the Master policy $\widetilde{\pi}^*(a|x,\boldsymbol{\mu}_n^{\mu_0,\widetilde{\pi}^*})$ is the same as the one obtained with a best response policy $\hat{\boldsymbol{\pi}}_n^{\mu_0}(a|x)$ starting from μ_0.

Theorem 1. *Assume that, for all $\mu_0 \in M$, the MFG admits an equilibrium consistent with μ_0 and that the equilibrium MF flow is unique. Then there exists a Master policy $\widetilde{\pi}^*$.*

Existence and uniqueness of the MFG-NE for a given μ_0 can be proved under a mild monotonicity condition, see *e.g.* Perrin et al. (2020). Thm. 1 is proved in detail in the Appx. We check, step by step, that the MF flow generated by $\widetilde{\pi}^*$ and the associated population-agnostic policy as defined in (2) form a MFG-NE. The key idea is to use dynamic programming relying on the Master value function V and the uniqueness of the associated equilibrium MF flow.

Algorithm

We have demonstrated above that the Master policy is well-defined and allows to recover Nash equilibria. We now propose a method to compute such a policy.

Fictitious Play

We introduce an adaptation of the Fictitious Play algorithm to learn a Master policy. This extends to the case of population-dependent policies the algorithm introduced by Cardaliaguet and Hadikhanloo (2017). In the same fashion, at every iteration k, it alternates three steps: (1) computing a best response policy $\widetilde{\pi}_k$ against the current averaged MF flows $\bar{\mathcal{M}}_k$, (2) computing $(\boldsymbol{\mu}^{\mu_0,\widetilde{\pi}_k})_{\mu_0\in\mathcal{M}}$, the MF flows induced by $\widetilde{\pi}_k$, and (3) updating $\bar{\mathcal{M}}_{k+1}$ with $(\boldsymbol{\mu}^{\mu_0,\widetilde{\pi}_k})_{\mu_0\in\mathcal{M}}$. In contrast with Cardaliaguet and Hadikhanloo (2017), we learn policies that are randomized and not deterministic.

We choose Fictitious Play rather than a simpler fixed-point approach because it is generally easier to check that an MFG model satisfies the assumptions used to prove convergence (monotonicity condition rather than contraction properties, as *e.g.* in Huang et al. (2006); Guo et al. (2019)).

Ideally, we would like to train the population-dependent policy on every possible distribution, but this is not feasible. Thus, we take a finite training set $\mathcal{M}$ of initial distributions. Each training distribution is used at each iteration of Fictitious Play. Another possibility would have been to swap these two loops, but we chose not to do this because of *catastrophic forgetting* (French 1999; Goodfellow et al. 2014), a well-known phenomenon in cognitive science that also occurs in neural networks, describing the tendency to forget previous information when learning new information. Our proposed algorithm is summarized in Alg. 1 and we refer to it as *Master Fictitious Play*.

Alg. 1 returns $\overline{\widetilde{\pi}}_K$, which is the uniform distribution over past policies. We use it as follows. First, let: $\boldsymbol{\mu}_{k,0}^{\mu_0} = \mu_0$, $k = 1,\dots,K$, $\bar{\boldsymbol{\mu}}_{K,0}^{\mu_0} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\mu}_{k,0}^{\mu_0}$, and then, for $n \geq 0$,

$$\begin{cases} \boldsymbol{\mu}_{k,n+1}^{\mu_0} = \phi(\boldsymbol{\mu}_{k,n}^{\mu_0}, \widetilde{\pi}_k(\cdot|\cdot, \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0})), & k = 1,\dots,K \\ \bar{\boldsymbol{\mu}}_{K,n+1}^{\mu_0} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\mu}_{k,n+1}^{\mu_0}. \end{cases}$$

Note that $\overline{\widetilde{\pi}}_K$ is used in the same way for every μ_0. We will show numerically that this average distribution and the associated average reward are close to the equilibrium ones.

Define the average exploitability as:

$$\bar{\mathcal{E}}_{\mathcal{M}}(\overline{\widetilde{\pi}}_K) = \mathbb{E}_{\mu_0\sim\text{UNIFORM}(\mathcal{M})}\big[\bar{\mathcal{E}}(\mu_0, \overline{\widetilde{\pi}}_K)\big], \qquad (3)$$

where: $\bar{\mathcal{E}}(\mu_0, \overline{\widetilde{\pi}}_K) = \max_{\boldsymbol{\pi}'} J(\mu_0, \boldsymbol{\pi}'; \bar{\boldsymbol{\mu}}_K^{\mu_0}) - \frac{1}{K}\sum_{k=1}^K J(\mu_0, \widetilde{\pi}_k; \bar{\boldsymbol{\mu}}_K^{\mu_0})$. We expect $\bar{\mathcal{E}}(\mu_0, \overline{\widetilde{\pi}}_K) \to 0$ as $K \to +\infty$. We show that this indeed holds under suitable conditions in the idealized setting with continuous time updates, where $\overline{\widetilde{\pi}}_k$, $k = 0,1,2,\dots$, is replaced by $\overline{\widetilde{\pi}}_t$, $t \in [0,+\infty)$ (see Appx for details).

Theorem 2. *Assume the reward is separable and monotone, i.e., $r(x,a,\mu) = r_A(x,a) + r_M(x,\mu)$ and $\sum_{x\in X}(r_M(x,\mu) - r_M(x,\mu'))(\mu-\mu')(x) < 0$ for every $\mu \neq \mu'$. Assume the transition depends only on x and a: $p(\cdot|x,a,\mu) = p(\cdot|x,a)$. Then $\bar{\mathcal{E}}_{\mathcal{M}}(\overline{\widetilde{\pi}}_t) = O(1/t)$, where $\overline{\widetilde{\pi}}_t$ is the average policy at time t in the continuous time version of Master Fictitious Play.*

The details of continuous time Master Fictitious Play and the proof of this result are provided in the Appx, following the lines of (Perrin et al. 2020) adapted to our setting. Studying continuous time updates instead of discrete ones enables us to use calculus, which leads to a simple proof. To the best of our knowledge, there is no rate of convergence for discrete time Fictitious Play in the context of MFG except for potential or linear-quadratic structures, see (Geist et al. 2021) and (Delarue and Vasileiadis 2021).

Algorithm 1: Master Fictitious Play

input : Initial $\widetilde{\pi}_0 \in \widetilde{\Pi}$, training set of initial distributions $\mathcal{M}$, number of Fictitious Play steps K

Let $\bar{\pi} = \widetilde{\pi}_0$; let $\bar{\boldsymbol{\mu}}_{0,n}^{\mu_0} = \mu_0$ for all $\mu_0 \in M$, all $n \geq 0$
Let $\bar{\mathcal{M}}_0 = (\bar{\boldsymbol{\mu}}_0^{\mu_0})_{\mu_0\in\mathcal{M}}$
for $k = 1,\dots,K$ **do**
  Train $\widetilde{\pi}_k$ against $\bar{\mathcal{M}}_k = (\bar{\boldsymbol{\mu}}_k^{\mu_0})_{\mu_0\in\mathcal{M}}$, to maximize Eq. (4)
  for $\mu_0 \in \mathcal{M}$ **do**
    Compute $\boldsymbol{\mu}_k^{\mu_0}$, the MF flow starting from μ_0 induced by $\widetilde{\pi}_k$ against $\bar{\boldsymbol{\mu}}_k^{\mu_0}$
    Let $\bar{\boldsymbol{\mu}}_k^{\mu_0} = \frac{k}{k+1}\bar{\boldsymbol{\mu}}_{k-1}^{\mu_0} + \frac{1}{k+1}\boldsymbol{\mu}_k^{\mu_0}$
  Update $\overline{\widetilde{\pi}}_k = \text{UNIFORM}(\widetilde{\pi}_0,\dots,\widetilde{\pi}_k)$
return $\overline{\widetilde{\pi}}_K = \text{UNIFORM}(\widetilde{\pi}_0,\dots,\widetilde{\pi}_K)$

Deep RL to Learn a Population-dependent Policy

In Alg. 1, a crucial step is to learn a population-dependent best response against the current averaged MF flows $\bar{\mathcal{M}}_k =$

$(\bar{\boldsymbol{\mu}}_k^{\mu_0})_{\mu_0 \in \mathcal{M}}$, *i.e.*, $\widetilde{\pi}_k^*$ maximizing

$$\widetilde{\pi} \mapsto \tfrac{1}{|\mathcal{M}|} \textstyle\sum_{\mu_0 \in \mathcal{M}} J(\mu_0, \widetilde{\pi}; \bar{\boldsymbol{\mu}}_k^{\mu_0}). \quad (4)$$

Solving the optimization problem (4) can be reduced to solving a standard but non-stationary MDP. Since we aim at optimizing over population-dependent policies, the corresponding Q-function is a function of not only an agent's state-action pair (x, a) but also of the population distribution: $\widetilde{Q}(x, \mu, a)$. Adding the current mean field state μ to the Q-function allows us to recover a stationary MDP. As we know that the optimal policy is stationary, we now have a classical RL problem with state (x, μ) (instead of x only), and we can use Deep RL methods such as DQN (Mnih et al. 2013) to compute $\widetilde{Q}_k$. The policy $\widetilde{\pi}_k$ can then be recovered easily by applying the argmax operator to the Q-function.

Various algorithms could be used, but we choose DQN to solve our problem because it is sample-efficient. An algorithm detailing our adaptation of DQN to our setting is provided in Alg. 2 in the Appx. For the numerical results presented below, we used the default implementation of RLlib (Liang et al. 2017).

The neural network representing the Q-function takes as inputs the state x of the representative player and the current distribution μ of the population, which can simply be represented as a histogram (the proportion of agents in each state). In practice, μ is a mean-field state coming from one of the averaged MF flows $\bar{\boldsymbol{\mu}}_k^{\mu_0}$ and is computed in steps 7 and 8 of Alg. 1 with a Monte-Carlo method, *i.e.* by sampling a large number of agents that follow the last population-dependent best response $\widetilde{\pi}_k$ and averaging it with $\bar{\boldsymbol{\mu}}_{k-1}^{\mu_0}$. Then, the Q-function can be approximated by a feedforward fully connected neural network with these inputs. In the examples considered below, the finite state space comes from the discretization of a continuous state space in dimension 1 or 2. The aforementioned simple approximation gives good results in dimension 1. However, in dimension 2, the neural network did not manage to learn a good population-dependent policy in this way. This is probably because passing a histogram as a flat vector ignores the geometric structure of the problem. We thus resort to a more sophisticated representation. We first create an *embedding* of the distribution by passing the histogram to a convolutional neural network (ConvNet). The output of this embedding network is then passed to a fully connected network which outputs probabilities for each action (see Fig.1). The use of a ConvNet is motivated by the fact that the state space in our examples has a clear geometric interpretation and that the population can be represented as an image.

On the Theoretical vs. Experimental Settings

Theoretically, we expect the algorithm Alg. 1 to converge perfectly to a Master policy. This intuition is supported by Thm. 2 and comes from the fact that Fictitious Play has been proved to converge to population-agnostic equilibrium policies when the initial distribution is fixed (Cardaliaguet and Hadikhanloo 2017; Perrin et al. 2020). However, from a practical viewpoint, here we need to make several approximations. The main one is related to the challenges of conditioning on a MF state. Even though the state space X is finite, the space of MF states $M = \Delta_X$ is infinite and of dimension equal to the number of states, which is potentially very large. This is why we need to rely on function approximation (*e.g.*, by neural networks as in our implementation) to learn an optimal population-dependent policy. Furthermore, the training procedure uses only a finite (and relatively small) set of training distributions. On top of this, other more standard approximations are to be taken into account, in particular due to the use of a Deep RL subroutine.

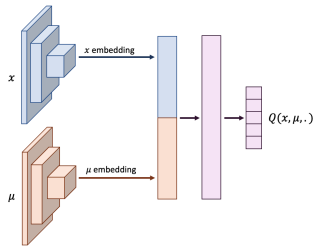

Figure 1: Neural network architecture of the Q-function for the 2D beach bar experience.

Numerical Experiments

Experimental Setup

We now illustrate the efficiency and generalization capabilities of the Master policy learned with our proposed method.

Procedure. To demonstrate experimentally the performance of the learned Master policy trained by Alg. 1, we consider: several initial distributions, several benchmark policies and several metrics. For each metric, we illustrate the performance of each policy on each initial distribution. The initial distributions come from two sets: the training set $\mathcal{M}$ used in Alg. 1 and a testing set. For the benchmark policies, in the absence of a population-dependent baseline of reference (since, to the best of our knowledge, our work is the first to deal with Master policies), we focus on natural candidates that are population-agnostic. The metrics are chosen to give different perspectives: the population distribution and the policy performance in terms of reward.

Training set of initial distributions. In our experiments, we consider a training set $\mathcal{M}$ composed of Gaussian distributions such that the union of all these distributions sufficiently covers the whole state space. This ensures that the policy learns to behave on any state $x \in X$. Furthermore, although we call "training set" the set of initial distributions, the policy actually sees more distributions during the training episodes. Each distribution visited could be considered as an initial distribution. Note however that it is very different from training the policy on all possible *population distributions* (which is a simplex with dimension equal to the number of states, *i.e.*, 32 or $16^2 = 256$ in our examples).

Testing set of initial distributions. It is composed of two types of distributions. First, random distributions generated by sampling uniformly a random number in $[0, 1]$ for each state independently, and then normalizing the distribution.

Second, Gaussian distributions with means located between the means of the training set, and various variances (see the Appx for a representation of the training and testing sets).

Benchmark type 1: Specialized policies. For a given initial distribution μ_0^i with $i \in \{1, \dots, |\mathcal{M}|\}$, we consider a Nash equilibrium starting from this MF state, *i.e.*, a population-agnostic policy $\hat{\boldsymbol{\pi}}^i$ and a MF flow $\hat{\boldsymbol{\mu}}^i$ satisfying Def. 1 with μ_0 replaced by μ_0^i. In the absence of analytical formula, we compute such an equilibrium using Fictitious Play algorithm with backward induction (Perrin et al. 2020). We then compare our learned Master policy with each $\hat{\boldsymbol{\pi}}^i$, either on μ_0^i or on another μ_0^j. In the first case, it allows us to check the correctness of the learned Master policy, and in the second case, to show that it generalizes better than $\hat{\boldsymbol{\pi}}^i$.

Benchmark type 2: Mixture-reward policy. Each (population-agnostic) policy discussed above is specialized for a given μ_0^i but our Alg. 1 trains a (population-dependent) policy on various initial distributions. It is thus natural to see how the learned Master policy fares in comparison with a population-agnostic policy trained on various initial distributions. We thus consider another benchmark, called *mixture-reward policy*, which is a population-agnostic policy trained to optimize an average reward. It is computed as the specialized policies described above but we replace the reward definition with an average over the training distributions. For $1 \leq i \leq |\mathcal{M}|$, recall $\hat{\boldsymbol{\mu}}^i$ is a Nash equilibrium MF flow starting with MF state μ_0^i. We consider the average reward: $\bar{r}_n(x, a) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} r(x, a, \hat{\boldsymbol{\mu}}_n^i)$. The mixture-reward policy is an optimal policy for the MDP with this reward function. In our experiments, we compute it as for the specialized policies described above.

Benchmark type 3: Unconditioned policy. Another meaningful comparison is to use the same algorithm while removing the population input. This amounts to running Alg. 1 where, in the DQN subroutine, the Q-function neural network is a function of x and a only. So in Fig. 1, we replace the μ input embedding by zeros. We call the resulting policy *unconditioned policy* because it illustrates the performance when removing the conditioning on the MF term. This benchmark will be used to illustrate that the success of our approach is not only due to combining Deep RL with training on various μ_0: conditioning the Q-function and the policy on the MF term plays a key role.

Metric 1: Wasserstein distance between MF flows. We first measure how similar the policies are in terms of induced behavior at the scale of the population. Based on the Wasserstein distance W between two distributions (see Appx for details), we compute the following distance between MF flows truncated at some horizon N_T: $W_{i,j} := \frac{1}{N_T+1} \sum_{n=0}^{N_T} W(\boldsymbol{\mu}_n^{\boldsymbol{\pi}^i, \mu_0^i}, \boldsymbol{\mu}_n^{\boldsymbol{\pi}^j, \mu_0^j})$. Note that $W_{i,i} = 0$. The term $\boldsymbol{\mu}^{\boldsymbol{\pi}^j, \mu_0^j} = \hat{\boldsymbol{\mu}}^j$ is the equilibrium MF flow starting from μ_0^j, while $\boldsymbol{\mu}^{\boldsymbol{\pi}^i, \mu_0^j}$ is the MF flow generated by starting from μ_0^j and using policy $\boldsymbol{\pi}^i$.

Metric 2: Exploitability. We also assess the performance of a given policy by measuring how far from being a Nash it is. To this end, we use the exploitability. We compute for each i, j: $E_{i,j} = \mathcal{E}(\mu_0^j, \hat{\boldsymbol{\pi}}^i)$. When $i = j$, $E_{i,i} = 0$ because $(\hat{\boldsymbol{\pi}}^i, \hat{\boldsymbol{\mu}}^i)$ is a Nash equilibrium starting from μ_0^i. When $i \neq j$, $E_{i,j}$ measures how far from being optimal $\hat{\boldsymbol{\pi}}^i$ is when the population also uses $\hat{\boldsymbol{\pi}}^i$, but both the representative player and the population start with μ_0^j. If $E_{i,j} = 0$, then $\hat{\boldsymbol{\pi}}^i$ is a Nash equilibrium policy even when starting from μ_0^j.

Experiment 1: Pure exploration in 1D

We consider a discrete 1D environment inspired by Geist et al. (2021). Transitions are deterministic, the state space is $X = \{1, \dots, |X| = 32\}$. The action space is $A = \{-1, 0, 1\}$: agents can go left, stay still or go right (as long as they stay in the state space). The reward penalizes the agent with the amount of people at their location, while discouraging them from moving too much: $r(x, a, \mu) = -\log(\mu(x)) - \frac{1}{|X|}|a|$. The training set of initial distributions $\mathcal{M}$ consists of four Gaussian distributions with the same variance but different means. The testing set is composed of random and Gaussian distributions with various variances (see Appx, the distributions are represented in the same order as they are used in Fig. 2). We can see that the Master policy is still performing well on these distributions, which highlights its generalization capacities. Note that the white diagonal is due to the fact that the Wasserstein distance and exploitability is zero for specialized baselines evaluated on their corresponding μ_0. We also observe that the random policy is performing well on random distributions, and that exact solutions that have learned on a randomly generated distribution seem to perform quite well on other randomly generated distributions. We believe this is due to this specific environment, because a policy that keeps enough entropy will have a good performance.

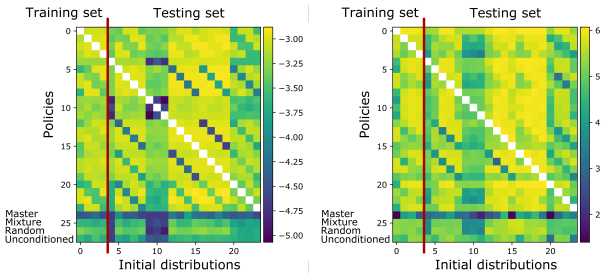

Figure 2: **Exploration 1D: Performance matrices when the training set is made of Gaussian distributions.** From left to right: **(a)** Log of Wasserstein distances to the exact solution (average over time steps); **(b)** Log of exploitabilities.

Experiment 2: Beach bar in 2D

We consider the 2D beach bar problem of Perrin et al. (2020) to highlight that the method can scale to larger environments. The state space is a discretization of a 2-dimensional

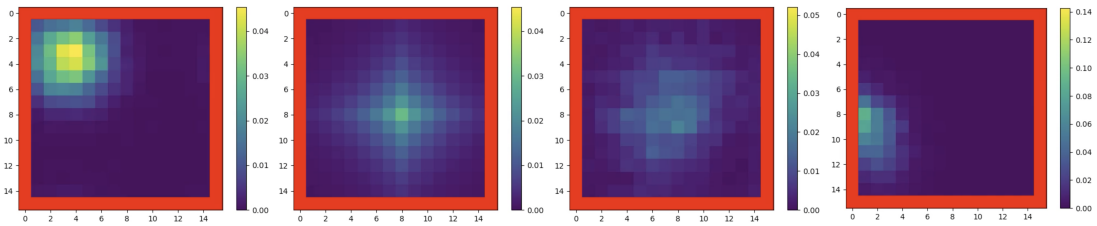

Figure 3: **Beach bar 2D: Environment.** From left to right: **(a)** an initial distribution $\mu_0 \in \mathcal{M}$; **(b)** MF state at equilibrium (specialized policy); **(c)** MF state at equilibrium (learned Master policy); **(d)** MF state at equilibrium (specialized policy of another initial distribution). Note that the scale is very different for the last figure.

square. The agents can move by one state in the four directions: up, down, left, right, but there are walls on the boundaries. The instantaneous reward is: $r(x, a, \mu) = d_{\text{bar}}(x) - \log(\mu(x)) - \frac{1}{|X|}\|a\|_1$, where d_{bar} is the distance to the bar, located at the center of the domain. Starting from an initial distribution, we expect the agents to move towards the bar while spreading a bit to avoid suffering from congestion.

We use the aforementioned architecture (Fig. 1) with one fully connected network following two ConvNets: one for the agent's state, represented as a one-hot matrix, and one for the MF state, represented as a histogram. Having the same dimension (equal to the number $|X|$ of states) and architecture for the position and the distribution makes it easier for the deep neural network to give an equal importance to both of these features. Deep RL is crucial to cope with the high dimensionality of the input. Here $|X| = 16^2 = 256$.

Fig. 4 illustrates the performance of the learned Master policy. Once again, it outperforms the specialized policies as well as the random, mixture-reward, and unconditioned policies. An illustration of the environment and of the different policies involved is available in Fig. 3.

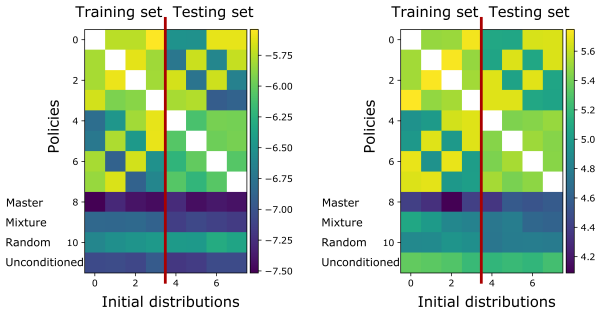

Figure 4: **Beach bar 2D: Performance matrices with Gaussian distributions.** From left to right: **(a)** Log of Wasserstein distances to the exact solution (average over time steps); **(b)** Log of exploitabilities.

Conclusion

Motivated by the question of generalization in MFGs, we extended the notion of policies to let them depend explicitly on the population distribution. This allowed us to introduce the concept of Master policy, from which a representative player is able to play an optimal policy against any population distribution, as we proved in Thm. 1. We then proved that a continuous time adaptation of Fictitious Play can approximate the Master policy at a linear rate (Thm. 2). However, implementing this method is not straightforward because policies and value functions are now functions of the population distribution and, hence, out of reach for traditional computational methods. We thus proposed a Deep RL-based algorithm to compute an approximate Master policy. Although this algorithm trains the Master policy using a small training set of distributions, we demonstrated numerically that the learned policy is competitive on a variety of unknown distributions. In other words, for the first time in the RL for MFG literature, our approach allows the agents to generalize and react to many population distributions. This is in stark contrast with the existing literature, which focuses on learning population-agnostic policies, see *e.g.* (Guo et al. 2019; Anahtarci, Kariksiz, and Saldi 2020; Fu et al. 2019; Elie et al. 2020; Perrin et al. 2021). To the best of our knowledge, the only work considering population-dependent policies is (Mishra, Vasal, and Vishwanath 2020), but it relies on solving a fixed point at each time step for every distribution, which is infeasible except for very small state spaces. Deep learning for finite-state Master equations has been proposed in (Laurière 2021, Section 7.2), but it is based on the full knowledge of the model. Some numerical aspects of Bellman equations involving population distributions have also been discussed in the context of mean field control, with knowledge of the model (Germain et al. 2021) or without (Carmona, Laurière, and Tan 2019; Gu et al. 2020; Motte and Pham 2019). However, these approaches deal only with optimal control and not Nash equilibria and do not treat the question of generalization in MFGs.

Our approach opens many directions for future work. First, the algorithm we proposed is a proof of concept and we plan to investigate other methods, such as Online Mirror Descent (Hadikhanloo 2017; Perolat et al. 2021). For high-dimensional examples, the question of distribution embedding is crucial. Second, the generalization capabilities of the learned Master policy offers many new possibilities. We plan to investigate how it can be used when the agent can only access a partial observation of the population. Last, theoretical properties (e.g., approximation and generalization theory) are also left for future work. An interesting question is choosing the training set so as to optimize generalization capabilities of the learned Master policy.

References

Al-Aradi, A.; Correia, A.; Naiff, D.; Jardim, G.; and Saporito, Y. 2018. Solving nonlinear and high-dimensional partial differential equations via deep learning. *arXiv preprint arXiv:1811.08782*.

Anahtarci, B.; Kariksiz, C. D.; and Saldi, N. 2020. Q-learning in regularized mean-field games. *arXiv preprint arXiv:2003.12151*.

Bensoussan, A.; Frehse, J.; and Yam, S. C. P. 2015. The Master equation in mean field theory. *Journal de Mathématiques Pures et Appliquées*, 103(6): 1441–1474.

Brown, N.; and Sandholm, T. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374): 418–424.

Campbell, M.; Hoane Jr, A. J.; and Hsu, F.-h. 2002. Deep blue. *Artificial intelligence*, 134(1-2): 57–83.

Cardaliaguet, P.; Delarue, F.; Lasry, J.-M.; and Lions, P.-L. 2019. The master equation and the convergence problem in mean field games.

Cardaliaguet, P.; and Hadikhanloo, S. 2017. Learning in mean field games: the fictitious play. *ESAIM Cont. Optim. Calc. Var.*

Cardaliaguet, P.; and Lehalle, C.-A. 2018. Mean field game of controls and an application to trade crowding. *Mathematics and Financial Economics*, 12(3): 335–363.

Carmona, R.; and Delarue, F. 2018. *Probabilistic theory of mean field games with applications. I*, volume 83 of *Probability Theory and Stochastic Modelling*. Springer, Cham. ISBN 978-3-319-56437-1; 978-3-319-58920-6. Mean field FBSDEs, control, and games.

Carmona, R.; and Laurière, M. 2021. Convergence Analysis of Machine Learning Algorithms for the Numerical Solution of Mean Field Control and Games I: The Ergodic Case. *SIAM Journal on Numerical Analysis*, 59(3): 1455–1485.

Carmona, R.; Laurière, M.; and Tan, Z. 2019. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv preprint arXiv:1910.12802*.

Cui, K.; and Koeppl, H. 2021. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *proc. of ICML*, 1909–1917. PMLR.

Delarue, F.; and Vasileiadis, A. 2021. Exploration noise for learning linear-quadratic mean field games. *arXiv preprint arXiv:2107.00839*.

Elie, R.; Perolat, J.; Laurière, M.; Geist, M.; and Pietquin, O. 2020. On the Convergence of Model Free Learning in Mean Field Games. In *proc. of AAAI*.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135.

Fu, Z.; Yang, Z.; Chen, Y.; and Wang, Z. 2019. Actor-Critic Provably Finds Nash Equilibria of Linear-Quadratic Mean-Field Games. In *proc. of ICLR*.

Geist, M.; Pérolat, J.; Laurière, M.; Elie, R.; Perrin, S.; Bachem, O.; Munos, R.; and Pietquin, O. 2021. Concave Utility Reinforcement Learning: the Mean-field Game viewpoint. arXiv:2106.03787.

Germain, M.; Laurière, M.; Pham, H.; and Warin, X. 2021. DeepSets and their derivative networks for solving symmetric PDEs. *arXiv preprint arXiv:2103.00838*.

Goodfellow, I. J.; Mirza, M.; Xiao, D.; Courville, A.; and Bengio, Y. 2014. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *proc. of ICLR*.

Gu, H.; Guo, X.; Wei, X.; and Xu, R. 2020. Mean-Field Controls with Q-learning for Cooperative MARL: Convergence and Complexity Analysis. *arXiv preprint arXiv:2002.04131*.

Guo, X.; Hu, A.; Xu, R.; and Zhang, J. 2019. Learning mean-field games. In *proc. of NeurIPS*.

Guo, X.; Hu, A.; Xu, R.; and Zhang, J. 2020. A General Framework for Learning Mean-Field Games. *CoRR*, abs/2003.06069.

Guo, X.; Xu, R.; and Zariphopoulou, T. 2020. Entropy Regularization for Mean Field Games with Learning. arXiv:2010.00145.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *proc. of ICML*.

Hadikhanloo, S. 2017. Learning in anonymous nonatomic games with applications to first-order mean field games. *arXiv preprint arXiv:1704.00378*.

Huang, M.; Malhamé, R. P.; Caines, P. E.; et al. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3): 221–252.

Lanctot, M.; Waugh, K.; Zinkevich, M.; and Bowling, M. 2009. Monte Carlo sampling for regret minimization in extensive games. volume 22, 1078–1086.

Lanctot, M.; Zambaldi, V.; Gruslys, A.; Lazaridou, A.; Tuyls, K.; Pérolat, J.; Silver, D.; and Graepel, T. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4193–4206.

Lasry, J.-M.; and Lions, P.-L. 2007. Mean field games. *Japanese journal of mathematics*, 2(1): 229–260.

Laurière, M. 2021. Numerical Methods for Mean Field Games and Mean Field Type Control. *2020 AMS short course lecture notes. arXiv preprint arXiv:2106.06231*.

Liang, E.; Liaw, R.; Nishihara, R.; Moritz, P.; Fox, R.; Gonzalez, J.; Goldberg, K.; and Stoica, I. 2017. Ray RLLib: A Composable and Scalable Reinforcement Learning Library. *CoRR*, abs/1712.09381.

Lillicrap, T.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2016. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971.

Lions, P.-L. ???? Lecture at the Cours au Collège de France.

Mguni, D.; Jennings, J.; and Munoz de Cote, E. 2018. Decentralised Learning in Systems With Many, Many Strategic Agents. In *proc. of AAAI*.

Mishra, R. K.; Vasal, D.; and Vishwanath, S. 2020. Model-free reinforcement learning for non-stationary mean field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 1032–1037. IEEE.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *In NIPS Workshop on Deep Learning and Unsupervised Feature Learning. arXiv preprint arXiv:1312.5602*.

Moravčík, M.; Schmid, M.; Burch, N.; Lisỳ, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337): 508–513.

Motte, M.; and Pham, H. 2019. Mean-field Markov decision processes with common noise and open-loop controls. *arXiv preprint arXiv:1912.07883*.

Perolat, J.; Perrin, S.; Elie, R.; Laurière, M.; Piliouras, G.; Geist, M.; Tuyls, K.; and Pietquin, O. 2021. Scaling up Mean Field Games with Online Mirror Descent. *arXiv preprint arXiv:2103.00623*.

Perrin, S.; Laurière, M.; Pérolat, J.; Geist, M.; Élie, R.; and Pietquin, O. 2021. Mean Field Games Flock! The Reinforcement Learning Way. In *proc. of IJCAI*.

Perrin, S.; Pérolat, J.; Laurière, M.; Geist, M.; Elie, R.; and Pietquin, O. 2020. Fictitious play for mean field games: Continuous time analysis and applications. In *proc. of NeurIPS*.

Rezende, D.; and Mohamed, S. 2015. Variational Inference with Normalizing Flows. In *proc. of ICML*.

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3).

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust Region Policy Optimization. In *proc. of ICML*.

Shannon, C. E. 1959. Programming a Computer Playing Chess. *Philosophical Magazine*, Ser.7, 41(312).

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587).

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; and Hassabis, D. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 632(6419).

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676).

Subramanian, J.; and Mahajan, A. 2019. Reinforcement learning in stationary mean-field games. In *proc. of AAMAS*, 251–259.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2007. Regret minimization in games with incomplete information. volume 20, 1729–1736.

Notations used in the text

The main notations used in the text are summarized in the following table. Please note that $\boldsymbol{\pi}$, $\overline{\pi}$ and $\hat{\pi}$ are population-agnostic policies, while $\widetilde{\pi}$, $\overline{\widetilde{\pi}}$ and $\widetilde{\pi}^*$ are population-dependent policies.

Policy	$\boldsymbol{\pi} \in \boldsymbol{\Pi}$
Average policy	$\overline{\pi} \in \boldsymbol{\Pi}$
Equilibrium policy	$\hat{\pi} \in \boldsymbol{\Pi}$
Population-dependent policy	$\widetilde{\pi} \in \widetilde{\Pi}$
Average population-dependent policy	$\overline{\widetilde{\pi}} \in \widetilde{\Pi}$
Master policy	$\widetilde{\pi}^* \in \widetilde{\Pi}$
Mean field state	$\mu \in M$
Mean field flow	$\boldsymbol{\mu} \in \mathbf{M}$
Training set of initial distributions	$\mathcal{M} \subset M$

Details on the Experiments

Wasserstein distance. The Wasserstein distance W (or earth mover's distance) measures the minimum cost of turning one distribution into another: for $\mu, \mu' \in M = \Delta_X$,

$$W(\mu, \mu') = \inf_{\nu \in \Gamma(\mu, \mu')} \sum_{(x,x') \in X \times X} d(x, x')\nu(x, x'),$$

where $\Gamma(\mu, \mu')$ is the set of probability distributions on $X \times X$ with marginals μ and μ'. This notion is well defined if the state space has a natural notion of distance d, which is the case in our numerical examples because they come from the discretization of 1D or 2D Euclidean domains.

Initial distributions. We provide here a representation of the initial distributions used in the experiments.

For the pure exploration model in 1D, the training and testing sets are represented in Fig. 5 and Fig. 6 respectively.

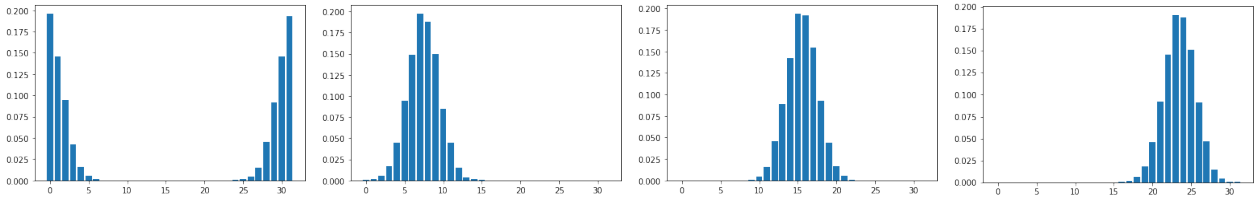

Figure 5: Pure exploration 1D: Training set

For the beach bar model in 2D, the training and testing sets are represented in Fig. 7 and Fig. 8 respectively.

Learning a population-dependent policy with Deep RL

Recall that in line 4 of Alg. 1, we want solve an MDP which is stationary because we have put the distribution μ as an input together with the agent's state x. To this end, we use DQN and, as described in Alg. 2, we use a finite horizon approximation N_T. This approximation is common in the literature and is not problematic as we set the horizon high enough so that the stationary population distribution can be (approximately) reached.

Proof of Theorem 1

Proof of Theorem 1. By assumption, for every μ_0, there is a unique equilibrium MF flow $\hat{\boldsymbol{\mu}}^{\mu_0}$. We also consider an associated equilibrium (population-agnostic) policy $\hat{\boldsymbol{\pi}}^{\mu_0}$ (if there are multiple choices of such policies, we take one of them). The superscript is used to stress the dependence on the initial MF state. Let us define the following population dependent policy:

$$\widetilde{\pi}(x, \mu_0) := \hat{\boldsymbol{\pi}}_0^{\mu_0}(x). \tag{5}$$

We prove that any population-dependent policy defined in the above way is a master policy, *i.e.*, for each μ_0 it gives an equilibrium policy not only at initial time but at all time steps.

Fix μ_0. Let $\tilde{\boldsymbol{\mu}}^{\mu_0}$ and $\tilde{\boldsymbol{\pi}}^{\mu_0}$ be the MF flow and the population-agnostic policy induced by using $\widetilde{\pi}$ starting from μ_0, *i.e.*, for $n \geq 0$,

$$\tilde{\boldsymbol{\pi}}_n^{\mu_0}(x) = \widetilde{\pi}(x, \tilde{\boldsymbol{\mu}}_n^{\mu_0}), \qquad \tilde{\boldsymbol{\mu}}_{n+1}^{\mu_0} = \phi(\tilde{\boldsymbol{\mu}}_n^{\mu_0}, \tilde{\boldsymbol{\pi}}_n^{\mu_0}). \tag{6}$$

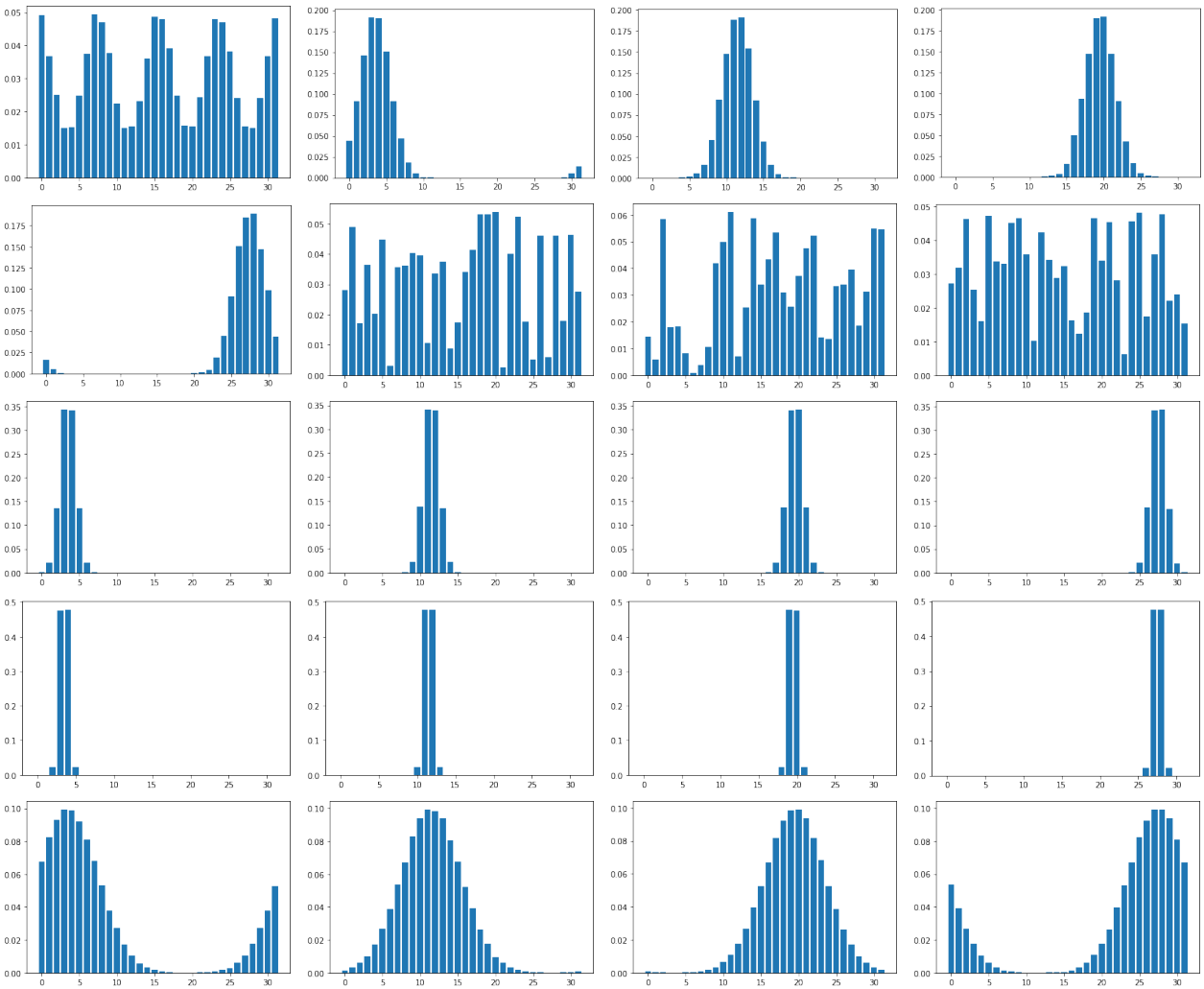

Figure 6: Pure exploration 1D: Testing set

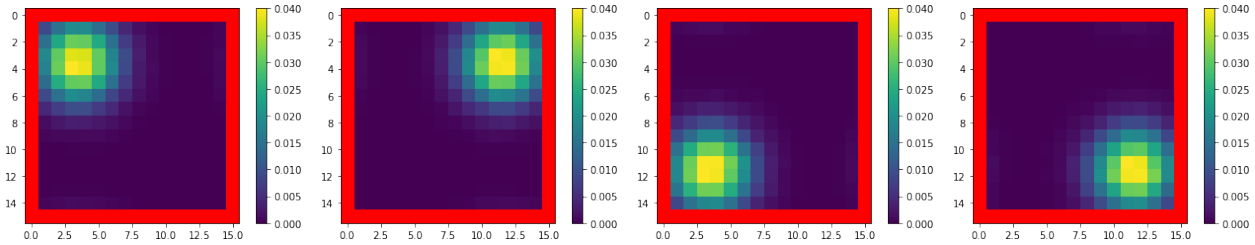

Figure 7: Beach bar 2D: Training set

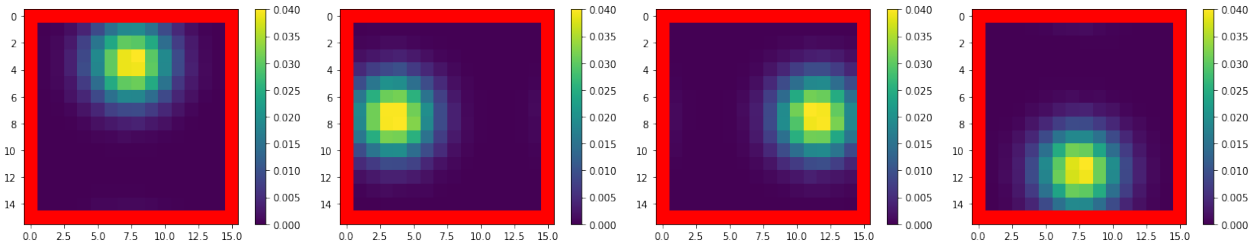

Figure 8: Beach bar 2D: Testing set

Algorithm 2: DQN for a population-dependent Best Response

input : Initial weights θ_k and θ'_k for network $\widetilde{Q}_{\theta_k}$ and target network $\widetilde{Q}_{\theta'_k}$; training set $\mathcal{M}$ of initial distributions; set $\bar{\mathcal{M}}_k$ of average MF flows; number of episodes N_{episodes}; number of inner steps N; horizon N_T for estimation; number of steps C between synchronization of the two networks; parameter $\epsilon \in [0,1]$ for exploration

Initialize weights θ_k of network $\widetilde{Q}_{\theta_k}$ and weights θ'_k of target network $\widetilde{Q}'_{\theta'_k}$

Initialize replay memory B

for $e = 1, \dots, N_{\text{episodes}}$ **do**

  Sample initial $\mu_0 \in \mathcal{M}$ and get the associated $\bar{\boldsymbol{\mu}}_k^{\mu_0}$ from $\bar{\mathcal{M}}_k$

  Sample $x_0 \sim \mu_0$

  for $n = 0, \dots, N-1$ **do**

    With probability ϵ select random action a_n, otherwise select $a_n \in \operatorname{argmax}_a \widetilde{Q}'_k(a|x_n, \bar{\boldsymbol{\mu}}_{k,n}^{\mu_0})$

    Execute action a_n, observe reward r_n and state x_{n+1}

    Add the transition $(x_n, a_n, \bar{\boldsymbol{\mu}}_{k,n}^{\mu_0}, r_n, \bar{\boldsymbol{\mu}}_{k,n+1}^{\mu_0})$ to B

  Sample a random minibatch of N_T transitions $\{(x_n, a_n, \mu_n, r_n, \mu_{n+1}), n = 1, \dots, N_T\}$ from B

  Let $v_n = r_n + \gamma \max_{a'} \widetilde{Q}_{\theta_k}(x_{n+1}, \mu_{n+1}, a')$ for $n = 1, \dots, N_T$

  Update θ_k by performing a gradient step in the direction of minimizing w.r.t. θ:

$$\frac{1}{N_T} \sum_{n=1}^{N_T} \left| v_n - \widetilde{Q}_\theta(x_n, \mu_n, a_n) \right|^2$$

  Every C steps, copy weights θ_k of $\widetilde{Q}_{\theta_k}$ to the weights θ'_k of $\widetilde{Q}'_{\theta'_k}$

return $\widetilde{Q}_{\theta_k}$

We check that it is a Nash equilibrium starting with μ_0. The second condition in Def. 1 is automatically satisfied by definition of $\tilde{\boldsymbol{\mu}}_{n+1}^{\mu_0}$, see (6). For the optimality condition, we proceed by induction to show that for every $n \geq 0$, $\tilde{\boldsymbol{\mu}}_n^{\mu_0} = \hat{\boldsymbol{\mu}}_n^{\mu_0}$, which is the unique equilibrium MF flow starting from μ_0. Note first that, by (5) and dynamic programming,

$$\tilde{\boldsymbol{\pi}}_0^{\mu_0}(x) = \widetilde{\pi}(x, \mu_0) = \hat{\boldsymbol{\pi}}_0^{\mu_0}(x) \in \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}\Big[r(x, a, \mu_0) + \gamma \hat{V}(x_1; \hat{\boldsymbol{\mu}}_1^{\mu_0}) \;\Big|\; x_1 \sim p(.|x, a, \mu_0),\ a \sim \pi(.|x) \Big],$$

where $\hat{V}$ is the stationary and population-dependent value function for a representative agent facing a population playing according to a Nash equilibrium starting from a given distribution. Moreover,

$$\tilde{\boldsymbol{\mu}}_1^{\mu_0} = \phi(\tilde{\boldsymbol{\mu}}_0^{\mu_0}, \tilde{\boldsymbol{\pi}}_0^{\mu_0}) = \phi(\mu_0, \widetilde{\pi}(\cdot, \mu_0)) = \phi(\mu_0, \hat{\boldsymbol{\pi}}_0^{\mu_0}) = \hat{\boldsymbol{\mu}}_1^{\mu_0},$$

where we used (6) for the first and second equalities, and (5) for the third equality. The last equality holds because $(\hat{\boldsymbol{\mu}}^{\mu_0}, \hat{\boldsymbol{\pi}}^{\mu_0})$ is an MFG Nash equilibrium consistent with μ_0. So:

$$\tilde{\boldsymbol{\pi}}_0^{\mu_0}(x) \in \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}\Big[r(x, a, \mu_0) + \gamma \hat{V}(x_1; \tilde{\boldsymbol{\mu}}_1^{\mu_0}) \;\Big|\; x_1 \sim p(.|x, a, \mu_0),\ a \sim \pi(.|x) \Big],$$

At time $n \geq 1$, for the sake of induction, assume $\tilde{\boldsymbol{\mu}}_i^{\mu_0} = \hat{\boldsymbol{\mu}}_i^{\mu_0}$ for all $i \leq n$. Then

$$\tilde{\boldsymbol{\pi}}_n^{\mu_0}(x) = \widetilde{\pi}(x, \tilde{\boldsymbol{\mu}}_n^{\mu_0}) = \hat{\boldsymbol{\pi}}_0^{\tilde{\boldsymbol{\mu}}_n^{\mu_0}}(x) \in \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}\Big[r(x, a, \tilde{\boldsymbol{\mu}}_n^{\mu_0}) + \gamma \hat{V}(x_1; \hat{\boldsymbol{\mu}}_1^{\tilde{\boldsymbol{\mu}}_n^{\mu_0}}) \;\Big|\; x_1 \sim p(.|x, a, \tilde{\boldsymbol{\mu}}_n^{\mu_0}),\ a \sim \pi(.|x) \Big]. \quad (7)$$

Moreover,

$$\tilde{\boldsymbol{\mu}}_{n+1}^{\mu_0} = \phi(\tilde{\boldsymbol{\mu}}_n^{\mu_0}, \tilde{\boldsymbol{\pi}}_n^{\mu_0}) = \phi(\hat{\boldsymbol{\mu}}_n^{\mu_0}, \hat{\boldsymbol{\pi}}_0^{\hat{\boldsymbol{\mu}}_n^{\mu_0}}) \underbrace{=}_{(\star)} \phi(\hat{\boldsymbol{\mu}}_n^{\mu_0}, \hat{\boldsymbol{\pi}}_n^{\mu_0}) = \hat{\boldsymbol{\mu}}_{n+1}^{\mu_0},$$

where the first equality is by (6) and the second equality is by the induction hypothesis and (7). Equality $(\star)$ means that the population distributions generated at the next time step by $\hat{\boldsymbol{\pi}}_0^{\hat{\boldsymbol{\mu}}_n^{\mu_0}}$ and $\hat{\boldsymbol{\pi}}_n^{\mu_0}$ when starting from $\hat{\boldsymbol{\mu}}_n^{\mu_0}$ are the same (although these two policies could be different). This is because both of them are best responses to this population distribution and because we assumed uniqueness of the equilibrium MF flow. Indeed, by definition, $\hat{\boldsymbol{\pi}}_0^{\hat{\boldsymbol{\mu}}_n^{\mu_0}}$ is the initial step of a policy which is part of an MFG Nash equilibrium consistent with $\hat{\boldsymbol{\mu}}_n^{\mu_0}$. Furthermore, $(\hat{\boldsymbol{\pi}}_n^{\mu_0}, \hat{\boldsymbol{\mu}}_n^{\mu_0})_{n\geq 0}$ is an MFG Nash equilibrium consistent with μ_0 and, as a consequence, for any $n_0 \geq 0$, $(\hat{\boldsymbol{\pi}}_n^{\mu_0}, \hat{\boldsymbol{\mu}}_n^{\mu_0})_{n\geq n_0}$ is an MFG Nash equilibrium consistent with $\hat{\boldsymbol{\mu}}_{n_0}^{\mu_0}$.

We conclude that $(\star)$ holds by using the fact that we assumed uniqueness of the equilibrium MF flow so these two policies must have the same result in terms of generated population distribution.

So we proved that $\tilde{\boldsymbol{\mu}}_n^{\mu_0} = \hat{\boldsymbol{\mu}}_n^{\mu_0}$ for all $n \geq 0$ and $(\tilde{\boldsymbol{\pi}}_n^{\mu_0})_{n\geq 0}$ is an associated equilibrium policy.

□

On the Convergence of Master Fictitious Play

In this section we study the evolution of the averaged MF flow generated by the Master Fictitious Play algorithm, see Alg. 1. We then introduce a continuous time version of this algorithm and prove its convergence at a linear rate.

On the mixture of policies. Given $\bar{\widetilde{\pi}}_K = \text{UNIFORM}(\widetilde{\pi}_1, \dots, \widetilde{\pi}_K)$ and given an initial μ_0, we first compute an average population distribution composed of K subpopulations where subpopulation k uses $\widetilde{\pi}_k$ to react to the current average population. Formally, recall that we define:

$$\begin{cases} \boldsymbol{\mu}_{k,0}^{\mu_0} = \mu_0, & k = 1, \dots, K \\ \bar{\boldsymbol{\mu}}_{K,0}^{\mu_0} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\mu}_{k,0}^{\mu_0} \end{cases}$$

and for $n \geq 0$,

$$\begin{cases} \boldsymbol{\mu}_{k,n+1}^{\mu_0} = \phi(\boldsymbol{\mu}_{k,n}^{\mu_0}, \widetilde{\pi}_k(\cdot|\cdot, \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0})), & k = 1, \dots, K \\ \bar{\boldsymbol{\mu}}_{K,n+1}^{\mu_0} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\mu}_{k,n+1}^{\mu_0}. \end{cases}$$

We recall that the notation $\boldsymbol{\mu}_{k,n+1}^{\mu_0} = \phi(\boldsymbol{\mu}_{k,n}^{\mu_0}, \widetilde{\pi}_k(\cdot|\cdot, \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}))$ means:

$$\boldsymbol{\mu}_{k,n+1}^{\mu_0}(x) = \phi(\boldsymbol{\mu}_{k,n}^{\mu_0}, \widetilde{\pi}_k(\cdot|x, \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0})) = \sum_{x'} \boldsymbol{\mu}_{k,n}^{\mu_0}(x') \sum_a \widetilde{\pi}_k(a|x', \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}) p(x|x', a, \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}), \qquad x \in X. \tag{8}$$

Hence: for all $x \in X$,

$$\bar{\boldsymbol{\mu}}_{K,n+1}^{\mu_0}(x) = \frac{1}{K}\sum_{k=1}^K \sum_{x'} \boldsymbol{\mu}_{k,n}^{\mu_0}(x') \sum_a \widetilde{\pi}_k(a|x', \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}) p(x|x', a, \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}) \tag{9}$$

$$= \sum_{x'} \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}(x') \sum_a \underbrace{\left(\frac{1}{K}\sum_{k=1}^K \frac{\boldsymbol{\mu}_{k,n}^{\mu_0}(x')}{\bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}(x')} \widetilde{\pi}_k(a|x', \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}) \right)}_{=: \bar{\widetilde{\pi}}_{K,n}^{\mu_0}(a|x', \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0})} p(x|x', a, \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}), \tag{10}$$

where, in the last expression, the first sum over $\{x' \in X : \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}(x') > 0\}$. So the evolution of the average population can be interpreted as the fact that all the agents use the policy $\bar{\widetilde{\pi}}_{K,n}^{\mu_0}(a|x', \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0})$ given by the terms between parentheses above. Note that this policy depends on μ_0 and n.

We then consider the reward obtained by an infinitesimal player from the average population. This player belongs to subpopulation k with probability $1/K$. So the reward can be expressed as:

$$\frac{1}{K}\sum_{k=1}^K J(\mu_0, \widetilde{\pi}_k; \bar{\boldsymbol{\mu}}_K^{\mu_0}).$$

We expect that when $K \to +\infty$ (*i.e.*, we run more iterations of the Master Fictitious Play algorithm, see Alg. 1), then this quantity converges to the one obtained by a typical player in the Nash equilibrium starting from μ_0, *i.e.*:

$$J(\mu_0, \hat{\boldsymbol{\pi}}; \hat{\boldsymbol{\mu}}^{\mu_0})$$

where $\hat{\boldsymbol{\mu}}^{\mu_0} = \Phi(\mu_0, \hat{\boldsymbol{\pi}})$ with $\hat{\boldsymbol{\pi}} \in \operatorname{argmax}_{\boldsymbol{\pi}} J(\mu_0, \boldsymbol{\pi}; \hat{\boldsymbol{\mu}}^{\mu_0})$.

Note that $\bar{\widetilde{\pi}}_{K,n}^{\mu_0}(a|x', \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0})$ takes $\bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}$ as an input. However, this dependence is superfluous because $\bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}$ can be derived from μ_0 and $(\bar{\widetilde{\pi}}_{K,m}^{\mu_0}(a|x', \bar{\boldsymbol{\mu}}_{K,m}^{\mu_0}))_{m\leq n}$. Proceeding by induction, we can show that there exists $\bar{\boldsymbol{\pi}}_K^{\mu_0} \in \boldsymbol{\Pi}$ s.t.

$$\bar{\widetilde{\pi}}_{K,n}^{\mu_0}(a|x', \bar{\boldsymbol{\mu}}_{K,n}^{\mu_0}) = \bar{\boldsymbol{\pi}}_{K,n}^{\mu_0}(a|x')$$

Continuous Time Master Fictitious Play. We now describe the Continuous Time Master Fictitious Play (CTMFP) scheme in our setting. Here the iteration index $k \in \{1, 2, 3, \dots\}$ is replaced by a time t, which takes continuous values in $[1, +\infty)$. Intuitively, it corresponds to the limiting regime where the updates happen continuously.

Based on (9), we introduce the CTMFP mean-field flow defined for all $t \geq 1$ by: $\bar{\boldsymbol{\mu}}^{\mu_0}_{t,n} = \boldsymbol{\mu}^{\mu_0,\text{BR}}_{t,n} = \mu_0$, and for $n = 1, 2, \dots,$

$$\bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x) = \frac{1}{t}\int_{s=0}^{t} \boldsymbol{\mu}^{\mu_0,\text{BR}}_{s,n}(x)ds, \quad \text{or in differential form:} \quad \frac{d}{dt}\bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x) = \frac{1}{t}\left(\boldsymbol{\mu}^{\mu_0,\text{BR}}_{t,n}(x) - \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x)\right), \tag{11}$$

where $\boldsymbol{\mu}^{\mu_0,\text{BR}}_{t,n}$ denotes the distribution induced by a best response policy $(\boldsymbol{\pi}^{\mu_0,\text{BR}}_{t,n})_{n\geq 0}$ against $\bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x)$.

As in (10) for the discrete update case, the distribution $\bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}$ corresponds to the population distribution induced by the averaged policy $(\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n})_n$ defined as follows: for all $n = 1, 2, \dots,$ and all $t \geq 1$:

$$\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}(a|x)\int_{s=0}^{t} \boldsymbol{\mu}^{\mu_0,\text{BR}}_{s,n}(x)ds = \int_{s=0}^{t} \boldsymbol{\mu}^{\mu_0,\text{BR}}_{s,n}(x)\boldsymbol{\pi}^{\mu_0,\text{BR}}_{s,n}(a|x)ds \tag{12}$$

$$\text{or in differential form: } \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x)\frac{d}{dt}\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}(a|x) = \frac{1}{t}\boldsymbol{\mu}^{\mu_0,\text{BR}}_{t,n}(x)[\boldsymbol{\pi}^{\mu_0,\text{BR}}_{t,n}(a|x) - \bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}(a|x)]. \tag{13}$$

The CTMFP process really starts from time $t = 1$, but it is necessary to define what happens just before this starting time. For $t \in [0, 1)$, we define $\bar{\boldsymbol{\pi}}^{\mu_0}_{t<1} = (\bar{\boldsymbol{\pi}}^{\mu_0}_{t<1,n})_n = (\boldsymbol{\pi}^{\mu_0,\text{BR}}_{t<1,n})_n$, where $\boldsymbol{\pi}^{\mu_0,\text{BR}}_{t<1}$ is constant and equal to an arbitrary policy. The induced distribution between time 0 and 1 is $\bar{\boldsymbol{\mu}}^{\mu_0}_{t<1} = \boldsymbol{\mu}^{\mu_0}_{t<1} = \boldsymbol{\mu}^{\mu_0,\boldsymbol{\pi}^{\mu_0}_{t<1}} = (\boldsymbol{\mu}^{\mu_0,\boldsymbol{\pi}^{\mu_0}_{t<1}}_n)_{n\geq 0}$.

Proof of convergence. We assume the transition p is independent of the distribution: $x_{n+1} \sim p(.|x_n, a_n)$, and we assume the reward can be split as:

$$r(x, a, \mu) = r_A(x, a) + r_M(x, \mu). \tag{14}$$

A useful property is the so-called monotonicity condition, introduced by Lasry and Lions (2007).

Definition 3. The MFG is said to be *monotone* if: for all $\mu \neq \mu' \in M$,

$$\sum_x (\mu(x) - \mu'(x))(r_M(x, \mu) - r_M(x, \mu')) < 0. \tag{15}$$

This condition intuitively means that the agent gets a lower reward if the population density is larger at its current state. Monotonicity implies that for every μ_0, there exists at most one MF Nash equilibrium consistent with μ_0; see (Lasry and Lions 2007). This can be checked by considering the exploitability.

Here, we are going to use the average exploitability as introduced in (3):

$$\bar{\mathcal{E}}_{\mathcal{M}}(\bar{\boldsymbol{\pi}}_t) = \mathbb{E}_{\mu_0\sim\text{UNIFORM}(\mathcal{M})}\left[\bar{\mathcal{E}}(\mu_0, \bar{\boldsymbol{\pi}}^{\mu_0}_t)\right],$$

where $\bar{\boldsymbol{\pi}}_t = (\bar{\boldsymbol{\pi}}^{\mu_0}_t)_{\mu_0\in\mathcal{M}}$ is the uniform distribution over past best responses $(\boldsymbol{\pi}^{\mu_0,\text{BR}}_s)_{s\in[0,t],\mu_0\in\mathcal{M}}$, and we define in the continuous-time setting:

$$\bar{\mathcal{E}}(\mu_0, \bar{\boldsymbol{\pi}}^{\mu_0}_t) = \max_{\boldsymbol{\pi}'} J(\mu_0, \boldsymbol{\pi}'; \bar{\boldsymbol{\mu}}^{\mu_0}_t) - \frac{1}{t}\int_{s=0}^{t} J(\mu_0, \boldsymbol{\pi}^{\mu_0,\text{BR}}_t; \bar{\boldsymbol{\mu}}^{\mu_0}_t).$$

Theorem 3 (Theorem 2 restated)**.** *Assume the reward is separable, the MFG is monotone, and the transition is independent of the population. Then, for every $\mu_0 \in \mathcal{M}$, $\bar{\mathcal{E}}(\bar{\boldsymbol{\pi}}_t) \in O(1/t)$.*

Proof. We follow the proof strategy of Perrin et al. (2020), adapted to our setting. To alleviate the notation, we denote $\langle f, g\rangle_A = \sum_{a\in A} f(a)g(a)$ for two functions f, g defined on A, and similarly for $\langle\cdot,\cdot\rangle_X$. We also denote: $r^{\pi}(x, \mu) = \langle\pi(\cdot|x), r(x, \cdot, \mu)\rangle_A$.

We first note that, by the structure of the reward function given in (14),

$$\nabla_\mu r^{\boldsymbol{\pi}^{\mu_0,\text{BR}}_{t,n}}(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}) = \nabla_\mu r_M(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}) \text{ and } \nabla_\mu r^{\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}}(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}) = \nabla_\mu r_M(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}).$$

Moreover, using (13) and (11) respectively, we have, for every $x \in X$,

$$-\langle\frac{d}{dt}\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}(.|x), r(x, ., \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n})\rangle_A\, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x) = -\frac{1}{t}r^{\boldsymbol{\pi}^{\mu_0,\text{BR}}_{t,n}}(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n})\boldsymbol{\mu}^{\mu_0,\text{BR}}_{t,n}(x) + \frac{1}{t}r^{\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}}(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n})\boldsymbol{\mu}^{\mu_0,\text{BR}}_{t,n}(x),$$

$$-r^{\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}}(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n})\frac{d}{dt}\bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x) = \frac{1}{t}r^{\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}}(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n})\bar{\boldsymbol{\mu}}^{\mu_0}_{t,n}(x) - \frac{1}{t}r^{\bar{\boldsymbol{\pi}}^{\mu_0}_{t,n}}(x, \bar{\boldsymbol{\mu}}^{\mu_0}_{t,n})\boldsymbol{\mu}^{\mu_0,\text{BR}}_{t,n}(x).$$

Using the definition of exploitability together with the above remarks, we deduce:

$$\begin{aligned}
\frac{d}{dt}\bar{\mathcal{E}}(\mu_0, \bar{\boldsymbol{\pi}}_t^{\mu_0}) &= \frac{d}{dt}\Big[\max_{\boldsymbol{\pi}'} J(\mu_0, \boldsymbol{\pi}'; \boldsymbol{\mu}^{\bar{\boldsymbol{\pi}}_t^{\mu_0}, \mu_0}) - J(\mu_0, \bar{\boldsymbol{\pi}}_t^{\mu_0}; \boldsymbol{\mu}^{\bar{\boldsymbol{\pi}}_t^{\mu_0}, \mu_0})\Big] \\
&= \sum_{n=0}^{+\infty} \gamma^n \sum_{x\in X} \Big[\langle \nabla_\mu r^{\boldsymbol{\pi}_{t,n}^{\mu_0,\text{BR}}}(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}), \frac{d}{dt}\bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}\rangle_X \, \boldsymbol{\mu}_{t,n}^{\mu_0,\text{BR}}(x) \\
&\qquad - \langle \nabla_\mu r^{\bar{\boldsymbol{\pi}}_{t,n}^{\mu_0}}(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}), \frac{d}{dt}\bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}\rangle_X \, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}(x) \\
&\qquad - \langle \frac{d}{dt}\bar{\boldsymbol{\pi}}_{t,n}^{\mu_0}(\cdot|x), r(x, \cdot, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0})\rangle_A \, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}(x) - r^{\bar{\boldsymbol{\pi}}_{t,n}^{\mu_0}}(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}) \frac{d}{dt}\bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}(x)\Big] \\
&= \sum_{n=0}^{+\infty} \gamma^n \sum_{x\in X} \Big[t\langle \nabla_\mu r_M(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0})), \frac{d}{dt}\bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}\rangle_X \, \frac{1}{t}\Big(\boldsymbol{\mu}_{t,n}^{\mu_0,\text{BR}}(x) - \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}(x)\Big)\Big] \\
&\quad + \sum_{n=0}^{+\infty} \gamma^n \sum_{x\in X} \Big[\frac{1}{t} r^{\bar{\boldsymbol{\pi}}_{t,n}^{\mu_0}}(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}) \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}(x) - \frac{1}{t} r^{\boldsymbol{\pi}_{t,n}^{\mu_0,\text{BR}}}(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}) \boldsymbol{\mu}_{t,n}^{\mu_0,\text{BR}}(x)\Big] \\
&= -\frac{1}{t}\bar{\mathcal{E}}(\mu_0, \bar{\boldsymbol{\pi}}_t^{\mu_0}) + \sum_{n=0}^{+\infty} \gamma^n \sum_{x\in X} \Big[t\langle \nabla_\mu r_M(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}), \frac{d}{dt}\bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}\rangle_X \, \frac{d}{dt}\bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}(x)\Big],
\end{aligned}$$

where the last term is non-positive. Indeed, the monotonicity condition (15) implies that, for all $\tau \geq 0$, we have:

$$\sum_{x\in X} (\bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}(x) - \bar{\boldsymbol{\mu}}_{t+\tau,n}^{\mu_0}(x))(r_M(x, \bar{\boldsymbol{\mu}}_{t,n}^{\mu_0}) - r_M(x, \bar{\boldsymbol{\mu}}_{t+\tau,n}^{\mu_0})) \leq 0.$$

The result follows after dividing by τ^2 and letting τ tend to 0.

□