Scaling Mean Field Games by Online Mirror Descent

Julien Perolat
DeepMind
France
perolat@google.com

Sarah Perrin
Univ. Lille, CNRS, Inria, Centrale Lille,
UMR 9189 CRIStAL
France

Romuald Elie
DeepMind
France

Mathieu Laurière
Google Research, Brain team
France

Georgios Piliouras
SUTD
Singapore

Matthieu Geist
Google Research, Brain team
France

Karl Tuyls
DeepMind
France

Olivier Pietquin
Google Research, Brain team
France

ABSTRACT

We address the scaling of equilibrium computation in Mean Field Games (MFGs) by using Online Mirror Descent (OMD). We show that continuous-time OMD provably converges to a Nash equilibrium under a natural and well-motivated set of monotonicity assumptions. A thorough experimental investigation on various single and multi-population MFGs shows that OMD outperforms traditional algorithms such as Fictitious Play. We empirically show that OMD scales and converges significantly faster than Fictitious Play by solving, for the first time to our knowledge, examples of MFGs with hundreds of billions states.

KEYWORDS

Mean Field Games; Game Theory

ACM Reference Format:
Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. 2022. Scaling Mean Field Games by Online Mirror Descent. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 22 pages.

1 INTRODUCTION

Solving decision making problems involving multiple agents has been the topic of intensive research in Artificial Intelligence for decades [69, 78]. Such research finds applications in a wide variety of domains such as (amongst others): economics [2, 33, 56], resource management [34, 39], crowd motion modeling [5] or even animal behaviour analysis [13, 60]. Despite the vast literature on Game Theory and numerous fundamental results, application to real-world problems remains a challenge. Recent successes of combining Game Theory and Machine Learning (especially Deep Learning [42] and Reinforcement Learning (RL) [75]) led to solutions for large scale games such as chess [21], Go [70–72], Poker [18, 19, 53] and even complex video games like StarCraft II [77]. Although this allowed for tackling problems involving large state spaces, the number of agents still remains limited and scaling up to large populations of players remains intractable, which prevents a real-world impact.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

To address this challenge of scaling to many agents, the Mean Field Game (MFG) theory was introduced in [46, 48] to study a category of games that involves an infinite population of agents. By considering the limit case of a continuous distribution of identical agents (*i.e.*, anonymous and with symmetric interests), the MFG framework allows the learning problem to be reduced to the characterization of the optimal behavior of a single representative agent in its interactions with the full population. Given this asymptotic formulation, traditional solutions to MFGs entail a coupled system of differential equations: one capturing the forward dynamics of the population and a second being the dynamic programming optimality equation of the representative player. Despite important progress in the area, such approaches are based on numerical approximation schemes for partial differential equations [3, 4, 7, 16, 17, 25, 26], or for stochastic differential equations [11, 31], which do not easily scale to large state spaces. Also, given the sensitivity to limit conditions, only simple configurations of the state space can be considered. Consequently, until recently, we were left with solutions that either scale in terms of the state space dimension (deep RL), or scale in terms of large populations of agents (MFGs).

By introducing solutions inspired by game theory (*e.g.*, Fictitious Play [63, 68]) into MFGs [24, 37, 59], recent research leverages the generalization capacity of Machine Learning to compute a Nash equilibrium (NE) in very large games. Fictitious Play is a generic algorithm that alternates two steps starting from an arbitrary strategy for the representative player: i) computing the best response of this agent against the rest of the population, ii) compute the mixture of that best response with its previous strategy. [59] proposes to make use of recent RL methods to learn the best response and solve problems with millions of states with non-trivial topology. Unfortunately, Fictitious Play seems hard to scale further for several reasons. Firstly, the computation of the best response remains a hard problem even if RL is promising. Secondly, its computational efficiency seems very low in practice. Finally, Fictitious Play requires storing multiple quantities (*e.g.*, averaged policies and induced distributions, etc.), which contributes to cap scalability.

In this context, our main contribution is the introduction of a new algorithm that can tackle a large number of agents as well as

large state spaces. This algorithm, namely Online Mirror Descent (OMD) [67], computes a NE in a large class of MFGs. Inspired by convex optimization and the Mirror Descent algorithm [54], our method doesn't require the computation of a best response. It rather alternates a step of evaluation of the current strategy with a step of improvement of that strategy. The evaluation is done through the computation of the expected accumulated pay-offs of the strategy over time in the shape of a so-called Q-function. The improvement step reduces to computing the soft-max of the quantity obtained by integrating the Q-functions over iterations (like the MD algorithm suggests). Quantities that need to be stored by OMD (the strategy and the integrated Q-function) are thus limited compared to Fictitious Play. As a second contribution, we provide a proof of convergence for continuous time OMD to a NE for MFGs under reasonable assumptions (common in the field). These theoretical results naturally extend to multi-population MFGs as well as to settings where noise is commonly shared by all agents. Our third contribution is an extensive empirical evaluation of OMD on different tasks involving single or multiple populations, in the presence of common noise or not, with non trivial topologies. We highlight that the scale of the considered problems reaches 10^{11} states and trillions of state-action pairs, surpassing by four or five orders of magnitudes existing results. These experiments demonstrate that OMD's computational efficiency is much stronger than state-of-the-art Fictitious Play, which results in faster convergence. Furthermore we provide a proof of convergence under a monotonicity assumption which improve over the more widely used contraction assumption used in the literature (see related work for more detail in Sec. 5).

2 PRELIMINARIES ON MEAN FIELD GAMES

In a Multi-Population Mean Field Game (MP-MFG), an infinite number of players from N_p different populations interact with each other in a temporally and spatially extended game (the case $N_p = 1$ corresponds to a standard MFG). MP-MFG are easily encompassed within MFGs on an extended state space (including the population type), but we use this setting for sake of clarity and completeness. Let $\mathcal{X}$ be the finite discrete state space and $\mathcal{A}$ be the finite discrete action space of the MP-MFG. We denote by $\Delta\mathcal{X}$ and $\Delta\mathcal{A}$ respectively the spaces of probability distributions over states and actions. In this sequential decision problem, a representative player of population $i \in \{1, \dots, N_p\}$ starts at a state $x_0^i \in \mathcal{X}$ according to a distribution $\mu_0^i \in \Delta\mathcal{X}$. We consider a finite time horizon $N > 0$. At each time step $n \in \{0, \dots, N\}$, the representative player of population i is in state x_n^i and takes an action according to $\pi_n^i(.|x_n^i)$, where $\pi_n^i \in (\Delta\mathcal{A})^{\mathcal{X}}$ is a policy. Given this action a_n^i, the representative player moves to a next state x_{n+1}^i with probability $p(.|x_n^i, a_n^i)$ and receives a reward $r^i(x_n^i, a_n^i, \mu_n^1, \dots, \mu_n^{N_p})$, where μ_n^j is the distribution of the population j at time n. Here $p \in (\Delta\mathcal{X})^{\mathcal{X}\times\mathcal{A}}$ and $r^i : \mathcal{X} \times \mathcal{A} \times (\Delta\mathcal{X})^{N_p} \to \mathbb{R}$. Observe that the transition kernel does not depend on the Multi-population distribution as in most classical MFG examples, see *e.g.*, the original work [48].

For the reader's convenience, we denote $\pi^i = \{\pi_n^i\}_{n\in\{0,\dots,N\}}$, $\mu^i = \{\mu_n^i\}_{n\in\{0,\dots,N\}}$, $\pi = \{\pi^i\}_{i\in\{1,\dots,N_p\}}$, $\mu = \{\mu^i\}_{i\in\{1,\dots,N_p\}}$, $\pi_n = \{\pi_n^i\}_{i\in\{1,\dots,N_p\}}$ and $\mu_n = \{\mu_n^i\}_{i\in\{1,\dots,N_p\}}$.

During the game and given a fixed multi-population distributions sequence μ, a representative player of population i accumulates the following sum of rewards:

$$J^i(\pi^i, \mu) = \mathbb{E}\Big[\sum_{n=0}^{N} r^i(x_n^i, a_n^i, \mu_n) \;\Big|\; x_0^i \sim \mu_0^i, a_n^i \sim \pi_n^i(.|x_n^i), \\ x_{n+1}^i \sim p(.|x_n^i, a_n^i)\Big].$$

Backward Equation: Given a population i, a time n, a state x^i, an action a^i, a policy π^i and a multi-population distribution sequence μ, we define the Q-**function**:

$$Q_n^{i,\pi^i,\mu}(x^i, a^i) = \mathbb{E}\Big[\sum_{k=n}^{N} r^i(x_k^i, a_k^i, \mu_k)\Big| x_n^i = x^i, a_n^i = a^i, \\ a_k^i \sim \pi_k^i(.|x_k^i), x_{k+1}^i \sim p(.|x_k^i, a_k^i)\Big]$$

and the **value function**:

$$V_n^{i,\pi^i,\mu}(x^i) = \mathbb{E}\Big[\sum_{k=n}^{N} r^i(x_k^i, a_k^i, \mu_k)\Big| x_n^i = x^i, \\ a_k^i \sim \pi_k^i(.|x_k^i), x_{k+1}^i \sim p(.|x_k^i, a_k^i)\Big].$$

These two quantities can be computed recursively with the following backward equations:

$$\begin{aligned} &Q_N^{i,\pi^i,\mu}(x^i, a^i) = r^i(x^i, a^i, \mu_N) \\ &Q_{n-1}^{i,\pi^i,\mu}(x^i, a^i) = r^i(x^i, a^i, \mu_{n-1}) \\ &\quad + \sum_{x'^i\in\mathcal{X}} p(x'^i|x^i, a^i)\mathbb{E}_{b^i\sim\pi_n^i(.|x'^i)}\Big[Q_n^{i,\pi^i,\mu}(x^i, b^i)\Big], \\ &V_n^{i,\pi^i,\mu}(x^i) = \mathbb{E}_{a^i\sim\pi_n^i(.|x'^i)}\Big[Q_n^{i,\pi^i,\mu}(x^i, a^i)\Big]. \end{aligned}$$

Finally, the sum of rewards is $J^i(\pi^i, \mu) = \mathbb{E}_{x^i\sim\mu_0^i}[V_n^{i,\pi^i,\mu}(x^i)]$.

Forward Equation: If all the agents of a population i follow the policy π^i, the induced population distribution defines recursively via the following forward equation: $\mu_0^{i,\pi^i} = \mu_0^i$ and, for all $x'^i \in \mathcal{X}$,

$$\mu_{n+1}^{i,\pi^i}(x'^i) = \sum_{(x^i,a^i)\in\mathcal{X}\times\mathcal{A}} \pi_n^i(a^i|x^i)p(x'^i|x^i, a^i)\mu_n^{i,\pi^i}(x^i)\,, \quad (1) \\ \text{for } n \le N-1\,.$$

We denote $\mu^\pi = (\mu^{i,\pi^i})_{i\in\{1,\dots,N_p\}}$ and emphasize the following property for the cumulative sum of rewards:

$$J^i(\pi^i, \mu) = \sum_{n=0}^{N}\sum_{(x^i,a^i)\in\mathcal{X}\times\mathcal{A}} \mu_n^{i,\pi^i}(x^i)\pi_n^i(a^i|x^i)r^i(x^i, a^i, \mu_n)$$

.

Best Response and Exploitability: A **best response policy** $\pi^{i,br,\mu}$ to a multi-population distribution sequence μ verifies the following property $\max_{\pi^i} J^i(\pi^i, \mu) = J^i(\pi^{i,br,\mu}, \mu)$. It can be computed

recursively by finding the best responding Q-function $Q^{i,br,\mu}$:

$$Q_N^{i,br,\mu}(x^i,a^i) = r^i(x^i,a^i,\mu_N)$$
$$Q_{n-1}^{i,br,\mu}(x^i,a^i) = r^i(x^i,a^i,\mu_{n-1}) + \sum_{x'^i\in\mathcal{X}} p(x'^i|x^i,a^i)\max_{b^i}\left[Q_n^{i,br,\mu}(x^i,b^i)\right].$$

Finally, $\pi_n^{i,br,\mu}(.|x^i) \in \arg\max Q_n^{i,br,\mu}(x^i,.)$.

The **exploitability** measures the distance to an equilibrum and is defined as $\phi(\pi) = \sum_{i=1}^{N_p}\phi^i(\pi)$ where, for each i, $\phi^i(\pi) = \max_{\pi'^i} J^i(\pi'^i,\mu^\pi) - J^i(\pi^i,\mu^\pi)$.

Monotonicity: A multi-population game is said to be **weakly monotone** if, for any $\rho_n^i, {\rho'}_n^i \in \Delta(\mathcal{X}\times\mathcal{A})$ and $\mu_n^i, {\mu'}_n^i \in \Delta\mathcal{X}$ satisfying

$$\mu_n^i = \sum_{a^i\in\mathcal{A}}\rho_n^i(.,a^i) \text{ and } {\mu'}_n^i = \sum_{a^i\in\mathcal{A}}{\rho'}_n^i(.,a^i)$$

for all i,n, we have

$$\sum_i \sum_{(x^i,a^i)\in\mathcal{X}\times\mathcal{A}} (\rho_n^i(x^i,a^i) - {\rho'}_n^i(x^i,a^i)) \times (r^i(x^i,a^i,\mu_n) - r^i(x^i,a^i,{\mu'}_n)) \le 0\,.$$

It is **strictly weakly monotone** if the inequality is strict whenever $\rho_n \neq \rho'_n$. This condition means that the players are discouraged from taking similar state-action pairs as the rest of the population. Intuitively, it can be interpreted as an aversion to crowded areas.

We have the following property, whose proof is postponed to Appendix. E.

Lemma 1. *The weak monotonicity property implies that for any π,π' with $\pi\neq\pi'$,*

$$\tilde{\mathcal{M}}(\pi,\pi') := \sum_{i=1}^{N_p}\left[J^i(\pi^i,\mu^\pi) + J^i(\pi'^i,\mu^{\pi'}) - J^i(\pi^i,\mu^{\pi'}) - J^i(\pi'^i,\mu^\pi)\right] \le 0. \quad (2)$$

Strictly weak monotonicity implies a strict inequality above.

Moreover, the weak monotonicity condition is met in the following classical setting (see Appx. A).

Lemma 2. *Assume the reward is* **separable**, *i.e.* $r^i(x^i,a^i,\mu) = \bar{r}^i(x^i,a^i) + \tilde{r}^i(x^i,\mu)$ *and the following* **monotonicity condition** *holds: for all $\mu\neq\mu'$, $\sum_i\sum_{x\in\mathcal{X}}(\mu^i(x^i)-{\mu'}^i(x^i))(\tilde{r}^i(x^i,\mu)-\tilde{r}^i(x^i,\mu')) \le 0$ (resp. <0). Then the game is weakly monotone (resp. strictly weakly monotone).*

An example of such a separable and monotone reward can be found in multi-population predator prey models where the reward can be expressed as a network zero-sum game:

$$r^i(x^i,a^i,\mu) = \bar{r}^i(x^i,a^i) + \underbrace{\hat{r}^i(x^i,\mu^i) + \sum_{j\neq i}\mu^j(x^i)\hat{r}^{i,j}(x^i)}_{=\tilde{r}^i(x^i,\mu)} \quad (3)$$

if $\hat{r}^{i,j} = -\hat{r}^{j,i}$ and $\hat{r}$ satisfies the previous monotonicity condition.

Nash Equilibrium (NE): A NE is a vector of policies for all populations that has 0 exploitability. The existence of a NE in MFGs has been studied in many settings [14, 23, 27]. In our framework, it is a consequence of the convergence of the Fictitious Play dynamics in monotone games, which is detailed in Appx. C.

Proposition 1 (Existence and uniqueness of Nash). *Any weakly monotone MP-MFG admits a NE. Besides, if the weak monotonicity is strict, the NE is unique.*

Proof. The existence result follows from Theorem 2 and uniqueness is proven in Appx. F. □

3 ONLINE MIRROR DESCENT: ALGORITHM AND CONVERGENCE RESULT

We now turn to the Online Mirror Descent Algorithm and introduce a regularizer $h:\Delta\mathcal{A}\to\mathbb{R}$, that is assumed to be ρ-strongly convex for some constant $\rho>0$. Furthermore, we will assume from this point forward that the regularizer h is *steep*, i.e., $\|\nabla h(\pi)\|\to\infty$ whenever π approaches the border of $\Delta\mathcal{A}$; The classic negentropy regularizer, which results to replicator dynamics is the prototypical example of this class. Denote by $h^*:\mathbb{R}^{|\mathcal{A}|}\to\mathbb{R}$ its convex conjugate defined by $h^*(y) = \max_{p\in\Delta\mathcal{A}}[\langle y,p\rangle - h(\pi)]$. Since h is differentiable almost everywhere, we have, for almost every y,

$$\Gamma(y) := \nabla h^*(y) = \arg\max_{p\in\Delta\mathcal{A}}[\langle y,p\rangle - h(\pi)].$$

Discrete Time Online Mirror Descent: The OMD algorithm is implemented as described in Algorithm 1. At each iteration, the first step consists in computing, for each population, the evolution of the population's distribution by using the current policy, see (1). In the second step, each population's policy is updated with learning rate α. This update is done by first updating the corresponding y variable and then obtaining the policy thanks to the function Γ. We have for all $t>0, i\in\{1,\dots,N_p\}, n\in\{0,\dots,N\}$,

$$y_{n,t+1}^i(x^i,a^i) = \sum_{s=0}^{t}\alpha Q_n^{i,\pi_s^i,\mu^{\pi_s}}(x^i,a^i),$$
$$\pi_{n,t+1}^i(.|x^i) = \Gamma(y_{n,t+1}^i(x^i,.)).$$

Algorithm 1 Online Mirror Descent (OMD)

Input: learning rate α, $y_{n,0}^i = 0$ for all i,n; t_{max}.
repeat
  Forward Update: Compute for all i, $\mu^{i,\pi^{(t)}}$
  Backward Update: Compute for all i, Q^{i,π_t^i}, μ^{π_t}
  Update for all i,n,x,a,

$$y_{n,t+1}^i(x,a) = y_{n,t}^i(x,a) + \alpha Q_n^{i,\pi_t^i,\mu^{\pi_t}}(x,a)$$
$$\pi_{n,t+1}^i(.|x) = \Gamma(y_{n,t+1}^i(x,.))$$

until $t = t_{max}$

Continuous Time Online Mirror Descent: We study the theoretical convergence of the continuous time version of Alg. 1.

Namely, the Continuous Time Online Mirror Descent (CTOMD) algorithm [50] is defined as: for all $i \in \{1, \dots, N_p\}, n \in \{0, \dots, N\}$, $y^i_{n,0} = 0$, and for all $t \in \mathbb{R}_+$,

$$y^i_{n,t}(x^i, a^i) = \int_0^t Q_n^{i,\pi^i_s,\mu^{\pi_s}}(x^i, a^i)ds,$$
$$\pi^i_{n,t}(.|x^i) = \Gamma(y^i_{n,t}(x^i, .)). \tag{4}$$

From here on, unless otherwise specified, we assume that the weak monotonicity condition holds and denote by π^* a NE, whose existence follows from Proposition 1. We let $y^{i,*} : (x^i, a^i) \mapsto y^{i,*}(x^i, a^i)$ be the corresponding dual variable such that $\pi^{i,*}(.|x^i) = \Gamma(y^{i,*}(x^i, .))$ for every i.

Measure of similarity with the NE π^*: Based on the regularizer h, we define in the dual space the following measure of similarity $H : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ with the NE π^*:

$$H(y) := \sum_{i=1}^{N_p} \sum_{n=0}^{N} \sum_{x^i \in \mathcal{X}} \mu_n^{i,\pi^*}(x^i) \Big[h^*(y^i_{n,t}(x^i, .)) - h^*(y^{i,*}(x^i, .)) - \langle \pi^{i,*}_{n,t}, y^i_{n,t}(x^i, .) - y^{i,*}_{n,t}(x^i, .) \rangle \Big].$$

As detailed below, this quantity will be decreasing through the iterations of CTOMD. Observe that since the regularizer is steep and thus always maps in the interior of the simplex, it can also be expressed in terms of Bregman divergence as:

$$H(y) = \sum_{i=1}^{N_p} \sum_{n=0}^{N} \sum_{x^i \in \mathcal{X}} \mu_n^{i,\pi^*}(x^i)[D_h(\pi_n^{i,*}(x^i, \cdot), \pi_n^i(x^i, \cdot))].$$

which is always non-negative. Here D_F denotes the Bregman divergence associated with a map F and defined as :

$$D_F(p, q) := F(p) - F(q) - \langle \nabla F(q), p - q \rangle.$$

In this derivation we have used known relations between Fenchel couplings and Bregman divergences (e.g., [51]) and denoted $\pi^i_n := \Gamma(y^i_n)$. Thus, the similarity measure H can also be expressed in terms of proximity between policies.

We are now in position to characterize the dynamics of the similarity to the Nash mapping via the following lemma, whose proof is provided in Appendix D.

Lemma 3 (Similarity dynamics). *In CTOMD, the measure of similarity H to the Nash π^* satisfies*

$$\frac{d}{dt} H(y_t) = \Delta J(\pi_t, \pi^*) + \tilde{\mathcal{M}}(\pi_t, \pi^*)$$

where $\Delta J(\pi_t, \pi^*) := \sum_{i=1}^{N_p} J^i(\pi^i_t, \mu^{\pi^*}) - J^i(\pi^{i,*}, \mu^{\pi^*})$ is always non-positive, and the weak monotonicity metric $\tilde{\mathcal{M}}$ is defined in (2).

Convergence to the Nash for MP-MFGs: We now turn to the main theoretical contribution of the paper, by deriving the convergence of CTOMD to the set of NE for MP-MFGs (proof in Appx G).

Theorem 1 (Convergence of CTOMD). *If a MP-MFG satisfies $\tilde{\mathcal{M}}(\pi, \pi') < 0$ if $\mu^\pi \neq \mu^{\pi'}$ and 0 otherwise, then $(\pi_t)_{t \geq 0}$ generated by CTOMD given in* (4) *converges to the set of Nash equilibria of the game as $t \to +\infty$.*

Proof. The assumption $\frac{dH(y_t)}{dt} < 0$ is enough to guarantee convergence of $H(y_t)$ to 0. This relies on the so-called strict Lyapunov condition, which is classical in Lyapunov theory. It can be found in non-linear system books such as [47] or more recently in discrete time in [58]. Let's briefly sketch the main argumentation that relies on a contradiction argument and divides in the following steps in the context of our problem:

- First, in order to have a one to one mapping between π and y, one can rewrite an equivalent dynamical system on the policy

$$y^i_{n,t}(x^i, a^i) = \int_0^t Q_n^{i,\pi^i_s,\mu^{\pi_s}}(x^i, a^i)ds - \int_0^t Q_n^{i,\pi^i_s,\mu^{\pi_s}}(x^i, a^i_{x^i})ds$$

  where $a^i_{x^i}$ is a fixed action for state x^i without changing the trajectory of the policy.

- Second, if $\frac{d}{dt}H(y_t) = \Delta J(\pi_t, \pi^*) + \tilde{\mathcal{M}}(\pi_t, \pi^*) = 0$, we have $\tilde{\mathcal{M}}(\pi_t, \pi^*) = 0$ as $\Delta J(\pi_t, \pi^*) \leq 0$ which is only true if π_t is a Nash (under Theorem 1 conditions).

- Then, assume that $H(y_t)$ is bounded from below by $c > 0$. Given the sign of the derivative, it is also bounded from above by $C = H(y_0)$.

- The set $\{y | H(y) \leq C\}$ must be bounded which is true in our case as h is *steep* (H goes to infinity as π gets close to the boundary) and :

$$H(y) = \sum_{i=1}^{N_p} \sum_{n=0}^{N} \sum_{x^i \in \mathcal{X}} \mu_n^{i,\pi^*}(x^i)[D_h(\pi_n^{i,*}(x^i, \cdot), \pi_n^i(x^i, \cdot))]$$

- Hence, the set $A_{Cc} = \{y | c \leq H(y) \leq C\}$ is compact as a closed bounded set (recall that H is continuous in y).

- As

$$\frac{dH(y_t)}{dt} = \Delta J(\Gamma(y_t), \pi^*) + \tilde{\mathcal{M}}(\Gamma(y_t), \pi^*)$$

  , while $\Delta J(\Gamma(y), \pi^*) + \tilde{\mathcal{M}}(\Gamma(y), \pi^*) < 0$ for all y in the compact set A_{Cc}, we deduce the existence of a constant k_{max} such that $\Delta J(\Gamma(y), \pi^*) + \tilde{\mathcal{M}}(\Gamma(y), \pi^*) \leq k_{max} < 0$ for $y \in A_{Cc}$ (the image of a compact through a continuous function is a compact).

- As $H(y_t) = H(y_0) + \int_0^t \frac{dH(y_\tau)}{d\tau} d\tau$, this implies that $H(y_t) \leq C + t \times k_{max}$, so there will be a time t when $H(y_t) < c$. This provides a contradiction and implies that $H(y_t)$ must converge to 0.

This concludes the sketch of the proof. □

Thanks to Lemma 1 together with Proposition 1, we easily deduce the convergence to the unique NE in some more stringent classes of MP-MFGs. It is worth noticing that our line of argument differs from the usual approaches on regret minimization arguments as e.g. in [82].

Environment	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{X}\rvert \times \lvert\mathcal{A}\rvert$	OMD	Fictitious Play
Garnet	$2 \times 10^3 - 2 \times 10^4$	$2 \times 10^4 - 4 \times 10^5$	84 – 229KB	168 – 458KB
Building	8×10^9	5.6×10^{10}	0.21TB	0.42TB
Common noise	2.73×10^{11}	1.092×10^{12}	5.0TB	10TB
Multi-Pop. medium	5×10^7	2×10^8	0.93GB	1.9GB
Multi-Pop. large	8×10^8	3.2×10^9	73GB	146GB

Table 1: Number of states, action-states pairs & RAM memory required for the experiments. $|\mathcal{X}|$ = positions × timesteps × common noise × number of populations (KB stands for Kilo Byte, G stands for Giga and T stands for Tera).

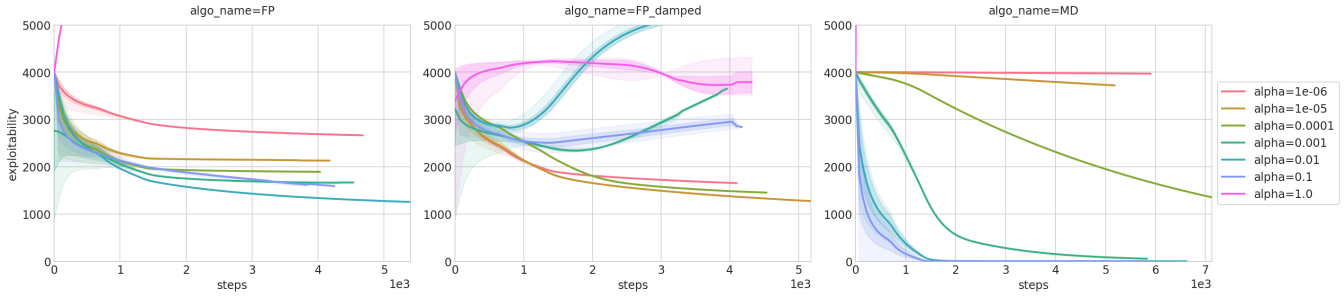

Figure 1: 5 Garnet sampled with param $n_x = 20000$, $n_a = 10$, $t = 2000$, $s_f = 10$

Corollary 1 (Convergence of CTOMD for weakly monotone MFG). *For any strictly weakly monotone MP-MFG, $(\pi_t)_{t\geq 0}$ generated by CTOMD given in* (4) *converges to the unique NE, as $t \longrightarrow +\infty$.*

Restriction to single population MFG: Finally, considering the number of populations N_p equal to 1, the convergence of CTOMD to the NE for single population strictly weakly monotone MFG follows.

Corollary 2 (Convergence of CTOMD for Single Population MFG). *For any single population MFG satisfying the strictly weak monotonicity assumption, $(\pi^{(t)})_{t\geq 0}$ generated by CTOMD given in* (4) *converges to the unique NE of the game, as $t \longrightarrow +\infty$.*

Remark 1. *Our proofs only give an asymptotic result and we don't think anything better is achievable in general. However if one can upper bound the monotony coefficient $\tilde{\mathcal{M}}(\pi_t, \pi^*)$ by the Lyapunov function $H(y_t)$ for example, the Gronwall inequality would give an exponential convergence rate.*

4 NUMERICAL EXPERIMENTS

We illustrate the theoretical convergence of CTOMD with an extensive empirical evaluation of OMD described in Algorithm 1 within various settings involving single or multiple populations as well as non trivial topologies (videos available here). These settings are typically hardly tractable using classical numerical approximation schemes for partial differential equations. Besides, the scale of the numerical experiments grows up to 10^{12} states, establishing a new scalability benchmark in the MFG literature. We emphasize the diversity of tractable environments by considering (randomized MDP) Garnet settings, a twenty-storey high building evacuation, a crowd movement example in the presence of common noise and finally an essentially zero sum multi-population chasing game.

Experimental setup: We compare OMD and Fictitious Play with different learning rates α. In discrete-time OMD, α appears in the backward update of y:

$$y^i_{n,t+1}(x,a) = y^i_{n,t}(x,a) + \alpha Q^{i,\pi^i_t,\mu^{\pi_t}}_n(x,a),$$

whereas in discrete-time Fictitious Play, it corresponds to the weight for updating the average policy with the new best response $\pi^i_{n,t+1}(x^i,a^i)$ given by

$$\frac{(1-\alpha_t)\mu^{i,\pi_t}_n(x^i)\pi^i_{n,t+1}(x^i,a^i) + \alpha_t \mu^{i,br}_{n,t}(x^i)\pi^{i,br}_{n,t}(x^i,a^i)}{(1-\alpha_t)\mu^{i,\pi_t}_n(x^i) + \alpha_t \mu^{i,br}_{n,t}(x^i)}.$$

Fictitious Play is experimented with decreasing $\alpha_t = \alpha/(2+t)$ or constant $\alpha_t = \alpha$ learning rate. This latter is referred to hereafter as *Fictitious Play damped*, while $\alpha = 1$ corresponds to the fixed point iteration algorithm, *i.e.* the population applies the last best response policy. The theoretical proof of convergence relies on restrictive conditions which only hold for a small class of games. We provide a thorough evaluation in Table 1 of the complexity of the environments along with the memory required to compute our results. For OMD, we only need to store y of size $|\mathcal{X}| \times |\mathcal{A}|$ and the distributions, of size $|\mathcal{X}|$. For Fictitious Play, we need to store the last best response, the average policy, the last distribution and the average distribution, requiring a total of $2\times(|\mathcal{X}|\times|\mathcal{A}|)+2\times|\mathcal{X}|$. In all the experiments, h is the entropy: $h = -\sum_{a\in\mathcal{A}} \pi(a)\log(\pi(a))$. This implies that $h^*(y) = \log(\sum_a \exp(y(a)))$, and we find that Γ is

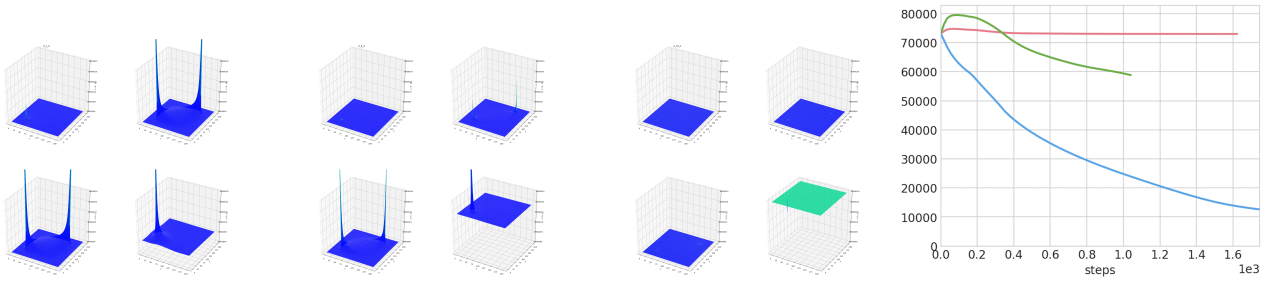

Figure 2: Population distribution at consecutive dates (three first figures on the left). Each plot of a subfigure is a different floor, the bottom floor is the bottom-right plot, the top floor is the top-left plot. The figure on the right displays the exploitability of: Fictitious Play (red, $\alpha = 10^{-5}$), Fictitious Play damped (green, $\alpha = 10^{-3}$) and OMD (blue, $\alpha = 10^{-4}$).

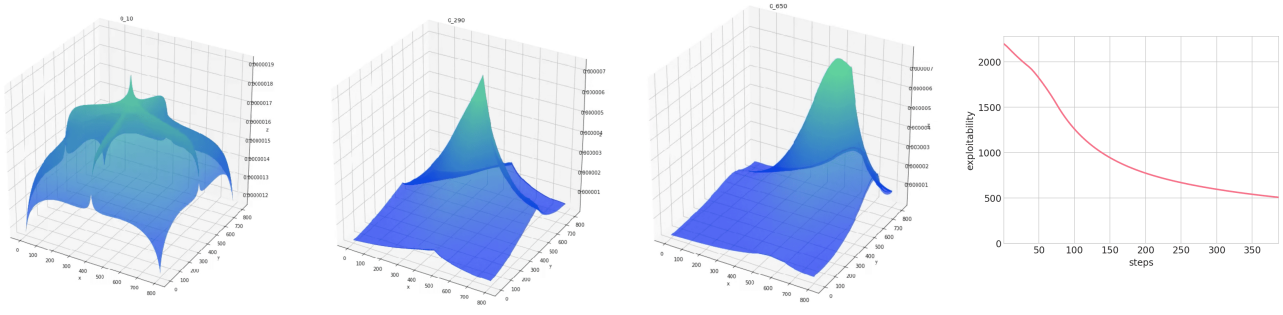

Figure 3: Crowd position at different consecutive dates when the point of interest is randomly shifted to the right by a common noise. The fourth graph is displaying the exploitability of MD.

a softmax if we take the gradient of h^*.

4.1 Garnet

We first evaluate our algorithm on a set of randomly generated problems (repeatability of our results for varying sizes).

Environment: A garnet is an abstract and randomly generated MDP [12]. We adapt this concept to single-population MFGs by modifying the reward. In our case, a Garnet is built from the set of parameters $(n_x, n_a, n_b, s_f, \eta)$, with n_x and n_a respectively the numbers of states and actions. The term n_b is a branching factor, and the transition kernel (independent of μ) is built as follows: n_b transiting states are drawn randomly without replacement, and the associated transition probabilities are obtained by partitioning the unit interval with $n_x - 1$ uniformly sampled random points. The reward term $\bar{r}(x, u)$ is set to 0 for s_f states sampled randomly without replacement, for each of the remaining states it is set for all actions to a random value sampled uniformly in the unit interval. We set $\bar{r}(s, \mu) = -\eta \log(\mu(x))$. This reward encourages the agents to spread out accross the MDP states and can model social distancing. This process generates a monotone MFG.

Numerical results: Fig. 1 (main text) and 10 (Appx. H.1) shows various Garnet experiments. We fix $s_f = 10$, $t = 2000$, $\eta = 1$ and $n_b = 1$ (deterministic dynamics) and vary $n_x \in \{2.10^3; 2.10^4\}$ and $n_a \in \{10, 20\}$. In each case, results are averaged over 5 randomly generated Garnets. We compare OMD to Fictitious Play, damped or not. We observe that OMD consistently converges faster for the right choice of α. $\alpha = 1$ might lead to unstable results while $\alpha = 0.1$ consistently provides fast convergence to the Nash. In all cases, the number of states influences the convergence rate, but much less for OMD.

4.2 Building evacuation

Environment: We now turn to a single-population crowd modeling problem, namely a building evacuation. This kind of problem has been considered in several studies on MFG (see e.g. [5, 6] for a single room and [36] for a multilevel building). The building consists of 20 floors, each of dimension 200×200. At each floor, two staircases are located at two opposite corners, such as the crowd has to cross the whole floor to take the next staircase. Each agent can remain in place, move in the 4 directions (up, down, right, left) as well as go up or down when on a staircase location. The initial distribution is uniform over all the floors. Each agent of the crowd wants to go downstairs as quickly as possible - as it gets a reward

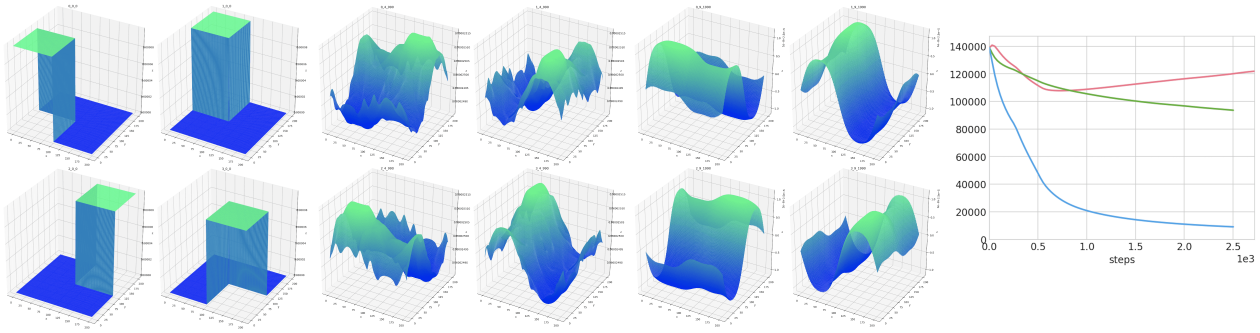

Figure 4: 4-population chasing. Right figure : Fictitious Play (red, $\alpha = 10^{-3}$), Fictitious Play damped (green, $\alpha = 10^{-5}$) and OMD (blue, $\alpha = 10^{-5}$). From left to right, 3 picture chowing the distribution evolving through time and a fourth one displaying the exploitability.

of 10 at the bottom floor - while favoring social distancing:

$$r(x, a, \mu) = -\eta \log(\mu(x)) + 10 \times \mathbb{1}_{floor=0}$$

Numerical results: We compute this problem with a horizon of

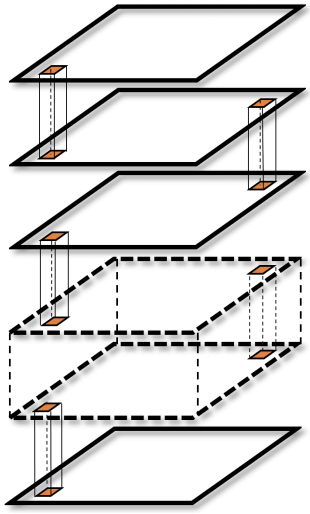

Figure 5: Building environment.

10000, so $|\mathcal{X}| = 8^{10}$. We take $\eta = 1$. To ensure that the reward stays bounded, we clip the first part $-\eta \log(\mu(x))$ to -40. As expected, we observe in Fig. 2 that the agents go downstairs and do not concentrate on the shortest path but rather spread mildly. OMD converges faster than both Fictitious Play and Fictitious Play damped.

4.3 Crowd motion with randomly shifted point of interest

Environment: We consider a second crowd modeling MFG, extending the Beach Bar problem of [59] in two dimensions. The environment is a 2D torus of dimensions 1000×1000, with a point of interest initially located at the center of the square. After 200 timesteps, the point of interest changes location, moving randomly in the direction of one of the corner. This process repeats itself 5 times. This random location change adds common noise to the environment and increases exponentially the number of states. Considering MFG with common noise can be encompassed in our previous study by simply increasing the state space with the common noise and adding time to the reward and the transition kernel. For every random movement, four possible directions are possible, making the total number of states $|\mathcal{X}| = 2 \times 10^8 \times \sum_{k=0}^{4} 4^k = 2.73 \times 10^{11}$ states. The reward is: $r(x, a, \mu) = C \times (1 - \frac{\|bar-(i,j)\|_1}{2 \times N_{side}}) - \log(\mu(x))$.

Numerical results: We set $C = 10$. We observe in Fig. 3 that the population is organizing itself with respect to the point of interest and follows it closely as it randomly moves within the dedicated square region. In the common noise setting, we get more than a trillion states, making it hard for Fictitious Play to scale. More plots with a smaller state space are available in Appx. H for a comparison of OMD and Fictitious Play.

4.4 Multi-population chasing

Environment: We finally look at MP-MFGs, where the populations are chasing each other in a cyclic manner. For the sake of clarity, we explain the reward structure with 3 populations, but more populations are considered in the experiments. With three populations, the game closely relates to the well known Hens-Foxes-Snakes outdoor game for kids. Hens are trying to catch snakes, while snakes are chasing foxes, who are willing to eat hens. It can also be interpreted as a control version of the spatially extended Rock-Paper-Scissors, where patterns of travelling waves appear under certain conditions [62]. The interplay between nontransitive interactions and biodiversity has been the subject of extensive, mostly experimental, research showing that the setting details critically affect the emergent behavior [76]. To ensure $\bar{r}^{i,j} = -\bar{r}^{j,i}$ we implement MP-MFGs with the reward structure defined in Table 6 (ex. with 3 populations).

The reward of population i is monotone (*cf.* Appx. H.4.1) and follows the definition (3):

$r^i(x, a, \mu^1, \dots, \mu^N) = -\log(\mu^i(x)) + \sum_{j \neq i} \mu^j(x) \bar{r}^{i,j}(x)$.

The distributions are initialized either randomly or in different corners. The number of agents of each population is fixed, but the reward encourages the agent to chase the population that it

	R	P	S
R	0	-1	1
P	1	0	-1
S	-1	1	0

Figure 6: $\bar{r}^{i,j}$ for three-population.

dominates. For example, if an agent is Rock, the second term of the reward is proportional to the amount μ^S of Scissors agents where the Rock agent is located, and inversely proportional to the proportion μ^P of Paper agents, making the Rock agent to flee from places populated by Paper agents.

Numerical results: We present a four-population example, each is initially located at a corner of the environment. We observe that the populations are chasing each other in a cyclic fashion. Fig. 4 highlights that OMD algorithm outperforms Fictitious Play in terms of exploitability minimization (full comparison with different values of α in Appx. H.4). It demonstrates the robustness of the OMD algorithm within the different topologies considered. Topologies of the environment are a torus, a basic square or the 'donut' topology (an environment where the agent gets a negative reward if it goes inside a large zone at the center of the square).

5 RELATED WORK

OMD in Normal Form Games : OMD dynamics have been studied extensively within the field of multi-agent games [30, 55]. Leveraging the well known advantageous regret properties of such dynamics [73], one can prove strong time-average convergence results both in zero-sum games (and network variants thereof) [20, 40] as well as in smooth-games [64]. Recently, there has been explicit focus on understanding their day-to-day behavior which has been shown to be non-equilibrating even in standard bilinear zero-sum games [50, 61]. Moreover, even in simple games the behavior of such dynamics can become formally chaotic [32, 57, 66]. Nevertheless, sufficient conditions have been established under which converge to NE is guaranteed even in the sense of the day-to-day behavior [15, 81]. We find sufficient conditions for convergence in the more demanding setting of MP-MFG.

Learning in Mean Field Games : Related to the question of learning in MFGs, [80] studied a MF oscillator game, while [24] initiated the study of Fictitious Play in MFGs, which has been further studied in [45]. Recently, these ideas have been combined with RL by [37, 59]. These methods allow solving MFGs under a monotonicity assumption, which is at the same time easier to check and less restrictive than the ones used to ensure convergence for fixed point iterations [9, 43] or single-loop Fictitious Play iterations [10, 79]. In our work, we also prove convergence under such a weak monotonicity condition, which enables us to cover a large class of MFGs. Furthermore, we consider time-dependent problems (as e.g. in [52]) instead of stationary equilibria. Mirror Descent for MFGs has been introduced in [44] for first-order, single-population MFG, while our results cover second order MP-MFG. As far as we know, our work is the first one to provide a well-suited monotonicity condition for MP-MFG. Traditional numerical methods for solving MFGs typically rely on a finite difference scheme introduced in [4]. This approach can be extended to solve MP-MFG, see [1]. However, to the best of our knowledge, there is no general convergence guarantees, nor has it been tested on examples with as many states as we consider.

Contraction hypothesis in Mean Field Games : In contrast to this work, a large part of learning algorithms in Mean Field Games assume some form of a contraction assumption ([9, 43, 79]). Despite being a source of great convergence properties, it has been shown in [35] that Mean Field Games examples in the literature generally fail to satisfy this assumption. Furthermore, in no way these papers exhibit examples that check the necessary contraction property they assume. To circumvent this problem [35] introduce a form of smoothing having the effect of biasing the solution found by the algorithm.

Numerical methods : More recently, several numerical methods to solve MFGs based on machine learning tools have been proposed using either an analytical viewpoint [8, 22, 28, 49, 65] or a stochastic viewpoint [29, 38, 41]. To the best of our knowledge, these algorithms have not been proved to converge, are applicable only under rather stringent conditions (on the structure or the regularity of the problem) and do not seem to be directly applicable to complex geometries due to boundary conditions. Last, the question of learning with multiple infinite populations of agents has also been studied recently in [74]. The authors consider several groups where the agents cooperate among each group, which differs from our setting where all the agents compete.

6 CONCLUSION

We proposed Online Mirror Descent for MP-MFGs and proved that, under appropriate monotonicity assumptions, OMD converges to a NE. Moreover, we considered multiple experimental benchmarks, some with hundreds of billions states, and have extensively compared OMD to state-of-the-art Fictitious Play. OMD scales empirically remarkably well, and consistently converges significantly faster than Fictitious Play. An interesting direction of future work would be to study the rate of convergence of OMD. Fictitious Play benefits from a $O(1/t)$ rate of convergence (see Appx. C) but the corresponding line of argument does not extend to OMD. Empirically, we envision to extend this approach to a model-free setting with function approximation and address even larger problems.

ACKNOWLEDGMENTS

This research-project is supported in part by the National Research Foundation, Singapore under NRF 2018 Fellowship NRF-NRFF2018-07, AI Singapore Program (AISG Award No: AISG2-RP-2020-016), NRF2019-NRF-ANR095 ALIAS grant, AME Programmatic Fund (Grant No. A20H6b0151) from the Agency for Science, Technology and Research (A*STAR), grant PIE-SGP-AI-2018-01 and Provost's Chair Professorship grant RGEPPV2101.

REFERENCES

[1] Yves Achdou, Martino Bardi, and Marco Cirant. 2017. Mean field games models of segregation. *Math. Models Methods Appl. Sci.* 27, 1 (2017), 75–113. https://doi.org/10.1142/S0218202517400036

[2] Yves Achdou, Francisco Buera, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. 2014. PDE Models in Macroeconomics. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* (2014).

[3] Yves Achdou, Fabio Camilli, and Italo Capuzzo-Dolcetta. 2012. Mean field games: numerical methods for the planning problem. *SIAM Journal on Control and Optimization* 50, 1 (2012).

[4] Yves Achdou and Italo Capuzzo-Dolcetta. 2010. Mean field games: numerical methods. *SIAM J. Numer. Anal.* 48, 3 (2010). https://doi.org/10.1137/090758477

[5] Yves Achdou and Jean-Michel Lasry. 2019. Mean field games for modeling crowd motion. In *Contributions to partial differential equations and applications*. Springer.

[6] Yves Achdou and Mathieu Laurière. 2015. On the system of partial differential equations arising in mean field type control. *Discrete Contin. Dyn. Syst.* 35, 9 (2015). https://doi.org/10.3934/dcds.2015.35.3879

[7] Yves Achdou and Mathieu Laurière. 2020. Mean Field Games and Applications: Numerical Aspects. In *Mean Field Games*. C.I.M.E. Foundation Subseries, Vol. 2281. Springer International Publishing.

[8] Ali Al-Aradi, Adolfo Correia, Danilo Naiff, Gabriel Jardim, and Yuri Saporito. 2018. Solving nonlinear and high-dimensional partial differential equations via deep learning. *arXiv preprint arXiv:1811.08782* (2018).

[9] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. 2020. Q-learning in regularized mean-field games. *arXiv preprint arXiv:2003.12151* (2020).

[10] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. 2020. Unified reinforcement Q-learning for mean field game and control problems. *arXiv preprint arXiv:2006.13912* (2020).

[11] Andrea Angiuli, Christy V Graves, Houzhi Li, Jean-François Chassagneux, François Delarue, and René Carmona. 2019. Cemracs 2017: numerical probabilistic approach to MFG. *ESAIM: Proceedings and Surveys* 65 (2019).

[12] TW Archibald, KIM McKinnon, and LC Thomas. 1995. On the generation of markov decision processes. *Journal of the Operational Research Society* 46, 3 (1995), 354–361.

[13] Martino Bardi and Pierre Cardaliaguet. 2020. Convergence of some Mean Field Games systems to aggregation and flocking models. *arXiv:2004.04403* (2020).

[14] Alain Bensoussan, Jens Frehse, and Sheung Chi Phillip Yam. 2013. *Mean Field Games and Mean Field Type Control Theory*. Springer, New York.

[15] Mario Bravo, David S Leslie, and Panayotis Mertikopoulos. 2018. Bandit learning in concave N-person games. *arXiv preprint arXiv:1810.01925* (2018).

[16] Luis M. Briceño Arias, Dante Kalise, Ziad Kobeissi, Mathieu Laurière, Álvaro Mateos González, and Francisco J. Silva. 2019. On the implementation of a primal-dual algorithm for second order time-dependent Mean Field Games with local couplings. *ESAIM: Proceedings* 65 (2019). https://doi.org/10.1051/proc/201965330

[17] Luis M. Briceño Arias, Dante Kalise, and Francisco J. Silva. 2018. Proximal methods for stationary mean field games with local couplings. *SIAM Journal on Control and Optimization* 56, 2 (2018). https://doi.org/10.1137/16M1095615

[18] Noam Brown and Tuomas Sandholm. 2017. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 360, 6385 (December 2017).

[19] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365, 6456 (2019). https://doi.org/10.1126/science.aay2400 arXiv:https://science.sciencemag.org/content/365/6456/885.full.pdf

[20] Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. 2016. Zero-Sum Polymatrix Games: A Generalization of Minmax. *Mathematics of Operations Research* 41, 2 (2016), 648–655.

[21] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep Blue. *Artificial intelligence* 134, 1-2 (2002).

[22] Haoyang Cao, Xin Guo, and Mathieu Laurière. 2020. Connecting GANs, MFGs, and OT. *arXiv preprint arXiv:2002.04112* (2020).

[23] Pierre Cardaliaguet. 2012. Notes on mean field games. *P.-L. Lions' Lectures at Collège de France* (2012).

[24] Pierre Cardaliaguet and Saeed Hadikhanloo. 2017. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations* 23, 2 (2017).

[25] Elisabetta Carlini and Francisco J. Silva. 2014. A fully discrete semi-Lagrangian scheme for a first order mean field game problem. *SIAM J. Numer. Anal.* 52, 1 (2014). https://doi.org/10.1137/120902987

[26] Elisabetta Carlini and Francisco J. Silva. 2015. A semi-Lagrangian scheme for a degenerate second order mean field game system. *Discrete and Continuous Dynamical Systems* 35, 9 (2015). https://doi.org/10.3934/dcds.2015.35.4269

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

[28] René Carmona and Mathieu Laurière. 2019. Convergence Analysis of Machine Learning Algorithms for the Numerical Solution of Mean Field Control and Games: I–The Ergodic Case. *arXiv preprint arXiv:1907.05980* (2019).

[29] René Carmona and Mathieu Laurière. 2019. Convergence Analysis of Machine Learning Algorithms for the Numerical Solution of Mean Field Control and Games: II–The Finite Horizon Case. *arXiv preprint arXiv:1908.01613* (2019).

[30] Nicolò Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

[31] Jean-François Chassagneux, Dan Crisan, François Delarue, et al. 2019. Numerical method for FBSDEs of McKean–Vlasov type. *The Annals of Applied Probability* 29, 3 (2019).

[32] Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. 2019. The route to chaos in routing games: When is Price of Anarchy too optimistic? *arXiv preprint arXiv:1906.02486* (2019).

[33] Vincent Conitzer and Tuomas Sandholm. 2011. Expressive markets for donating to charities. *Artif. Intell.* 175, 7-8 (2011), 1251–1271.

[34] Romain Couillet, Samir M Perlaza, Hamidou Tembine, and Mérouane Debbah. 2012. Electrical vehicles in the smart grid: A mean field game analysis. *IEEE Journal on Selected Areas in Communications* 30, 6 (2012).

[35] Kai Cui and Heinz Koeppl. 2021. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1909–1917.

[36] Boualem Djehiche, Alain Tcheukam, and Hamidou Tembine. 2017. A mean-field game of evacuation in multilevel building. *IEEE Trans. Automat. Control* 62, 10 (2017).

[37] Romuald Elie, Julien Pérolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. 2020. On the convergence of model free learning in mean field games. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (2020), 7143–7150.

[38] Jean-Pierre Fouque and Zhaoyu Zhang. 2020. Deep Learning Methods for Mean Field Control Problems With Delay. *Frontiers in Applied Mathematics and Statistics* 6 (2020). https://doi.org/10.3389/fams.2020.00011

[39] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. 2020. Adapting a kidney exchange algorithm to align with human values. *Artif. Intell.* 283 (2020), 103261.

[40] Yoav Freund and Robert E Schapire. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior* 29, 1-2 (1999), 79–103.

[41] Maximilien Germain, Joseph Mikael, and Xavier Warin. 2019. Numerical resolution of McKean-Vlasov FBSDEs using neural networks. *arXiv preprint arXiv:1909.12678* (2019).

[42] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[43] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2019. Learning mean-field games. In *Proceedings of NeurIPS*.

[44] Saeed Hadikhanloo. 2017. Learning in anonymous nonatomic games with applications to first-order mean field games. *arXiv preprint arXiv:1704.00378* (2017).

[45] Saeed Hadikhanloo and Francisco J. Silva. 2019. Finite mean field games: fictitious play and convergence to a first order continuous mean field game. *Journal de Mathématiques Pures et Appliquées (9)* 132 (2019). https://doi.org/10.1016/j.matpur.2019.02.006

[46] 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 and Systems* 6, 3 (2006). http://projecteuclid.org/euclid.cis/1183728987

[47] Hassan K Khalil. 2002. *Nonlinear systems; 3rd ed.* Prentice-Hall. https://cds.cern.ch/record/1173048

[48] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese Journal of Mathematics* 2, 1 (2007). https://doi.org/10.1007/s11537-007-0657-8

[49] Alex Tong Lin, Samy Wu Fung, Wuchen Li, Levon Nurbekyan, and Stanley J Osher. 2020. APAC-Net: Alternating the Population and Agent Control via Two Neural Networks to Solve High-Dimensional Stochastic Mean Field Games. *arXiv preprint arXiv:2002.10113* (2020).

[50] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. 2018. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2703–2717.

[51] Panayotis Mertikopoulos and William H Sandholm. 2016. Learning in games via reinforcement and regularization. *Mathematics of Operations Research* 41, 4 (2016), 1297–1324.

[52] Rajesh K Mishra, Deepanshu Vasal, and Sriram Vishwanath. 2020. Model-free Reinforcement Learning for Non-stationary Mean Field Games. In *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 1032–1037.

[53] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356, 6337 (2017).

[54] AS Nemirovsky and DB Yudin. 1979. Problem Complexity and Optimization Method Efficiency. *M.: Nauka* (1979).

[55] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory*. Cambridge University Press, USA.

[56] Abraham Othman, David M. Pennock, Daniel M. Reeves, and Tuomas Sandholm. 2013. A Practical Liquidity-Sensitive Automated Market Maker. *ACM Trans. Economics and Comput.* 1, 3 (2013), 14:1–14:25.

[57] Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. 2017. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *Advances in Neural Information Processing Systems*. 5872–5882.

[58] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. 2021. From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization. In *Proc. of ICML*, Marina Meila and Tong Zhang (Eds.). 8525–8535.

[59] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. 2020. Fictitious play for mean field games: Continuous time analysis and applications. *Proc. of NeurIPS* (2020).

[60] Steve Phelps, Wing Lon Ng, Mirco Musolesi, and Yvan I. Russell. 2018. Precise time-matching in chimpanzee allogrooming does not occur after a short delay. *PLOS One* 13, 9 (2018).

[61] Georgios Piliouras and Jeff S Shamma. 2014. Optimization despite chaos: Convex relaxations to complex limit sets via Poincaré recurrence. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 861–873.

[62] C. M. Postlethwaite and A. M. Rucklidge. 2017. Spirals and heteroclinic cycles in a spatially extended Rock-Paper-Scissors model of cyclic dominance. *EPL (Europhysics Letters)* 117, 4 (Feb 2017), 48006. https://doi.org/10.1209/0295-5075/117/48006

[63] Julia Robinson. 1951. An iterative method of solving a game. *Annals of mathematics* (1951).

[64] Tim Roughgarden. 2009. Intrinsic robustness of the price of anarchy. In *Proc. of STOC*. 513–522.

[65] Lars Ruthotto, Stanley J Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. 2020. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences* 117, 17 (2020).

[66] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. 2002. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences* 99, 7 (2002), 4748–4751. https://doi.org/10.1073/pnas.032086299 arXiv:http://www.pnas.org/content/99/7/4748.full.pdf+html

[67] Shai Shalev-Shwartz et al. 2011. Online learning and online convex optimization. *Foundations and trends in Machine Learning* 4, 2 (2011), 107–194.

[68] Harold N. Shapiro. 1958. Note on a computation method in the theory of games. In *Communications on Pure and Applied Mathematics*.

[69] Yoav Shoham. 1993. Agent-oriented programming. *Artificial Intelligence* 60, 1 (1993), 51–92. https://doi.org/10.1016/0004-3702(93)90034-9

[70] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016).

[71] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 632, 6419 (2018).

[72] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017).

[73] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. 2011. On the universality of online mirror descent. *arXiv preprint arXiv:1107.4080* (2011).

[74] Jayakumar Subramanian, Raihan Seraj, and Aditya Mahajan. 2018. Reinforcement learning for mean-field teams. In *Workshop on Adaptive and Learning Agents at International Conference on Autonomous Agents and Multi-Agent Systems*.

[75] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press.

[76] A Szolnoki, BF de Oliveira, and D Bazeia. 2020. Pattern formations driven by cyclic interactions: A brief review of recent developments. *EPL (Europhysics Letters)* 131, 6 (2020), 68001.

[77] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019).

[78] Michael Wooldridge and Nicholas R. Jennings. 1995. Agent theories, architectures, and languages: A survey. In *Intelligent Agents*, Michael J. Wooldridge and Nicholas R. Jennings (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–39.

[79] Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. 2020. Provable Fictitious Play for General Mean-Field Games. *arXiv preprint arXiv:2010.04211* (2020).

[80] Huibing Yin, Prashant G Mehta, Sean P Meyn, and Uday V Shanbhag. 2010. Learning in mean-field oscillator games. In *49th IEEE Conference on Decision and Control (CDC)*. IEEE.

[81] Zhengyuan Zhou, Panayotis Mertikopoulos, Aris L Moustakas, Nicholas Bambos, and Peter Glynn. 2017. Mirror descent learning in continuous games. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 5776–5783.

[82] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2008. Regret minimization in games with incomplete information. In *Proceedings of NeurIPS*.