Learning Equilibria in Mean-Field Games: Introducing Mean-Field PSRO

Paul Muller
DeepMind - Paris
pmuller@deepmind.com

Mark Rowland
DeepMind - London

Romuald Elie
DeepMind - Paris

Georgios Piliouras
Singapore UTD

Julien Perolat
DeepMind - Paris

Mathieu Lauriere
Google Brain - Paris

Raphael Marinier
Google Brain - Paris

Olivier Pietquin
Google Brain - Paris

Karl Tuyls
DeepMind - Paris

ABSTRACT

Recent advances in multiagent learning have seen the introduction of a family of algorithms that revolve around the population-based training method PSRO, showing convergence to Nash, correlated and coarse correlated equilibria. Notably, when the number of agents increases, learning best-responses becomes exponentially more difficult, and as such hampers PSRO training methods. The field of mean-field games provides an asymptotic solution to this problem when the considered games are anonymous-symmetric. Unfortunately, the mean-field approximation introduces non-linearities which prevent a straightforward adaptation of PSRO. Building upon optimization and adversarial regret minimization, this paper sidesteps this issue and introduces mean-field PSRO, an adaptation of PSRO which learns Nash, coarse correlated and correlated equilibria in mean-field games. The key is to replace the exact distribution computation step by newly-defined mean-field no-adversarial-regret learners, or by black-box optimization. We compare the asymptotic complexity of the approach to standard PSRO, greatly improve empirical bandit convergence speed by compressing temporal mixture weights, and ensure it is theoretically robust to payoff noise. Finally, we illustrate the speed and accuracy of mean-field PSRO on several mean-field games, demonstrating convergence to strong and weak equilibria.

KEYWORDS

Multiagent, mean-field, Reinforcement Learning, Game Theory, correlated equilibrium, coarse correlated equilibrium

ACM Reference Format:
Paul Muller, Mark Rowland, Romuald Elie, Georgios Piliouras, Julien Perolat, Mathieu Lauriere, Raphael Marinier, Olivier Pietquin, and Karl Tuyls. 2022. Learning Equilibria in Mean-Field Games: Introducing Mean-Field PSRO. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 18 pages.

1 INTRODUCTION

This paper introduces a new mean-field reinforcement learning algorithm, Mean-Field Policy Space Response Oracles (MF-PSRO), guaranteed to converge to Nash, correlated and coarse-correlated equilibria in a large variety of games, without any hypothesis thereupon. Policy Space Response Oracles (PSRO) [20] is originally a two-player zero-sum game algorithm meant to be a generalization of double-oracle [25], fictitious play [6], and independent reinforcement learning [24]. The algorithm's main loop is composed of two steps: given a policy set, compute an optimal distribution of play. Then, compute a best-response to this distribution, add it to the set and re-iterate. Remarkably, recent years have shown the algorithm's versatility by demonstrating great advances in learning N-player equilibria using PSRO-derived approaches, managing to converge towards α-Rank [30, 34] optimal strategy cycles [28], or towards (coarse) correlated equilibria[1] [23]. However, both the latter two approaches' convergence results rely on potentially fully exploring the space of deterministic strategies, which grows exponentially in the number of players. Computing a best response in the general case of randomized opponent strategies also becomes exponentially more complex as the number of players increases, even with symmetric simplifications such as anonymity [37], centralized settings [19], or fully cooperative settings [31]. Although anonymity can allow Polynomial-time Approximation Schemes for computing approximate Nash equilibria [9, 10], in practice such algorithms are typically too slow for real life applications. A more promising way to address such complexity issues is by approximation in the case of symmetric games by considering asymptotic versions thereof, where the number of players is infinite and only their distribution matters: mean-field games [17, 22].

The question of learning Nash equilibria in mean-field games has been receiving a growing amount of attention, and many methods have been recently proposed. Among these, we can distinguish those relying on fixed-point contraction [2, 14, 38], fictitious-play [8, 12, 33] or online mirror descent [32]. Comparatively, learning correlated and coarse correlated equilibria in mean-field games has not yet, to the best of our knowledge, been studied. The literature has only started introducing notions of mean-field correlated and coarse correlated equilibria [7, 11]. However, learning (coarse) correlated equilibria, notably in Mean-Field Games, is a promising way to tackle very difficult assignment problems (Routing of car or network traffic, energy prosumer trade storage strategies, and even, more broadly, mechanism design problems in general) in a fair and

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.

[1] A broad relaxation of Nash, correlated equilibria are closely connected to regret minimization. They are sometimes referred to as Hannan consistency [16, 27, 29].

non-coercive way (Every actor has interest in following the central authority's recommendation). These equilibria are also easier to learn than Nash equilibria in N-player games, and can be straightforwardly approximated using adversarial no-regret learners [5] or PSRO-like algorithms [23].

Our central question is: *What are the modifications required for PSRO to successfully converge towards Nash, correlated and coarse correlated equilibria in mean-field games?*

In order to answer it, after introducing the framework of interest (Section 2), we expand on the obstacles encountered when attempting to adapt PSRO to mean-field games (Section 3), identify and treat the cases where a straightforward adaptation is possible, then consider all cases without hypothesis on games. Note that the general treatment is fundamentally different for Mean-Field Nash equilibria (Section 4), and for Mean-Field (coarse) correlated equilibria (Section 5). Finally, we test our algorithms on a number of OpenSpiel [21] games in Section 6, demonstrating convergence, and, where possible, comparing with alternative benchmarks.

2 BACKGROUND

2.1 Definitions

A game is a set $(\mathcal{X}, \mathcal{A}, r, P, \mu_0)$ where $\mathcal{X}$ is the finite set of states, $\mathcal{A}$ is the finite set of actions, $r : \mathcal{X} \times \mathcal{A} \times \Delta(\mathcal{X}) \to \mathbb{R}$ is the reward function where $\Delta(\mathcal{X})$ is the set of distributions over $\mathcal{X}$, $p : \mathcal{X} \times \mathcal{A} \times \Delta(X) \to \mathcal{X}$ is the state transition function, $\mu_0 \in \Delta(\mathcal{X})$ is the initial state occupancy measure. Given a set $\mathcal{Y}$, we name $\Delta(\mathcal{Y})$ the set of distributions over $\mathcal{Y}$.

A policy is a function $\pi : \mathcal{X} \to \Delta(\mathcal{A})$. We write $\pi(x, a)$ the probability of playing action a under policy π at state x. We also consider the special case of deterministic policies, which are of the form $\forall x \in \mathcal{X},\ \exists a \in \mathcal{A},\ \pi(x) = \delta_a$, or $\pi(x, a') = \mathbb{1}_{a=a'}$. We take Π to be the set of deterministic policies, which is *finite* and whose convex hull spans all policies.

We name J the expected payoff function

$$J(\pi, \mu) := \sum_{x \in \mathcal{X}, a \in \mathcal{A}} \mu^{\pi}(x)\pi(x, a)r(x, a, \mu)$$

where μ^{π} is the expected state occupancy measure of a representative player playing policy π. State occupancy measures can be defined in several ways, and our derivations apply to all:

- γ-discounted: $\mu^{\pi}(x) = \mu_0(x) + \gamma \sum_{x' \in \mathcal{X}} \sum_{a \in \mathcal{A}} p(x|x', a, \mu^{\pi})\pi(x', a)\mu^{\pi}(x')$
- Finite-horizon: $\mu^{\pi}_{t+1}(x) = \sum_{x' \in \mathcal{X}} \sum_{a \in \mathcal{A}} p(x|x', a, \mu^{\pi}_t)\pi(x', a)\mu^{\pi}_t(x')$ with $\mu^{\pi}_0 = \mu_0$ (in which case another summation term over t appears in J).

Given policies $\pi_1, ..., \pi_n \in \Pi$, we call **restricted game** the stateless game where players choose one policy among $\{\pi_i | 1 \leq i \leq n\}$ at the beginning of the game, then keep playing it until the end.

We also define **meta-games**, which are normal-form games whose payoff matrix for player 1 is, at row i and column j, $J(\pi_i, \mu^{\pi_j})$ - and the transpose thereof for player 2. The complex relationship between these notions, which are equivalent in N-player games, is explored in Section 3.

A **correlation device** ρ is a distribution over distributions of policies: $\rho \in \Delta(\Delta(\Pi))$, where $\Delta(\Pi)$ is the set of distribution over Π. It is used to sample population distributions $\nu \in \Delta(\Pi)$, from which individual population recommendations π are in turn sampled: the distribution of policies over the whole mean-field population follows ν with probability $\rho(\nu)$. Given a sequence of distributions $(\nu_t)_t$ and a distribution $(\rho_t)_t$ over them, we write $(\rho_t, \nu_t)_t$ the correlation device recommending ν_t with probability ρ_t.

The **empirical play** of a sequence $\nu_1, ..., \nu_T$ is the correlation device which uniformly selects one of the joint members of the sequence: $\forall 1 \leq t \leq T,\ \rho(\nu_t) = \frac{1}{T}$.

We write $\mu(\nu)$ the **state occupancy measure** of the population when policies are distributed according to ν. We also write $\pi(\nu)$ the stochastic policy resulting from sampling an initial policy according to ν and playing it until the end of the game. Directly, $\mu^{\pi(\nu)} = \mu(\nu)$.

2.2 Mean-field equilibria

We define three notions of mean-field equilibrium: Nash, coarse-correlated, and correlated equilibrium.

Definition 1 (mean-field Nash equilibrium). *A **mean-field Nash equilibrium (MFNE)** is a policy π such that, when the whole population plays π, no agent has an incentive to deviate, ie.*

$$J(\pi', \mu^{\pi}) - J(\pi, \mu^{\pi}) \leq 0\,,\ \ \forall \pi' \in \Pi\,.$$

Our notions of correlated and coarse correlated equilibria are different from that of [7] by the fact that our correlation device samples population distributions $\nu \in \Delta(\Pi)$, ie. a distribution of policies which the population plays, and each agent the gets a sampled recommendation from ν.

Definition 2 (mean-field coarse-correlated equilibrium). *A **mean-field coarse-correlated equilibrium (MFCCE)** is a correlation device ρ from which players do not have an incentive to deviate before being given their recommendations, ie.*

$$\mathbb{E}_{\nu \sim \rho, \pi \sim \nu}\left[J(\pi', \mu(\nu)) - J(\pi, \mu(\nu))\right] \leq 0\,,\ \ \forall \pi' \in \Pi.$$

Definition 3 (mean-field correlated equilibrium). *A **mean-field correlated equilibrium (MFCE)** is a correlation device ρ such that players do not have an incentive to deviate even after being given their recommendations, ie.*

$$\rho(\pi)\mathbb{E}_{\nu \sim \rho(\cdot|\pi)}\left[J(\pi', \mu(\nu)) - J(\pi, \mu(\nu))\right] \leq 0\,,\ \ \forall \pi, \pi' \in \Pi$$

where $\rho(\pi) = \sum_{\nu} \nu(\pi)\rho(\nu)$.

ϵ-variants of these equilibria are defined by changing 0 in the above inequalities by $\epsilon > 0$: these are approximate equilibria where one may only gain up to ϵ by deviating.

When applied to restricted games, we call these equilibria restricted MFNE, restricted MFCE and restricted MFCCE respectively.

2.3 PSRO in N-player games

PSRO [20] is a generalization of Double Oracle [26], and as such is an iterated best-response algorithm for computing Nash equilibria in N-player games. The algorithm, presented in Algorithm 1 initiates with sets containing random policies. At each iteration, an optimal policy distribution is computed over the policy sets, and a best response to this distribution is computed for each player. If all best responses were already in each player's policy set, the algorithm terminates; it continues otherwise.

The original PSRO paper introduced several different meta-solvers (Uniform, Exact Nash and PRD, an approximate Nash solver), all of

Algorithm 1: PSRO(Meta-Solver) (N-player games)

Result: Policy sets $(\Pi_k^* = \{\pi_1^k, ..., \pi_n^k\})_{k=1..K}$ for all K players, policy distributions $(\nu_k^*)_{k=1..N}$

$\forall k,\ \Pi_k^1 = \{\pi_k^1\}$ with π_k^1 any policy, $\nu_k(\pi_k^1) = 1.0$, $n = 1$;

while $(\Pi_{n+1} \setminus \Pi_n) \neq \emptyset$ **do**

  $\forall k,\ \Pi_k^{n+1} = \Pi_k^n \cup \{BR_k(\nu)\}$;

  $n = n + 1$;

  Fill payoff tensors $(T_k)_{k=1..K}$:

  $\forall x_1, ..., x_K, T_k(x_1, ..., x_K) = \text{Payoff}_k(\pi_{x_1}, ..., \pi_{x_K})$;

  ν = Meta-Solver$((T_k)_{k=1..K})$

end

which were proven to make PSRO converge to a Nash equilibrium in two-player zero-sum games. Recent work has extended convergence to Alpharank [30]-optimal subsets [28] and to correlated and coarse correlated equilibria [23] in N-player games when using the right meta-solvers and best-responders. Crucially, the game specified by the payoff tensors that the meta-solver computes an equilibrium form is a normal-form matrix game. This yields a 'linearity of evaluation' property; specifically, the payoffs when players make use of mixed strategies are straightforwardly computed from the payoff tensors specifying the payoffs of the pure strategies in the game, non-rigorously, we have $\text{Payoff}(\alpha\pi_1 + (1-\alpha)\pi_2) = \alpha\text{Payoff}(\pi_1) + (1-\alpha)\text{Payoff}(\pi_2)$, where $\alpha\pi_1 + (1-\alpha)\pi_2$ is a joint policy playing π_1 with frequency α, and π_2 with frequency $(1-\alpha)$.

In the rest of this paper, unless otherwise directly specified, we consider n to be the current PSRO iteration.

3 CHALLENGES IN SCALING TO MEAN-FIELD GAMES

Our central proposal in this paper is a generalisation of PSRO to the mean-field setting. We introduce two distinct algorithms for the computation of either MFNE or MFCE/MFCCE. Both MF-PSRO algorithms are described as Algorithms 2 and 3 below.

Algorithm 2: MF-PSRO(Nash)

Result: Policy set $\Pi^* = \{\pi_1, ..., \pi_n\}$, Policy Distribution $\nu^* \in \Delta(\Pi^*)$ yielding game Nash $\pi(\nu^*)$

$\Pi_1 = \{\pi_1\}$ with π_1 any policy, $\nu_1(\pi_1) = 1.0$, $n = 1$;

while $(\Pi_{n+1} \setminus \Pi_n) \neq \emptyset$ **do**

  $\Pi_{n+1} = \Pi_n \cup \{BR(\mu^{\pi(\nu_n)})\}$;

  $n = n + 1$;

  $\nu_n = \underset{\nu \in \Delta(\Pi_n)}{\arg\min} \max_{i=1,...,n} J(\pi_i, \mu(\nu)) - J(\pi(\nu), \mu(\nu))$;

end

These two algorithms have a very similar structure to the PSRO as described for N-player games in Section 2.3; within the inner loop, a distribution is computed for the restricted game under consideration (either a Nash equilibrium, or a (coarse) correlated equilibrium), and new policies are derived as certain types of best response against the computed equilibrium. Keeping the same insight as [23], we define two different Best Responder functions BR_{CE} and BR_{CCE}, for use with MF-PSRO in computing CEs and CCEs, respectively:

Algorithm 3: MF-PSRO((C)CE)

Result: Policy set $\Pi^* = \{\pi_1, ..., \pi_n\}$, ϵ-mean-field correlated equilibrium $\rho^* \in \Delta(\Delta(\Pi^*))$

$\Pi_0 = \emptyset$, $\Pi_1 = \{\pi_1\}$ with π_1 any policy, $\rho(\delta_{\pi_1}) = 1.0$, $n = 1$;

while $(\Pi_{n+1} \setminus \Pi_n) \neq \emptyset$ **do**

  (If CE) $\Pi_{n+1} = \Pi_n \cup \{BR_{CE}(\pi_i, \rho_n) \mid \pi_i,\ \rho_n(\pi_i) > 0\}$;

  (If CCE) $\Pi_{n+1} = \Pi_n \cup BR_{CCE}(\rho_n)$;

  $n = n + 1$;

  (If CE) $\rho_n =$

  $\underset{\rho \in \Delta(\Delta(\Pi_n))}{\arg\min}\ \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}[\max_{i=1..n} J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$;

  (If CCE) $\rho_n =$

  $\underset{\rho \in \Delta(\Delta(\Pi_n))}{\arg\min} \max_{i=1,...,n} \mathbb{E}_{\nu\sim\rho,\pi\sim\nu}[J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$;

end

- $BR_{CCE}(\rho) := \underset{\pi^* \in \Pi}{\arg\max} \sum_\nu \rho(\nu) J(\pi^*, \mu(\nu))$;
- $BR_{CE}(\pi_k, \rho) := \underset{\pi^* \in \Pi}{\arg\max} \sum_\nu \rho(\nu|\pi_k) J(\pi^*, \mu(\nu))$.

We note that $BR_{CCE}(\rho)$ is the Best Response corresponding to a unilateral deviation from ρ, ie. deviating before having been given a recommendation, whereas $BR_{CE}(\pi_k, \rho)$ is the best response generated by deviating from recommendation π_k.

Given these proto-algorithms, several important questions are immediately raised. First, are these algorithms guaranteed to return instances of the equilibria they seek to find? This is a purely mathematical question. Second, how should the restricted game equilibria in the inner loop be computed? As described in Section 2.3, the restricted game in usual applications of PSRO satisfies a 'linearity of evaluation' property. However, this linearity property is lost in the case of mean-field games, in which the representative player's payoff is generally non-linear as a function of the population occupancy measure; to take the same example as Section 2.3, in general, $J(\alpha\pi_1 + (1-\alpha)\pi_2, \mu) = \alpha J(\pi_1, \mu) + (1-\alpha) J(\pi_2, \mu)$, but due to potential non-linearity of r, and thus of J, in μ, and to dependence of p in μ, $J(\pi, \mu^{\alpha\pi_1 + (1-\alpha)\pi_2}) \neq \alpha J(\pi, \mu^{\pi_1}) + (1-\alpha) J(\pi, \mu^{\pi_2})$. We provide an example of this non-linearity in Appendix C. This presents a serious barrier in directly applying PSRO to mean-field games, and an important contribution of this paper is how to circumvent this barrier. We do however note that for a limited class of mean-field games, linearity is preserved; we describe the details of this case in Appendix D.

The next two sections treat the theoretical and implementation questions raised above for Nash equilibria, and for (coarse) correlated equilibria, in turn.

4 CONVERGENCE TO NASH EQUILIBRIA

4.1 Existence and computation of restricted game equilibria

In the inner loop of MF-PSRO(Nash), an important subroutine is the computation of a mean-field Nash equilibrium for the restricted

game; namely, a distribution $\nu \in \Delta(\Pi_n)$ such that

$$J(\pi', \mu(\nu)) - J(\pi(\nu), \mu(\nu)) \leq 0\,, \quad \forall \pi' \in \{\pi_1, ..., \pi_n\}.$$

We note that if at least one such ν exists, then the following optimization problem in the inner loop of MF-PSRO(Nash), which minimizes exploitability, will return a Nash equilibrium

$$\nu^* = \arg\min_{\nu \in \Delta_n} \max_{i=1...n} J(\pi_i, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\,. \tag{1}$$

Fortunately, the conditions of existence for a Nash equilibrium of the restricted game - so called restricted Nash equilibrium - only require continuity of r with respect to μ, as shown in the following theorem.

Theorem 4 (Existence of restricted Nash equilibria). *If the reward function of the game is continuous with respect to μ, then there always exists a restricted game Nash equilibrium.*

Proof. Let $\phi : \Delta(\Pi_n) \to 2^{\Delta(\Pi_n)}$ be the best-response map in the restricted game characterized by policies in the set Π_n:

$$\forall \nu \in \Delta(\Pi_n), \quad \phi(\nu) := \arg\max_{\nu' \in \Delta(\Pi_n)} J(\pi(\nu'), \mu(\nu)).$$

$\Delta(\Pi_n)$ is non-empty and convex, together with closed and bounded in a finite-dimensional space, and therefore compact.

For all $\nu \in \Delta(\Pi_n)$, $\arg\max_{\nu' \in \Delta(\Pi_n)} J(\pi(\nu'), \mu(\nu)) \subseteq \Delta(\Pi_n)$ because $\Delta(\Pi_n)$ is closed, and $\phi(\nu)$ is therefore non-empty.

Let $\nu_1,\ \nu_2 \in \phi(\nu)$, $t \in [0, 1]$.

$$J(\pi(t\nu_1 + (1-t)\nu_2), \mu(\nu)) = tJ(\pi(\nu_1), \mu(\nu)) + (1-t)J(\pi(\nu_2), \mu(\nu))$$

so $t\nu_1 + (1-t)\nu_2 \in \phi(\nu)$ and $\phi(\nu)$ is therefore convex.

The proof of Graph(ϕ) being closed is provided in Appendix A. It relies on the fact that since r is continuous in μ, so is J, and since the function $\nu \to J(\pi(\nu), \mu)$ is linear for all $\nu \in \Delta(\Pi_n)$, the function $(\nu_1, \nu_2) \to J(\pi(\nu_1), \mu(\nu_2))$ is bicontinuous, which is enough to ensure Graph closedness. We have all the hypotheses required to apply Kakutani's fixed point theorem [18]: there thus exists $\nu^* \in \Delta(\Pi_n)$ such that $\nu^* \in \phi(\nu^*)$, ie. $\nu^* = \arg\max_{\nu'} J(\pi(\nu'), \mu(\nu^*))$, which means that $\forall \nu' \in \Delta(\Pi_n), J(\pi(\nu'), \mu(\nu^*)) \leq J(\pi(\nu^*), \mu(\nu^*))$, in other words: ν^* is a Nash equilibrium of the restricted game. □

Having established the existence of Nash equilibria for the restricted mean-field game in the inner loop of MF-PSRO(Nash), we now turn to the problem of how such an equilibrium can be (approximately) computed. As remarked earlier, due to the non-linearity of the restricted game, this problem is a non-linear (and potentially non-convex) optimisation problem over $\Delta(\Pi_n)$. Thus, the optimal solution of Equation (1) can be, in the absence of any additional assumptions on the game, found via Black-Box optimization approaches, such as random search [35], Bayesian optimization [13], evolutionary search (our experiments use CMA-ES [15]), or any other appropriate method for the considered game.

4.2 Convergence to Nash

The termination condition of PSRO is the following: if at step $N+1$, the new policy π_{n+1} produced by the algorithm is in Π_n, then the algorithm terminates. Given that each π_i is a deterministic policy, and that the set of deterministic policies is finite, PSRO will therefore necessarily terminate. We must only prove one thing:

Proposition 5 (Termination-optimality). *If MF-PSRO(Nash) terminates, it stops at a Nash equilibrium of the true game.*

Proof. If MF-PSRO(Nash) terminates at step n, then

$$\pi^* = \arg\max_{\pi \in \Pi} J(\pi, \mu(\nu)) \in \Pi_n\,.$$

Since ν is a Nash equilibrium of the restricted game by assumption, then necessarily $J(\pi^*, \mu(\nu)) \leq J(\pi(\nu), \mu(\nu))$, and thus $\forall \pi \in \Pi, J(\pi, \mu(\nu)) \leq J(\pi(\nu), \mu(\nu))$, which concludes the proof. □

Using the former discussion and this property, we deduce

Theorem 6 (mean-field PSRO convergence to Nash equilibria). *MF-PSRO(Nash) converges to a Nash equilibrium of the true game.*

5 CONVERGENCE TO (COARSE) CORRELATED EQUILIBRIA

We now turn our attention to the versions of MF-PSRO that aim to compute mean-field correlated equilibria and mean-field coarse correlated equilibria.

5.1 Overview

Computing restricted MF(C)CEs is potentially more involved than computing restricted MFNE; while the optimisation problem defining restricted Nash equilibria is over the finite-dimensional space $\Delta(\Pi_n)$, the optimisation problem defining restricted MF(C)CEs is over the infinite-dimensional space $\Delta(\Delta(\Pi_n))$. One could resort to computing an approximate MFNE (a special case of both MFCE and MFCCE) using the black-box optimisation approach described in the previous section, but it is possible to exploit the structure of the mean-field game to compute approximate MF(C)CEs more efficiently. The approach we pursue is fundamentally based on no-regret learning; we also find opportunities to increase the quality of the approximate equilibrium by post-processing the output of the regret-minimisation algorithm via linear programming; see Figure 2 for an overview of the techniques at play.

5.2 Approximate (coarse) correlated equilibria via regret minimisation

Our goal is to approximate an MF(C)CE for the restricted MFG based on the policy set $\Pi_n = \{\pi_1, \ldots, \pi_n\}$, as required within the inner loop of Algorithm 3. Recall that this amounts to solving the optimisation problem

$$\rho_n = \arg\min_{\rho \in \Delta(\Delta(\Pi_n))} \max_{i=1,\ldots,n} \mathbb{E}_{\nu \sim \rho, \pi \sim \nu}[J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$$

in the case of coarse correlated equilibria, and

$$\rho_n = \arg\min_{\rho \in \Delta(\Delta(\Pi_n))} \mathbb{E}_{\nu \sim \rho, \pi \sim \nu}[\max_{i=1..n} J(\pi_i, \mu(\nu)) - J(\pi, \mu(\nu))]$$

in the case of correlated equilibria. In principle, similar black-box techniques described for approximating Nash equilibria in the previous section may be applied to solve these problems too. However, such an approach is likely to be inefficient in practice, and instead we build on regret-minimisation theory, a classical approach to computing (C)CEs in N player games.

The overall approach relies on the fact that if the population distribution μ is fixed, the payoff function $\mathbb{E}_{\pi\sim\nu}[J(\pi,\mu)]$ is linear in the distribution $\nu \in \Delta(\Pi_n)$, and we are in fact considering online linear optimisation problems. Focusing first on the case of coarse correlated equilibria, we will make use of Algorithms **A** achieving $O(\sqrt{T})$ external regret in online linear optimisation, of the form described in Algorithm 4.

We may apply such an algorithm for MF(C)CE computation as shown in Algorithm 5.

Algorithm 4: Generic form of regret-minimisation algorithm for online linear optimisation on the domain $\Delta(\Pi_n)$.

Result: A sequence of predictions $(\nu_t)_{t=1}^T$ such that $\max_{\nu\in\Delta(\Pi_n)} \sum_{t=1}^T R_t(\nu) - \sum_{t=1}^T R_t(\nu_t) = O(\sqrt{T})$.
for $t = 1, 2, \ldots, T$ **do**
  Algorithm makes a prediction $\nu_t \in \Delta(\Pi_n)$;
  Algorithm observes a linear reward function $R_t : \Delta(\Pi_n) \to \mathbb{R}$;
  Algorithm receives the reward $R_t(\nu_t)$;
end

Algorithm 5: Protocol for computing an approximate MF(C)CE via regret-minimisation

for $t = 1, 2, \ldots, T$ **do**
  Representative player selects distribution $\nu_t \in \Delta(\Pi_n)$ using a regret-minimisation algorithm $\mathbb{A}$ based on past loss function $(R_s)_{s=1}^{t-1}$;
  Player observes reward function $R_t(\nu) = \mathbb{E}_{\pi\sim\nu}[J(\pi, \mu(v_t)]$;
  Representative player receives reward $R_t(\nu_t) = \mathbb{E}_{\pi\sim\nu_t}[J(\pi, \mu(v_t)]$;
end
Return empirical average $\rho = \frac{1}{T}\sum_{t=1}^T \delta_{\nu_t}$.

Algorithm 5 returns the empirical average $\frac{1}{T}\sum_{t=1}^T \delta_{\nu_t}$, which is in fact an approximate MF(C)CE for the restricted game, as the following result shows.

Proposition 7. *The empirical average $\rho = \frac{1}{T}\sum_{t=1}^T \delta_{\nu_t}$ returned by Algorithm 5 using a regret-minimisation algorithm $\mathbb{A}$ of the form described in Algorithm 4, is a $O(1/\sqrt{T})$-MF(C)CE for the restricted mean-field game.*

Proof. This is a direct computation. The benefit of the representative player deviating to π_i under the correlation device ρ is

$$\begin{aligned}
&\mathbb{E}_{\nu\sim\rho}[J(\pi_i, \mu(\nu)) - \mathbb{E}_{\pi\sim\nu}[J(\pi, \mu(\nu))]] \\
=&\frac{1}{T}\sum_{t=1}^T \left(J(\pi_i, \mu(\nu_t)) - \mathbb{E}_{\pi\sim\nu_t}[J(\pi, \mu(\nu_t))]\right) \\
=&\frac{1}{T} O\left(\sqrt{T}\right) = O\left(1/\sqrt{T}\right),
\end{aligned}$$

where the penultimate equality follows from the regret-minimising property of algorithm $\mathbb{A}$. The proof for CEs is similar. □

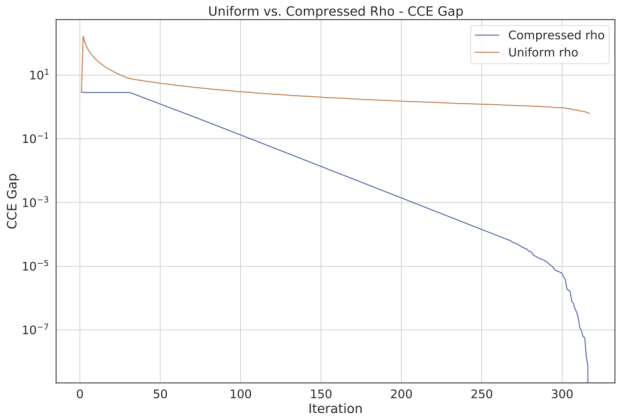

Figure 1: Uniform vs. Compressed ρ - CCE Gap / Time

This result establishes a rigorous means of approximating an MF(C)CE in the restricted game considered within the inner loop of mean-field PSRO, and therefore provides an implementable version of mean-field PSRO. By strengthening the regret minimisation algorithm described above to minimise *internal* regret, we obtain a time-average strategy that is an approximate MFCE. In both cases, we have the following correctness guarantee for MF-PSRO.

Theorem 8 (MF-PSRO Convergence to MF(C)CEs). *MF-PSRO using a no-internal-regret (Respectively no-external-regret) algorithm to compute its MFCE (Respectively MFCCE) with average regret threshold ϵ and Best-Response Computation BR_{CE} (Respectively BR_{CCE}) converges to an ϵ-MFCE (Respectively an ϵ-MFCCE).*

Proof. Based on previous discussions, we know that PSRO must necessarily terminate.

If PSRO terminates when using a restricted MFCCE, we must have

$$\pi^* = \arg\max_\pi \sum_\nu \rho(\nu) J(\pi, \mu(\nu)) \in \Pi_n .$$

By definition of ρ, $\sum_\nu \rho(\nu)\Big(J(\pi^*, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\Big) \leq \epsilon$, and therefore $\forall \pi \in \Pi$, $\sum_\nu \rho(\nu)\Big(J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\Big) \leq \epsilon$, ergo: ρ is a mean-field ϵ-coarse correlated equilibrium.

The proof for mean-field correlated equilibria follows a similar line of arguments and is detailed in Appendix B. □

As we will see in the next section, it is often possible to improve upon the uniform mixture of $(\nu_t)_{t=1}^T$ output by the regret-minimisation algorithm to obtain a more accurate approximation to an MF(C)CE.

5.3 Improving the Bandit: Speed

5.3.1 The No-Regret Speedup Algorithm: Bandit Compression. One could use No-regret learners directly to converge towards MF(C)CE, but their equilibrium contains T different distributions. This potentially means a very high amount of different ν_t recommended by our (C)CE, which can lead to learning difficulties on the part of

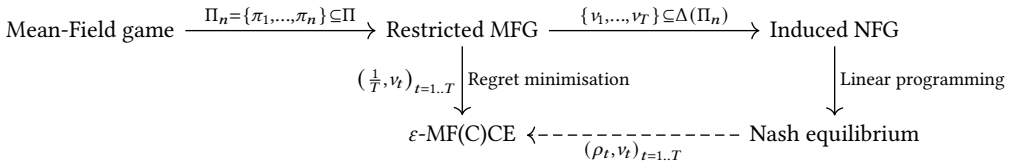

Figure 2: Reductions involved in approximation equilibrium computation in MF-PSRO.

best-responders (since every separate ν must be taken into account), implementation difficulties of equilibria in the real world, and inefficiencies: Indeed, changing per-timestep weights $\frac{1}{T}$ to potentially non-uniform ρ_t can lead to converging to ϵ'-MF(C)CE instead of ϵ ones, with $\epsilon' \ll \epsilon$, which is illustrated in figure 1, computed at the first iteration of PSRO, in the Crowd Modelling [33] game. We define $(\rho_t)_t$ as the optimal solution of the following optimization problem:

$$\min_{\rho} \max_{i} \rho^t \text{Regret}_i \quad \text{s.t. } \forall t \ \ \rho_t \geq 0, \ \ \sum_t \rho_t = 1 \tag{2}$$

with $\text{Regret}_i[t] := J(\pi_i, \mu(\nu_t)) - J(\pi(\nu_t), \mu(\nu_t))$.

We note that Problem (2) can be interpreted as finding the row player's Nash equilibrium distribution in a zero-sum normal-form game whose payoff matrix for player 1 is Regret. We note that this objective can be expressed linearly.

A similar problem can be solved to find better restricted mean-field correlated equilibria. First, define

$$\text{Regret}_{i,j}(t) = \nu_t(i)\Big(J(\pi_j, \mu(\nu_t)) - J(\pi_i, \mu(\nu_t))\Big)$$

The following problem gives optimal temporal weights ρ for restricted mean-field correlated equilibria

$$\min_{\rho} \max_{i,j} \rho^t \text{Regret}_{i,j} \quad \text{s.t. } \forall t \ \ \rho_t \geq 0, \ \ \sum_t \rho_t = 1. \tag{3}$$

This problem can similarly be expressed linearly. The following theorem confirms the optimality of ρ, the solution of Problem (2) or Problem (3):

Theorem 9 (Optimality of ρ). *If $\rho = \frac{1}{T}\sum_{t=1}^{T} \delta_{\nu_t}$ is a restricted ϵ-MFCCE (respectively ϵ-MFCE), then $(\rho_t^*, \nu_t)_t$, with ρ^* the optimal solution of Problem 2 (respectively 3), yields a restricted ϵ'-MF(C)CE of the restricted game, with $\epsilon' \leq \epsilon$; and no other ρ distribution over $(\nu_t)_t$ can yield an ϵ''-MF(C)CE with $\epsilon'' < \epsilon'$.*

Proof. For restricted MFCCEs, the deviation incentive against the correlation device sampling ν_t with probability ρ_t in the restricted game is

$$\mathbb{E}_{\nu\sim\rho,\pi\sim\nu}\left[J(\pi', \mu(\nu)) - J(\pi, \mu(\nu))\right] = \max_i \rho^t \text{Regret}_i \,.$$

Since the uniform distribution is a possible value for ρ, we necessarily have $\max_i \rho^t \text{Regret}_i \leq \max_i \frac{1}{T}\sum_t \text{Regret}_i[t] = \epsilon$, which concludes that part of the proof. The proof for restricted MFCEs follows the same line of arguments, and is detailed in Appendix I.4.

Optimality of the solutions of problems (2) and (3) directly follows from their definitions together with the above derivations. □

Given the empirical tendency of this approach to compress temporal distribution, we name it **bandit compression**. Empirically, it allows us to find much more accurate (figure 1) and sparser (Appendix E) distributions than uniformly averaging over $(\nu_t)_t$, and in a much lower number of steps. Yet, this algorithm is only exact in the case where the regret used by the algorithm is noiseless. The next question is therefore, how sensitive is bandit compression to noise in the regret matrix?

5.3.2 On the value-continuity of min-max problems. We provide bounds on computed Average Regrets differences when J is perturbed by an additive random variable ϵ: $\tilde{J}(\pi, \mu) = J(\pi, \mu) + \epsilon$, giving rise to notation Regret_i^ϵ, and to the identity, if we write $\tilde{\epsilon}_t = \epsilon_t - (\nu_t)^t \epsilon_t$, $\text{Regret}_i^\epsilon = \text{Regret}_i + \tilde{\epsilon}_i$.

We write

$$\text{Regret}_* = \min_\rho \max_i \rho^t \text{Regret}_i, \ \ \text{Regret}_*^\epsilon = \min_\rho \max_i \rho^t \text{Regret}_i^\epsilon$$

We name i_* and ρ_* the terms such that $\text{Regret}_* = (\rho_*)^t \text{Regret}_{i_*}$, and i_*^ϵ and ρ_*^ϵ the same values for Regret_*^ϵ.

The quantity we wish to bound is how much additional regret we experience in expectation (ie. without noise) when using the noisy mixture weight ρ_*^ϵ instead of ρ_*, which we name $\Delta_O = \max_i (\rho_*^\epsilon)^t \text{Regret}_i - (\rho_*)^t \text{Regret}_{i_*}$.

Proposition 10 (Value-continuity of min-max). *The optimality gap Δ_O is bounded in the following way:*

$$0 \leq \Delta_O \leq (\rho_*)^t \tilde{\epsilon}_{i_*^\epsilon} - \min_i (\rho_*^\epsilon)^t \tilde{\epsilon}_i \leq 2||\tilde{\epsilon}||_\infty \leq 4||\epsilon||_\infty \,.$$

Proof. By optimality of ρ_*, we already have that $\Delta_O \geq 0$.

$$\begin{aligned}
\Delta_O &= \max_i (\rho_*^\epsilon)^t \text{Regret}_i - (\rho_*)^t \text{Regret}_{i_*} \\
&= \max_i (\rho_*^\epsilon)^t (\text{Regret}_i + \tilde{\epsilon}_i) - (\rho_*^\epsilon)^t \tilde{\epsilon}_i - (\rho_*)^t \text{Regret}_{i_*} \\
&\leq (\rho_*^\epsilon)^t (\text{Regret}_{i_*^\epsilon} + \tilde{\epsilon}_{i_*^\epsilon}) - \min_i (\rho_*^\epsilon)^t \tilde{\epsilon}_i - (\rho_*)^t (\text{Regret}_{i_*^\epsilon} + \tilde{\epsilon}_{i_*^\epsilon}) + (\rho_*)^t \tilde{\epsilon}_{i_*^\epsilon} \\
&\leq (\rho_*^\epsilon - \rho_*)^t (\text{Regret}_{i_*^\epsilon} + \tilde{\epsilon}_{i_*^\epsilon}) + (\rho_*)^t \tilde{\epsilon}_{i_*^\epsilon} - \min_i (\rho_*^\epsilon)^t \tilde{\epsilon}_i \\
&\leq (\rho_*)^t \tilde{\epsilon}_{i_*^\epsilon} - \min_i (\rho_*^\epsilon)^t \tilde{\epsilon}_i \leq 2||\tilde{\epsilon}||_\infty
\end{aligned}$$

$\forall t$, $\tilde{\epsilon}_t = \epsilon_t - (\nu_t)^t \epsilon_t$, and $\epsilon_t \leq ||\epsilon||_\infty$ and $-(\nu_t)^t \epsilon_t \leq ||\epsilon||_\infty$, therefore $||\tilde{\epsilon}||_\infty \leq 2||\epsilon||_\infty$, which concludes the proof. □

The tightness of this bound can be verified via noting that if $\rho_* = \rho_*^\epsilon$ and the minimum of $(\rho_*)^t \epsilon_i$ is reached for $i = i_*^\epsilon$, then the optimality gap is null.

We discuss this bound in more details in Appendix F, where we compute its value on several examples.

5.3.3 The improved PSRO algorithm. We add bandit compression onto Algorithm 5, accompanied with a few optimization criteria, yielding Algorithm 6. The improvements and their motivations are discussed in Appendix I.1.

Remark 11 (Use of the Algorithm for Nash-Convergence). *We note that one can also use Algorithm 6 for convergence towards MFNE if one uses an iterative solver for computing the Nash equilibrium - in that case, $\mathbb{A}$ is the Nash solver, and $Regret_*$ is the exploitability. Since a Nash equilibrium only uses a single distribution, one can either bypass solving Problem 2, or solve it trivially with $\rho(\nu_*) = 1$.*

Algorithm 6: Sped-up mean-field PSRO((C)CE)

Result: Policy set $\Pi^* = \{\pi_1, ..., \pi_n\}$, ϵ-MF(C)CE ρ^*
$\Pi_0 = \emptyset$, $\Pi_1 = \{\pi_1\}$ with π_1 any policy, $\rho(\delta_{\pi_1}) = 1.0$, $N = 1$;
while $(\Pi_{n+1} \setminus \Pi_n) \neq \emptyset$ *or* $\rho_{tol} > \rho_{lim}$ **do**
  $\Pi_{n+1} = \Pi_n \cup \{BR_{(C)CE}(\pi_i, \rho_T) \mid \pi_i,\ \rho(\pi_i) > 0\}$;
  if $\Pi_{n+1} == \Pi_n$ **then**
    $\rho_{tol} = \frac{\rho_{tol}}{2}$
  end
  $N = N + 1$;
  Initialize $\mathbb{A}(\Pi_{n+1})$;
  Step Count = 0;
  while $Regret_* > \rho_{tol}$ **do**
    Step Count += 1 ;
    Do one step of $\mathbb{A}(\Pi_{n+1})$;
    if *Step Count* $\equiv \tau_{Compress} == 0$ **then**
      Compute ρ_* optimal solution of Problem 2 (CCE) / 3 (CE) ;
      Compute ρ_*'s associated regret Regret_* ;
    end
  end
  $\rho_{n+1} = \rho_*$
end

6 EXPERIMENTAL RESULTS

To demonstrate the viability of our approach, we use three different metrics presented in Section 6.1, which we evaluate when running MF-PSRO on four different mean-field games, which are described in Section 6.2. Evaluation methods are detailed in Section 6.3, and evaluation results are discussed in Section 6.4.

6.1 Evaluation metrics

For a given correlation device ρ, we define

$$\text{CCEGap}(\rho) := \max_{\pi} \sum_{\nu} \rho(\nu)\big(J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu))\big)$$

By construction, we directly have that $\text{CCEGap}(\rho) = 0$ is equivalent to ρ being an MFCCE. In the same fashion, we define

$$\text{CEGap}(\rho) := \max_{\pi'} \max_{\pi | \rho(\pi) > 0} \sum_{\nu} \rho(\nu|\pi)\big(J(\pi', \mu(\nu)) - J(\pi(\nu), \mu(\nu))\big)$$

for MFCE characterisation. Finally, for a given population distribution $\nu \in \Delta(\Pi)$, we introduce

$$\text{Exploitability}(\nu) := \max_{\pi} J(\pi, \mu(\nu)) - J(\pi(\nu), \mu(\nu))$$

so that $\text{CCEGap}(\rho) = 0$, which reaches 0 if and only if ν is an MFNE.

6.2 Evaluation games

The four games we use to evaluate convergence include atwo complex games available in OpenSpiel [21], Predator-Prey [32] and Crowd Modeling [33], and two new small normal-form mean-field games, *Coop / Betray / Punish* and *mean-field biased indirect Rock-Paper-Scissors*, which are described in detail and motivated in Appendix G.1. Summarily, *Coop / Betray / Punish* is a 3-action normal-form game where agents can choose to either Cooperate, and all get a good reward; betray and take advantage of others; or punish the betrayers. But punishing agents also take some reward away from cooperators (they must support the punishers). Payoffs are non-linear (quadratic) in distributions. *mean-field biased indirect Rock-Paper-Scissors* is a classic biased Rock-Paper-Scissors game - except playing Rock yields the reward of playing Paper; Paper, that of Scissors, and so on, hence the *Indirect* of the name.

6.3 Evaluation Methods

The regret minimizer used by mean-field PSRO((C)CE) is Regret Matching [36], and the Black-Box Optimization method used by mean-field PSRO(Nash) is CMA-ES [15]. As per Remark 11, we use Algorithm 6 for both mean-field PSRO((C)CE) and mean-field PSRO(Nash), since the Nash solver CMA-ES is iterative.

Regarding convergence to MF(C)CE, since there exists, to the best of our knowledge, no other algorithm known to converge towards these weaker equilibria we investigate the convergence behavior of mean-field PSRO((C)CE) with additional payoff noise.

Regarding convergence towards MFNE, we compare mean-field PSRO to Online Mirror Descent (OMD) with several different learning rates, and Fictitious Play, both algorithms available on OpenSpiel.

6.4 Evaluation Results

Figure 3 presents the CCE-Gap of mean-field PSRO(CCE), 4, the CE-Gap of mean-field PSRO(CE), while Figure 5 exposes the Exploitability of mean-field PSRO(Nash) on the four mean-field game environments described above. We note that in both normal-form games, mean-field PSRO converges within numerical precision towards mean-field correlated, coarse correlated and Nash equilibria after only a few iterations.

Nash-wise, OMD seems capable to follow PSRO at a similar speed on *Coop / Betray / Punish*, but fails utterly to converge on *mean-field biased indirect Rock-Paper-Scissors*. We note that OMD's convergence is strongly affected by its learning rate. Fictitious play does not manage to find good equilibria in these games. We note that Mean-Field Biased Indirect Rock-Paper-Scissors is a non-monotonic game, and see here one of the strengths of our approach: Where traditional approaches such as OMD or FP *do* require such properties to converge, Mean-Field PSRO manages to find equilibria even in their absence.

On more complex games, mean-field PSRO quickly converges towards very good correlated (CCE Gap $\approx 10^{-1}$), coarse correlated equilibria (CE Gap $\approx 10^{-1}$), and mean-field PSRO(Nash) seems to quickly minimize exploitability - but it does much more slowly (time-wise) than both OMD and FP. This hints at a strong potential

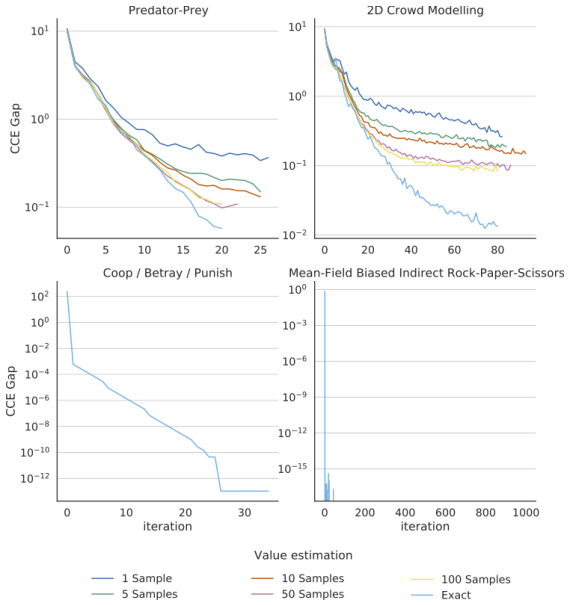

Figure 3: CCE Gap of mean-field PSRO(CCE).

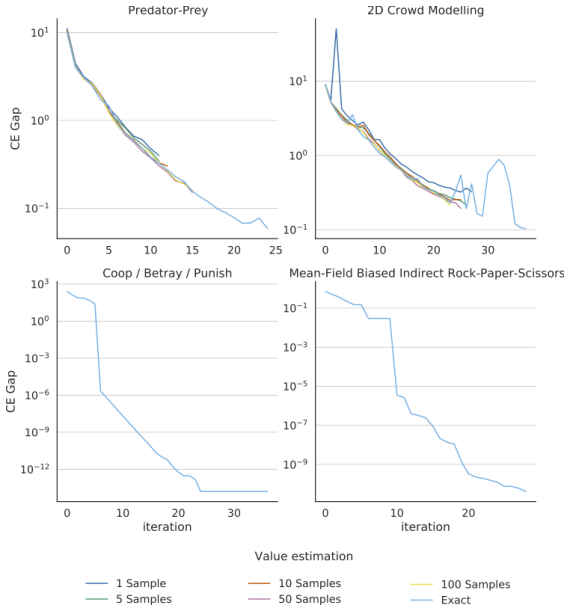

Figure 4: CE Gap of mean-field PSRO(CE).

direction of improvement for mean-field PSRO. We note that in this zoomed-in plot, FP seems to outperform OMD. We provide a zoomed-out version in Appendix G.3 where we see that OMD, with the correct learning rate, outperforms Fictitious Play as expected.

7 DISCUSSION

Despite its modularity, several improvements on our approach are envisioned for further research. First, our approach cannot efficiently select higher-welfare (C)CEs over lower ones. This problem

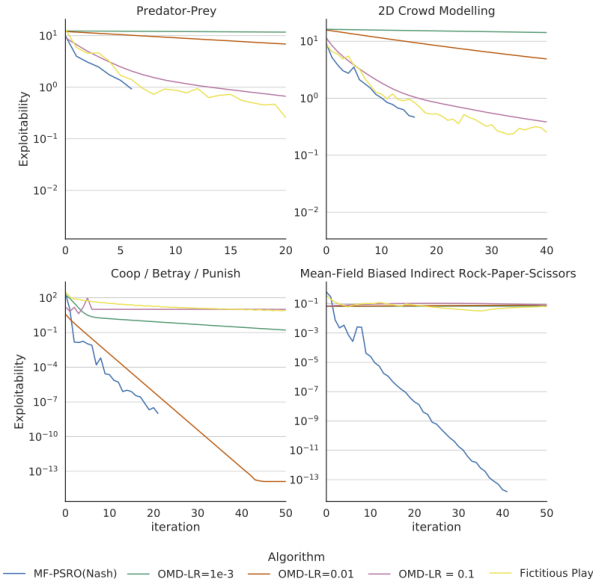

Figure 5: Exploitability of mean-field PSRO(Nash).

is known to be NP-Hard in general but learning approaches could hold the key to unlocking these possibilities (see the more detailed discussion in Appendix I.3). Second, mean-field PSRO(Nash) relies on a black-box algorithm, whose characteristics strongly impacts the speed and equilibrium accuracy of the algorithm. Finding a principled, general and fast Nash solver in complex restricted games, like we have for mean-field (C)CEs, could yield great improvements, both theoretically and performance-wise. Finally, our method is much slower than OMD or Fictitious Play on large games. This is largely due to a combination of slow payoff evaluation (be it sampled payoff or exact payoff) and relatively large amounts of steps needed to find a restricted equilibrium.

However, Mean-Field PSRO also shines in several ways: it is for now the only algorithm capable of finding Mean-Field (coarse) correlated equilibria, is not overly sensitive to hyperparameters, and finally, it is able to converge towards equilibria even in the absence of monotonicity.

8 CONCLUSION

We have introduced a new mean-field Multi-Agent Reinforcement Learning algorithm, Mean-Field PSRO, and demonstrated its ability to converge to Nash, correlated and coarse correlated equilibria both theoretically and empirically in various benchmark games. Additionally, the approach was successfully sped up using a new method named bandit compression, which is motivated by noise robustness and empirical speed. The approach has only been tested so far using the computation of exact best-responses. We expect Reinforcement-Learning algorithms to work out of the box, and answering this question would unlock (C)CE convergence in very large and complex games.

ACKNOWLEDGMENTS

M. Geist and S. Perrin for thoughtful discussions, and grant [1].

REFERENCES

[1] [n.d.]. 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), grants PIE-SGP-AI-2018-01 and RGEPPV2101.

[2] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. 2020. Q-learning in regularized mean-field games. In *arXiv*.

[3] Siddharth Barman and Katrina Ligett. 2015. Finding Any Nontrivial Coarse Correlated Equilibrium Is Hard. arXiv:1504.06314 [cs.GT]

[4] Avrim Blum and Yishay Mansour. 2005. From External to Internal Regret.

[5] A. Blum and Y. Mansour. 2007. Learning, Regret minimization, and Equilibria. In *Algorithmic Game Theory*, Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani (Eds.). Cambridge University Press, Chapter 4, 79–102.

[6] George W Brown. 1951. Iterative solution of games by fictitious play. *Activity analysis of production and allocation* 13, 1 (1951), 374–376.

[7] Luciano Campi and Markus Fischer. 2021. Correlated equilibria and mean field games: a simple model. arXiv:2004.06185 [math.OC]

[8] Pierre Cardaliaguet and Saeed Hadikhanloo. 2015. Learning in Mean Field Games: the Fictitious Play. arXiv:1507.06280 [math.OC]

[9] Constantinos Daskalakis and Christos Papadimitriou. 2007. Computing equilibria in anonymous games. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 83–93.

[10] Constantinos Daskalakis and Christos H Papadimitriou. 2008. Discretized multinomial distributions and Nash equilibria in anonymous games. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, 25–34.

[11] Laura Degl'Innocenti. 2018. Correlated equilibria in static mean-field games.

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

[13] Peter I. Frazier. 2018. A Tutorial on Bayesian Optimization. arXiv:1807.02811 [stat.ML]

[14] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2020. A General Framework for Learning Mean-Field Games. arXiv:2003.06069 [cs.LG]

[15] Nikolaus Hansen. 2016. The CMA Evolution Strategy: A Tutorial. arXiv:1604.00772 [cs.LG]

[16] Sergiu Hart and Andreu Mas-Colell. 2013. *Simple adaptive strategies: from regret-matching to uncoupled dynamics*. Vol. 4. World Scientific.

[17] Minyi Huang, Roland P Malhamé, Peter E Caines, 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 (2006), 221–252.

[18] Shizuo Kakutani. 1941. A generalization of Brouwer's fixed point theorem. *Duke mathematical journal* 8, 3 (1941), 457–459.

[19] Arbaaz Khan, Clark Zhang, Daniel D. Lee, Vijay Kumar, and Alejandro Ribeiro. 2018. Scalable Centralized Deep Multi-Agent Reinforcement Learning via Policy Gradients. arXiv:1805.08776 [cs.LG]

[20] Marc Lanctot et al. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. arXiv:1711.00832 [cs.AI]

[21] Marc Lanctot et al. 2020. OpenSpiel: A Framework for Reinforcement Learning in Games. arXiv:1908.09453 [cs.LG]

[22] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean Field Games. *Japanese Journal of Mathematics* 2 (03 2007), 229–260. https://doi.org/10.1007/s11537-007-0657-8

[23] Luke Marris, Paul Muller, et al. 2021. Multi-Agent Training beyond Zero-Sum with Correlated Equilibrium Meta-Solvers. arXiv:2106.09435 [cs.MA]

[24] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. 2012. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review* 27, 1 (2012), 1–31.

[25] H Brendan McMahan, Geoffrey J Gordon, and Avrim Blum. 2003. Planning in the presence of cost functions controlled by an adversary. In *International Conference on Machine Learning (ICML)*.

[26] H. Brendan McMahan, Geoffrey J Gordon, and Avrim Blum. 2003. Planning in the presence of cost functions controlled by an adversary.

[27] Barnabé Monnot and Georgios Piliouras. 2017. Limits and limitations of no-regret learning in games. *The Knowledge Engineering Review* 32 (2017).

[28] Paul Muller et al. 2020. A Generalized Training Approach for Multiagent Learning. arXiv:1909.12823 [cs.MA]

[29] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. 2007. *Algorithmic game theory*. Cambridge University Press.

[30] Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. 2019. α-Rank: Multi-Agent Evaluation by Evolution. arXiv:1903.01373 [cs.MA]

[31] Afshin OroojlooyJadid and Davood Hajinezhad. 2021. A Review of Cooperative Multi-Agent Deep Reinforcement Learning. arXiv:1908.03963 [cs.LG]

[32] Julien Perolat et al. 2021. Scaling up Mean Field Games with Online Mirror Descent. arXiv:2103.00623 [cs.AI]

[33] Sarah Perrin et al. 2020. Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications. arXiv:2007.03458 [math.OC]

[34] Mark Rowland, Shayegan Omidshafiei, Karl Tuyls, Julien Perolat, Michal Valko, Georgios Piliouras, and Remi Munos. 2019. Multiagent Evaluation under Incomplete Information. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/510f2318f324cf07fce24c3a4b89c771-Paper.pdf

[35] Francisco J Solis and Roger J-B Wets. 1981. Minimization by random search techniques. *Mathematics of operations research* 6, 1 (1981), 19–30.

[36] Oskari Tammelin. 2014. Solving Large Imperfect Information Games Using CFR+. arXiv:1407.5042 [cs.GT]

[37] Tanvi Verma, Pradeep Varakantham, and Hoong Chuin Lau. 2020. Entropy based Independent Learning in Anonymous Multi-Agent Settings. arXiv:1803.09928 [cs.LG]

[38] Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. 2020. Provable fictitious play for general mean-field games. *arXiv preprint arXiv:2010.04211* (2020).