Decentralised Learning in Systems
with Many, Many Strategic Agents

David Mguni,[1] **Joel Jennings,**[1] **Enrique Munoz de Cote**[1,2]
[1]PROWLER.io, Cambridge, UK
[2]Department of Computer Science, INAOE, Mexico

Abstract

Although multi-agent reinforcement learning can tackle systems of strategically interacting entities, it currently fails in scalability and lacks rigorous convergence guarantees. Crucially, learning in multi-agent systems can become intractable due to the explosion in the size of the state-action space as the number of agents increases. In this paper, we propose a method for computing closed-loop optimal policies in multi-agent systems that scales *independently* of the number of agents. This allows us to show, for the first time, successful convergence to optimal behaviour in systems with an *unbounded* number of interacting adaptive learners. Studying the asymptotic regime of $N-$player stochastic games, we devise a learning protocol that is guaranteed to converge to equilibrium policies even when the number of agents is extremely large. Our method is *model-free* and completely decentralised so that each agent need only observe its local state information and its realised rewards. We validate these theoretical results by showing convergence to Nash-equilibrium policies in applications from economics and control theory with thousands of strategically interacting agents.

Introduction

Multi-agent reinforcement learning (MARL) provides the potential to systematically analyse environments with strategically interacting agents. Despite the fundamental relevance of multi-agent systems (MASs) with appreciably large populations, learning stable, best-response policies in MASs with more than a few agents remains a significant challenge due to growth in complexity as the number of agents increases (Shoham and Leyton-Brown 2008). Consequently, the task of understanding agent behaviour in many systems of interest has been left unaddressed.

This paper seeks to address the problem of learning stable, best-response policies within non-cooperative[1] MASs when the size of the population is large, therefore expanding the range of applications of multi-agent technology.

In a non-cooperative MAS, selfish agents compete to obtain a sequence of rewards within an unknown environment.

[1]In game theory, the term non-cooperative implies that each agent seeks to pursue its own objectives and agreements between agents over their actions cannot be enforced.

A stochastic (dynamic) game (SG) is a mathematical framework that analyses the behaviour of strategically interacting entities in non-cooperative settings. Stochastic games enable stable policy outcomes in which agents respond optimally to one another (know as *equilibria*), to be fully described. In SGs, it is assumed that agents have either fixed knowledge of their environment or can acquire knowledge of any missing data by simply observing other agents with which agents can compute best-response actions. However, in many multi-agent systems, agents do not have full information of the environment from the outset and direct computation of optimal behaviour is often prohibitively complex.

Naturally, integrating stochastic game theory with reinforcement learning (RL) - a framework that enables agents to learn optimal behaviour within an unknown environment through direct interaction and exploration - suggests the potential to learn stable policies in multi-agent systems. Although this approach has led to fruitful analysis in multi-agent systems with few interacting agents (Leibo et al. 2017), current methods of computing multi-agent equilibria using RL (e.g. Nash Q-Learning (Hu and Wellman 2004), Friend-or-Foe Q-learning (Littman 2001), minimax-Q (Littman 1994)) have computational complexity that increases exponentially with the number of agents (Busoniu, Babuska, and De Schutter 2008; Tuyls and Weiss 2012). This renders the task of using RL to learn equilibrium policies intractable for many systems of interest.

In this paper, we introduce an approach that enables equilibrium policies of multi-agent systems to be computed even when the size of the population is extremely large. Unlike current multi-agent learning procedures, our method scales independently of the number of interacting agents. In contrast to approaches that compute equilibria in large population games (Cardaliaguet and Hadikhanloo 2017), our method is a model-free, fully decentralised learning procedure that only requires agents to observe local state information and their realised rewards.

Our main result demonstrates that the equilibria of $N-$player SGs can be computed by solving an optimal control problem (OCP) using a model-free learning procedure under very mild assumptions. To do this, we prove a series of theoretical results: first, we establish a novel link between reinforcement learning in MASs and a class of games known as *discrete-time mean field games* - $N-$player SGs in an

asymptotic regime as $N \to \infty$. Secondly, we demonstrate that in the asymptotic regime, the resultant game belongs to a class of games known as *potential games*. These are reducible to a single objective OCP leading to a vast reduction in the problem complexity. Our last result proves that the equilibria of SGs in an asymptotic regime are in fact approximate equilibria of the $N-$ player SG with an approximation error that vanishes as N increases. Finally, we validate our theoretical results by application to a series of problems within economics and optimal control theory. Our approach is based on a variant of the fictitious play - a belief-based learning rule for static games introduced by Brown (Brown 1951) and generalised to adaptive play in (Leslie and Collins 2006) for games with finite action sets.

After formulating the problem as an SG, the paper is organised as follows: first, we provide a formal description of a discrete-time mean field game and show that the game is a potential game (Theorem 1). We show that given the potentiality property, the problem is reducible to a single objective OCP. We then introduce the learning protocol and show that under this protocol, the game has strong convergence guarantees to equilibrium policies (Theorem 2). We lastly show that the equilibria generated for the mean field game are approximate equilibria of the $N-$player SG with an approximation error that vanishes as N increases (Theorem 3). Taking benchmark examples from economics and the multi-agent literature, we demonstrate our method within a number of examples with large populations of interacting agents.

Background

We now give the background for SG theory by introducing the $N-$player SG formalism. In order to handle multi-agent systems with large populations, we consider the $N-$player SG formalism when the number of agents tends to infinity - we therefore introduce the notion of mean field games - SGs studied in the asymptotic regime in the number of agents.

Problem formulation: $N-$Player Stochastic Games

The canonical framework to describe multi-agent systems in which agents behave rationally and non-cooperatively is a stochastic (dynamic) game (SG). Let us therefore introduce a formal description of an SG:

Let $\mathcal{N} \triangleq \{1, \dots, N\}$ denote the set of agents where $N \in \mathbb{N}$. At each time step $k \in 1, 2, \dots T \in \mathbb{N}^2$, the state of agent $i \in \mathcal{N}$ is $x_k^i \in \mathcal{S}^i$ where $\mathcal{S}^i \subset \mathbb{R}^d$ is a $d-$dimensional state space. The state of the system at time $k \leq T$ is given by $x_k \triangleq (x_k^i)_{i \in \mathcal{N}}$ where $x_k \in \mathcal{S} \triangleq \times_{j \in \mathcal{N}} \mathcal{S}^j$, which is the Cartesian product of space of states for each agent. Let Π^i be the set of stochastic closed-loop policies[3] for agent i where $\pi^i : \mathcal{S} \to \Delta \mathbb{A}^i$ where $\mathbb{A}^i \subseteq \mathbb{R}^d$ is a compact, non-empty action set for each agent $i \in \mathcal{N}$. Denote by $\mathbf{\Pi}$ the set of policies for all agents i.e. $\mathbf{\Pi} \triangleq \times_{j \in \mathcal{N}} \Pi^j$ where each Π^j is a

non-empty compact set. We denote by $\Pi^{-i} \triangleq \times_{j \in \mathcal{N} \backslash \{i\}} \Pi^j$, the Cartesian product of the policy sets for all agents except agent $i \in \mathcal{N}$.

At each time step, each agent $i \in \mathcal{N}$ exercises its policy π^i, the agent's state then transitions according to the following[4]:

$$x_{k+1}^i = f(x_k^i, a_k^i, \zeta_k), \quad a_k^i \sim \pi^i, \quad k = 0, 1, \dots \quad (1)$$

where $\{\zeta_k\}_{0 \leq k \leq T}$ is a collection of i.i.d. random variables that introduce randomness in the agent's state transition. Each agent i has a cumulative reward function $J^i : \mathcal{S} \times \Pi^i \times \Pi^{-i} \to \mathbb{R}$ that it seeks to maximise given by the following:

$$J^i[x_t, \pi^i, \pi^{-i}] = \mathbb{E}_{\pi^i} \left[\sum_{k=t}^T L(x_k^i, x_k^{-i}, a_k^i) \Big| a_k^i \sim \pi^i \right] \quad (2)$$

where $x_k^i \in \mathcal{S}^i$ and $x_k^{-i} \in \mathcal{S}^{-i}$ are the state for agent i and the collection of states for agents $j \in \mathcal{N} \backslash \{i\}$ at time $k \leq T$ respectively and π_k^i is the policy for agent i. The function L is the *instantaneous reward function* which measures the reward received by the agent at each time step. We refer to the system of equations (1) - (2) as game (A).

We now formalise the notion of optimality within an SG, in particular, a *Markov-Nash-equilibrium* for the game (A) is the solution concept when every agent plays their best response to the policies of other agents. Formally, we define the notion of equilibrium for this game by the following:

Definition 1. *The strategy profile $\pi \in \mathbf{\Pi}$ is said to be a Markov-Nash equilibrium (M-NE) strategy if, for any policy for agent i, $\pi'^i \in \Pi^i$ and $\forall y \in \mathcal{S}$, we have that:*

$$J^i[y, \pi^i, \pi^{-i}] \geq J^i[y, \pi'^i, \pi^{-i}], \quad \forall i \in \mathcal{N}. \quad (3)$$

The M-NE condition identifies strategic configurations in which no agent can improve their rewards by a unilateral deviation from their current strategy.

We will later consider approximate solutions to the game (A). In order to formalise the notion of an approximate solution, we introduce $\epsilon-Markov-Nash equilibria$ ($\epsilon-$M-NE) which extends the concept of M-NE to strategy profiles in which the incentive to deviate never exceeds some fixed constant. The notion of an $\epsilon-$M-NE can be described using an analogous condition to (3). Formally, the strategy profile $\boldsymbol{\pi} \in \mathbf{\Pi}$ is an $\epsilon-$Markov-Nash equilibrium strategy profile if for a given $\epsilon > 0$ and for any individual strategy for agent i, $\pi'^i \in \Pi^i$ we have that $\forall y \in \mathcal{S}$:

$$J^i[y, \pi^i, \pi^{-i}] \geq J^i[y, \pi'^i, \pi^{-i}] - \epsilon, \quad \forall i \in \mathcal{N} \quad (4)$$

Although in principle, methods within RL such as TD learning can be used to compute the equilibrium policies, learning takes place in the product space of the state space and the set of actions across agents, so the problem complexity grows exponentially with the number of agents.

---

[2]The formalism can be straightforwardly extended to infinite horizon cases by appropriate adjustment of the reward function.

[3]*Closed-loop* policies are maps from states to actions and are likely to be the only policies that produce optimal behaviour in stochastic systems. *Open-loop* policies simply specify pre-computed (state-independent) sequences of actions.

[4]With this specification, agents do not influence each others' transition dynamics directly - this is a natural depiction of various systems e.g. a portfolio manager's modification to their own market position. This does not limit generality since prohibited state transitions (e.g. collisions) can be disallowed with a reward function that heavily penalises such joint action behaviour.

A second issue facing RL within MASs is the appearance of non-stationarity produced by other adaptive agents. During the learning phase, agents update their policies and thus the way they influence the system. In a non-cooperative MAS with even just a few agents learning independently, the presence of other adaptive agents induces the appearance of a non-stationary environment from the perspective of an individual agent. This in turn may severely impair the agent's own reinforcement learning process and lead to complex and non-convergent dynamics (Tuyls and Weiss 2012).

With these concerns, we present an alternative approach which involves studying the game (A) in an asymptotic regime as $N \to \infty$. This results in a mean field game - an SG with an infinite population which, as we shall show is both reducible to a single OCP and has M-NE that are $\epsilon-$M-NE for $N-$player SGs where $\epsilon \sim \mathcal{O}(\frac{1}{\sqrt{N}})$.

In order to show the discrete-time mean field game is reducible to single OCP, we demonstrate that they belong to a class of games known as dynamic potential games (PGs). Before considering the mean-field game case, let us formally define a PG in the context of an $N-$player SG:

Definition 2. *An SG is called a (dynamic) potential game (PG) if for each agent $i \in \mathcal{N}$ and for any given strategy profile $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ there exists a potential function $\Omega : \Pi^i \times \Pi^{-i} \to \mathbb{R}$ that satisfies the following condition $\forall \pi'^i \in \Pi^i$:*

$$
J^i[\cdot, (\pi^i, \pi^{-i})] - J^i[\cdot, (\pi'^i, \pi^{-i})]
$$
$$
= \Omega[\cdot, (\pi^i, \pi^{-i})] - \Omega[\cdot, (\pi'^i, \pi^{-i})]. \quad (5)
$$

A PG has the property that any agent's change in reward produced by a unilateral deviation in their strategy is exactly expressible through a single global function. In PGs - the (Nash) equilibria can be found by solving an OCP (Monderer and Shapley 1996). This is a striking result since obtaining the solution to an OCP is, in general, an easier task than standard methods to obtain equilibria which rely on fixed point arguments.

Mean Field Games

In this section, we introduce a mean field game (MFG) which is a central framework to our approach. Mean field game theory is a mathematical formalism that handles large-population systems of non-cooperative rational agents. MFGs are formulated as SGs in the form (A) analysed at the asymptotic limit as the number of agents tends to infinity. This formulation enables the collective behaviour of agents to be jointly represented by a probability distribution over the state space (Lasry and Lions 2007).

The MFG formulation results in a description of the agents' optimal behaviour that is compactly characterised by a coupled system of partial differential equations. However, obtaining closed analytic solutions (or even approximations by tractable numerical methods) for the system of equations is often unachievable but for specific cases.

This work offers a solution to this problem; in particular we introduce a learning procedure by which the equilibria of MFG can be learned by adaptive agents. Beginning with the case in which the number of agents is finite, we introduce the

following empirical measure which describes the N agents' joint state at time k:

$$
m_{x_k} \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_k^i}, \quad (6)
$$

where $x_k^i \in \mathcal{S}$ and δ_x is the Dirac-delta distribution evaluated at the point $x \in \mathcal{S}$.

We now study a game with an infinite number of agents by considering the formalism the $N-$player game (A) in the asymptotic reime as $N \to \infty$ which allows us to treat the ensemble in (6) as being continuously distributed over $\mathcal{S}$. We call this limiting behaviour the *mean field limit* which is an application of the law of large numbers for first order (strategic) interactions in the game (A). We observe that by taking the limit as $N \to \infty$ and using de Finetti's theorem[5], we can replace the empirical measure (6) with a probability distribution $m_{x_k} \in \mathbb{P}(\mathcal{H})$ where $\mathbb{P}(\mathcal{H})$ is a space of probability measures. The distribution m_{x_k} describes the joint locations of all agents in terms of a distribution.

With this structure, instead of agents responding to the actions of other agents individually, each agent now performs its actions in response to the mass which jointly represents the collection of states for all agents.

As is standard within the MFG framework (Lasry and Lions 2007), we assume that the MFG satisfies the *indistinguishably property* - that is the game is invariant under permutation of the agents' indices.

The following concept will allow us to restrict our attention to games with a single M-NE:

Definition 3. *The function $v : \mathcal{S} \times \cdot \to \mathbb{R}$ is said to be* strictly monotone *in the L^2- norm given $m_1, m_2 \in \mathbb{P}(\mathcal{H})$ if the following condition is satisfied:*

$$
\int_{\mathcal{S}} (v(m_1, \cdot) - v(m_2, \cdot))(m_1 - m_2) dx \geq 0 \implies m_1 \equiv m_2.
$$

The strict monotonicity condition means that in any given state, agents prefer a lower concentration of neighbouring agents. This property is a natural feature within many practical applications in which the presence of others reduces the available rewards for a given agent e.g. spectrum sharing (Ahmad et al. 2010).

We make use of the following result which is proved in (Lasry and Lions 2007):

Proposition 1 (Lasry & Lions, 2007)**.** *If the instantaneous reward function of a MFG is strictly monotone in $m \in \mathbb{P}(\mathcal{H})$, then there exists a unique M-NE for the MFG.*[6]

---

[5]Given a sequence of indexed random variables $x_1, x_2, \ldots$ which are invariant under permutations of the index, De Finetti's theorem (de Finetti 1931) ensures the existence of the random variable m_{x_k} in (6) in the limit as $N \to \infty$.

[6]In (Lasry and Lions 2007) the result is proven for mean field games with continuous action and state spaces in continuous time. The corresponding results for discrete games (discrete state space, time and action set) is given in (Gomes, Mohr, and Souza 2010) (see theorem 2, pg 6).

MAS with Infinite Agents

To develop a learning procedure that scales with the number of agents, we now consider the game (A) in the mean-field limit. We shall demonstrate that this procedure allows us to reduce the game (A) to a strategic interaction between an agent and an entity that represents the collection of other agents. This plays a key role in reducing the problem complexity and collapsing it to a single OCP.

We shall firstly define the N−player stochastic game (A) in the asymptotic regime. We note that in light of the *indistingushability criterion*, we can drop the agent indices:

Definition 4 (Discrete-Time Mean Field Game). *We call the following system a discrete-time mean field game if its dynamics can be represented by the following system:*

$$x_{k+1} = f(x_k, a_k, \zeta_k) \qquad (7)$$
$$m_{x_{k+1}} = g(m_{x_k}, a_k), \qquad (8)$$

where $a_k \sim \pi$ for some $\pi \in \Pi$, $k \in 0, 1, \ldots, T$ for a given time horizon $T < \infty$, $m_{x_k} \in \mathbb{P}(\mathcal{H})$ is the agent density corresponding to the asymptotic distribution (6) evaluated at $x_k \in \mathcal{S}$, π_k is a policy exercised at each time step $k \leq T$ and ζ_k is an i.i.d. variable which captures the system stochasticity. We refer to the system (7) - (8) as game (B).
Given $m_{x_t} \in \mathbb{P}(\mathcal{H})$ and $\pi \in \Pi$ as above with the agent index removed, we consider games where each agent has the following reward function:

$$J[x_t, \pi, m_{x_t}] = \mathbb{E}\Big[\sum_{k=t}^{T} L(x_k, m_{x_k}, a_k) \Big| a_k \sim \pi \Big], \quad (9)$$

where $x_t \in \mathcal{S}$ is some initial state.

We are now in a position to describe a MFG system at equilibrium i.e. when each agent plays a best-response to the actions of other agents.

Given some initial state $x_t \in \mathcal{S}$, the joint solution $(\tilde{\pi}, \tilde{m})$ to the game (B) is described by the following triplet of equations which describes the M-NE:

$$\tilde{\pi} \in \underset{\pi \in \Pi}{\mathrm{argmax}}\, J[x_t, \pi, \tilde{m}_{x_t}], \qquad (10)$$

$$\tilde{x}_{k+1} = f(\tilde{x}_k, \tilde{a}_k, \zeta_k), \;\; \tilde{m}_{\tilde{x}_{k+1}} = g(\tilde{m}_{\tilde{x}_k}, \tilde{a}_k), \;\; \tilde{a}_k \sim \tilde{\pi}, \tag{11}$$

where as before, $\{\zeta_k\}_{0 \leq k \leq T}$ is a collection of i.i.d random variables and, $m_{\tilde{x}}$ is the agent density induced when the policy $\tilde{\pi}$ is exercised by each agent.

An important feature of the system (10) - (11) is that the agent's problem is reduced to a strategic interaction between itself and a single entity $\tilde{m}_x$. This property serves a crucial role in overcoming the appearance of non-stationarity in an environment with many adaptive learners since the influence of all other agents on the system is now fully captured a single entity $\tilde{m}_x$ which, influences the system dynamics in a way that an adaptive agent can learn its optimal policy.

Existing methods of computing equilibria in MFGs however rely on the agents having full knowledge of the environment to compute best responses and involve solving to non-linear partial differential equations (Cardaliaguet and Hadikhanloo 2017; Cardaliaguet et al. 2015) which, in a number of cases leads to intractability of the framework. MFGs are closely related to anonymous games - games in which the agents' rewards do not depend on the identity of the agents they interact with (but do depend on the interacting agents' strategies). Multi-agent learning has been studied for anonymous games (Kash, Friedman, and Halpern 2011) however, this approach requires agents to fix their policies over *stages* and to explicitly compue approximuate best-responses. In the following sections of the paper, we develop a technique which enables equilibrium policies of MFGs to be computed by adaptive learners in an unknown environment without solving partial differential equations.

Theoretical Contribution

Mean Field Games are Potential Games

We now demonstrate that the discrete-time MFG problem (B) is reducible to an objective maximisation problem. By proving that the discrete-time mean field game is a PG, the following theorem enables us to reduce the problem to a single OCP:

Theorem 1. *The discrete-time mean field game (B) is a PG.*

We defer the proof of the theorem to the appendix. The key insight of Theorem 1 is that the M-NE of MFGs can be computed by considering a general form of a *team game* in which each agent seeks to maximise the potential function. Crucially, thanks to Theorem 1, the problem of computing the equilibrium policy is reduced to solving a control problem for the potential function.

Learning in Large Population MAS

We now develop a model-free decentralised learning procedure based on a variant of fictitious play using the potentiality property. This generates a sequence of polices that converges to the M-NE of the discrete-time MFG.

Firstly, it is necessary to introduce some concepts relating to convergence to equilibria:

Definition 5. *Let $\{\pi^{i,k}\}_{k \geq 1}$ be a set of policies for agent $i \in \mathcal{N}$. We define a path by a sequence of strategy profiles $\rho_\pi^i \triangleq (\pi^{i,k})_{k \geq 1} \in \Pi^i \times \Pi^i \times \ldots$ where the strategy $\pi^{i,k+1}$ is obtained from an update of the strategy $\pi^{i,k}$ using some given learning rule.*

Definition 6. *The path $\rho_\pi^i \in \Pi^i \times \Pi^i \times \ldots$ is called an improvement path for agent i if after every update the agent's expected reward increases, formally an improvement path satisfies the following condition:*

$$J^i[\cdot, \pi^{i,k+1}, \pi^{-i}] \geq J^i[\cdot, \pi^{i,k}, \pi^{-i}], \quad \forall i \in \mathcal{N}. \quad (12)$$

Definition 7. *A path converges to equilibrium if each limit point is an equilibrium.*

We now describe a 'belief-based' learning rule introduced by Brown (Brown 1951) known as fictitious play of which our method is a variant:

Let $\rho_\pi^i \in \Pi^i \times \Pi^i \times \ldots$ be a path, then the learning rule is a *fictitious play process* (FPP) if the update in the sequence $\{\pi^{i,k}\}_{k \geq 1}$ is performed in the following way

$\forall x' \in \mathcal{S}, i \in \mathcal{N}$:

$$\sup_{\pi' \in \Pi} J[x', \pi', \rho_\pi^{-i}] = J[x', \pi^{i,k+1}, \rho_\pi^{-i}], \qquad (13)$$

so that $\pi^{i,k+1}$ is a best-response policy against ρ_π^{-i}.

If the FPP converges to equilibrium then we say that the game has the *fictitious play property*.

We now apply these definitions to the case of MFGs. We note that by the indistinguishability assumption for MFGs, we can drop the agent indices in each of the above definitions. We define the FPP for the MFG (B) by the following learning procedure $\forall x' \in \mathcal{S}$:

$$\sup_{\pi' \in \Pi} J[x', \pi', \bar{m}_{x'}^k] = J[x', \pi^{k+1}, \bar{m}_{x'}^k], \quad \bar{m}_{x'}^k \triangleq \frac{1}{k} \sum_{j=1}^k m_{x'}^j$$

so that π^{k+1} is a best-response policy against $\bar{m}_{x'}^k$ which summarises each agent's belief of the joint state of all agents after the kth update.

In order to solve game (B) we therefore seek a learning process that produces sequence $\{\pi^n, m^n\}_{n \to \infty}$ for which $\{\pi^n\}_{n \geq 1}$ is an improvement path for the policy π^n.

For the discrete-time MFG, we shall seek a pair $(\pi^n, m^n) \in \Pi \times \mathbb{P}(\mathcal{H})$ that converges to equilibrium as $n \to \infty$ so that the sequence $\{\pi^n, m^n\}_{n \in \mathbb{N}}$ converges to a cluster point $(\tilde{\pi}^n, \tilde{m}^n)$ which is a solution to (B).

Our convergence result is constructed using results that we now establish. Before proving the result we report an important result in the model-based setting:

Proposition 2 (Cardaliaguet, Hadikhanloo; 2017). *Mean Field Games have the fictitious play property.*

This result was established in (Cardaliaguet and Hadikhanloo 2017) within a continuous-time and model-based setting. Here, given some initial belief about the distribution m^n and some initial value function v^n associated to each agent's problem, the agents update the pair (v^n, m^n) according to a (model-based) fictitious play procedure. This produces a paired sequence $\{v^n, m^n\}$ for which $\lim_{n \to \infty}(v^n, m^n) = (\tilde{v}, \tilde{m})$ where $(\tilde{v}, \tilde{m})$ is joint solution to the continuous-time MFG.

In order to compute the best responses at each step, the FPP discussed in (Cardaliaguet and Hadikhanloo 2017) requires agents to use knowledge of their reward functions. Moreover, the agents' update procedure involves solving a system of partial differential equations at each time step. Obtaining closed solutions to this system of equations is generally an extremely difficult task and often no method of obtaining closed solutions exists.

We are therefore interested in procedures for which agents can achieve their M-NE policies by simple adaptive play with no prior knowledge of the environment. We are now in position to state our main result:

Theorem 2. *There exists a fictitious-play improvement path process such that the sequence $(\pi^n, m^n) \in \Pi \times \mathbb{P}(\mathcal{H})$ converges to an ϵ−M-NE of the game (A).*

The following corollary demonstrates that we can construct a learning procedure that leads to an improvement path, the limit point of which is a solution of the MFG (B).

Corollary 1. *Let $\{u^n, m^n\}_{n \geq 1}$ be a mean field improvement path generated by an actor-critic fictitious play method, then $\{u^n, m^n\}$ converges to a cluster point $\{u, m\}$. Moreover, the cluster point $\{u, m\}$ is a solution to (B).*

Corollary 1 immediately leads to our method which computes the optimal policies for the MFG (B). The method uses an actor-critic framework with TD learning on the critic and policy gradient on the learner. An episode is simulated using some initial belief over the distribution m over $\mathcal{S}$. The agent then updates its policy using an actor-critic and updates the distribution m^k fictitiously.

The Approximation Error

In this section, we show that the Nash equilibria generated by the game (B) are approximate equilibria for the N−player stochastic game (A).

Theorem 3. *Let $\bar{\pi} \in \Pi$, $\tilde{\pi} \in \Pi$ be the NE strategy profile for the game (A) and game (B) respectively and let be a NE strategy profile for the MFG (B). Let $\tilde{m}_x$ and $\bar{m}_x$ be the distributions generated by the agents in the mean field game and the N−player SG respectively; then there exists a constant $c > 0$ s.th $\forall x \in \mathcal{S}$:*

$$|J(x, \tilde{m}_x, \tilde{\pi}) - J(x, \bar{m}_x, \bar{\pi})| < cN^{-\frac{1}{2}}. \qquad (14)$$

Theorem 3 says that the solution to the MFG (B) is in fact an ϵ−M-NE to the N−agent SG (A). Moreover, the approximation error from using a MFG to approximate the N−player game is $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

As a direct consequence of theorem 3, we can use the actor-critic fictitious play method on the MFG formulation (B) to compute near-optimal policy solutions of the stochastic game (A). Moreover, the error produced by the mean field approximation vanishes asymptotically as we consider systems with *increasing* numbers of agents.

Experiments

To investigate convergence of our method, we present three experiments drawn from benchmark problems within economics and control theory that involve large populations of strategically interacting agents. In each case, we show that our method converges to an M-NE policy.

We firstly demonstrate the use of our technique in a (stochastic) congestion game, testing the convergence to a stationary policy in a large-population system with a complicated reward structure. The second problem is a supply and demand problem that we formulate as an SG allowing us to test convergence to optimal policies in dynamic problems requiring long-term strategic planning in the presence of other learning agents. This demonstrates that our method is able to overcome the non-stationary interference of other adaptive agents. Lastly, we apply our method to study a multi-agent generalisation of a fundamental problem within optimal control theory, namely the linear quadratic control (LQC) problem. The analytic solution of the LQC problem allows us to verify that our method converges to a known M-NE policy.

Experiment 1: *Spatial Congestion Game*

In the spatial congestion game the rewards are dependent

on the agents' use of a shared resource (a sub-region of $\mathcal{S} \subset \mathbb{R}^2$) and the number agents using that resource. Games of this type are known as *congestion games* and represent a large class of interactions e.g. spectrum sharing problems.

In the spatial congestion (SC) game there are N agents, given some initial position $x_0 \in \mathcal{S}$, each agent chooses an action in order to move to a desired location $x_T \in \mathcal{S}$ which is a terminal state. Certain areas of $\mathcal{S}$ are more desirable to occupy than others, however the agents are averse to occupying crowded areas - they receive the largest rewards for occupying parts of $\mathcal{S}$ that are both desirable and have relatively low concentrations of agents. The agents simultaneously select a movement vector $u \in \mathbb{R}^{2 \times 1}$ resulting in movement to a terminal state x_T. Each agent then receives its reward L which depends on the desirability of the location and the concentration of agents m_{x_T} at x_T.

Formally, we model the desirability of a region $x_t \in \mathcal{S}$ at time t as[7]:

$$L(x_t, m_{x_t}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \frac{e^{-(x_t - \mu)^T \Sigma^{-1}(x_t - \mu)}}{(1 + m_{x_t})^\alpha},$$

where $m_{x_t} \in \mathbb{P}(\mathcal{H})$ is the density of agents at the point x_t and $\mu \in \mathbb{R}^2, \Sigma \propto 1_{2 \times 2}$ are given parameters representing the mean and spread of the distribution of the rewards over $\mathcal{S}$. The map $L : \mathcal{S} \times \mathbb{P}(\mathcal{H}) \to \mathbb{R}$ measures the instantaneous reward for an agent at x_t with a local agent density m_{x_t}. The parameter $\alpha > 0$ is a measure of each agent's averseness to occupying the same region as other agents with higher values representing greater averseness.

Given some initial position $x_0 \in \mathcal{S}$ and $u \in \mathbb{R}^{2 \times 1}$, the transition dynamics are given by the following expression:

$$x_T = f_1(x_0, u, \epsilon) \triangleq A_1 x_0 + B_1 u + \sigma_1 \epsilon, \qquad (15)$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon); \sigma_1, A_1, B_1 \propto 1_{(2 \times 2)}$ and $c, \sigma_\epsilon \in \mathbb{R}^+$. The reward function for an agent is then given by the following expression:

$$J[x_0, \pi, m_{x_0}] = \mathbb{E}_{x_T \sim f_1}\left[L(x_T, m_{x_T}) - \frac{1}{2} u^T R u | u \sim \pi\right],$$

where $R = \eta 1_{(2 \times 2)}$ is a control weight matrix and $\eta \in \mathbb{R}$ is the marginal control cost (cost of movement). Using the indistingishability condition, we have omitted agent indices. At equilibrium each agent optimally trades-off state-dependent rewards with its proximity to nearby agents.

The problem generalises the beach domain problem studied in (Devlin et al. 2014) since we now consider a reward function with state dependency. In particular, the desirability over $\mathcal{S}$ is described by a Gaussian function over $\mathcal{S}$. Moreover the problem we now consider consists of a system with 1000 interacting agents. The problem is also closely related to the spectrum sharing problem, see (Ahmad et al. 2010).

In accordance with the theory, our method converges to a stable policy - after 2,000 episodes of training we find that the agent's policy and rewards stabilise Figure 1.a). Figure 1.b) shows the terminal distribution of agents over $\mathcal{S}$ for

---

[7]We note that the function L is continuously differentiable in $\mathcal{S}$ so that assumption 2 is *a fortiori* satisfied, moreover, it can be easily verified that assumption 3 holds.
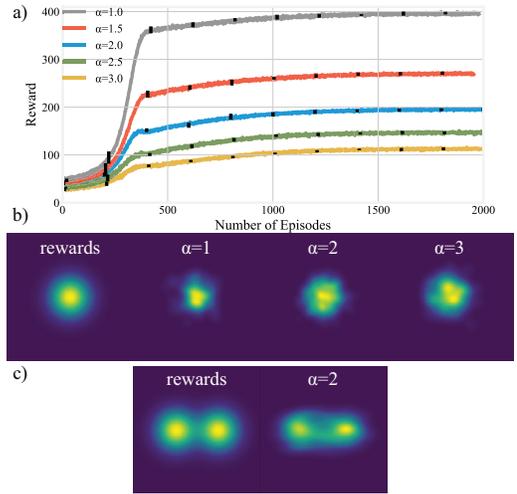

Figure 1: a) Rewards over 2000 episodes of training for a Gaussian distribution of desirability. b) agent distributions for different averseness parameter α. c) agent distribution for bimodal Gaussian function reward function.

$\alpha \in \{1.0, 2.0, 3.0\}$. We observe that the agents learn to optimally trade-off state-dependent rewards with distance from neighbouring agents resulting in a fixed terminal distribution of agents. As expected, the agents disperse themselves further as the value of α is increased - in all cases converging to a stable distribution over $\mathcal{S}$.

Figure 1.c) shows the distribution of agents for a more complicated reward structure specified by a mixture of two Gaussians over $\mathcal{S}$ with peaks at $(-1, 0)$ and $(0, 0)$). We initialise the agents at the point $(1, 0)$. In this case, an individual agent's reward is given by the following expression:

$$J[x_0, \pi, m_{x_0}] = \mathbb{E}_{x_T \sim f_1}\left[\sum_{n=1}^{2} L_n(x_T, m_{x_T}) - \frac{1}{4} u_n^T R u_n\right]$$

where $L_n(x_t, m_{x_t}) \triangleq [16\pi^2 |\Sigma_n|]^{-\frac{1}{2}} e^{-(x_t - \mu_n)^T \Sigma_n^{-1}(x_t - \mu_n)} \cdot (1 + m_{x_t})^{-\alpha}$ and $u_n \sim \pi$. Since learning is internal to each agent, a possible (suboptimal) outcome is for the agents to cluster at the nearest peak of rewards. However, using our method, the agents learn to spread themselves across the state space and distribute themselves across both peaks.

Experiment 2: *Supply with Uncertain Demand*
Optimally distributing goods and services according to demand is a fundamental problem within logistics and industrial organisation. In order to maximise their revenue, firms must strategically locate their supplies given some uncertain future demand whilst considering the actions of rival firms which may reduce the firm's own prospects.

We now apply our method to a supply and demand problem in which individual firms seek to maximise their revenue by strategically placing their goods when the demand process has future uncertainty. The demand process, which quantifies the level of demand associated with each point in space, is *a priori* unknown and is affected by the actions of thousands of rival firms. Each firm directs
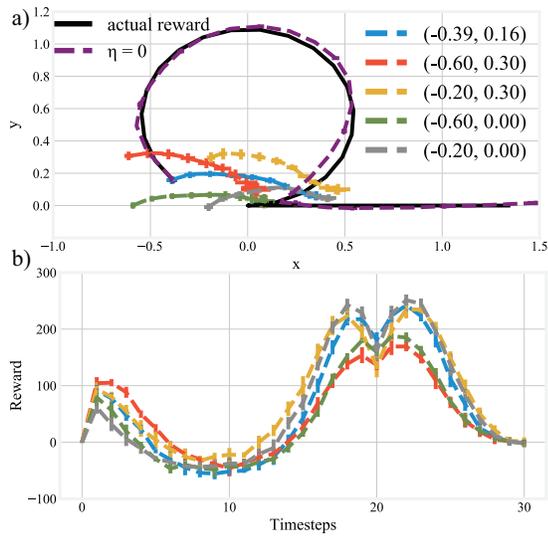
Figure 2: a) Paths of agents with different initial points (co-ordinates indicated) for cost of movement $\eta = 2$. Shown in black is the path of the rewards (demand) and in purple is the path of agents with $\eta = 0$. b) The intra-episode rewards for the paths show in a). Note incursion of negative rewards.

supply of its goods to regions in time and space however, the firms face transport costs so that each firm seeks to optimally trade-off transportation costs and tracking the demand. As firms begin to concentrate on a particular area of demand, the sale opportunities diminish, reducing the rewards associated to that region of demand.

We model this problem as an episodic problem with a distribution of rewards traversing a path through the state space $\mathcal{S} \in \mathbb{R}^2$ (illustrated in Figure 2.a)). Agents seek to locate themselves in areas of high concentrations of rewards for a fixed number (30) time steps. The agents are penalised for both movement and occupying areas with a high density of other agents. The experiment tests the ability of the method to avoid convergence to suboptimal outcomes. In particular, movement costs are highly convex so traversing the path of rewards leads to low overall rewards.

Figure 2.a) illustrates the path of the agents after training for several different initial positions. An interesting outcome is that with non-zero movement costs, the agents learn to move horizontally to intersect the path of the rewards at a later time. This behaviour conforms with intuition - to maximise long-term rewards the agents must choose a path that initially incurs higher costs, forgoing immediate rewards whilst they traverse the regions with sparse rewards. Figure 2.b) shows that the agents are able to learn to initially incur negative rewards to maximise their cumulative pay-offs. Without long-term planning each agent would attempt to trace the same path as the rewards. Such a strategy would lead to reduced overall rewards since attempting to match the locations of the rewards is prohibitively costly. Setting $\eta = 0$ we see that the agents trace the path of the rewards (represented by the dashed purple line in Figure 2.a)).

Experiment 3: *Mean-Field Linear Quadratic Regulator*

The linear quadratic control (LQC) problem is a funda-mental problem within optimal stochastic control theory (OSCT). It concerns a system whose transition dynamics evolves according to a stochastic process that is linearly controlled subject to quadratic costs. The LQC problem es-sentially captures the local problem of a large class of prob-lems in OSCT and can be solved analytically, the solution being given by the linear-quadratic regulator (Bardi 2012; Xu 2007). LQC models have been extended to mean field interactions in which a large population of agents affect the dynamics of a system using linear controls subject to quadratic control costs and a cost term which depends on the actions of other agents.

The reward function for the LQC problem is given by:

$$J[x_0, u, m_{x_0}] = \mathbb{E}_{x_T \sim f_2}\Big[\sum_{t=0}^{T_n} \{C(x_t, \bar{m}_{x_t}) - \frac{1}{2} u_t^T R u_t\} \Big],$$

where $C(x_t, m_{x_t}) \triangleq -(x_t - \alpha)^T Q_t (x_t - \alpha)$.
At time $k \leq T_n$, given some position $x_k \in \mathcal{S}$, each agent then chooses an vector control parameter $u_t \in \mathbb{R}^{2 \times 1}$. The transition are given by the following expression:

$$x_{k+1} = f_2(x_k, u_k, \epsilon_k) \triangleq A_1 x_k + B_1 u_k + \sigma_1 \epsilon_k, \quad (16)$$

where $\epsilon_k \sim \mathcal{N}(0, \sigma_{\epsilon_k}) \, \forall k < T_n, A_1, B_1 \propto 1_{(2 \times 2)}, \sigma_1 = c 1_{(2 \times 2)}$ where $c \in \mathbb{R}^+$ is some constant that measures the magnitude of the stochasticity in each agent's transition.

In (Bardi 2012) the distribution m_x is reported after con-vergence to the stationary M-NE. We compare our results (E3) with this stationary policy (B1) in the following table:

	B1	E3
μ	(0.50000, -0.50000)	(0.50717, -0.50537)
σ^2	0.14060	0.16100

Clearly, our results converge to values that closely replicate the analytic solution.

Conclusion

We develop an approach to MARL with large numbers of agents. This is the first paper to prove convergence results to best-response policies in multi-agent systems with an un-bounded number of agents. This allows RL to be applied across a broader range of applications with large agent pop-ulations, in contrast to current methods (Hu and Wellman 2004; Littman 2001; 1994). Our approach advances existing work in MFGs as in (Cardaliaguet and Hadikhanloo 2017) that require both full knowledge of the environment and to perform involved analytic computation. In contrast, by de-veloping a connection between RL in MASs and MFGs, we demonstrate a procedure that is model-free, enabling agents to learn best-response policies solely through adaptive play which overcomes the problem of non-stationarity. In our ex-periments we provide a novel approach of analysing prob-lems in control theory and economics.

Acknowledgements

References

Ahmad, S.; Tekin, C.; Liu, M.; Southwell, R.; and Huang, J. 2010. Spectrum sharing as spatial congestion games. *Preprint arXiv:1011.5384*.

Bardi, M. 2012. Explicit solutions of some Linear-Quadratic Mean Field Games. *Networks and Heterogeneous Media* 7(2):243 – 261.

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

Busoniu, L.; Babuska, R.; and De Schutter, B. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 38 (2), 2008.*

Cardaliaguet, P., and Hadikhanloo, S. 2017. Learning in mean field games: The fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations* 23(2):569–591.

Cardaliaguet, P.; Delarue, F.; Lasry, J.-M.; and Lions, P.-L. 2015. The master equation and the convergence problem in mean field games. *Preprint arXiv:1509.02505*.

de Finetti, B. 1931. Funzione caratteristica di un fenomeno aleatorio. In *Atti della R. Accademia Nazionale dei Lincei*, volume 4, 251 – 299.

Devlin, S.; Yliniemi, L.; Kudenko, D.; and Turner, K. 2014. Potential-based difference rewards for multiagent reinforcement learning. In *Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, volume 1, 165–172.

Gomes, D. A.; Mohr, J.; and Souza, R. R. 2010. Discrete time, finite state space mean field games. *Journal de Mathématiques Pures et Appliquées* 93(3):308 – 328.

Hu, J., and Wellman, M. 2004. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research* 4(6):1039 – 1069.

Kash, I. A.; Friedman, E. J.; and Halpern, J. Y. 2011. Multiagent learning in large anonymous games. *Journal of Artificial Intelligence Research* 40:571 – 598.

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

Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proc. Conf. on Autonomous Agents and MultiAgent Systems (AAMAS)*, 464 – 473.

Leslie, D. S., and Collins, E. J. 2006. Generalised weakened fictitious play. *Games and Economic Behavior* 56:285 – 298.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, volume 157, 157 – 163.

Littman, M. L. 2001. Friend-or-foe Q-learning in general-sum games. In *Proc. Int. Conf. on Machine Learning (ICML)*, volume 1, 322 – 328.

Monderer, D., and Shapley, L. S. 1996. Potential games. *Games and Economic Behavior* 14(1):124 – 143.

Shoham, Y., and Leyton-Brown, K. 2008. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

Tuyls, K., and Weiss, G. 2012. Multiagent learning: Basics, challenges, and prospects. *AI Magazine* 33(3):41 – 52.

Xu, Y. 2007. Uniformly Lipschitz feedback optimal controls in a linear-quadratic framework. *Journal of Mathematical Analysis and Applications* 327(1):665 – 678.