Approximately Solving Mean Field Games via Entropy-Regularized Deep Reinforcement Learning

Kai Cui
Technische Universität Darmstadt
kai.cui@bcs.tu-darmstadt.de

Heinz Koeppl
Technische Universität Darmstadt
heinz.koeppl@bcs.tu-darmstadt.de

Abstract

The recent mean field game (MFG) formalism facilitates otherwise intractable computation of approximate Nash equilibria in many-agent settings. In this paper, we consider discrete-time finite MFGs subject to finite-horizon objectives. We show that all discrete-time finite MFGs with non-constant fixed point operators fail to be contractive as typically assumed in existing MFG literature, barring convergence via fixed point iteration. Instead, we incorporate entropy-regularization and Boltzmann policies into the fixed point iteration. As a result, we obtain provable convergence to approximate fixed points where existing methods fail, and reach the original goal of approximate Nash equilibria. All proposed methods are evaluated with respect to their exploitability, on both instructive examples with tractable exact solutions and high-dimensional problems where exact methods become intractable. In high-dimensional scenarios, we apply established deep reinforcement learning methods and empirically combine fictitious play with our approximations.

1 Introduction

The framework of mean field games (MFG) was introduced independently by the seminal works of Huang et al. (2006) and Lasry and Lions (2007) in the fully continuous setting of stochastic differential games. In the meantime, it has sparked great interest and investigation both in the mathematical community, where interests lie in the theoretical properties of MFGs, and in the applied research communities as a framework for solving and analyzing large-scale multi-agent problems.

At its core lies the idea of reducing the classical, intractable multi-agent solution concept of Nash equilibria to the interaction between a representative agent and the 'mass' of infinitely many other agents – the so-called mean field. The solution to this limiting problem is the so-called mean field equilibrium (MFE), characterized by a forward evolution equation for the agent's state distributions, and a backward optimality equation of representative agent optimality. Importantly, the MFE constitutes an approximate Nash equilibrium in the corresponding finite agent game of sufficiently many agents (Huang et al. (2006)), which would otherwise be intractable to compute (Daskalakis et al. (2009)).

Nonetheless, computing an MFE remains difficult in the general case. Standard assumptions in existing literature are MFE uniqueness and operator contractivity (Huang et al. (2006), Anahtarcı et al. (2020), Guo et al. (2019)) to obtain convergence via simple fixed point iteration. While these assumptions hold true for some games, we address the case where such restrictive assumptions fail. Applications for such mean field models are manifold and include e.g. finance (Guéant et al. (2011)), power control (Kizilkale and Malhame (2016)), wireless communication (Aziz and Caines (2016)) or public health models (Laguzet and Turinici (2015)).

A motivating example. Consider the following trivial situation informally: Let a large number of agents choose simultaneously between going left (L) or right (R). Afterwards, each agent shall be punished proportional to the number of agents that chose the same action. If we had infinitely many independent, identically acting agents, the only stable solution would be to have all agents pick uniformly at random.

The MFG formalism models this problem by picking one representative agent and abstracting all other agents into their state distribution. Unfortunately, analytically obtaining fixed points in general proves difficult and existing computational methods can fail.

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

Licence: CC BY 4.0 / Creative Commons Attribution 4.0 International

Our contribution. We begin by formulating the mean field analogue to finite games in game theory. In this setting we give simplified proofs for both existence and the approximate Nash equilibrium property of mean field equilibria. Moreover, we show that in finite MFGs, all non-constant fixed point operators are non-contractive, necessitating a different approach than naive fixed point iteration as in Anahtarcı et al. (2020).

Consequently, we approximate the fixed point operator by introducing relative entropy regularization and Boltzmann policies. We prove guaranteed convergence for sufficiently high temperatures, while remaining arbitrarily exact for sufficiently low temperatures. Furthermore, repeatedly iterating on the prior policy allows us to perform an iterative descent on exploitability, successively improving the equilibrium approximation.

Finally, our methods are extensively evaluated and compared to other methods such as fictitious play (FP, see Perrin et al. (2020)), which in general fail to converge to a fixed point. We outperform existing state-of-the-art methods in terms of exploitability in our problems, allowing us to find approximate mean field equilibria in the general case and paving the way to practical application of mean field games. In otherwise intractable problems, we apply deep reinforcement learning techniques together with particle-based simulations.

2 Finite mean field games

2.1 Finite agent games

Consider a discrete-time N-agent stochastic game with finite agent state space $\mathcal{S}$ and finite agent action space $\mathcal{A}$, equipped with the discrete metric. Let $\mathcal{T} = \{0, 1, \dots, T-1\}$ denote the time index set. Denote by $\mathcal{P}(\mathcal{X})$ the set of all Borel probability measures on a metric space $\mathcal{X}$. Since we work with finite spaces, we abuse notation and denote both a measure ν and its probability mass function by $\nu(\cdot)$. For each agent, the dynamical behavior is described by the state transition function $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S}) \to [0, 1]$ and the initial state distribution $\mu_0 : \mathcal{S} \to [0, 1]$. For agents $i = 1, \dots, N$ at times $t \in \mathcal{T}$, their states S^i_t and actions A^i_t are random variables with values in $\mathcal{S}$ and $\mathcal{A}$ respectively. Let $\mathbb{G}^N_s \equiv \frac{1}{N} \sum_{i=1}^N \delta_{s_i}$ denote the empirical measure of agent states $s = (s_1, \dots, s_N) \in \mathcal{S}^N$, where δ is the Dirac measure. Consider for each agent i a Markov policy $\pi^i = (\pi^i_t)_{t \in \mathcal{T}} \in \Pi$, where $\pi^i_t : \mathcal{A} \times \mathcal{S} \to [0, 1]$ and Π is the space of all Markov policies. The state evolution of agent i begins with $S^i_0 \sim \mu_0$ and subsequently for all applicable times t follows

$$\mathbb{P}(A^i_t = a \mid S^i_t = s_i) \equiv \pi^i_t(a \mid s_i) \,,$$
$$\mathbb{P}(S^i_{t+1} = s'_i \mid S_t = s, A^i_t = a) \equiv p(s'_i \mid s_i, a, \mathbb{G}^N_s) \,,$$

for arbitrary $s_i, s'_i \in \mathcal{S}$, $a \in \mathcal{A}$, $s = (s_1, \dots, s_N) \in \mathcal{S}^N$ and $S_t = (S^1_t, \dots, S^N_t)$. Finally, define agent i's finite horizon objective function

$$J^N_i(\pi^1, \dots, \pi^N) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(S^i_t, A^i_t, \mathbb{G}^N_{S_t})\right]$$

to be maximized, where $r : \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S}) \to \mathbb{R}$ is the agent reward function. With this, we can give the notion of optimality used by Saldi et al. (2018).

Definition 1. *A Markov-Nash equilibrium is a* 0*-Markov-Nash equilibrium. For $\varepsilon \geq 0$, an ε-Markov-Nash equilibrium (approximate Markov-Nash equilibrium) is defined as a tuple of policies $(\pi^1, \dots, \pi^N) \in \Pi^N$ such that for any $i = 1, \dots, N$, we have*

$$J^N_i(\pi^1, \dots, \pi^N) \geq \max_{\pi \in \Pi} J^N_i(\pi^1, \dots, \pi^{i-1}, \pi, \pi^{i+1}, \dots, \pi^N) - \varepsilon \,.$$

Since analyzing policies acting on joint state information or the state history is difficult, optimality has been restricted to the set of Markov policies Π acting on the agent's own state. Although this may seem like a significant restriction, in the $N \to \infty$ limit, the evolution of all other agents – the mean field – becomes deterministic and therefore non-informative.

2.2 Mean field games

The $N \to \infty$ limit of the N-agent game constitutes its corresponding finite mean field game (i.e. with a finite state and action space). It consists of the same elements $\mathcal{T}, \mathcal{S}, \mathcal{A}, p, r, \mu_0$. However, instead of modeling N separate agents, it models a single representative agent and collapses all other agents into their common state distribution, i.e. the mean field $\mu = (\mu_t)_{t \in \mathcal{T}} \in \mathcal{M}$ with $\mu_t : \mathcal{S} \to [0, 1]$, where $\mathcal{M}$ is the space of all mean fields and μ_0 is given. The deterministic mean field μ replaces the empirical measure of the finite game. Consider a Markov policy $\pi \in \Pi$ as before. For some fixed mean field μ, the evolution of random states S_t and actions A_t begins with $S_0 \sim \mu_0$ and subsequently for all applicable times t follows

$$\mathbb{P}(A_t = a \mid S_t = s) \equiv \pi_t(a \mid s),$$
$$\mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) \equiv p(s' \mid s, a, \mu_t) \,,$$

and the objective analogously becomes

$$J^\mu(\pi) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(S_t, A_t, \mu_t)\right] .$$

The mean field μ induced by some fixed policy π begins with the given μ_0 and is defined recursively by

$$\mu_{t+1}(s') \equiv \sum_{s \in \mathcal{S}} \mu_t(s) \sum_{a \in \mathcal{A}} \pi_t(a \mid s) p(s' \mid s, a, \mu_t) \,.$$

By fixing a mean field $\mu \in \mathcal{M}$, we obtain an induced Markov Decision Process (MDP) with time-dependent transition function $p(s' \mid s, a, \mu_t)$ and reward function $r(s, a, \mu_t)$. Denote the set-valued map from mean field to optimal policies π of the induced MDP as $\hat{\Phi} : \mathcal{M} \to 2^\Pi$ (i.e. such that π is optimal at any time and state). Analogously, define the map from a policy to its induced mean field as $\Psi : \Pi \to \mathcal{M}$. Finally, we can define the $N \to \infty$ analogue to Markov-Nash equilibria.

Definition 2. *A mean field equilibrium (MFE) is a pair $(\pi, \mu) \in \Pi \times \mathcal{M}$ such that $\pi \in \hat{\Phi}(\mu)$ and $\mu = \Psi(\pi)$ holds.*

By defining any single-valued map $\Phi : \mathcal{M} \to \Pi$ to an optimal policy, we obtain a composition $\Gamma = \Psi \circ \Phi : \mathcal{M} \to \mathcal{M}$, henceforth MFE operator. Shown by Saldi et al. (2018) for general Polish $\mathcal{S}$ and $\mathcal{A}$, the MFE exists and constitutes an approximate Markov-Nash equilibrium for sufficiently many agents under technical conditions. In the Appendix, we give simplified proofs for finite MFGs under the following standard assumption.

Assumption 1. *The functions $r(s, a, \mu_t)$ and $p(s' \mid s, a, \mu_t)$ are continuous, therefore bounded.*

Note that we metrize probability measure spaces $\mathcal{P}(\mathcal{X})$ with the total variation distance d_{TV}. For probability measures ν, ν' on finite spaces $\mathcal{X}$, d_{TV} simplifies to

$$d_{TV}(\nu, \nu') = \frac{1}{2} \sum_{x \in \mathcal{X}} |\nu(x) - \nu'(x)| \,.$$

Accordingly, we equip $\Pi, \mathcal{M}$ with sup metrics, i.e. for policies $\pi, \pi' \in \Pi$ and mean fields $\mu, \mu' \in \mathcal{M}$ we define the metric spaces (Π, d_Π) and $(\mathcal{M}, d_\mathcal{M})$ with

$$\begin{aligned} d_\Pi(\pi, \pi') &\equiv \max_{t \in \mathcal{T}} \max_{s \in \mathcal{S}} d_{TV}(\pi_t(\cdot \mid s), \pi'_t(\cdot \mid s)) \,, \\ d_\mathcal{M}(\mu, \mu') &\equiv \max_{t \in \mathcal{T}} d_{TV}(\mu_t, \mu'_t) \,. \end{aligned}$$

Proposition 1. *Under Assumption 1, there exists at least one MFE $(\pi^*, \mu^*) \in \Pi \times \mathcal{M}$.*

Proof. See Appendix.

Theorem 1. *Under Assumption 1, if (π^*, μ^*) is an MFE, then for any $\varepsilon > 0$ there exists $N' \in \mathbb{N}$ such that for all $N > N'$, the policy $(\pi^*, \dots, \pi^*)$ is an ε-Markov-Nash equilibrium in the N-agent game.*

Proof. See Appendix.

Importantly, finding Nash equilibria in large-N games is hard (Daskalakis et al. (2009)), whereas an MFE can be significantly more tractable to compute. Accordingly, solving the limiting MFG approximately solves the finite-N game for large N in a tractable manner.

3 Exact fixed point iteration

Repeated application of the MFE operator constitutes the exact fixed point iteration approach to finding MFE. The standard assumption for convergence in the literature is contractivity and thereby MFE uniqueness (e.g. Caines and Huang (2019); Guo et al. (2019)).

Proposition 2. *Let Φ, Ψ be Lipschitz with constants c_1, c_2, fulfilling $c_1 c_2 < 1$. Then, the fixed point iteration $\mu^{n+1} = \Psi(\Phi(\mu^n))$ converges to the mean field of the unique MFE for any initial $\mu^0 \in \mathcal{M}$.*

Proof. Let $\mu, \mu' \in \mathcal{M}$ arbitrary, then

$$\begin{aligned} d_\mathcal{M}(\Gamma(\mu), \Gamma(\mu')) &= d_\mathcal{M}(\Psi(\Phi(\mu)), \Psi(\Phi(\mu'))) \\ &\le c_2 \cdot d_\Pi(\Phi(\mu), \Phi(\mu')) \\ &\le c_2 \cdot c_1 \cdot d_\mathcal{M}(\mu, \mu') \,. \end{aligned}$$

Since μ, μ' are arbitrary, Γ is Lipschitz with constant $c_1 \cdot c_2 < 1$. (Π, d_Π) and $(\mathcal{M}, d_\mathcal{M})$ are complete metric spaces (see Appendix). Therefore, Banach's fixed point theorem implies convergence to the unique fixed point for any starting $\mu^0 \in \mathcal{M}$. □

Unfortunately, it remains unclear how to proceed if multiple optimal policies of an induced MDP exist, or if contractivity fails, e.g. when multiple MFE exist. In the following, consider again the illuminating example from the introduction.

3.1 Toy example

Consider $\mathcal{S} = \{C, L, R\}$, $\mathcal{A} = \mathcal{S} \setminus \{C\}$, $\mu_0(C) = 1$, $r(s, a, \mu_t) = -\mathbf{1}_{\{L\}}(s) \cdot \mu_t(L) - \mathbf{1}_{\{R\}}(s) \cdot \mu_t(R)$ and $\mathcal{T} = \{0, 1\}$. The transition function allows picking the next state directly, i.e. for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$,

$$\mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) = \mathbf{1}_{\{s'\}}(a) \,.$$

Clearly, any MFE (π^*, μ^*) must fulfill $\pi^*_0(L \mid C) = \pi^*_0(R \mid C) = 1/2$, while π^*_1 can be arbitrary. Even if the operator Φ chooses suitable optimal policies, the fixed point operator Γ remains non-contractive, as the mean field will necessarily alternate between left and right for any non-uniform starting $\mu^0 \in \mathcal{M}$.

We observe that the example has infinitely many MFE, but no deterministic MFE, i.e. an MFE such that for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ either $\pi_t(a \mid s) = 0$ or $\pi_t(a \mid s) = 1$ holds, similar to the classical game-theoretical insight of mixed Nash equilibrium existence (cf. Fudenberg and Tirole (1991)). Therefore, choosing optimal, deterministic policies will typically fail.

Most existing work assumes contractivity, which is too restrictive. In many scenarios, agents need to "coordinate" with each other. For example, a herd of hunting animals may collectively choose one of multiple hunting grounds, allowing for multiple MFEs. Hence, it can be difficult to apply existing MFG methodologies in practice, as many problems automatically fail contractivity.

3.2 General non-contractivity

From the previous example, we may be led to believe that non-contractivity is a general property of finite MFGs. And indeed, regardless of number of MFEs, it turns out that in any finite MFG with non-constant MFE operator, a policy selection operator Φ with finite image Π_Φ will lead to non-contractivity. Note that this includes both the conventional $\arg\max$ and the $\arg\max$-e (cf. Guo et al. (2019)) choice of actions.

Theorem 2. *Let the image of Φ be a finite set $\Pi_\Phi \subseteq \Pi$. Then, either it holds that $\Gamma = \Psi \circ \Phi$ is a constant, or Γ is not Lipschitz continuous and thus not a contraction.*

Proof. See Appendix.

Therefore, typical discrete-time finite MFGs have non-contractive fixed point operators and we must change our approach. Note that although non-contractivity does not imply non-convergence, the trivial example from before strongly suggests that non-convergence is the case for many finite MFGs.

4 Approximate mean field equilibria

Exact fixed point iteration fails to solve most finite MFGs. Therefore, a different solution approach is necessary. In the following, we present two related approaches that guarantee convergence while plausibly remaining approximate Nash equilibria in the finite-N case. For our results, we require a stronger Lipschitz assumption that implies Assumption 1.

Assumption 2. *The functions $r(s, a, \mu_t)$ and $p(s' \mid s, a, \mu_t)$ are Lipschitz continuous, therefore bounded.*

4.1 Relative entropy mean field games

A straightforward idea is regularization by replacing the objective by the well-known (see e.g. Abdolmaleki et al. (2018)) relative entropy objective

$$\tilde{J}^\mu(\pi) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(S_t, A_t, \mu_t) - \eta \log \frac{\pi_t(A_t \mid S_t)}{q_t(A_t \mid S_t)}\right]$$

with temperature $\eta > 0$ and positive prior policy $q \in \Pi$, i.e. $q_t(a \mid s) > 0$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$. Shown in the Appendix, the unique optimal policy $\tilde{\pi}_t^{\mu,\eta}$ fulfills

$$\tilde{\pi}_t^{\mu,\eta}(a \mid s) = \frac{q_t(a \mid s) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a' \mid s) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,s,a')}{\eta}\right)}$$

for the MDP induced by fixed $\mu \in \mathcal{M}$, with the soft action-value function $\tilde{Q}_\eta(\mu, t, s, a)$ given by the smooth-maximum Bellman recursion

$$\tilde{Q}_\eta(\mu, t, s, a) = r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \cdot \eta \log\left(\sum_{a' \in \mathcal{A}} q_{t+1}(a' \mid s') \exp \frac{\tilde{Q}_\eta(\mu, t+1, s', a')}{\eta}\right)$$

of the MDP induced by fixed $\mu \in \mathcal{M}$, with terminal condition $\tilde{Q}_\eta(\mu, T-1, s, a) \equiv r(s, a, \mu_{T-1})$. Note that we recover optimality as $\eta \to 0$, see Theorem 4. Define the relative entropy MFE operator $\tilde{\Gamma}_\eta \equiv \Psi \circ \tilde{\Phi}_\eta$ with policy selection $\tilde{\Phi}_\eta(\mu) \equiv \tilde{\pi}^{\mu,\eta}$ for all $\mu \in \mathcal{M}$.

Definition 3. *An η-relative entropy mean field equilibrium (η-RelEnt MFE) for some positive prior policy $q \in \Pi$ is a pair $(\pi^E, \mu^E) \in \Pi \times \mathcal{M}$ such that $\pi^E = \tilde{\Phi}_\eta(\mu^E)$ and $\mu^E = \Psi(\pi^E)$ hold. An η-maximum entropy mean field equilibrium (η-MaxEnt MFE) is an η-RelEnt MFE with uniform prior policy q.*

RelEnt MFE are guaranteed to exist for any $\eta > 0$ by Proposition 3. Furthermore, convergence to the regularized solution is guaranteed for large η by Theorem 3.

4.2 Boltzmann iteration

Since only deterministic policies fail, a derivative approach is to use softmax policies directly with the unregularized action-value function, also called Boltzmann policies. Assume that the action-value function Q^* fulfilling the Bellman equation

$$Q^*(\mu, t, s, a) = r(s, a, \mu_t) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a, \mu_t) \cdot \max_{a' \in \mathcal{A}} Q^*(\mu, t+1, s', a') \,.$$

of the MDP induced by fixed $\mu \in \mathcal{M}$ with terminal condition $Q^*(\mu, T-1, s, a) \equiv r(s, a, \mu_{T-1})$ is known. Define the map $\Phi_\eta(\mu) \equiv \pi^{\mu,\eta}$ for any $\mu \in \mathcal{M}$, where

$$\pi_t^{\mu,\eta}(a \mid s) \equiv \frac{q_t(a \mid s) \exp\left(\frac{Q^*(\mu,t,s,a)}{\eta}\right)}{\sum_{a' \in \mathcal{A}} q_t(a' \mid s) \exp\left(\frac{Q^*(\mu,t,s,a')}{\eta}\right)}$$

for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$ and temperature $\eta > 0$.

Definition 4. *An η-Boltzmann mean field equilibrium (η-Boltzmann MFE) for some positive prior policy $q \in \Pi$ is a pair $(\pi^B, \mu^B) \in \Pi \times \mathcal{M}$ such that $\pi^B = \Phi_\eta(\mu^B)$ and $\mu^B = \Psi(\pi^B)$ hold.*

4.3 Theoretical properties

Both η-RelEnt MFE and η-Boltzmann MFE are guaranteed to exist for any temperature $\eta > 0$.

Proposition 3. *Under Assumption 1, η-Boltzmann and η-RelEnt MFE exist for any temperature $\eta > 0$.*

Proof. See Appendix.

Contractivity of both η-Boltzmann MFE operator $\Gamma_\eta \equiv \Psi \circ \Phi_\eta$ and η-RelEnt MFE operator $\tilde{\Gamma}_\eta \equiv \Psi \circ \tilde{\Phi}_\eta$ is guaranteed for sufficiently high temperatures, even if all possible original Φ are not Lipschitz continuous.

Theorem 3. *Under Assumption 2, $\mu \mapsto Q^*(\mu, t, s, a)$, $\mu \mapsto \tilde{Q}_\eta(\mu, t, s, a)$ and $\Psi(\pi)$ are Lipschitz continuous with constants K_{Q^*}, $K_{\tilde{Q}}$ and K_Ψ for arbitrary $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}, \eta > \eta', \eta' > 0$. Furthermore, Γ_η and $\tilde{\Gamma}_\eta$ are a contraction for*

$$\eta > \max\left(\eta', \frac{|\mathcal{A}|\left(|\mathcal{A}|-1\right)K_Q K_\Psi q_{\max}^2}{2q_{\min}^2}\right)$$

where $K_Q = K_{Q^*}$ for Γ_η, $K_Q = K_{\tilde{Q}}$ for $\tilde{\Gamma}_\eta$, $q_{\max} \equiv \max_{t\in\mathcal{T},s\in\mathcal{S},a\in\mathcal{A}} q_t(a \mid s) > 0$ and $q_{\min} \equiv \min_{t\in\mathcal{T},s\in\mathcal{S},a\in\mathcal{A}} q_t(a \mid s) > 0$.

Proof. See Appendix.

Sufficiently large η hence implies convergence via fixed point iteration. On the other hand, for sufficiently low temperatures η, both η-Boltzmann and η-RelEnt MFE will also constitute an approximate Markov-Nash equilibrium of the finite-N game.

Theorem 4. *Under Assumption 2, if $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ is a sequence of η_n-Boltzmann or η_n-RelEnt MFE with $\eta_n \to 0$, then for any $\varepsilon > 0$ there exist $n', N' \in \mathbb{N}$ such that for all $n > n', N > N'$, the policy $(\pi_n^*, \dots, \pi_n^*) \in \Pi^N$ is an ε-Markov-Nash equilibrium of the N-agent game, i.e.*

$$\begin{aligned} &J_i^N(\pi_n^*, \dots, \pi_n^*) \geq \\ &\quad \max_{\pi_i \in \Pi} J_i^N(\pi_n^*, \dots, \pi_n^*, \pi_i, \pi_n^*, \dots, \pi_n^*) - \varepsilon\,. \end{aligned}$$

Proof. See Appendix.

If we can obtain contractivity for sufficiently low η, we can find good approximate Markov-Nash equilibria. As it is impossible to have both $\eta \to 0$ and $\eta \to \infty$, it depends on the problem and prior whether we can converge to a good solution. Nonetheless, we find that it is often possible to empirically find low η that provide convergence as well as a good approximate MFE.

4.4 Prior descent

In principle, we can insert arbitrary prior policies $q \in \Pi$. Under Assumption 1, by boundedness of both $\tilde{Q}_\eta$ and Q^* (see Appendix), both η-RelEnt and η-Boltzmann MFE policies converge to the prior policy as $\eta \to \infty$. Therefore, in principle we can show that for any $\varepsilon > 0$, for sufficiently large η and N, the η-RelEnt and η-Boltzmann MFE under q will be at most an ε-worse approximate Nash equilibrium than the prior policy. Furthermore, we obtain guaranteed contractivity by Theorem 3. Thus, any prior policy gives a worst-case bound on the performance achievable over all $\eta > 0$. On the other hand, if we obtain better results for sufficiently low η, we may iteratively improve our policy and thus our equilibrium quality.

5 Related work

The original work of Huang et al. (2006) introduces contractivity and uniqueness assumptions into the continuous MFG setting. Analogously, Guo et al. (2019) and Caines and Huang (2019) assume contractivity for discrete-time MFGs and dense graph limit MFGs respectively. Further existing work on discrete-time MFGs similarly assumes uniqueness of the MFE, which includes Saldi et al. (2018) and Gomes et al. (2010) for approximate optimality and existence results, and Anahtarcı et al. (2020) for an analysis on contractivity requirements. Mguni et al. (2018) solve discrete-time continuous state MFG problems under the classical uniqueness conditions of Lasry and Lions (2007). Further extensions of the MFG formula include partial observability (Saldi et al. (2019)) or major agents (Nourian and Caines (2013)).

The work of Anahtarci et al. (2020) is related and studies theoretical properties of finite-N regularized games and their limiting MFG. In their work, the existence and approximate Nash property of MFE in stationary regularized games is shown, and Q-Learning error propagation is investigated. In comparison, we consider the original, unregularized finite-N game in a transient setting and perform extensive empirical evaluations. Guo et al. (2019) and Yang et al. (2018) previously proposed to apply Boltzmann policies. The former applies the approximation heuristically, while the latter focuses on directly solving finite-N games.

An orthogonal approach to computing MFE is fictitious play. Rooted in game-theory and classical economic works (Brown (1951)), it has since been adapted to MFGs. In fictitious play, all past mean fields (Cardaliaguet and Hadikhanloo (2017)) and policies (Perrin et al. (2020)) are averaged to produce a new mean field or policy. Importantly, convergence is guaranteed in certain special cases only (cf. Elie et al. (2019)). Although introduced in a differentiable setting, we evaluate fictitious play empirically in our setting and find that both our regularization and fictitious play may be combined successfully.

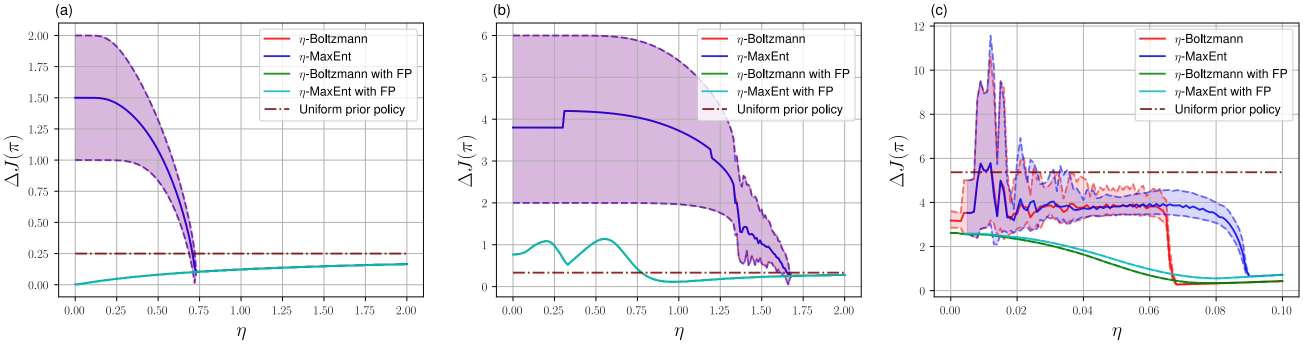

Figure 1: Mean exploitability over the final 10 iterations. Dashed lines represent maximum and minimum over the final 10 iterations. (a) LR, 10000 iterations; (b) RPS, 10000 iterations; (c) SIS, 10000 iterations. Maximum entropy (MaxEnt) results begin at higher temperatures due to limited floating point accuracy. Temperature zero depicts the exact fixed point iteration for both η-MaxEnt and η-Boltzmann MFE. In LR and RPS, η-MaxEnt and η-Boltzmann MFE coincide both with and without fictitious play (FP), here averaging both policy and mean field over all past iterations. The exploitability of the prior policy is indicated by the dashed horizontal line.

6 Evaluation

In practice, we find that our approaches are capable of generating solutions of lower exploitability than otherwise obtained. Unless stated otherwise, we compute everything exactly, use the maximum entropy objective (MaxEnt) with the uniform prior policy q where $q_t(a \mid s) = 1/|\mathcal{A}|$ for all $t \in \mathcal{T}, s \in \mathcal{S}, a \in \mathcal{A}$, and initialize with $\mu^0 = \Psi(q)$ generated by q. As the main evaluation metric, we define the exploitability of a policy $\pi \in \Pi$ with induced mean field $\mu \equiv \Psi(\pi)$ as

$$\Delta J(\pi) \equiv \max_{\pi^*} J^\mu(\pi^*) - J^\mu(\pi) \ .$$

Clearly, the exploitability of π is zero if and only if (π, μ) is an MFE. Indeed, for any $\varepsilon > 0$, any policy $\pi \in \Pi$ is a $(\Delta J(\pi) + \varepsilon)$-Markov Nash equilibrium if N sufficiently large, i.e. the exploitability translates directly to the limiting equilibrium quality in the finite-N game, see also Theorem 4 and its proof.

We evaluate the algorithms on the LR, RPS, SIS and Taxi problems, ordered in increasing complexity. Details of the algorithms, hyperparameters, problems and experiment configurations as well as further experimental results can be found in the Appendix.

6.1 Exploitability

In Figure 1, we plot the minimum, maximum and mean exploitability for varying temperatures η during the last 10 fixed point iterations, i.e. a single value when the exploitability (and usually mean field) converges. Observe that the lowest convergent temperature outperforms not only the exact fixed point iteration (drawn at temperature zero), but also the uniform prior policy. Although developed for a different setting, we also show results of fictitious play similar to the version from Perrin et al. (2020), i.e. both policies and mean fields are averaged over all past iterations. It can be seen that fictitious play only converges to the optimal solution in the LR problem. In the other examples, supplementing fictitious play with entropy regularization is effective at producing better results. A non-existent fictitious play variant averaging only the policies finds the exact MFE in RPS, but nevertheless fails in SIS. See the Appendix for further results.

Evaluating and solving finite-N games is highly intractable by the curse of dimensionality, as the local state is no longer sufficient to perform dynamic programming in the presence of the random empirical state measure. Since it has already been proven that the exploitability for $N \to \infty$ will converge to the exploitability of the corresponding mean field game, we refrain from evaluating on finite-N games.

Note that the plots are entirely deterministic and not stochastic as it would seem at first glance, since the depicted shaded area visualizes the non-convergence of exploitability and is a result of the fixed point updates running into a limit cycle (cf. Figure 2).

6.2 Convergence

In Figure 2, the difference between the exploitability of the current policy and the minimal exploitability reached during the final 10 iterations is shown for η-Boltzmann MFE. As the temperature η decreases, time to convergence increases until non-convergence is reached in form of a limit cycle. Analogous results for η-RelEnt MFE can be found in the Appendix.

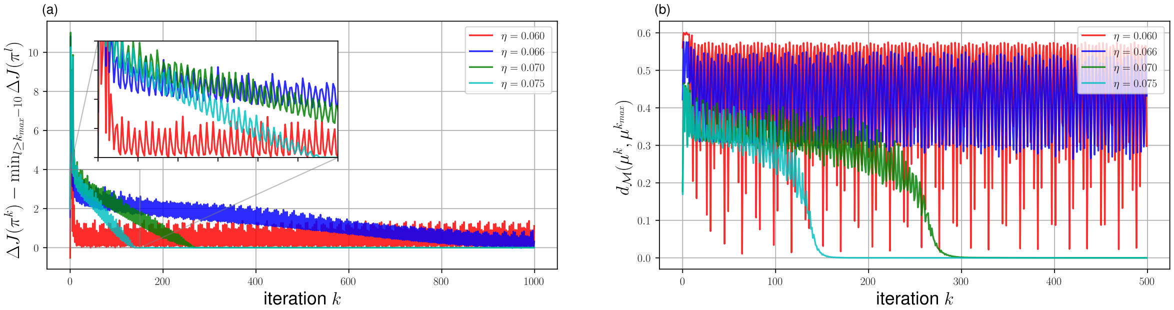

Figure 2: (a) Difference between current and final minimum exploitability over the last 10 iterations; (b) Distance between current and final mean field. Plotted for the η-Boltzmann MFE iterations in SIS for different indicated temperature settings. Note the periodicity of the lowest temperature setting, indicating a limit cycle.

Note also that in LR, we can analytically find $K_Q = 1$ and $K_\Psi = 1$. Thus, we obtain guaranteed convergence via η-Boltzmann MFE iteration if $\eta > 1$. In Figure 1, we see convergence already for $\eta \geq 0.7$. Note further that the non-converged regime can allow for lower exploitability. However, it is unclear a priori when to stop, and for approximate solutions where DQN is used for evaluation, the evaluation of exploitability may become inaccurate.

6.3 Deep reinforcement learning

For problems with intractably large state spaces, we adopt the DQN algorithm (Mnih et al. (2013)), using the implementation of Shengyi et al. (2020) as a base. Particle-based simulations are used for the mean field, and stochastic performance evaluation on the induced MDP is performed (see Appendix). Note that the approximation introduces three sources of stochasticity into the otherwise deterministic algorithms, i.e. stochastic evaluation, mean field simulation and DQN. To counteract the randomness, we average our results over multiple runs. The hyperparameters and architectures used are standard and can be found in the Appendix.

Fitting the soft action-value function directly using a network is numerically problematic, as the log-exponential transformation of approximated action-values quickly fails due to limited floating point accuracy. Thus, we limit ourselves to the classical Bellman equation with Boltzmann policies only.

In Figure 3, we evaluate the exploitability of Boltzmann DQN iteration, evaluated exactly in SIS and RPS, and stochastically in Taxi over 2000 realizations. Minimum, maximum and mean exploitability are taken over the final 5 iterations and averaged over 5 seeds. Note that it is very time-consuming to solve a full reinforcement learning problem using DQN repeatedly in every iteration. Nonetheless, we observe that a temperature larger than zero appears to improve exploitability and convergence in the SIS example. Both due to the noisy nature of approximate solutions and the lower number of iterations, it can be seen that a higher temperature is required to converge than in the exact case.

In the intractable Taxi environment, the policy oscillates between two modes as in exact LR, and regularization fails to obtain better results, see also the Appendix. An important reason is that the prior policy performs extremely bad (exploitability of ~ 35) as most states require specific actions for optimality. Hence we cannot find an $\eta > 0$ for which the algorithm both converges and performs well. Using prior descent and iteratively refining a better prior policy would likely increase performance, but is deferred to future investigations as the required computations grow very large.

Fictitious play is expensive in combination with approximate Q-Learning and particle simulations, as policies and particles of past iterations must be kept to perform exact fictitious play. For this reason, we do not attempt approximate fictitious play with approximate solution methods. In theory, supervised learning for fitting summarizing policies and randomly sampling particles may help, but is out of scope of this paper.

6.4 Prior descent

In Figure 4, we repeatedly perform outer iterations consisting of 100 η-RelEnt MFE iterations each with the indicated fixed temperature parameters in SIS. After each outer iteration, the prior policy is updated to the newest resulting policy. Note again that the results are entirely deterministic.

Searching for a suitable η dynamically every iteration would keep the exploitability from increasing, as for $\eta \to \infty$ we obtain the original prior policy. Since it is expensive to scan over all temperatures in each outer iteration, we use a heuristic. Intuitively, since the prior will become increasingly good, it will be increasingly

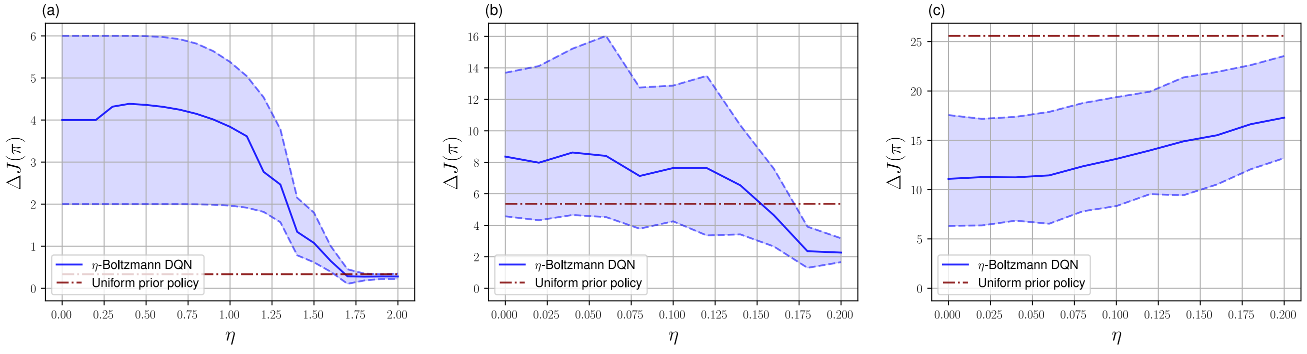

Figure 3: Mean exploitability over the final 5 iterations using DQN, averaged over 5 seeds. Dashed lines represent the averaged maximum and minimum exploitability over the last 5 iterations. (a) RPS, 1000 iterations; (b) SIS, 50 iterations; (c) Taxi, 15 iterations. Evaluation of exploitability is exact except in Taxi, which uses DQN and averages over 1000 episodes. The point of zero temperature depicts fixed point iteration using exact DQN policies.

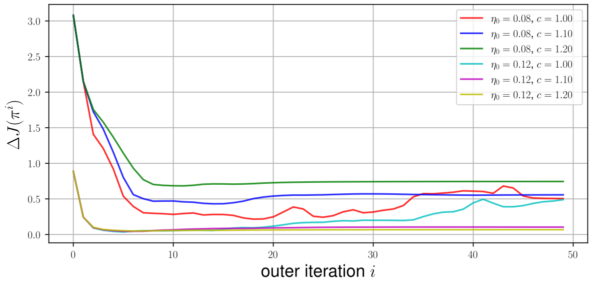

Figure 4: Exploitability over outer iterations in SIS, using 100 η-RelEnt MFE iterations per outer iteration. Note that the results are deterministic. Not shown: Running the fixed temperature settings $c = 1$ for longer does not converge for at least 1000 iterations.

difficult to obtain a better policy. Thus, increasing the temperature will help sticking close to the prior and converge. Consequently, we use the simple heuristic

$$\eta_{i+1} = \eta_i \cdot c$$

for each outer iteration i, where $c \geq 1$ adjusts the temperature after each outer iteration.

Importantly, even for our simple heuristic, prior descent already achieves an exploitability of ~ 0.068, whereas the best results for the fixed uniform policy from Figure 1 show an optimal mean exploitability of ~ 0.281. Furthermore, repeated prior policy updates succeed in computing the exact MFE in RPS and LR under a fixed temperature (see Appendix).

Note that prior descent creates a double loop around solving the optimal control problem, becoming highly expensive under deep reinforcement learning. Hence, we refrain from prior descent with DQN. Automatically adjusting temperatures to monotonically improve exploitability is left for potential future work.

7 Conclusion

In this work, we have investigated the necessity and feasibility of approximate MFG solution approaches – entropy regularization, Boltzmann policies and prior descent – in the context of finite MFGs. We have shown that the finite MFG case typically cannot be solved by exact fixed point iteration or fictitious play alone. Entropy regularization and Boltzmann policies in combination with deep reinforcement learning may enable feasible computation of approximate MFE. We believe that lifting the restriction of inherent contractivity is an important step in ensuring applicability of MFG models in practical problems. We hope that entropy regularization and the insight for finite MFGs can help transfer the MFG formalism from its so-far mostly theory-focused context into real world application scenarios. Nonetheless, there still remain many restrictions to the applicability of the MFG formalism.

For future work, an efficient, automatic temperature adjustment for prior descent could be fruitful. Furthermore, it would be interesting to generalize relative entropy MFGs to infinite horizon discounted problems, continuous time, and continuous state and action spaces. Moreover, it could be of interest to investigate theoretical properties of fictitious play in finite MFGs in combination with entropy regularization. For non-Lipschitz mappings from policy to induced mean field, the proposed approach does not provide a solution. It could nonetheless be important to consider problems with threshold-type dynamics and rewards, e.g. majority vote problems. Most notably, the current formalism precludes common noise entirely, i.e. any games with common observations. In practice, many problems will allow for some type of common observation between agents, leading to non-independent agent distributions and stochastic as opposed to deterministic mean fields.

Acknowledgements

This work has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center. The authors acknowledge the Lichtenberg high performance computing cluster of the TU Darmstadt for providing computational facilities for the calculations of this research.

References

Minyi Huang, Roland P Malhamé, Peter E Caines, et al. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

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

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Berkay Anahtarcı, Can Deha Karıksız, and Naci Saldi. Value iteration algorithm for mean-field games. *Systems & Control Letters*, 143:104744, 2020.

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

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

Arman C Kizilkale and Roland P Malhame. Collective target tracking mean field control for markovian jump-driven models of electric water heating loads. In *Control of Complex Systems*, pages 559–584. Elsevier, 2016.

Mohamad Aziz and Peter E Caines. A mean field game computational methodology for decentralized cellular network optimization. *IEEE transactions on control systems technology*, 25(2):563–576, 2016.

Laetitia Laguzet and Gabriel Turinici. Individual vaccination as nash equilibrium in a sir model with application to the 2009–2010 influenza a (h1n1) epidemic in france. *Bulletin of Mathematical Biology*, 77(10):1955–1984, 2015.

Sarah Perrin, Julien Perolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *arXiv preprint arXiv:2007.03458*, 2020.

Naci Saldi, Tamer Basar, and Maxim Raginsky. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Peter E Caines and Minyi Huang. Graphon mean field games and the gmfg equations: ε-nash equilibria. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 286–292. IEEE, 2019.

Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.

Diogo A Gomes, Joana Mohr, and Rafael Rigao Souza. Discrete time, finite state space mean field games. *Journal de mathématiques pures et appliquées*, 93(3): 308–328, 2010.

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

Naci Saldi, Tamer Başar, and Maxim Raginsky. Approximate nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*, 44(3):1006–1033, 2019.

Mojtaba Nourian and Peter E Caines. Epsilon-nash mean field game theory for nonlinear stochastic dynamical systems with major and minor agents. *SIAM Journal on Control and Optimization*, 51(4):3302–3331, 2013.

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

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5571–5580, 2018.

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

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

Romuald Elie, Julien Pérolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. Approximate fictitious play for mean field games. *arXiv preprint arXiv:1907.02633*, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Huang Shengyi, Dossa Rousslan, and Chang Ye. Cleanrl: High-quality single-file implementation of deep reinforcement learning algorithms. `https://github.com/vwxyzjn/cleanrl/`, 2020.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003, 2016.

Lloyd Shapley. Some topics in two-person games. *Advances in game theory*, 52:1–29, 1964.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361, 2017.

Boris Belousov and Jan Peters. Entropic regularization of markov decision processes. *Entropy*, 21(7):674, 2019.