A policy gradient algorithm to compute boundedly rational stationary mean field equilibria

Jayakumar Subramanian
McGill University
Montreal, Quebec
jayakumar.subramanian@mail.mcgill.ca

Aditya Mahajan
McGill University
Montreal, Quebec
aditya.mahajan@mcgill.ca

ABSTRACT

In this paper, we define a bounded rationality based generalization of stationary mean field equilibrium that we call gradient based stationary mean field equilibrium. Unlike Nash equilibrium and its variants, where each agent plays a best response policy given the policy of others, in a gradient based equilibrium, each plays a policy such that the performance gradient with respect to the policy parameters is zero. We then propose a policy gradient based algorithm to compute gradient-based stationary mean field equilibrium. We demonstrate the performance of this algorithm using a numerical experiment based on malware spread in networks.

KEYWORDS

Mean field games, multi-agent systems, stationary equilibria, policy gradient, bounded rationality

1 INTRODUCTION

Several problems such as automated trading in large financial markets, control of smart grids, decision making in industries with many firms involve systems with large number of agents. Another characteristic feature of such problems is that each agent is usually small, i.e., a single agent cannot meaningfully affect the system. Such systems are called mean field games and the corresponding solution concept is called mean field equilibrium (MFE) [11, 16, 23]. A mean field equilibrium is a trajectory of policies and a trajectory of mean field distributions satisfying sequential rationality and consistency. Here, sequential rationality requires that the specified policy trajectory of each agent is a best response to the given trajectory of mean field distributions. Consistency requires that the given trajectory of the mean field distributions is generated when agents play the specified policy trajectory.

Further simplification is possible when the system dynamics are such that the mean field becomes stationary. The relevant solution concepts in this case are stationary mean field equilibrium (SMFE) (also called stationary equilibrium) [1] and oblivious equilibrium [47]. A stationary mean field equilibrium is a time-homogeneous policy and a time-homogeneous mean field distribution satisfying sequential rationality and consistency. Here, sequential rationality requires that the specified time-homogeneous policy is a best response to the given time-homogeneous mean field distribution. Consistency requires that the given time-homogeneous mean field distribution is stationary when agents play the specified time-homogeneous policy.

However, even in such cases, computing the best response might be a challenging task for each agent. An agent with limited computational capability, or bounded rationality might choose to satisfice rather than optimize [40]. In this paper, motivated by bounded rationality, we define a generalization of SMFE that is easier to compute. We call this solution concept as gradient based stationary mean field equilibrium (∇-SMFE). ∇-SMFE is a time-homogeneous oblivious policy and a time-homogeneous (stationary) mean field distribution (belief) satisfying gradient based sequential rationality and consistency. Here, gradient based sequential rationality requires that the specified time-homogeneous policy is a local best response (i.e., the policy at which the gradient of the agent's assessment of performance is zero) to the given time-homogeneous mean field distribution. Consistency requires that the given time-homogeneous mean field distribution is generated when agents play the specified time-homogeneous policy. In this paper, we present an algorithm to compute ∇-SMFE that is a stochastic approximation algorithm consisting of two components: (i) a particle filter based approach to compute the stationary distribution corresponding to a policy; and (ii) a policy gradient based approach to compute the gradient of the performance of a policy. We illustrate the algorithm using an example considering a stylized model of malware spread in networks.

Policy gradient based methods are one of the key approaches in reinforcement learning [2, 17, 21, 34, 44]. Recent extensions of policy gradient algorithms such as trust region based methods [36, 37, 49] have met with several successes. Reinforcement learning methods have also been used in game theory and multi-agent systems [6, 7, 10, 22, 24–30, 32, 33, 45, 46]. See [4, 38, 39] for an overview. Reinforcement learning in mean field games has also been presented in [20, 50, 51]. However, in this paper, we pursue a different approach to determining equilibria in stationary mean field games.

1.1 Notation

The letter n denotes the number of agents and $N = \{1, \dots, n\}$ denotes the set of agents. $\mathcal{X}$ denotes the state space and $\mathcal{A}$ denotes the action space. In general capital letters denote random variables such as X for state and A for action, while corresponding small letters such as x and a denote their values respectively. For any discrete set $\mathcal{S}$, $\Delta(\mathcal{S})$ denotes the space of probability mass functions on $\mathcal{S}$. Policies are represented using μ and their parameters using θ. In general, we assume stochastic policies, i.e, $\mu : \mathcal{X} \to \Delta(\mathcal{A})$. Using a slight abuse of notation, we also denote the probability of a particular action $a \in \mathcal{A}$, under policy μ in state $x \in \mathcal{X}$ as $\mu(a|x)$. ξ denotes the empirical mean field (or population average) and π denotes the

ICML/IJCAI/AAMAS 2018 Workshop on Planning and Learning (PAL'18), July 2018, Stockholm, Sweden.

statistical mean field (or infinite population limit of population average). We mostly use subscripts to denote time and in some cases the policy or the mean field. We sometimes use a subscript range to denote a set of quantities, such as, $X_{0:t} = \{X_0, X_1, \dots, X_t\}$. Bold letters are used to denote profiles, for instance, $\boldsymbol{X_t} := \{X^i_t\}_{i \in \{1,\dots,n\}}$. $[\cdot]_\Theta$ is projection onto Θ. Other variables are defined when they are introduced in the text.

2 MODEL

Consider a non-zero sum stochastic dynamic game with imperfect information that runs for an infinite horizon. Let $N := \{1, \dots, n\}$ denote the set of agents. Each agent $i, i \in N$, has a local state $X^i_t \in \mathcal{X}$ and chooses actions $A^i_t \in \mathcal{A}$. The state space $\mathcal{X}$ and the action space $\mathcal{A}$ are finite and identical for all agents. Let $\boldsymbol{X_t} := \{X^1_t, \dots, X^n_t\}$ and $\boldsymbol{A_t} := \{A^1_t, \dots, A^n_t\}$ denote the state and action of all agents. The state of agent $i, i \in N$, evolves in a controlled Markovian manner; in particular, for any $x^i \in \mathcal{X}$, we have

$$\mathbb{P}[X^i_{t+1} = x^i \mid \boldsymbol{X}_{1:t}, \boldsymbol{A}_{1:t}] = \mathbb{P}[X^i_{t+1} = x^i \mid X^i_t, A^i_t] =: P(x^i \mid X^i_t, A^i_t). \tag{1}$$

All agents have identical dynamics with controlled transition matrices $\{P(\cdot|\cdot, a)\}_{a \in \mathcal{A}}$. The per-step payoff to agent i is given by a utility function $u(X^i_t, A^i_t, \xi_t)$, where $\xi_t \in \Delta(\mathcal{X})$ is the population average (or the empirical mean field), which is given by:

$$\xi_t(x) = \frac{1}{n} \sum_{i \in N} \mathbb{1}\{X^i_t = x\}, \quad \forall x \in \mathcal{X}. \tag{2}$$

Note that the utility function is the same for all agents. In general each agent $i, i \in N$, may use a history dependent behavioral policy $\mu^i_t : \mathcal{X}^t \to \Delta(\mathcal{A})$. Let $\mu^i = \{\mu^i_0, \mu^i_1, \dots\}$ denote the policy of agent i for all time. The collection $\boldsymbol{\mu} := \{\mu^1, \dots, \mu^n\}$ is called the policy profile for all agents. Given a policy profile $\boldsymbol{\mu}$, the payoff of agent $i, i \in N$, is given by:

$$U^i(x; \boldsymbol{\mu}) = \mathbb{E}_{A^i_t \sim \mu^i_t(X^i_{1:t})} \Big[\sum_{t=0}^{\infty} \gamma^t u(X^i_t, A^i_t, \xi_t) \Bigm| \boldsymbol{X_0} = x \Big]. \tag{3}$$

When the number of agents is large, identifying a perfect Bayesian equilibrium of the above game is quite challenging. Therefore, four simplifying assumptions have been considered in the literature [1, 5, 12, 15, 16, 23, 47]:

(1) First, attention is restricted to oblivious policies, i.e., it is assumed that an agent uses only its current state to pick a distribution of actions. Thus, $\mu^i_t : \mathcal{X} \to \Delta(\mathcal{A})$ and $A^i_t \sim \mu^i_t(X^i_t)$.
(2) Second, attention is restricted to time-homogeneous oblivious policies, i.e., it is assumed that μ^i_t does not depend on time.
(3) Third, attention is restricted to symmetric policies, i.e., it is assumed that all agents play identical (oblivious) policies. Thus $\boldsymbol{\mu} = \{\mu, \mu, \dots, \mu\}$. For ease of notation, we simply refer to the policy profile as μ. For a symmetric policy, we have that for any $y \in \mathcal{X}$:

$$\xi_{t+1}(y) = \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \xi_t(x) \mu(a \mid x) P(y \mid x, a). \tag{4}$$

Thus, we may write,

$$\xi_{t+1} = \Phi(\xi_t, \mu). \tag{5}$$

(4) Finally, while evaluating the payoff, each agent assumes that the population average is stationary. Thus agent i's assessment of its payoff is:

$$V^i_{\mu,\pi}(x) = \mathbb{E}_{A^i_t \sim \mu(X^i_t)} \Big[\sum_{t=0}^{\infty} \gamma^t u(X^i_t, A^i_t, \pi) \Bigm| X^i_0 = x \Big]. \tag{6}$$

Note that this assessment is identical for all agents. So, in the sequel, we simply denote it by $V_{\mu,\pi}$.

Under these conditions, the following refinement of Nash equilibrium is used as a solution concept [1, 47].

Definition 2.1 (Stationary mean field equilibrium (SMFE)). A stationary mean field equilibrium (SMFE) is a pair of a belief $\pi \in \Delta(\mathcal{X})$ and a time-homogeneous oblivious policy $\mu : \mathcal{X} \to \Delta(\mathcal{A})$ which satisfies the following two properties:

(1) *Sequential Rationality*: For any other time-homogeneous oblivious policy $\tilde{\mu} : \mathcal{X} \to \Delta(\mathcal{A})$,

$$V_{\mu,\pi}(x) \geq V_{\tilde{\mu},\pi}(x), \quad \forall x \in \mathcal{X}. \tag{7}$$

(2) *Consistency*: The belief π is stationary under policy μ, i.e.,

$$\pi = \Phi(\pi, \mu). \tag{8}$$

Note that it is possible to write a recursive expression for $V_{\mu,\pi}$ using the standard dynamic programming decomposition. In particular, $V_{\mu,\pi}$ is given by the unique bounded solution of the following fixed point equation:

$$Q_{\mu,\pi}(x, a) = u(x, a, \pi) + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}[y \mid x, a) V_{\mu,\pi}(y)], \tag{9}$$

$$V_{\mu,\pi}(x) = \sum_{a \in \mathcal{A}} \mu(a|x) Q_{\mu,\pi}(x, a). \tag{10}$$

Given the above dynamic program, we may rewrite the sequential rationality condition of SMFE as follows:

(1') Given a $\pi \in \Delta(\mathcal{X})$, let V^*_π be the unique bounded solution of the following fixed point equation:

$$Q^*_\pi(x, a) = u(x, a, \pi) + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}[y \mid x, a) V^*_\pi(y)], \tag{11}$$

$$V^*_\pi(x) = \max_{a \in \mathcal{A}} \{Q^*_\pi(x, a)\}. \tag{12}$$

Then a policy $\mu : \mathcal{X} \to \Delta(\mathcal{A})$ is sequentially rational given π if and only if

$$\operatorname{supp}\{\mu(x)\} \in \arg\max_{a \in \mathcal{A}} \{Q^*_\pi(x, a)\}, \tag{13}$$

which means that every action which has a positive weight under $\mu(x)$ is a best response to π.

3 BOUNDEDLY RATIONAL STATIONARY MEAN FIELD EQUILIBRIA

SMFE provides a drastic simplification over perfect Bayesian equilibrium or even Markov perfect equilibrium for large population games. In spite of this simplification, numerically computing SMFE is still a formidable task. The obvious solution approach is to start with a guess $\pi^{(1)}$ for the stationary mean field distribution, find the best response $\mu^{(1)}$ by solving (12), and then find the stationary distribution $\pi^{(2)}$ corresponding to $\mu^{(1)}$ and then iterate. There are two difficulties in such an approach. The first is a conceptual difficulty:

it is not clear that such an iteration will converge. The second is a computational difficulty: each step of the iteration requires solving a dynamic program, which can be prohibitively difficult when the state space is large or continuous. Various sufficient conditions under which the above iteration converges have been identified in [1, 9, 13, 15, 47]. These partially resolve the first conceptual difficulty. In this paper, we propose a methodology to resolve the second computational difficulty.

To obtain a computationally tractable method to solve the dynamic program, we assume that the agents have bounded rationality [40] and restrict attention to a class of parametrized policies μ_θ, where the parameter θ belongs to a closed convex subset Θ of a Euclidean space. With a slight abuse of notation, we also sometimes use θ to denote μ_θ. In addition, we assume that instead of finding a global best response,

$$\theta \in \arg\max_{\theta \in \Theta} V_{\theta,\pi}, \tag{14}$$

the boundedly rational agent is satisfied with a local best response,

$$\theta \text{ is such that } \nabla_\theta V_{\theta,\pi} = 0. \tag{15}$$

Based on this, we define the following generalization of SMFE, which we call gradient based SMFE and denote by ∇-SMFE.

Definition 3.1 (∇-SMFE). A gradient based SMFE (∇-SMFE) is a pair of belief $\pi \in \Delta(\mathcal{X})$ and a parametrized policy $\mu_\theta : \mathcal{X} \to \Delta(\mathcal{A})$, where $\theta \in \Theta$, which satisfies the following two properties:

(1) *Gradient based sequential rationality*: Let $V_{\theta,\pi}$ be the fixed point of (9) and (10). Then,

$$\nabla_\theta V_{\theta,\pi} = 0. \tag{16}$$

(2) *Consistency*: The belief π is stationary under policy μ_θ, i.e.,

$$\pi = \Phi(\pi, \mu_\theta). \tag{17}$$

Proposition 3.2. *Given a stationary belief $\pi \in \Delta(\mathcal{X})$, an initial distribution $\xi_0 \in \Delta(\mathcal{X})$, and a policy μ_θ, $\theta \in \Theta$, define the agent's assessment of the payoff as:*

$$J_{\theta,\pi} := \mathbb{E}_{X\sim\xi_0}[V_{\theta,\pi}(X)] = \sum_{x \in \mathcal{X}} V_{\theta,\pi}(x)\xi_0(x). \tag{18}$$

Then the policy μ_θ is gradient based sequentially rational with respect to π if and only if

$$\nabla_\theta J_{\theta,\pi} = 0.$$

Proof. We first prove if μ_θ is gradient based sequentially rational then $\nabla_\theta J_{\theta,\pi} = 0$. Now, $J_{\theta,\pi} = \mathbb{E}_{X\sim\xi_0} V_{\theta,\pi}(X)$. Therefore, $\nabla_\theta J_{\theta,\pi} = \mathbb{E}_{X\sim\xi_0} \nabla_\theta V_{\theta,\pi}(X)$. So, if θ is gradient based sequentially rational (i.e. $\nabla_\theta V_{\theta,\pi} = 0$), then $\nabla_\theta J_{\theta,\pi} = 0$.

We now prove the other direction, i.e., $\nabla_\theta J_{\theta,\pi} = 0$ implies μ_θ is gradient based sequentially rational. Suppose ξ_0 is a delta distribution with $\xi_0(x) = 1$ and $\xi_0(y) = 0$ for all $y \neq x$, where $x, y \in \mathcal{X}$. Then $\nabla_\theta J_{\theta,\pi} = 0$ implies that $\nabla_\theta V_{\theta,\pi}(x) = 0$. Since the choice of x is arbitrary, we have $\nabla_\theta V_{\theta,\pi}(x) = 0$ for all $x \in \mathcal{X}$. Thus, μ_θ is gradient based sequentially rational. □

Based on Proposition 3.2, we propose an iterative algorithm to compute a ∇-SMFE. The main idea is as follows. For any $\theta \in \Theta$, let π_θ be the stationary distribution corresponding to policy μ_θ and let G_θ be an unbiased estimator of $\nabla_\theta J_{\theta,\pi_\theta}$. Then, we can start with any initial guess $\theta_0 \in \Theta$, and at each step, update the guess using stochastic gradient ascent:

$$\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k}]_\Theta, \tag{19}$$

where $\{\alpha_k\}_{k\geq 0}$ is a sequence of learning rates that satisfy the standard conditions: $\sum \alpha_k = \infty$ and $\sum \alpha_k^2 < \infty$. Then we have the following:

Algorithm 1: StationaryDistribution

```
input   : $\xi_0$ : Initial distribution
          $\theta$ : Policy parameter
          $B$ : Burn-in period
          $n_p$ : Number of particles
output  : $\pi$ : Stationary distribution
for $i = 1 : n_p$ do
    $x_0^i \sim \xi_0$
    for $t = 0 : B$ do
        $a_t^i \sim \mu_\theta$
        $x_{t+1}^i \sim P(\cdot|x_t^i, a_t^i)$
for $x \in \mathcal{X}$ do
    $\pi(x) = \frac{1}{n_p} \sum_{i=1}^{n_p} \mathbb{1}\{x_{B+1}^i = x\}$
return $\pi$
```

Proposition 3.3. *If the iteration* (19) *converges to a limit θ^* along any sample path, then $(\theta^*, \pi_{\theta^*})$ is a ∇-SMFE.*

Proof. At any step of the iteration, we have that π_θ is the stationary distribution corresponding to μ_θ; and therefore consistency (17) is satisfied. At convergence, we have that $\nabla_\theta J_{\theta^*,\pi_{\theta^*}} = 0$. Therefore by Proposition 3.2, gradient based sequential rationality (16) is satisfied by θ^*. □

To convert the above iteration to a complete algorithm, we need two components:

(1) Given a policy μ_θ, compute the stationary distribution π_θ.
(2) Given a mean field π and a policy μ_θ, compute an unbiased estimate of $\nabla_\theta J_{\theta,\pi}$.

For the first component, we use a particle filter based approach to compute the stationary distribution π_θ. The details are shown in Algorithm 1.

For the second component, we use standard policy gradient approaches from reinforcement learning: likelihood ratio based gradient estimate [21, 44] or simultaneous perturbation based gradient estimate [3, 18, 31, 41]. The details are given below.

3.1 Likelihood ratio based gradient estimation

One approach to estimate the performance gradient is to use likelihood radio based estimates [8, 35, 48]. Suppose the policy $\mu_\theta(X)$ is differentiable with respect to θ. For any time t, define the likelihood function

$$\Lambda_\theta^t = \nabla_\theta \log[\mu_\theta(A_t \mid X_t)], \tag{20}$$

where with a slight abuse of notation $\mu_\theta(A_t|X_t)$ denotes the probability of choosing action A_t in state X_t under policy μ_θ. Then

Algorithm 2: Likelihood ratio based algorithm to compute ∇-SMFE

```
input  : $\theta_0$ : Initial policy parameter
         $K$ : Number of iterations
         $\xi_0$ : Initial mean field distribution
         $B$ : Burn-in period
         $n_p$ : Number of particles
output : $\theta$ : Estimated ∇-SMFE parameter
for iterations $k = 1 : K$ do
    $\pi_k$ = StationaryDistribution($\xi_0, \mu_{\theta_k}, B, n_p$)
    $G_{\theta_k}$ = PolicyGradient($\theta_k, \xi_0, \pi_k$)
    Compute $\theta_{k+1}$ using (19)
return $\theta_{K+1}$
```

from [2, 44, 48] we know that:

$$\nabla_\theta V_{\theta,\pi}(x) = \mathbb{E}_{A_t \sim \mu_\theta(X_t)}\Big[\sum_{\sigma=0}^{\infty} \Lambda_\theta^\sigma V_{\theta,\pi}(X_\sigma) \,\Big|\, X_0 = x\Big]. \tag{21}$$

Thus,

$$\nabla_\theta J_{\theta,\pi} = \mathbb{E}_{X\sim\xi_0}[\nabla_\theta V_{\theta,\pi}(X)]. \tag{22}$$

An algorithm to compute ∇-SMFE based on the likelihood ratio approach is given in Algorithm 2. The `PolicyGradient` function in Algorithm 2 can be obtained by an actor only method such as Monte Carlo [43] or Renewal Monte Carlo [42], or using an actor critic method such as SARSA [43]. Additionally, variance reduction techniques such as subtracting a baseline or using mini-batch averaging may also be used.

3.2 Simultaneous perturbation based gradient estimation

Another approach to estimate the performance gradient is to use simultaneous perturbation based methods [3, 18, 31, 41]. This approach is useful when the policy μ is not a differentiable function of its parameters θ. Now, given any distribution ξ_0, we can estimate $J_{\theta,\pi}$ using $V_{\theta,\pi}$ as:

$$J_{\theta,\pi} = \mathbb{E}_{X\sim\xi_0}[V_{\theta,\pi}(X)]. \tag{23}$$

The two-sided form of simultaneous perturbation estimates are given as:

$$\widehat{\nabla} J_{\theta,\pi} = \eta(J_{\theta+\beta\eta,\pi} - J_{\theta-\beta\eta,\pi})/2\beta. \tag{24}$$

Thus,

$$\widehat{\nabla} J_{\theta,\pi} = \eta(\mathbb{E}_{X\sim\xi_0}[V_{\theta+\beta\eta,\pi}(X)] - \mathbb{E}_{X\sim\xi_0}[V_{\theta-\beta\eta,\pi}(X)])/2\beta. \tag{25}$$

where η is a random variable with the same dimension as θ and β is a small constant. The above method is called simultaneous perturbation stochastic approximation (SPSA) [31, 41], when $\eta_i \sim$ Rademacher(±1); and it is called smoothed functional stochastic approximation (SFSA) [3, 18] when $\eta \sim \text{Normal}(0, I)$.

An algorithm to compute ∇-SMFE based on the simultaneous perturbation approach is given in Algorithm 3. As in the case of the likelihood ratio based approach, the `PolicyEvaluation` function in Algorithm 3 may be obtained by an actor only method such as Monte Carlo [43] or Renewal Monte Carlo [42], or using an actor critic method such as SARSA [43].

Algorithm 3: Simultaneous perturbation based algorithm to compute ∇-SMFE

```
input  : $\theta_0$ : Initial policy parameter
         $K$ : Number of iterations
         $\xi_0$ : Initial mean field distribution
         $B$ : Burn-in period
         $\beta$ : Magnitude of perturbation
         $n_p$ : Number of particles
output : $\theta$ : Estimated ∇-SMFE parameter
for iterations $k = 1 : K$ do
    $\pi_k$ = StationaryDistribution($\xi_0, \mu_{\theta_k}, B, n_p$)
    Let $\eta \sim$ Uniform$\{-1, 1\}$ or $\eta \sim \mathcal{N}(0, 1)$
    $\theta_k^+ = \theta_k + \eta\beta$
    $\theta_k^- = \theta_k - \eta\beta$
    $\widehat{J}_{\theta+\beta\eta,\pi}$ = PolicyEvaluation($\theta_k^+, \xi_0, \pi_k$)
    $\widehat{J}_{\theta-\beta\eta,\pi}$ = PolicyEvaluation($\theta_k^-, \xi_0, \pi_k$)
    Compute $G_{\theta_k}$ as an estimate of $\widehat{\nabla}_\theta J_{\theta,\pi}$ using (24)
    Compute $\theta_{k+1}$ using (19)
return $\theta_{K+1}$
```

4 NUMERICAL STUDY

For our numerical study, we consider a stylized model of malware spread in networks [13, 14]. In this model, the state of an agent denotes its health. The state space is continuous with $\mathcal{X} = [0, 1]$, where $X = 0$ is the most healthy state and $X = 1$ is the least healthy state. The action space is $\mathcal{A} = \{0, 1\}$, where $A^i = 0$ means do nothing and $A^i = 1$ means take corrective action. The dynamics are given by:

$$X_{t+1}^i = \begin{cases} X_t^i + (1 - X_t^i)\eta_t, & \text{for } A_t^i = 0, \\ 0, & \text{for } A_t^i = 1, \end{cases} \tag{26}$$

where $\{\eta_t\}_{t\geq 0}$ is an i.i.d. process with probability density function f. The per-step payoff is:

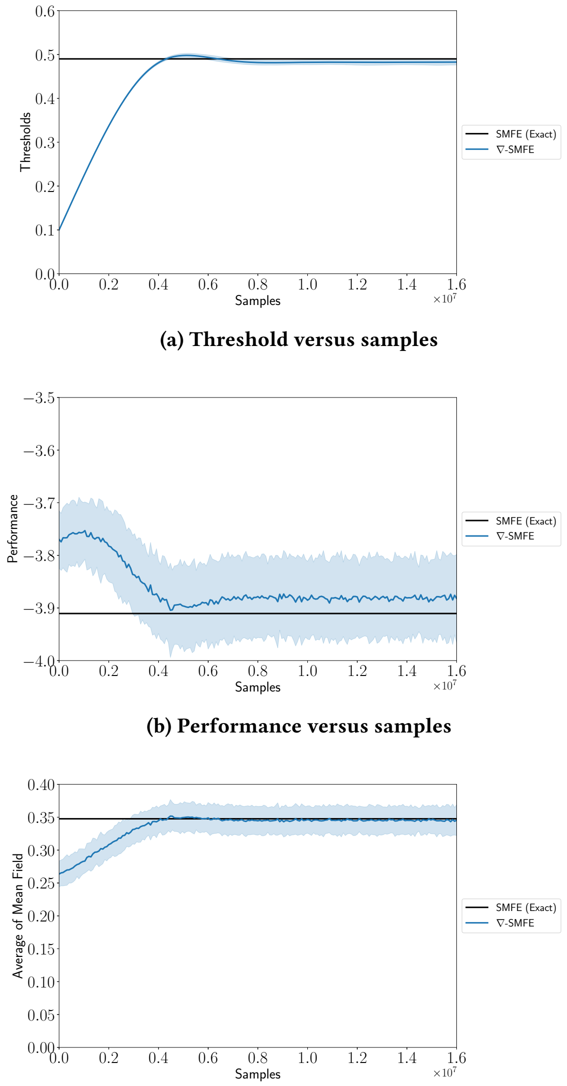
$$u(x, a, \xi) = -(k + \bar{\xi})x - \lambda a, \tag{27}$$

where $\bar{\xi}$ is the mean of ξ and k, λ are given constants. It is shown in [13, 14] that the SMFE policy has a threshold structure. So we restrict attention to threshold based policies with $\Theta = [0, 1]$, where:

$$\mu_\theta(x) = \begin{cases} 0, & \text{if } x < \theta, \\ 1, & \text{if } x \geq \theta. \end{cases} \tag{28}$$

In our experiments, we choose $f = \text{Uniform}[0, 1]$, $k = 0.2$, $\lambda = 0.5$, and we discretize the state space into 101 uniformly sized cells $\{0, 0.01, \dots, 1\}$. We use a discount factor of $\gamma = 0.9$. Since the policy is not differentiable, we estimate the gradient using simultaneous perturbation (Algorithm 3). We use the following parameters: $n_p = 1000$, $\theta_0 = 0.1$, $K = 200$, $B = 200$, $\beta = 0.1$, $\eta \sim \text{Uniform}\{-1, 1\}$ and $\xi_0 = \text{Uniform}(\mathcal{X})$, and choose the learning rates corresponding to ADAM [19] with the α parameter of ADAM as 0.01 and standard values for the other ADAM parameters.

The thresholds, performances and stationary mean fields versus samples over various iterations are given in figures 1a,1b and 1c. The exact values (SMFE) for these parameters obtained using the method specified in [13] are shown in black in these figures. We



(a) Threshold versus samples

(b) Performance versus samples

(c) Average of stationary mean field versus samples

Figure 1: ∇-SMFE computation. The solid lines show the mean values and the shaded regions show the ± two standard deviation region over 100 runs.

note that all these three parameters converge fairly quickly to the approximate values with small error.

5 CONCLUSION & FUTURE WORK

In this paper, we define a new equilibrium concept, gradient based stationary mean field equilibrium, for a class of large population games, based on stationary mean field equilibrium, which takes into account bounded rationality. We then develop an algorithm to compute this equilibrium and present a numerical example to illustrate computation of this equilibrium.

Although we presented only policy based algorithms (Actor only and Actor Critic), bounded rationality can also be modelled using a Critic only variant. Here, function approximation used in the Critic makes the agent boundedly rational. The detailed algorithm for this approach can be derived in a similar manner to the algorithms presented in this paper.

Another important point is the distinction of the algorithms presented here from reinforcement learning. The proposed algorithm is a simulation based algorithm but it is not an online reinforcement learning algorithm for the following reasons. Even though each agent can in-principle run these two algorithms individually (for Algorithm 3, the agent would need to know the per-step payoff function), prior to actually playing the game, each agent needs to make an assumption on all other agents' behaviour in the learning phase. This coordination in learning is not easily justified in a competitive game with strategic agents, where the agents can try and influence their opponents during learning. Since we do not explicitly account for this consideration, the proposed approach is not an online reinforcement learning algorithm. This implies that, though our algorithms are useful in computing a boundedly rational equilibrium, the iterates in our algorithm are not representative of the learning dynamics of individual agents.

REFERENCES

[1] Sachin Adlakha, Ramesh Johari, and Gabriel Y Weintraub. 2015. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory* 156 (2015), 269–316.

[2] Jonathan Baxter and Peter L Bartlett. 2001. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research* 15 (2001), 319–350.

[3] Shalabh Bhatnagar, HL Prasad, and LA Prashanth. 2013. *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods.* Vol. 434. Springer.

[4] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. 2010. Multi-agent reinforcement learning: An overview. In *Innovations in multi-agent systems and applications-1.* Springer, 183–221.

[5] Rene Carmona and François Delarue. 2017. *Probabilistic Theory of Mean Field Games with Applications I.* Springer.

[6] Jakob Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems.* 2137–2145.

[7] Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Philip Torr, Pushmeet Kohli, Shimon Whiteson, et al. 2017. Stabilising experience replay for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1702.08887* (2017).

[8] Peter Glynn. 1990. Likelihood Ratio Gradient Estimation for Stochastic Systems. *Commun. ACM* 33 (1990), 75–84.

[9] Diogo A. Gomes, Joana Mohr, and Rafael Rig ao Souza. 2010. Discrete time, finite state space mean field games. *Journal de Mathématiques Pures et Appliquées* 93, 3 (2010), 308 – 328.

[10] Matthew Hausknecht and Peter Stone. 2015. Deep reinforcement learning in parameterized action space. *arXiv preprint arXiv:1511.04143* (2015).

[11] Minyi Huang, Peter E Caines, and Roland P Malhamé. 2007. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized epsilon-Nash equilibria. *IEEE Trans. Automat. Control* 52, 9 (2007), 1560–1571.

[12] Minyi Huang, Peter E. Caines, and Roland P. Malhamé. 2007. The Nash certainty equivalence principle and McKean-Vlasov systems: An invariance principle and entry adaptation. In *46th IEEE Conference on Decision and Control, CDC 2007, New Orleans, LA, USA, December 12-14, 2007.* 121–126. https://doi.org/10.1109/CDC.2007.4434627

[13] M. Huang and Y. Ma. 2016. Mean field stochastic games: Monotone costs and threshold policies. In *2016 IEEE 55th Conference on Decision and Control (CDC).* 7105–7110.

[14] M. Huang and Y. Ma. 2017. Mean Field Stochastic Games with Binary Action Spaces and Monotone Costs. *ArXiv e-prints* (Jan. 2017). arXiv:math.OC/1701.06661

[15] M. Huang and Y. Ma. 2017. Mean field stochastic games with binary actions: Stationary threshold policies. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC).* 27–32.

[16] Minyi Huang, Roland P Malhamé, and Peter E Caines. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems* 6, 3

(2006), 221–252.
[17] Sham M Kakade. 2002. A natural policy gradient. In *Advances in Neural Information Processing Systems*. 1531–1538.
[18] V. Katkovnik and Y. Kulchitsky. 1972. Convergence of a class of random search algorithms. *Automation and Remote Control* 33, 8 (1972), 1321–1326.
[19] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[20] Arman C Kizilkale and Peter E Caines. 2013. Mean field stochastic adaptive control. *IEEE Trans. Automat. Control* 58, 4 (2013), 905–920.
[21] Vijay R Konda and John N Tsitsiklis. 2003. On actor-critic algorithms. *SIAM Journal on Control and Optimization* 42, 4 (2003), 1143–1166.
[22] Romain Laroche, Mehdi Fatemi, Joshua Romoff, and Harm van Seijen. 2017. Multi-Advisor Reinforcement Learning. *arXiv preprint arXiv:1704.00756* (2017).
[23] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese journal of mathematics* 2, 1 (2007), 229–260.
[24] Martin Lauer and Martin Riedmiller. 2000. An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, 535–542.
[25] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 464–473.
[26] Kaixiang Lin, Shu Wang, and Jiayu Zhou. 2017. Collaborative Deep Reinforcement Learning. *arXiv preprint arXiv:1702.05796* (2017).
[27] Michael L. Littman. 1994. Markov games as a framework for multi-agent reinforcement learning., In Proceedings of the eleventh International Conference on Machine Learning. *Proceedings of the eleventh international conference on machine learning*. Vol. 157.
[28] Michael L. Littman. 2001. Friend-or-foe Q-learning in general-sum games.. In *ICML*, Vol. Vol. 1.
[29] Michael L Littman. 2001. Value-function reinforcement learning in Markov games. *Cognitive Systems Research* 2, 1 (2001), 55–66.
[30] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *arXiv preprint arXiv:1706.02275* (2017).
[31] John L Maryak and Daniel C Chin. 2008. Global random optimization by simultaneous perturbation stochastic approximation. 53, 3 (April 2008), 780–783.
[32] Gregory Palmer, Karl Tuyls, Daan Bloembergen, and Rahul Savani. 2017. Lenient Multi-Agent Deep Reinforcement Learning. *arXiv preprint arXiv:1707.04402* (2017).
[33] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. *arXiv preprint arXiv:1707.06600* (2017).
[34] Jan Peters and Stefan Schaal. 2006. Policy gradient methods for robotics. In *International Conference on Intelligent Robots and Systems, 2006 IEEE/RSJ*. IEEE, 2219–2225.
[35] Reuven Y Rubinstein. 1989. Sensitivity analysis and performance extrapolation for computer simulation models. *Operations Research* 37, 1 (1989), 72–81.
[36] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 1889–1897.
[37] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
[38] Yoav Shoham, Rob Powers, and Trond Grenager. 2003. Multi-agent reinforcement learning: a critical survey. (2003).
[39] Yoav Shoham, Rob Powers, and Trond Grenager. 2007. If multi-agent learning is the answer, what is the question? *Artificial Intelligence* 171, 7 (2007), 365–377.
[40] Herbert A Simon. 1955. A behavioral model of rational choice. *The quarterly journal of economics* 69, 1 (1955), 99–118.
[41] James C Spall. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. 37, 3 (1992), 332–341.
[42] Jayakumar Subramanian and Aditya Mahajan. 2018. Renewal Monte Carlo: Renewal theory based reinforcement learning. *ArXiv e-prints* (April 2018). arXiv:cs.LG/1804.01116
[43] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. MIT Press.
[44] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*. 1057–1063.
[45] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2017. Multiagent cooperation and competition with deep reinforcement learning. *PloS one* 12, 4 (2017).
[46] Harm van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, and Jeffrey Tsang. 2017. Hybrid Reward Architecture for Reinforcement Learning. *arXiv preprint arXiv:1706.04208* (2017).
[47] Gabriel Y. Weintraub, C. Lanier Benkard, and Benjamin Van Roy. 2005. Oblivious Equilibrium: A Mean Field Approximation for Large-Scale Dynamic Games. In *Advances in Neural Information Processing Systems*. 1489–1496.
[48] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.
[49] Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. 2017. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. In *Advances in neural information processing systems*. 5285–5294.
[50] Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. 2018. Deep Mean Field Games for Learning Optimal Behavior Policy of Large Populations. In *International Conference on Learning Representations*.
[51] H. Yin, P. G. Mehta, S. P. Meyn, and U. V. Shanbhag. 2014. Learning in Mean-Field Games. *IEEE Trans. Automat. Control* 59, 3 (March 2014), 629–644.