Proceedings of Machine Learning Research (PMLR) vol 242:1–12, 2024

Learning ϵ-Nash Equilibrium Stationary Policies in Stochastic Games with Unknown Independent Chains Using Online Mirror Descent

Tiancheng Qin TQ6@ILLLNOIS.EDU
Department of Industrial and Systems Engineering, Coordinated Science Laboratory, University of Illinois Urbana-Champaign, Urbana, IL, 61801.

S. Rasoul Etesami ETESAMI1@ILLLNOIS.EDU
Department of Industrial and Systems Engineering, Department of Electrical and Computer Engineering, Coordinated Science Laboratory, University of Illinois Urbana-Champaign, Urbana, IL, 61801. *

Abstract

We study a subclass of n-player stochastic games, namely, stochastic games with independent chains and unknown transition matrices. In this class of games, players control their own internal Markov chains whose transitions do not depend on the states/actions of other players. However, players' decisions are coupled through their payoff functions. We assume players can receive only realizations of their payoffs, and that the players can not observe the states and actions of other players, nor do they know the transition probability matrices of their own Markov chain. Relying on a compact dual formulation of the game based on occupancy measures and the technique of *confidence set* to maintain high-probability estimates of the unknown transition matrices, we propose a fully decentralized mirror descent algorithm to learn an ϵ-Nash equilibrium stationary policy for this class of games. The proposed algorithm has the desired properties of *independence* and *convergence*. Specifically, assuming the existence of a *variationally stable* Nash equilibrium policy, we show that the proposed algorithm in which players make their decisions independently and in a decentralized fashion converges asymptotically to the stable ϵ-Nash equilibrium stationary policy with arbitrarily high probability.

Keywords: Stochastic games, independent and decentralized learning, stationary Nash equilibrium policy, occupancy measure, online mirror descent, variational stability.

1. Introduction

Learning Nash equilibrium (NE) points in noncooperative games is a fundamental problem that has emerged in many disciplines, such as control and game theory, operations research, and computer science (Zhang et al., 2021a; Cesa-Bianchi and Lugosi, 2006; Daskalakis et al., 2020). Typically, efficient and independent learning of NE is challenging, and it is known that computing NE is PPAD-hard (Daskalakis et al., 2009) for general-sum games. The learning task is even more complex for stochastic dynamic games (Shapley, 1953; Başar and Olsder, 1998) where the existence of state dynamics introduces additional nonstationarity to the environment. Expanding upon the existing literature, in this work, we study a subclass of noncooperative stochastic games, namely, stochastic games with independent chains and unknown transition matrices (Altman et al., 2007; Etesami, 2024), and our goal is to provide independent learning algorithms that converge to NE points. In this class of games, a set of n players, each with its own finite state and action space,

* This work is supported by the Air Force Office of Scientific Research (AFOSR YIP) under award number FA9550-23-1-0107 and the NSF CAREER Award under Grant No. EPCN-1944403.

controls its own Markov chain, whose transition does not depend on the states/actions of other players. However, the players are coupled through their payoff functions, which depend on the states and actions of all players. We also assume that the players can not observe the states and actions of other players, nor do they know the transition probability matrices of their own Markov chain. There are many interesting real-world problems that fit into this subclass of stochastic games, such as multi-agent robotic navigation (Zhang et al., 2021a), energy management in smart grids (Etesami et al., 2018), and power/bandwidth allocation in multi-agent wireless communication (Altman et al., 2007; Narayanan and Theagarajan, 2017; Altman et al., 2008).

To provide a more concrete motivating example from energy management in smart grids (Etesami et al., 2018), one can consider an energy market with one utility company and a set $[n] = \{1, \dots, n\}$ of users (players), which can both produce and consume energy. Each player generates energy using its solar panel or wind turbine and is equipped with a storage device that can store the remaining energy at the end of each day $t \in \mathbb{Z}_+$. Let s_i^t denote the (quantized) amount of stored energy of player i at the beginning of day t with maximum storage capacity C. Moreover, let g_i^t be a random variable denoting the amount of harvested energy for player i at the end of day t, whose distribution is determined by the unknown stochastic weather conditions on that day. Now if we use a_i^t to denote the total amount of energy consumed by player i during day t, then the stored energy at the end of day t (or the beginning of day $t+1$) is given by $s_i^{t+1} = \min\{C, g_i^t + (s_i^t - a_i^t)^+\}$, where $(s_i^t - a_i^t)^+ = \max\{0, s_i^t - a_i^t\}$. In particular, player i needs to purchase $(a_i^t - s_i^t)^+$ units of energy from the utility company on day t to satisfy its demand on that day. On the other hand, the utility company sets the energy price as a function of total demands $\{(a_i^t - s_i^t)^+, i \in [n]\}$, which is given by $p(a^t, s^t)$. If $u_i(a_i^t)$ denotes the utility that player i derives by consuming a_i^t units of energy, then the reward of player i at time t is given by $r_i(a^t, s^t) = u_i(a_i^t) - p(a^t, s^t) \times (a_i^t - s_i^t)^+$. In particular, if players are at distant locations, they likely experience independent weather conditions, so their transition probability models that are governed by stochasticity of $\{g_i^t, i \in [n]\}$ will be independent.

1.1. Related Work

For dynamic stochastic games, the prior work has largely focused on the special case of two-player zero-sum stochastic games (Zhao et al., 2022; Qiu et al., 2021; Zhang et al., 2021a; Tian et al., 2021; Sayin et al., 2021, 2022; Wei et al., 2021). While two-player zero-sum stochastic games constitute an important basic setting, there are many problems with a large number of players, a situation that hinders the applicability of the existing algorithms for computing a stationary NE. To address this issue, researchers have recently developed learning algorithms for finding NE in special structured stochastic games, e.g., mean-field and aggregative stochastic games (Zhang et al., 2021a; uz Zaman et al., 2020; Meigs et al., 2019). Moreover, Zhang et al. (2021b); Leonardos et al. (2021) show that n-player Markov potential games, an extension of static potential games to dynamic stochastic games, admit polynomial-time algorithms for computing their NE policies.

There has been a line of prior research on the study of decentralized stochastic games with independent chains (Altman et al., 2007; Singh and Hemachandra, 2014; Qiu et al., 2021; Zhang and Zou, 2022; Etesami, 2024). Specifically, Altman et al. (2007) showed the existence of a NE for the class of stochastic games with independent chains.[1] The work Singh and Hemachandra (2014) showed that the set of stationary NE for the class of games can be characterized via the

1. In their work, such a class of games is called constrained cost-coupled stochastic games with independent state processes that also include additional constraints.

global minimizers of a certain non-convex mathematical program. Recently, for n-player decentralized stochastic games with independent chains, relying on a dual formulation of the game based on occupancy measures, Etesami (2024) proposed polynomial-time learning algorithms based on dual averaging and dual mirror descent, which converge in terms of the averaged Nikaido-Isoda distance to the set of ϵ-NE policies. However, all of the aforementioned works assume players' prior knowledge of the transition probability matrices of their own Markov chain, which is somewhat restrictive in practice. Moreover, there was no algorithm with an asymptotic convergence guarantee to NE policies for the class of n-player stochastic games with independent chains.

1.2. Contributions

We consider the class of stochastic games with independent chains and unknown transition matrices. Relying on a dual formulation of the complete information stochastic game based on occupancy measures and introducing confidence sets to maintain high-probability estimates of the unknown transition matrices, we propose a Decentralized Mirror Descent algorithm to learn an ϵ-NE policy. The proposed algorithm has the desired properties of independence and convergence. Our contributions can be summarized as follows:

- We propose a learning algorithm that is simple, easy to implement, and works in a fully decentralized and independent manner. The only coordination needed is a simple signaling mechanism to indicate the end of each episode among players, which can be further relaxed by allowing an extra error term in the equilibrium computation.
- Under the assumption that the game admits a globally stable NE policy, which is a relaxation of the well-known monotonicity condition, we show that the proposed algorithm converges asymptotically to an ϵ-NE with arbitrarily high probability.

Due to space limitations, all the technical proofs can be found in the full version of this work available in Qin and Etesami (2023).

2. Problem Formulation

We consider an n-player stochastic game with independent and unknown state transitions, which is described by the tuple $(S_i, A_i, r_i, P_i)_{i=1}^n$, as follows.

- S_i is the finite set of states for player i with elements $s_i \in S_i$. We denote the joint state set of all the players by $S = \prod_{i=1}^n S_i$ with elements $\boldsymbol{s} \in S$, where $\boldsymbol{s} = (s_1, \dots, s_n)$.
- A_i is the finite set of actions for player i with elements $a_i \in A_i$. We denote the joint action set of all the players by $A = \prod_{i=1}^n A_i$, and the elements of A are denoted by $\boldsymbol{a} = (a_1, \dots, a_n)$.
- $r_i : S \times A \to [0, 1]$ is the reward function for player i, where $r_i(\boldsymbol{s}, \boldsymbol{a})$ is the immediate reward received by player i when the states of the players are $\boldsymbol{s} = (s_1, \dots, s_n)$, and the actions taken by them are given by the action profile $\boldsymbol{a} = (a_1, \dots, a_n)$.
- P_i is the transition probability matrix for player i, where $P_i(s_i'|s_i, a_i)$ is the probability that the state of player i moves from s_i to s_i' if she chooses action a_i. Crucial to this work, we assume that P_i is *unknown* to player i and is *independent* of other players' transition probability matrices.

Assumption 1 *We assume that the joint transition probability matrix $P(s'|s,a)$ can be factored into independent components $P(s'|s,a) = \prod_{i=1}^{n} P_i(s'_i|s_i,a_i)$, where $P_i(s'_i|s_i,a_i)$ is the transition probability matrix for player i.*

A sequence of probability measures $\pi_i = \{\pi_i^t, t = 0,1,\ldots\}$ over A_i that, at each time t selects an action $a_i \in A_i$ based on past observations $\mathcal{H}_i^t$ with probability $\pi_i^t(\cdot|\mathcal{H}_i^t)$, consists a general policy for player i.[2] However, use of general policies is often computationally expensive, and in practical applications, players are interested in the easily implementable *stationary* policies, as defined next.

Definition 1 *A policy π_i for player i is called stationary if the probability $\pi_i^t(a_i|\mathcal{H}_i^t)$ of choosing action a_i at time t depends only on the current state $s_i^t = s_i$, and is independent of the time t. In the case of the stationary policy, we use $\pi_i(a_i|s_i)$ to denote this time-independent probability.*

Given some initial state s^0, the objective for each player $i \in [n]$ is to choose a stationary policy π_i that maximizes its long-term expected average payoff given by

$$V_i(\pi_i, \pi_{-i}) = \mathbb{E}\Big[\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} r_i(s^t, a^t)\Big], \tag{1}$$

where $\pi_{-i} = (\pi_j, j \neq i)$,[3] and the expectation is with respect to the randomness introduced by players' internal chains $(P_1,\ldots,P_n)$ and their policies $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_n)$.

Next, in order to be able to establish meaningful convergence/learning results, we impose the following assumption througout this work.

Assumption 2 *For any player i and any stationary policy π_i chosen by that player, the induced Markov chain with transition probabilities $P^{\pi_i}(s'_i|s_i) = \sum_{a_i\in A_i} P_i(s'_i|a_i,s_i)\pi_i(a_i|s_i)$, is ergodic, and its mixing time is uniformly bounded above by some parameter τ; that is,*

$$\|(\nu-\nu')P^{\pi_i}\|_1 \leq e^{-1/\tau}\|\nu-\nu'\|_1, \qquad \forall i, \pi_i, \nu, \nu' \in \Delta(S_i).$$

In fact, Assumption 2 is a standard assumption used in the MDP literature and is much needed. Otherwise, if the transition probability matrix P_i of a player i is such that for some policy π_i the induced chain P^{π_i} takes an arbitrarily large time to mix, then there is no hope that player i can evaluate the performance of policy π_i in a reasonably short time. As is shown in the next section, under the ergodicity Assumption 2, for any stationary policy profile $\boldsymbol{\pi}$, the limit in (1) indeed exists. This fully characterizes an n-player stochastic game with initial state s^0, in which each player i wants to choose a stationary policy π_i to maximize its expected aggregate payoff $V_i(\pi_i,\pi_{-i})$. In the remainder of the paper, we shall refer to the above payoff-coupled stochastic game with independent chains and unknown transitions as $\mathcal{G} = ([n], \pi, \{V_i(\pi)\}_{i\in[n]})$.

Definition 2 *For a policy profile $\boldsymbol{\pi}^* = (\pi_1^*,\ldots,\pi_n^*)$, π_i^* is called a best response policy of π_{-i}^* if $V_i(\pi_i^*,\pi_{-i}^*) \geq V_i(\pi_i,\pi_{-i}^*)$ for any policy π_i. It is called an ϵ-best response policy if $V_i(\pi_i^*,\pi_{-i}^*) \geq V_i(\pi_i,\pi_{-i}^*) - \epsilon$ for any policy π_i. The policy profile $\boldsymbol{\pi}^* = (\pi_1^*,\ldots,\pi_n^*)$ is called a Nash equilibrium (NE) for the game $\mathcal{G}$ if for any i, π_i^* is a best response policy of π_{-i}^*. It is called an ϵ-NE if for any i, π_i^* is an ϵ-best response policy of π_{-i}^*.*

2. Here, $\mathcal{H}_i^t = \{s_i^l, a_i^l, r_i(s^l,a^l) : l = 0,1,\ldots,t-1\} \cup \{s_i^t\}$ denotes the history of player i's past observations, i.e., realized states, actions, and rewards.

3. More generally, given a vector v, we let $v_{-i} = (v_j, j \neq i)$ be the vector of all coordinates in v other than the ith one.

The main objective of this work is to develop a decentralized learning algorithm such that, if followed by the players independently, it brings the system to an ϵ-NE stationary policy.

3. A Dual Formulation and Preliminaries

In this section, we provide an alternative dual formulation for the stochastic game $\mathcal{G}$ based on *occupancy measures* (Etesami, 2024; Altman, 2021). Intuitively, from player j's point of view, its long-term expected average payoff depends on the proportion of time that player j spends in state s_j and takes action a_j, denoted by its occupancy measure. Thus, the policy optimization for player j can be viewed as an optimization problem in the space of occupancy measures, where players want to force their chains to spend most of their time in high-reward states.

3.1. Occupancy Measure

For a given MDP with a transition probability matrix P and any stationary policy π, one can associate with P and π notions of occupancy measures $\rho : S \times A \to [0,1]$, and $q : S \times A \times S \to [0,1]$:

$$\rho(s,a) = \lim_{t\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{P}(s^t = s, a^t = a), \tag{2}$$

$$q(s,a,s') = \lim_{t\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{P}(s^t = s, a^t = a, s^{t+1} = s'). \tag{3}$$

Intuitively, $\rho(s,a)$ and $q(s,a,s')$ are the long-term average proportion of time of encountering the state-action pair (s,a), and state-action-next-state triple (s,a,s'), when executing policy π in an MDP with transition probability matrix P. It can be readily shown that under Assumption 2, the limits in, (2), and (3) indeed exist.

In this work, we are primarily concerned with the occupancy measure $q(s,a,s')$ due to players not knowing their independent transition probability matrix P_i in the stochastic game $\mathcal{G}$. In the following, we provide conditions that fully characterize the set of feasible occupancy measures q.

Definition 3 *We define the polytope of feasible occupancy measures, denoted by Δ, as*

$$\Delta = \Big\{ q \in [0,1]^{|S\times A\times S|} \quad : \sum_{s,a,s'} q(s,a,s') = 1, \quad \sum_{s',a} q(s',a,s) = \sum_{a,s'} q(s,a,s'), \quad \forall s \in S \Big\}. \tag{4}$$

For any $q \in \Delta$, we define its induced transition probability matrix P^q and stationary policy π^q by

$$P^q(s'|s,a) = \frac{q(s,a,s')}{\sum_{s'} q(s,a,s')} \ \forall s,a,s', \qquad \pi^q(a|s) = \frac{\sum_{s'} q(s,a,s')}{\sum_{a',s'} q(s,a',s')} \ \forall s,a.$$

Moreover, for a fixed transition probability matrix P, we denote by $\Delta(P) \subset \Delta$ the set of occupancy measures whose induced transition probability matrix P^q is exactly P. Similarly, we denote by $\Delta(\mathcal{P}) \subset \Delta$ the set of occupancy measures whose induced transition probability matrix P^q belongs to a set of transition matrices $\mathcal{P}$.

Given the above definition, we have the following useful lemma that extends a result from Altman (2021) to the case of q occupancy measures.

Lemma 4 *If a function $q : S \times A \times S \to [0,1]$ belongs to the feasible occupancy polytope (4), then it is exactly the occupancy measure associated with its induced transition probability matrix P^q and stationary policy π^q. Specifically, we have (3) hold if one executes policy π^q in an MDP with transition probability matrix P^q.*

3.2. A Dual Formulation

It is shown in Etesami (2024) that under the ergodicity Assumption 2, due to the independency of players' internal chains, the payoff functions admit a simple decomposable form in terms of occupancy measures ρ. However, since we are interested in stochastic games with unknown transition probabilities, we first extend this result in terms of occupancy measures q. Specifically, assume that each player i is following a stationary policy π_i, and let q_i be the corresponding occupancy measures given in (3), that are induced by P_i and following the stationary policy π_i. Then, we have:

Proposition 5 *Let Assumptions 1 and 2 hold, and assume that each player i follows a stationary policy π_i. Then, we have*

$$V_i(\pi_i, \pi_{-i}) = V_i(q_i, q_{-i}) \triangleq \sum_{s,a} \prod_{j=1}^{n} \sum_{s'_j} q_j(s_j, a_j, s'_j) r_i(s,a) = \langle q_i, l_i(q_{-i}) \rangle, \tag{5}$$

where $l_i(q_{-i})$ is defined to be a vector of dimension $|A_i||S_i|^2$ whose (s_i, a_i, s'_i)-th coordinate equals

$$l_i(q_{-i})_{(s_i,a_i,s'_i)} = \sum_{s_{-i},a_{-i}} \prod_{j\neq i} \sum_{s'_j} q_j(s_j, a_j, s'_j) r_i(s,a). \tag{6}$$

We remark here that Proposition 5 lies at the core of our analysis, and it relies heavily on the assumption that players have independent internal Markov chains (Assumption 1). This is the reason why our analysis would not generalize to stochastic games without the independent chain assumption. Using Lemma 4 and Proposition 5, the problem of finding the optimal stationary policies for the players reduces to finding the optimal feasible occupancy measures for them.

Definition 6 *Let us define the feasible occupancy polytope for player i by*

$$\Delta_i = \Big\{ q_i \in [0,1]^{|A_i||S_i|^2} : \sum_{s_i,a_i,s'_i} q_i(s_i,a_i,s'_i) = 1, \sum_{s'_i,a_i} q_i(s'_i,a_i,s_i) = \sum_{s'_i,a_i} q_i(s_i,a_i,s'_i), \forall s_i \Big\}.$$

*Moreover, denote by $\Delta_i(P_i) \subset \Delta_i$ the set of feasible occupancy measures whose induced transition probability matrix P^{q_i} is exactly P_i. The **virtual game** $\mathcal{V} = ([n], q, \{V_i(q)\}_{i\in[n]})$ associated with the stochastic game $\mathcal{G}$ is an n-player continuous-action static game, where the action of player i is to choose an q_i from its action set $\Delta_i(P_i)$, and its payoff function is given by (5).*

Ideally, we would like every player to work with the virtual game $\mathcal{V}$ with action set $\Delta_i(P_i)$ as it admits a payoff function that is linear with respect to the player's action, hence making it amenable to the use of online learning algorithms. However, this can not be performed as P_i is

not known to player i so the player can not compute $\Delta_i(P_i)$. Nevertheless, observe that once each player i has decided on her occupancy measure $\hat{q}_i \in \Delta_i$ (which may not belong to $\Delta_i(P_i)$), then the game $\mathcal{G}$ is fully determined by the players' policies $\{\pi_i^{\hat{q}_i}\}_{i=1}^n$, where $\pi_i^{\hat{q}_i}$ is the stationary policy induced by $\hat{q}_i$. In this regard, with some abuse of notations, the payoff function of player i is given by $V_i(\hat{q}_i, \hat{q}_{-i}) = V_i(\pi_i^{\hat{q}_i}, \pi_{-i}^{\hat{q}_{-i}})$, where $\hat{q}_i \in \Delta_i$, and $V_i(\pi_i^{\hat{q}_i}, \pi_{-i}^{\hat{q}_{-i}})$ is as defined in (1). When $\hat{q}_i \in \Delta_i(P_i)$, we also have $V_i(\hat{q}_i, \hat{q}_{-i}) = V_i(\pi_i^{\hat{q}_i}, \pi_{-i}^{\hat{q}_{-i}}) = \langle \hat{q}_i, l_i(\hat{q}_{-i}) \rangle$ as defined in (5).

4. A Learning Algorithm for ϵ-NE Policies

In this section, we develop our learning algorithm for the stochastic game $\mathcal{G}$. The algorithm proceeds in different episodes, each containing a random number of time instances. The main idea is that each player i will use *confidence sets* and *online mirror descent* (OMD) to learn an occupancy measure $\hat{q}_i$ such that (i) its induced transition probability matrix $P_i^{\hat{q}_i}$ approximates the true transition probability matrix P_i, and (ii) its induced stationary policy $\pi_i^{\hat{q}_i}$ approximates player i's best response to $\pi_{-i}^{\hat{q}_{-i}}$. The complete pseudo-code of the proposed learning algorithm is presented in Algorithm 1. We first consider the following definition of a *shrunk polytope*.

Definition 7 (Shrunk Polytope) *Given $\delta_i \in (0, 1)$, we define*

$$\Delta_{i,\delta_i} \triangleq \Big\{ q_i \in \Delta_i : \sum_{s_i'} q_i(s_i, a_i, s_i') \geq \delta_i, \forall s_i, a_i \Big\}$$

to be the shrunk polytope of feasible occupancy measures for player i. Moreover, for a fixed transition probability matrix P_i or a set of transition probability matrices $\mathcal{P}_i$, we define $\Delta_{i,\delta_i}(P_i) \subseteq \Delta_{i,\delta_i}$ or $\Delta_{i,\delta_i}(\mathcal{P}_i) \subseteq \Delta_{i,\delta_i}$ as the set of occupancy measures q_i whose induced transition probability matrix P^{q_i} equals P_i or belongs to the set $\mathcal{P}_i$, respectively.

Restricting player i's occupancy measures to be in Δ_{i,δ_i} ensures that player i uses stationary policies that choose any action with probability at least δ_i, hence encouraging exploration during the learning process. Thanks to the continuity of the payoff functions, working with shrunk polytope Δ_{i,δ_i} with a sufficiently small threshold δ_i can only result in a negligible loss in players' payoffs, as shown in the following lemma (Etesami, 2024, Lemma 2).

Lemma 8 *For any $\epsilon > 0$, there exist polynomial-time computable thresholds $\{\delta_i > 0, i \in [n]\}$, such that*

$$\max_{q_i' \in \Delta_{i,\delta_i}(P_i)} V_i(q_i', q_{-i}) \geq \max_{q_i' \in \Delta_i(P_i)} V_i(q_i', q_{-i}) - \epsilon, \quad \forall q \in \Delta_1 \times \cdots \times \Delta_n. \tag{7}$$

Finally, we consider the following "nondegeneracy" assumption on players' internal chains, which requires that with some positive probability $\alpha_i > 0$, all states are reachable for each player i and under all policies. Assumption 3 serves to provide an upper bound for the length of each episode in our learning algorithm and can be viewed as a relaxation of that made in other works for the case of single-agent MDPs (Rosenberg and Mansour, 2019; Neu et al., 2010).

Assumption 3 *There exists some $\alpha > 0$ such that for every player i, $\sum_{a_i} P_i(s_i'|s_i, a_i) > \alpha$, $\forall s_i, s_i'$.*

Now, we are ready to describe our main distributed learning algorithm. Each player i performs two tasks in parallel: (i) maintains and updates a confidence set $\mathcal{P}_i$ of its own (unknown) transition probability matrix P_i, and (ii) uses an OMD rule to update the occupancy measure $\hat{q}_i \in \Delta(\mathcal{P}_i)$.

4.1. Confidence Set

For each player i, the algorithm maintains counters $N_i(s_i, a_i)$ and $M_i(s_i, a_i, s_i')$ to record the total number of visits of each state-action pair (s_i, a_i) and each state-action-state triple (s_i, a_i, s_i') so far, respectively. A *confidence set* $\mathcal{P}_i$, which includes all transition probability matrices that are close to P_i with high confidence, is maintained and updated for each episode. Specifically, at the end of each episode $k \geq 1$, player i will compute the empirical transition probability matrix $\bar{P}_i^k(s_i'|s_i, a_i) = \frac{M_i^k(s_i,a_i,s_i')}{\max\{1,N_i^k(s_i,a_i)\}}$ from the current counters $N_i^k(s_i, a_i)$ and $M_i^k(s_i, a_i, s_i')$, and will update the confidence set for episode k as

$$\mathcal{P}_i^k = \Big\{\hat{P} : |\hat{P}(s_i'|s_i, a_i) - \bar{P}_i(s_i'|s_i, a_i)| \leq \epsilon_i^k(s_i'|s_i, a_i),\ \forall s_i', s_i, a_i\Big\} \cap \mathcal{P}_i^{k-1}, \tag{8}$$

where $\epsilon_i^k(\cdot)$ is a parameter that will be determined later. Note that the confidence set in (8) is also a polytope with an efficient description in terms of the problem parameters.

4.2. Online Mirror Descent (OMD)

The OMD component of our algorithm is similar to Etesami (2024). Given any desired accuracy $\epsilon > 0$ for an ϵ-NE, each player first uses Lemma 8 to determine a threshold δ_i. During each episode k, player i takes actions according to the stationary policy $\pi_i^k := \pi_i^{\hat{q}_i^k}$. The episode continues until each player i has visited all its state-action pairs (s_i, a_i) at least once. At the end of the episode, player i will first update the confidence set $\mathcal{P}_i^k$ as in (8), and then will update its occupancy measure $\hat{q}_i^{k+1}$ using OMD:

$$\hat{q}_i^{k+1} = \underset{\hat{q}_i \in \Delta_{i,\delta_i}(\mathcal{P}_i^k)}{\operatorname{argmax}} \Big\{\eta^k \langle \hat{q}_i, R_i^k\rangle - D_{h_i}(\hat{q}_i||\hat{q}_i^k)\Big\},$$

where η^k is the stepsize, $D_{h_i}(p||q) \triangleq h_i(p) - h_i(q) - \langle \nabla h_i(q), p - q\rangle$ is the *Bregman divergence* induced by a μ-strongly convex regularizer $h_i(\cdot)$, and R_i^k is an estimator for the gradient of the payoff function $V_i(\pi_i^{\hat{q}_i^k}, \pi_{-i}^{\hat{q}_{-i}^k})$ constructed using the collected samples of the reward r_i during episode k. Since $\hat{q}_i^k \in \Delta_{i,\delta_i}$, from Assumption 3, at any time t, $\mathbb{P}(s_i^t = s_i) \geq \alpha\delta_i\ \forall s_i$. As a result, the expected length of each episode k is at most $\tilde{\mathcal{O}}\Big(\max_i \frac{|S_i|}{\alpha\delta_i^2}\Big)$, where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic terms.

5. Asymptotic Convergence to an ϵ-Nash Equilibrium Policy

In this section, we show that if Algorithm 1 is run with the choice of $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ satisfying Lemma 8, then the iterates generated by Algorithm 1 will converge asymptotically to a *globally stable* ϵ-NE policy of the game $\mathcal{G}$ (see Assumption 4) with high probability. Following Algorithm 1, every player i will hold a occupancy measure $\hat{q}_i^k$ during episode k, and will play according to policy $\pi_i^k = \pi^{\hat{q}_i^k}$ as defined in (9).[4] We denote the occupancy measure induced by π_i^k and the unknown transition probability matrix P_i over the space $S_i \times A_i \times S_i$, by q_i^k.

Definition 9 *Given $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$, we define $\mathcal{V}_{\boldsymbol{\delta}}$ to be the constrained version of the virtual game $\mathcal{V}$ in which the action set for each player i is given by $\Delta_{i,\delta_i}(P_i)$ (instead of $\Delta_i(P_i)$).*

4. We use bold symbols to denote aggregate variables of all the players, e.g., $\boldsymbol{\pi}^k = (\pi_1^k, \ldots, \pi_n^k)$, $\boldsymbol{q}^k = (q_1^k, \ldots, q_n^k)$, $\hat{\boldsymbol{q}}^k = (\hat{q}_1^k, \ldots, \hat{q}_n^k)$, $\boldsymbol{P} = \prod_{i=1}^n P_i$, $\boldsymbol{\Delta} = \prod_{i=1}^n \Delta_i$, and $\boldsymbol{\Delta}_{\boldsymbol{\delta}} = \prod_{i=1}^n \Delta_{i,\delta_i}$.

Algorithm 1 A Decentralized Online Mirror Descent Algorithm for Player i

Input: Initial occupancy measure $\hat{q}_i^1 = \frac{1}{|A_i||S_i|^2} \cdot \mathbf{1}$, counters $N_i(s_i, a_i) = 0$, $M_i(s_i, a_i, s_i') = 0$, step-size sequence $\{\eta^k\}_{k=1}^K$, mixing time thresholds $\{d^k\}_{k=1}^K$, and the regularizer $h_i : \Delta_{i,\delta_i} \to \mathbb{R}$.
For $k = 1, \ldots, K$, do the following:

- At the start of episode k, compute $\pi_i^k = \pi^{\hat{q}_i^k}$, i.e.,

$$\pi_i^k(a_i|s_i) = \frac{\sum_{s_i'} \hat{q}_i^k(a_i, s_i, s_i')}{\sum_{a_i', s_i'} \hat{q}_i^k(s_i, a_i', s_i')} \quad \forall s_i \in S_i, a_i \in A_i, \tag{9}$$

  and keep playing according to this stationary policy π_i^k during expisode k. Update counters $N_i(s_i, a_i)$ and $M_i(s_i, a_i, s_i')$ at each step.

- Let $\tau_i^k \geq d^k$ be the first (random) time such that all state-action pairs (s_i, a_i) are visited during steps $[d^k, \tau_i^k]$. Episode k terminates after $\tau^k = \max_i \tau_i^k$ steps.

- Let $X_i' = S_i \times A_i$, and $R_i^k \in \mathbb{R}_+^{|S_i||A_i|}$ be a random vector (initially set to zero), which is constructed sequentially during the sampling interval $[\tau^k + d, \tau^{k+1}]$ as follows:

  - **For** $t = d^k, \ldots, \tau^k$ and while $X_i \neq \emptyset$, player i picks an action a_i^t according to $\pi_i^k(\cdot|s_i^t)$, and observes the payoff $r_i(s^t, a^t)$ and its next state s_i^{t+1}. If $(s_i^t, a_i^t) \in X_i$, then update $X_i = X_i \setminus \{(s_i^t, a_i^t)\}$, and let $R_i^k = R_i^k + r_i(s^t, a^t)\, \mathbf{e}_{(s_i^t, a_i^t)}$, where $\mathbf{e}_{(s_i^t, a_i^t)}$ is the basis vector with all entries being zero except that the (s_i^t, a_i^t)-th entry is 1.

- Expand R_i^k from $\mathbb{R}^{|S_i \times A_i|}$ to $\mathbb{R}^{|S_i \times A_i \times S_i|}$, i.e., set $R_i^k(s_i, a_i, s_i') = R_i^k(s_i, a_i), \forall s_i, a_i, s_i'$.

- At the end of episode k, update the confidence set $\mathcal{P}_i^k$ as in (8), and the occupancy measure:

$$\hat{q}_i^{k+1} = \underset{\hat{q}_i \in \Delta_{i,\delta_i}(\mathcal{P}_i^k)}{\operatorname{argmax}} \left\{ \eta^k \langle \hat{q}_i, R_i^k \rangle - D_{h_i}(\hat{q}_i || \hat{q}_i^k) \right\}. \tag{10}$$

From Lemma 8, we know that a NE of $\mathcal{V}_{\boldsymbol{\delta}}$ is an ϵ-NE of the game $\mathcal{V}$, and so its induced policy is an ϵ-NE of the original complete information stochastic game $\mathcal{G}$. We make the following assumption on the constrained virtual game $\mathcal{V}_{\boldsymbol{\delta}}$.

Assumption 4 *(Mertikopoulos and Zhou (2019)) The constrained virtual game $\mathcal{V}_{\boldsymbol{\delta}}$ admits a unique NE $\boldsymbol{q}^*$ that is* globally stable*, i.e.,*

$$\langle \boldsymbol{v}(\boldsymbol{q}), \boldsymbol{q}^* - \boldsymbol{q} \rangle \geq 0, \qquad \forall \boldsymbol{q} \in \boldsymbol{\Delta}_{\boldsymbol{\delta}}(\boldsymbol{P}), \tag{11}$$

with equality if and only if $\boldsymbol{q} = \boldsymbol{q}^*$, where $\boldsymbol{v}(\boldsymbol{q}) = (v_i(q_{-i}), i \in [n])$ is the vector of players' payoff gradients with respect to their own strategies.

The notion of *variational stability* was first introduced in Mertikopoulos and Zhou (2019) as a relaxation of the well-known *monotonicity condition* (Rosen, 1965). Specifically, it is shown in Mertikopoulos and Zhou (2019) that the monotonicity condition implies that the game admits a unique NE that is *globally stable*. We are now ready to state the main result of this paper.

Theorem 10 *Assume Assumptions 1, 2, 3 and 4 hold. Given any $\epsilon > 0$, assume each player i follows Algorithm 1 with a choice of δ_i satisfying Lemma 8, and a nonincreasing sequence of step-sizes satisfying $\sum_{k=1}^{\infty} \eta^k = \infty$, $\sum_{k=1}^{\infty} (\eta^k)^2 < \infty$, and $\sum_{k=1}^{\infty} \eta^k \sqrt{\ln k/k} < \infty$. Then, for any $\gamma \in (0,1)$ and the choice of parameters $d^k = 2\tau \ln k$ and $\epsilon_i^k(s_i'|s_i,a_i) = \sqrt{\frac{\ln(2nk^2|A_i||S_i|^2) - \ln \gamma}{2\max(1, N_i^k(s_i,a_i))}}$, we have $\lim_{k\to\infty} \boldsymbol{\pi}^k = \boldsymbol{\pi}^*$ with probability at least $1-\gamma$, where $\boldsymbol{\pi}^*$ is an ϵ-NE policy of the game $\mathcal{G}$.*

Proof To prove Theorem 10, we first show in Lemma 11 that the event $\{P_i \in \mathcal{P}_i^k \ \forall k \in \mathbb{N}, i \in [n]\}$ happens with probability at least $1-\gamma$. Moreover, Lemma 12 shows the convergence of $\hat{\boldsymbol{q}}^k$ to $\boldsymbol{q}^k$.

Lemma 11 *Under the same assumptions as in Theorem 10, with probability at least $1-\gamma$, we have $P_i \in \mathcal{P}_i^k, \forall k \in \mathbb{N}$, $i \in [n]$.*

Lemma 12 *Under the same assumptions as in Theorem 10, and by conditioning on the event $\{P_i \in \mathcal{P}_i^k \ \forall k \in \mathbb{N}, i \in [n]\}$, we have $\lim_{k\to\infty} \boldsymbol{q}^k - \hat{\boldsymbol{q}}^k = \mathbf{0}$.*

It suffices to show with probability at least $1-2\gamma$, we have $\lim_{k\to\infty} \boldsymbol{q}^k = \boldsymbol{q}^*$, where $\boldsymbol{q}^*$ is the unique stable NE of the constrained virtual game $\mathcal{V}_{\boldsymbol{\delta}}$. Let us define $\boldsymbol{D_h}(\boldsymbol{q}^*||\hat{\boldsymbol{q}}^k) \triangleq \sum_{i=1}^n D_{h_i}(q_i^*||\hat{q}_i^k)$. From the definition of Bregman divergence as well as Lemma 12, we have

$$\lim_{k\to\infty} \boldsymbol{q}^k = \boldsymbol{q}^* \Leftrightarrow \lim_{k\to\infty} \hat{\boldsymbol{q}}^k = \boldsymbol{q}^* \Leftrightarrow \lim_{k\to\infty} \boldsymbol{D_h}(\boldsymbol{q}^*||\hat{\boldsymbol{q}}^k) = 0. \tag{12}$$

The rest of the proof proceeds in two steps. In the first step, we show that every neighborhood $U \subset \boldsymbol{\Delta}(\boldsymbol{P})$ of $\boldsymbol{q}^*$ is recurrent in $\{\boldsymbol{q}^k\}_{k=1}^{\infty}$. Then, we further show that for any $\epsilon, \delta > 0$, there exists $k_0 \in \mathbb{N}$, such that $\mathbb{P}\big\{\boldsymbol{D_h}(\boldsymbol{q}^*||\hat{\boldsymbol{q}}^k) \le \epsilon, \forall k > k_0\big\} \ge 1-\delta$.

Lemma 13 *Under the same assumptions as in Theorem 10, and by conditioning on the event $\{P_i \in \mathcal{P}_i^k \ \forall k \in \mathbb{N}, i \in [n]\}$, every open neighborhood $U \subset \boldsymbol{\Delta}(\boldsymbol{P})$ of $\boldsymbol{q}^*$ is recurrent in $\{\boldsymbol{q}^k\}_{k=1}^{\infty}$. More specifically, there exists a subsequence $\boldsymbol{q}^{k_m}$ of $\boldsymbol{q}^k$ such that $\boldsymbol{q}^{k_m} \to \boldsymbol{q}^*$ almost surely.*

Lemma 14 *Under the same assumptions as in Theorem 10, and by conditioning on the event $\{P_i \in \mathcal{P}_i^k \ \forall k \in \mathbb{N}, i \in [n]\}$, for any $\epsilon, \delta > 0$, there exists $k_0 \in \mathbb{N}$ such that $\mathbb{P}\big\{\boldsymbol{D_h}(\boldsymbol{q}^*||\hat{\boldsymbol{q}}^k) \le \epsilon, \forall k > k_0\big\} \ge 1-\delta$.*

To complete the proof Theorem 10, for any $\epsilon > 0$, let us consider the event $E_\epsilon := \{\exists k_0 \in \mathbb{N} : \boldsymbol{D_h}(\boldsymbol{q}^*||\hat{\boldsymbol{q}}^k) \le \epsilon, \ \forall k > k_0\}$. Then, we have $\mathbb{P}\{\lim_{k\to\infty} \boldsymbol{D_h}(\boldsymbol{q}^*||\hat{\boldsymbol{q}}^k) = 0\} = \mathbb{P}\big\{\cap_{r=1}^{\infty} E_{2^{-r}}\big\}$. Using Lemma 14 and conditioned on the event $\{P_i \in \mathcal{P}_i^k \ \forall k \in \mathbb{N}, i \in [n]\}$, we can show that $\mathbb{P}\big\{\lim_{k\to\infty} \boldsymbol{D_h}(\boldsymbol{q}^*||\hat{\boldsymbol{q}}^k) = 0\big\} = \mathbb{P}\big\{\cap_{r=1}^{\infty} E_{2^{-r}}\big\} = 1$. Therefore, using Lemma 11 and relation (12), we conclude that with probability at least $1-\gamma$, we have $\lim_{k\to\infty} \boldsymbol{q}^k = \boldsymbol{q}^*$. ■

6. Conclusion

In this work, we studied the class of stochastic games with unknown independent chains. Relying on a compact dual formulation of the game based on occupancy measures and the technique of confidence set to maintain high-probability estimates of the unknown transition matrices, we proposed a fully decentralized and independent online mirror descent algorithm to learn an ϵ-NE stationary policy for this class of stochastic games. The proposed algorithm has the desired properties of independence and convergence such that under the variational stability assumption of the game, it converges asymptotically to an ϵ-NE stationary policy with arbitrarily high probability.

References

Eitan Altman. *Constrained Markov Decision Processes*. Routledge, 2021.

Eitan Altman, Konstantin Avratchenkov, Nicolas Bonneau, Mérouane Debbah, Rachid El-Azouzi, and Daniel Sadoc Menasché. Constrained stochastic games in wireless networks. In *IEEE GLOBECOM 2007-IEEE Global Telecommunications Conference*, pages 315–320. IEEE, 2007.

Eitan Altman, Konstantin Avrachenkov, Nicolas Bonneau, Merouane Debbah, Rachid El-Azouzi, and Daniel Sadoc Menasche. Constrained cost-coupled stochastic games with independent state processes. *Operations Research Letters*, 36(2):160–164, 2008.

Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 5527–5540, 2020.

S. Rasoul Etesami. Learning stationary Nash equilibrium policies in n-player stochastic games with independent chains. *SIAM Journal on Control and Optimization*, 62(2):799–825, 2024.

S. Rasoul Etesami, Walid Saad, Narayan B Mandayam, and H Vincent Poor. Stochastic games for the smart grid energy management with prospect prosumers. *IEEE Transactions on Automatic Control*, 63(8):2327–2342, 2018.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Emily Meigs, Francesca Parise, and Asuman Ozdaglar. Learning in repeated stochastic network aggregative games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6918–6923. IEEE, 2019.

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173:465–507, 2019.

Prashant Narayanan and Lakshmi Narasimhan Theagarajan. Large player games on wireless networks. *arXiv preprint arXiv:1710.08800*, 2017.

Gergely Neu, András György, Csaba Szepesvári, et al. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pages 231–243. Citeseer, 2010.

Tiancheng Qin and S Rasoul Etesami. Scalable and independent learning of Nash equilibrium policies in n-player stochastic games with unknown independent chains. *arXiv preprint arXiv:2312.01587*, 2023.

Shuang Qiu, Xiaohan Wei, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. Provably efficient fictitious play policy optimization for zero-sum Markov games with structured transitions. In *International Conference on Machine Learning*, pages 8715–8725. PMLR, 2021.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32, 2019.

Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Başar, and Asuman Ozdaglar. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334, 2021.

Muhammed O Sayin, Francesca Parise, and Asuman Ozdaglar. Fictitious play in zero-sum stochastic games. *SIAM Journal on Control and Optimization*, 60(4):2095–2114, 2022.

Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953.

Vikas Vikram Singh and N Hemachandra. A characterization of stationary Nash equilibria of constrained stochastic games with independent state processes. *Operations Research Letters*, 42(1): 48–52, 2014.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown Markov games. In *International Conference on Machine Learning*, pages 10279–10288. PMLR, 2021.

Muhammad Aneeq uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2278–2284. IEEE, 2020.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on Learning Theory*, pages 4259–4299. PMLR, 2021.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in multi-agent Markov stochastic games: Stationary points and convergence. *arXiv preprint arXiv:2106.00198*, 2021b.

Wenzhao Zhang and Xiaolong Zou. Constrained average stochastic games with continuous-time independent state processes. *Optimization*, 71(9):2571–2594, 2022.

Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 2736–2761. PMLR, 2022.