IEEE Control Systems Letters paper presented at
2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Learning Nash in Constrained Markov Games with an α-Potential

Soham Das, *Student Member, IEEE* and Ceyhun Eksin, *Member, IEEE*

***Abstract*— We develop a best-response algorithm for solving constrained Markov games assuming limited violations for the potential game property. The limited violations of the potential game property mean that changes in value function due to unilateral policy alterations can be measured by the potential function up to an error α. We show the existence of stationary ϵ-approximate constrained Nash policy whenever the set of feasible stationary policies is non-empty. Our setting has agents accessing an efficient probably approximately correct solver for a constrained Markov decision process which they use for generating best-response policies against the other agents' former policies. For an accuracy threshold $\epsilon > 4\alpha$, the best-response dynamics generate provable convergence to ϵ-Nash policy in finite time with probability at least $1-\delta$ at the expense of polynomial bounds on sample complexity that scales with the reciprocal of ϵ and δ.**

I. Introduction

A stochastic game involves repeated interactions among several participants when the environment state is dynamic and evolves in response to the actions of the agents in a stochastic fashion. Each player optimizes its own objective function while considering the actions of others. For many applications, such as in the case of modeling safety critical behaviour for autonomous vehicles navigating crowded environments, constraints are additionally needed on the evolution of the game so that physical limitations (e.g., speed limits for vehicles) or safety requirements (e.g., collision avoidance) can be guaranteed (see [1]–[3]. Accordingly, here we are interested in constrained Markov games, i.e. we consider a stochastic dynamic game on an infinite time-horizon, with the system state evolving according to a transition kernel. The agents take actions after each transition of the system with the goal to maximize their discounted infinite horizon payoffs, while respecting constraints on potentially multiple other criteria. The transition kernel is unknown to the agents, but they can access a trajectory of sample paths and rewards by making subsequent calls to a given simulation oracle. Our goal is to design an algorithmic framework that can provably reach an approximate Nash policy in finite time, while maintaining strict feasibility throughout with bounded sample complexity.

Constrained multi-agent reinforcement learning (RL) is challenging. Most of the current results in multi-agent RL are for the unconstrained setting. For example, policy gradient (PG) methods are provably effective with good convergence characteristics [4]. Examples include entropy regularized natural PG [5] for Markov decision processes (MDPs), independent PG for Markov potential games (MPGs) [6], for zero-sum stochastic games [7], among others. Further, decentralized value-based methods, such as two-time scale Q-learning dynamics [8] are provably convergent to Nash in zero-sum discounted Markov games. Constrained stochastic games, however, present non-trivial challenges even in verifying the existence of a stationary Nash Equilibrium (NE) solution concept (see [9], [10], [11], [12] for existence results under different assumptions on the game).

Here, we first establish that a stationary constrained ϵ-Nash policy exists whenever the game has an α-potential [13], i.e, the maximum violation of the potential property is confined to some finite α and non-empty set of feasible policies (Theorem 1). Markov α-potential games is a new framework for studying Markov games, formulated in [13] and expanded upon in [14]. While the potential function furthers understanding of non-cooperative behavior between agents by relating the change in values to the change in potential as a result of policy changes by the agents [15], the framework does not extend easily to real world scenarios. Most Markov games with Markovian transition and policies do not admit such potentials. In addition, certifying whether a game is a MPG can be challenging. However, every Markov game, is a Markov α-potential game for some $\alpha \geq 0$.

Thereafter we design a best-response framework where agents consider their best feasible policy with respect to maximizing their value functions if the other agents' policies stay fixed to their current values. We hypothesize that agents have access to an efficient probably approximately correct (PAC) learner to solve the resulting single agent constrained MDP sub-problem (CMDP) with arbitrary accuracy and confidence, at the expense of a number of samples that grows polynomially in the reciprocal of the accuracy and confidence parameters. PAC learners have been successfully designed for the CMDP problem in [16]–[18], with suitable restrictions on the underlying constrained game. Under PAC accessibility, we show that our algorithm converges to an ϵ-Nash policy in finite time with a probability that can be chosen arbitrarily at the expense of a bounded sample complexity, the functional form of which would depend on the particular PAC learner's characteristics (Theorem 2). The analysis develops on ideas in [19], where Alatur et al. have a coordinate-ascent style algorithm (see Song et al. [20]), albeit, for the non-discounted finite horizon case with a stringent potential function assumption, different restrictions on a constrained MDP solver, and a single constraint on the game.

Notation: The notations $\mathbb{R}$, $\mathbb{N}$ and $\mathbb{Z}$ represent, respectively, the set of real numbers, natural numbers and integers. We define $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$ and $\mathbb{N}_{\geq 0} = \mathbb{N} \cup \{0\}$. We use $\Delta(\mathcal{X})$ to denote the space of probability distributions for any set $\mathcal{X}$ (the probability simplex). We use brackets around an

This work was supported by NSF ECCS-1953694, NSF CCF-2008855, and CAREER 2239410.

Soham Das and Ceyhun Eksin are with the Industrial and Systems Engineering Department, Texas A&M University, College Station, TX 77843. E-mail: soham.das@tamu.edu; eksinc@tamu.edu.

integer value k to refer to the set $[k] := \{1, 2, \ldots, k\}$. We will use $\mathcal{S}(\mathcal{G})$ to refer to the element $\mathcal{S}$ of a tuple $\mathcal{G}$, $\mathcal{N}(\mathcal{G})$ to refer to the element $\mathcal{N}$ of a tuple $\mathcal{G}$, and so on.

II. Constrained Markov games

A. Game definition

The constrained Markov game can be specified by the tuple $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \{\mathcal{A}_i, r_i\}_{i\in\mathcal{N}}, P, \{c_j, \beta_j\}_{j=1}^{k})$. Here $\mathcal{S}$ is a finite state space of size $S = |\mathcal{S}|$. We use $\mathcal{N} = [n]$ to denote the set of $n \geq 2$ agents in the game. $\mathcal{A}_i$ is the finite action space for agent $i \in \mathcal{N}$ with elements $a_i \in \mathcal{A}_i$. The notation $r_i : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the individual normalized reward function of agent $i \in \mathcal{N}$. The global dynamic state $s \in \mathcal{S}$ is driven by P, the transition probability kernel, i.e., $P(s'|s, a)$ is the probability of the state variable to move from state s to state s' when $a \in \mathcal{A}$ is the action profile of the agents. We define $\gamma \in (0, 1)$ as the discount factor for future rewards and costs incurred for agents. For each agent $i \in \mathcal{N}$, we consider a stochastic stationary policy $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ where $\pi_i \in \Pi_i := \Delta(\mathcal{A}_i)^{\mathcal{S}}$, that determines a probability distribution over the actions of agent i at each state $s \in \mathcal{S}$. The constrained Markov game $\mathcal{G}$ enforces k discounted cost constraints on the evolution of the game for any joint policy profile $\pi = (\pi_i)_{i\in\mathcal{N}} \in \Pi := \times_{i\in\mathcal{N}}\Pi_i$ as given by

$$\mathcal{U}_j(\pi) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c_j^t(s^t, a^t)|s^0 = s\right] \leq \beta_j \ \forall j \in [k], \quad (1)$$

where $s \in \mathcal{S}$ is the initial state, $s^t \in \mathcal{S}$ and $a^t \in \Delta(\mathcal{A})$ denote the state and action profile at time $t \in \mathbb{N}_{\geq 0} \cup \{\infty\}$, $c_j^t : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the j-th cost function at time step t, and $\beta_j \in (0, \frac{1}{1-\gamma}]$ is the right-hand-side of constraint j. A policy π is feasible, if it satisfies all k-constraints. We allow the constraints to be *coupled* across agents and time, that is they depend on the joint actions of all agents in the game, for all times. We use $\Pi^C \subseteq \Pi$ to refer to the set of stationary feasible policies in the game. Formally, $\Pi^C = \{\pi \in \Pi : \mathcal{U}_j(\pi) \leq \beta_j \ \forall j \in [k]\}$. We define $\Pi_i^C(\pi_{-i}) := \{\pi_i \in \Pi_i : (\pi_i, \pi_{-i}) \in \Pi^C\}$ to refer to the set of feasible policies available to agent $i \in \mathcal{N}$ when the remaining agents play π_{-i}. Similarly, we define $\Pi_{-i}^C(\pi_i) := \{\pi_{-i} \in \Pi_{-i} : (\pi_i, \pi_{-i}) \in \Pi^C\}$.

B. Value functions

For any policy $\pi \in \Pi$, the value function $V_i^s : \Pi \to \mathbb{R}$ gives the expected cumulative reward of agent $i \in \mathcal{N}$ when $s^0 = s$ and the agents draw their actions $a^t = (a_i^t, a_{-i}^t)$ for time $t \geq 0$ using the policies (π_i, π_{-i}). Define $r_i^t := r_i(s^t, a^t)$. Then

$$V_i^s(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_i^t|s^0 = s]. \quad (2)$$

V_i^s is a deterministic quantity, given fixed initial state s. Define $\tilde{V}_i^s(\pi)_H := \sum_{t=0}^{H} \gamma^t r_i^t$ as the random variable that captures the horizon-truncated value for agent i following policy $\pi \in \Pi^C$. We define the average horizon truncated value for agent i as the random variable

$$\hat{V}_i^s(\pi)_H = \frac{1}{M}\sum_{l=1}^{M} \tilde{V}_i^s(\pi)_H^{(l)} \quad (3)$$

where $M \in \mathbb{N}$ is the number of episodes selected to perform the average on, and $\tilde{V}_i^s(\pi)_H^{(l)}$ is $\tilde{V}_i^s(\pi)_H$ for the l-th episode. When $H = \infty$, we drop H from notation in (3). Whenever clear from the context, we drop the fixed initial state s from the superscript of the value functions.

C. The α-potential

Definition 1 *The function $\Phi : \mathcal{S} \times \Pi \to \mathbb{R}$ is an α-potential for game $\mathcal{G}$ if*

$$\big|(\Phi(s, \tilde{\pi}_i, \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i})) - (V_i^s(\tilde{\pi}_i, \pi_{-i}) - V_i^s(\pi_i, \pi_{-i}))\big| \leq \alpha \quad (4)$$

is satisfied for some $\alpha \geq 0$ for all $s \in \mathcal{S}$, $i \in \mathcal{N}$, $(\pi_i, \pi_{-i}) \in \Pi$, $(\tilde{\pi}_i, \pi_{-i}) \in \Pi$.

Definition 2 *A constrained Markov game is a constrained Markov α-potential game, if it admits an α-potential function for some $\alpha \geq 0$.*

In [13], the authors show that for a Markov game $\mathcal{G}$, an α-potential is guaranteed to exist, under some continuity conditions of the potential with respect to the policies. Thus, any Markov game $\mathcal{G}$ is a Markov α-potential game for some α. When $\alpha = 0$, the corresponding Φ is a candidate potential function (see Leonardos et al. [15] for more exposition on MPGs). In Markov games, verifying whether a potential function exists or not can be prohibitive, and the potential function assumption is not satisfied by most games. Thus we relax the potential function assumption as in [13].

Assumption 1 *The approximation value α is finite and known for the constrained Markov game $\mathcal{G}$.*

Remark 1 *In [13], the authors provide an optimization based approach to find the value of α for different Markov games. Clearly, from definition 2, every Markov game is a Markov α-potential game for some α. The focus of this work, however, is not to find the value of α for different games. Instead, we assume we are given a constrained Markov game with an α-potential. We need not necessarily know the exact functional form of the α-potential, but we know that such a potential exists for the given value of $\alpha \geq 0$.*

D. Solution concept: ϵ-NE policy profile

Assumption 2 *The initial state of the game is fixed to any arbitrary state $s \in \mathcal{S}$.*

The initial state serves as the boundary condition for the game $\mathcal{G}$. We define the solution concept that agents seek to achieve in game $\mathcal{G}$.

Definition 3 *(Constrained Stationary ϵ-NE) A policy profile $\pi^* = (\pi_i^*, \pi_{-i}^*) \in \Pi^C$ is a constrained stationary ϵ-Nash*

Equilibrium (ϵ-NE) of a constrained Markov game $\mathcal{G}$, for some $\epsilon > 0$ if for any $i \in \mathcal{N}$ and $\pi_i \in \Pi_i^C(\pi_{-i}^*)$ we have,

$$V_i^s(\pi_i^*, \pi_{-i}^*) + \epsilon \geq V_i^s(\pi_i, \pi_{-i}^*) \quad \forall s \in \mathcal{S}. \tag{5}$$

When $\epsilon = 0$, we retrieve the stationary NE definition. The NE is defined as a set of policies of the players which satisfy simultaneously all the constraints and for which, in addition, no player can improve his payoff when unilaterally modifying his policy while still satisfying its own constraints. In the following, we show that when the constrained Markov game is an α-potential game, for some $\alpha > 0$, we can claim that an approximate stationary NE policy must exist.

Theorem 1 *In a constrained Markov game $\mathcal{G}$ with α-potential Φ, a constrained stationary ϵ-Nash policy exists for all $\epsilon > \alpha$, when $\Pi^C \neq \emptyset$ and Φ is continuous in policies.*

Proof: Given $\Pi^C \neq \emptyset$, we have that $\mathcal{G}$ has stationary feasible policies. Since the constraints in (1) can be satisfied via equality, Π^C is closed. Moreover, since $\Pi^C \subseteq \Pi = \times_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)^{\mathcal{S}}$, Π^C is bounded. Therefore Π^C is compact, which implies that $\pi^* \in \operatorname{argmax}_{\pi \in \Pi^C} \Phi(s, \pi)$ must exist (using the extreme value theorem) for continuous Φ, where s is the arbitrary fixed initial state of the game. We claim that π^* is a constrained stationary ϵ-Nash policy profile for $\epsilon > \alpha$. Assume for the sake of contradiction that π^* is not an ϵ-Nash profile, for $\epsilon > \alpha$. Then, there exists $i \in \mathcal{N}$ such that $\hat{\pi}_i \in \operatorname{argmax}_{\pi_i \in \Pi_i^C(\pi_{-i}^*)} V_i^s(\pi_i, \pi_{-i}^*)$ such that

$$V_i^s(\hat{\pi}_i, \pi_{-i}^*) > V_i^s(\pi^*) + \epsilon \tag{6}$$

Now, because Φ is an α-potential for $\mathcal{G}$, we have

$$\Phi(s, \hat{\pi}_i, \pi_{-i}^*) - \Phi(s, \pi^*) \geq V_i^s(\hat{\pi}_i, \pi_{-i}^*) - V_i^s(\pi^*) - \alpha > \epsilon - \alpha \tag{7}$$

where the second-inequality is due to (6). Since $\epsilon > \alpha$ is given, we have a contradiction, as (7) indicates that π^* is not the maximizer for Φ. ∎

III. Sequential best-response dynamics

The goal of the agents is to produce a constrained stationary ϵ-NE policy profile for the game $\mathcal{G}$. We proceed to show that our prescribed Algorithm 1, a sequential best-response dynamic, is guaranteed to converge to ϵ-NE with high probability, under some restrictions on accuracy ϵ.

A. The constrained MDP sub-problem

Definition 4 *The Slater condition states that $\Pi^C \neq \emptyset$, and π^B exists in the relative interior of Π^C. Given $\mathcal{G}$, we define its Slater constant $\zeta = (\zeta_j)_{j=1}^k \in \mathbb{R}^k$ as follows*

$$\zeta = \min_{i \in \mathcal{N}} \min_{\pi_{-i} \in \Pi_{-i}} \max_{\pi_i \in \Pi_i} \{\beta - \mathcal{U}(\pi_i, \pi_{-i})\} \tag{8}$$

Assumption 3 *The constrained Markov game $\mathcal{G}$ is strictly feasible, and satisfies the Slater's condition.*

$\mathcal{G}$ is strictly feasible if and only if $\zeta_j > 0$ for all $j \in [k]$. Thus $\zeta_j \in (0, \frac{1}{1-\gamma}]$ for all $j \in [k]$, and $\frac{1}{\zeta} := (1/\zeta_j)_{j=1}^k$ is well-defined. We further assume that agents do not have access to the state-transition distributions and payoffs, but can learn by interacting with a sampling oracle of the game that returns a sample of the next state, when given a state-action pair as input. Define $I(\mathcal{G}) := (\mathcal{S}(\mathcal{G}), \mathcal{N}(\mathcal{G}), \{\mathcal{A}_i\}_{i \in \mathcal{N}}(\mathcal{G}), \{c_j, \beta_j\}_{j=1}^k(\mathcal{G}))$ as the information available to the agents $i \in \mathcal{N}$ in constrained Markov game $\mathcal{G}$. Further, the sampling oracle available to learning agents $\mathcal{M}$ takes input (s, a) and generates an immediate payoff $\hat{r}_i(s, a)$ and a state transition to next state $\hat{s}$ such that the next state is chosen with probabilities $P(\hat{s}|s, a)(\mathcal{G})$. Let $\mathcal{G}_{-i}$ refer to the constrained MDP (CMDP) obtained from the constrained Markov game $\mathcal{G}$ when all agents other than i fix their policies to $\pi_{-i} \in \Pi_{-i}$. Then the solution to the CMDP $\mathcal{G}_{-i}$ (the policy π_i^* that maximizes the value function for agent i when other agents play π_{-i}), under the assumption that agent i has access to information $I(\mathcal{G})$ and sampling oracle $\mathcal{M}$, is a RL problem.

Definition 5 *A learning algorithm L_i is a $(\hat{\epsilon}, \hat{\delta})$-efficient probably approximately correct (PAC) learner for the RL problem $\mathcal{G}_{-i}$, if for any approximation factor $\hat{\epsilon} > 0$ and confidence factor $\hat{\delta} \in (0, 1)$, $\zeta > 0$, L_i produces policy $\hat{\pi}_i \in \Pi_i^C(\pi_{-i})$ such that*

$$\mathbb{P}_{\mathcal{G}_{-i}}[V_i^s(\pi_i^*, \pi_{-i}) - V_i^s(\hat{\pi}_i, \pi_{-i}) \leq \hat{\epsilon}] \geq 1 - \hat{\delta} \tag{9}$$

where π^* is the optimal policy solution to $\mathcal{G}_{-i}$ and $\mathcal{L}_i$ produces $\hat{\pi}_i$ in time $poly(|\mathcal{S}|, |\mathcal{A}_i|, \frac{1}{\zeta}, \frac{1}{\hat{\epsilon}}, \frac{1}{\hat{\delta}}, \frac{1}{1-\gamma}, r_{\max})$ where $r_{\max}$ is the maximum immediate reward on any transition in the problem, and s is the arbitrary fixed initial state.

Assumption 4 *(PAC accessibility for agents) Agent i has access to a $(\hat{\epsilon}, \hat{\delta})$-efficient PAC learner L_i for solving RL problem $\mathcal{G}_{-i}$, for any $\hat{\epsilon} > 0$ and $\hat{\delta} \in (0, 1)$ and $i \in \mathcal{N}$.*

That is, L_i solves the problem $\mathcal{G}_{-i}$ with $\hat{\epsilon}$ accuracy, with probability at least $1 - \hat{\delta}$, while making at most a polynomial number of calls to the sampling oracle. A higher accuracy (smaller $\hat{\epsilon}$) or a higher confidence (smaller $\hat{\delta}$) causes the number of samples required to grow. Under the assumption that each call to the sampling oracle $\mathcal{M}$ can be resolved in $\mathcal{O}(1)$, the sample complexity of the PAC-learner L_i is $\mathcal{O}(poly(|\mathcal{S}|, |\mathcal{A}_i|, \frac{1}{\zeta}, \frac{1}{\hat{\epsilon}}, \frac{1}{\hat{\delta}}, \frac{1}{1-\gamma}, r_{\max})$.

Remark 2 *We remark here that not all constrained MDPs have a sample efficient PAC solver. In [16], [17], [18], sample efficient PAC solvers have been designed under suitable assumptions on the constrained MDP. Our main result (see Theorem 2, section IV) holds whenever the constrained MDP subproblem admits an efficient PAC solver.*

B. Algorithm 1 under the lens

We are now ready to highlight the core components of Algorithm 1. Starting with a feasible policy π^B, which we assume exists, the algorithm improves the policy through a sequence of best-response steps, where each agent i evaluates its best-response policy, assuming the policies of other agents remaining the same, via solving the RL problem $\mathcal{G}_{-i}$.

Algorithm 1 Sequential Best-Response Dynamics

Require: $\mathcal{G}$ is a constrained Markov game, initial state s
Ensure: $\epsilon > 0$, $T \in \mathbb{Z}_+$, $\pi^B \in \Pi^C$, $s \in \mathcal{S}$
1: **function** SEQUENTIAL-BR($\mathcal{G}, \epsilon, T, \pi^B$)
2: $\pi^0 \leftarrow \pi^B$
3: **for all** $t = 1, ..., T$ **do**
4: Estimate $\hat{V}_i^s(\pi^{t-1})_H$ for all $i \in \mathcal{N}$.
5: **for all** agent $i = 1, ..., n$ **do**
6: $\hat{\pi}_i^t \leftarrow L_i(\mathcal{G}_{-i}, I(\mathcal{G}), \mathcal{M})$
7: Estimate $\hat{V}_i^s(\hat{\pi}_i^t, \pi_{-i}^{t-1})_H$
8: $\Delta_i^t \leftarrow \hat{V}_i^s(\hat{\pi}_i^t, \pi_{-i}^{t-1})_H - \hat{V}_i^s(\pi^{t-1})_H$
9: **end for**
10: **if** $\max_{i \in \mathcal{N}} \Delta_i^t > \epsilon/2$ **then**
11: $j \leftarrow \text{argmax}_{i \in \mathcal{N}} \Delta_i^t$
12: $\pi^t \leftarrow (\hat{\pi}_j^t, \pi_{-j}^{t-1})$
13: **else**
14: $\pi^t \leftarrow \pi^{t-1}$
15: **return** π^t
16: **end if**
17: **end for**
18: **end function**

Following Line 6, agents estimate their value functions using the new policy $(\hat{\pi}_i, \pi_{-i})$ in Line 7, and store the improvement in the value function V_i^s in the variable Δ_i. To estimate the value function for a policy π, an agent simulates the system with actions sampled from the policy for $M \in \mathbb{N}$ episodes, starting from the fixed initial state $s \in \mathcal{S}$. Moreover, for each episode, the agent terminates the simulation of the policy after $H \in \mathbb{N}$ discrete time steps. After all agents calculate their individual improvements in estimated values, the agent whose update provides the best improvement in estimated values (Line 11) larger than $\epsilon/2$ gets to update the policy, where ϵ is the approximation parameter of the problem.

IV. Analysis of Best-Response

Theorem 2 (Main Result) *Let Assumptions 1-4 hold. Given any* $\epsilon > \max(4\alpha, \frac{8\gamma^{H+1}}{1-\gamma})$, $\delta \in (0,1)$, $M = \lceil \frac{128}{\epsilon'^2} log \frac{32n^2\alpha'}{\delta(\epsilon-4\alpha)} \rceil$, *we have that with probability at least* $1-\delta$, *Algorithm 1 converges to an* ϵ*-Nash policy in at most* $T = \lfloor \frac{4n\alpha'}{\epsilon-4\alpha} \rfloor$ *steps of the for-loop in Line 3, with a sample complexity of*

$$\sum_{i \in \mathcal{N}} poly(|S|, |\mathcal{A}_i|, \frac{1}{\zeta}, \frac{4}{\epsilon}, \frac{1}{\delta'}, \frac{1}{1-\gamma}, r_{\max}) + \frac{8n^2\alpha' MH}{\epsilon - 4\alpha} \tag{10}$$

where $\delta' = \frac{2\delta(\epsilon-4\alpha)}{n^2\alpha'}$, $\alpha' = \alpha + \frac{2}{1-\gamma}$ *and* $\epsilon' = \epsilon(1-\gamma) - 8\gamma^{H+1}$ *and* H *is the simulation episode length for agents.*

The theorem guarantees Algorithm 1 will converge to an ϵ-Nash in $t = T$ steps (see Line 3) with probability at least $1-\delta$, while consuming a bounded number of samples.

Remark 3 *The technical assumption* $\epsilon > \max(4\alpha, 8\frac{\gamma^{H+1}}{1-\gamma})$ *in Theorem 2 emphasizes the latent accuracy limitation of our proposed best-response dynamic. Given the known* α *for the game, Algorithm 1 is guaranteed to reach an* ϵ*-NE policy for* $\epsilon > 4\alpha$. *That is, the performance guarantee does not hold if agents want to converge to* $\epsilon \leq 4\alpha$. *The term* $8\frac{\gamma^{H+1}}{1-\gamma}$ *appears as a result of the restriction that agents must estimate the value by simulating the policy for a truncated horizon length of* H. *For sufficiently large* H, *the term* γ^{H+1} *approaches* 0 *for* $\gamma \in (0,1)$. *This implies that the restriction* $8\frac{\gamma^{H+1}}{1-\gamma}$ *on the accuracy threshold* ϵ *for Nash policy becomes negligible when agents use a sufficiently large horizon.*

Remark 4 *The sample complexity of Algorithm 1 is polynomial in* $1/\delta'$, *and* δ' *grows linearly with* δ. *This implies a higher confidence in convergence comes at the expense of a larger sample complexity requirement. Moreover,* $\frac{\partial \delta'}{\partial \alpha} = -2\delta(\epsilon + \frac{8}{1-\gamma})(n\alpha')^{-2}$ *which indicates the quadratic rate of decline of* δ' *with growth in* α. *Our provable guarantee therefore highlights that for a fixed confidence level* δ, *games which admit* α*-potentials for larger values of* α *would have a more stringent requirement in terms of the sample complexity of the best-response dynamic, as* δ' *declines quadratically with respect to* α *for the game.*

To prove theorem 2, we shall first show a few auxiliary results in what follows.

Lemma 1 *Assume agents have access to a feasible policy* $\pi^B \in \Pi^C$ *in game* $\mathcal{G}$. *Then* $\Phi(\pi) - \Phi(\pi^B) \leq n\alpha'$, *for all policies* π, *where* Φ *is an* α*-potential.*

Proof: Define $\{\tilde{\pi}^{(u)}\}_{u=0,..,n}$ as

$$\tilde{\pi}^{(u)} = (\pi_1^B, ..., \pi_u^B, \pi_{u+1}, ..., \pi_n) \tag{11}$$

for $u = 0, 1, ...n$. Then, $\tilde{\pi}^{(0)} = \pi$ and $\tilde{\pi}^{(n)} = \pi^B$. For any agent $i \in \mathcal{N}$, the policies $\tilde{\pi}^{(i-1)}$ and $\tilde{\pi}^{(i)}$ differ only in the index for agent i. Then, we know that

$$-\alpha \leq (\Phi(\tilde{\pi}^{(i-1)}) - \Phi(\tilde{\pi}^{(i)})) - (V_i(\tilde{\pi}^{(i-1)}) - V_i(\tilde{\pi}^{(i)})) \leq \alpha \tag{12}$$

since Φ is an α-potential. Note that we dropped the fixed, arbitrary initial state s from the superscript of the value functions above. Summing over $i = 1, ..., n$, we have the following upper bound

$$\Phi(\pi) \leq \alpha n + \sum_{i=1}^{n} (V_i(\tilde{\pi}^{(i-1)}) - V_i(\tilde{\pi}^{(i)})) + \Phi(\pi^B). \tag{13}$$

Now, see that

$$-2(1-\gamma)^{-1} \leq V_i(\tilde{\pi}^{(i-1)}) - V_i(\tilde{\pi}^{(i)}) \leq 2(1-\gamma)^{-1} \tag{14}$$

holds because we work with normalized rewards between $[-1, 1]$. Thus the right hand side of inequation (13) is at most $\alpha n + 2n(1-\gamma)^{-1} + \Phi(\pi^B)$, or $\Phi(\pi) - \Phi(\pi^B) \leq n\alpha'$ holds, substituting for $\alpha' = \alpha + 2(1-\gamma)^{-1}$. ■

Thus, the maximum increase in the potential value is bounded by $n\alpha'$, for any α-potential $\Phi \in \mathcal{F}$. We shall now show that for each step t in line 3 of Algorithm 1, there is a strict increase of the potential function value between π^{t+1} and π^t for all α-potentials. Recall here that $\hat{V}_i(\pi)_H$ is a random variable, while $V_i(\pi)$ is a deterministic quantity.

Definition 6 *(κ-accurate value estimation) Agents are capable of κ-accuracy for the value function estimation when for all policies $\pi \in \Pi$ encountered in the execution of Algorithm 1, the event $|\hat{V}_i(\pi)_H - V_i(\pi)| \leq \kappa$ for $\kappa \in \mathbb{R}_+$ happens with probability 1 for all agents $i \in \mathcal{N}$, for all steps t in Algorithm 1, line 3.*

Lemma 2 *When agents are capable of $\epsilon/8$-accuracy in value function estimation, then each step of line 3 in Algorithm 1 ensures a strict increase in the potential value of the current policy as*

$$\Phi(\pi^{t+1}) - \Phi(\pi^t) > \epsilon/4 - \alpha \tag{15}$$

for any α-potential $\Phi \in \mathcal{F}$, given finite α and $\epsilon > 4\alpha$.

Proof: If the Algorithm 1 does not terminate at step t, then there exists $j \in \mathcal{N}$ such that $\Delta_j^t > \epsilon/2$. Since Φ is an α-potential, we have $\Phi(\hat{\pi}_j^t, \pi_{-j}^{t-1}) - \Phi(\pi^{t-1}) \geq V_j(\hat{\pi}_j^t, \pi_{-j}^{t-1}) - V_j(\pi^{t-1}) - \alpha$. Conditioning on the event that agents have $\epsilon/8$-accurate value estimations in the course of execution of Algorithm 1, we have, $V_j(\hat{\pi}_j^t, \pi_{-j}^{t-1}) \geq \hat{V}_j(\hat{\pi}_j^t, \pi_{-j}^{t-1})_H - \epsilon/8$ holds. Again, following the same conditioning argument, we get $V_j(\pi^{t-1}) \leq \hat{V}_j(\pi^{t-1})_H + \epsilon/8$. Thus,

$$\begin{aligned} &V_j(\hat{\pi}_j^t, \pi_{-j}^{t-1}) - V_j(\pi^{t-1}) - \alpha \\ &\quad \geq \hat{V}_j(\hat{\pi}_j^t, \pi_{-j}^{t-1})_H - \epsilon/8 - \hat{V}_j(\pi^{t-1})_H - \epsilon/8 - \alpha. \end{aligned} \tag{16}$$

The right-hand-side of (16) is equal to $\Delta_j^t - \epsilon/4 - \alpha$ (See Algorithm 1, line 8). Since $\Delta_j^t > \epsilon/2$, we have (15) from (16). ∎

Corollary 1 *When agents have κ-accurate value estimations for $\kappa = \epsilon/8$, Algorithm 1 converges in at most $T = \lfloor \frac{n\alpha'}{\epsilon/4 - \alpha} \rfloor$ steps of the for-loop in line 3.*

Proof: Since the maximum increase in the potential value is bounded (Lemma 1), and there is a strict increase in the potential value for every iteration of the for-loop in line 3 (Lemma 2), we have the desired result. ∎

Definition 7 *(Event $\mathcal{E}_{\approx}$) We define $\mathcal{E}_{\approx}$ as the event where in Line 6 of Algorithm 1, an agent i uses L_i to produce an $\hat{\epsilon} = \epsilon/4$-optimal policy for $\mathcal{G}_{-i}$, for all $i \in \mathcal{N}$, for all time-steps $t \in [T]$ in Line 3.*

Lemma 3 *Conditioning on agents making $\epsilon/8$-accurate value estimations and event $\mathcal{E}_{\approx}$ during the execution of Algorithm 1, π^T is an ϵ-Nash policy where $T = \lfloor \frac{n\alpha'}{\epsilon/4 - \alpha} \rfloor$.*

Proof: For any $i \in \mathcal{N}$,

$$\begin{aligned} &V_i(\hat{\pi}_i^T, \pi_{-i}^{T-1}) - V_i(\pi^{T-1}) \\ &\quad \leq (\hat{V}_i(\hat{\pi}_i^T, \pi_{-i}^{T-1})_H + \epsilon/8) - (\hat{V}_i(\pi^{T-1})_H - \epsilon/8) \\ &\quad = \hat{V}_i(\hat{\pi}_i^T, \pi_{-i}^{T-1})_H - \hat{V}_i(\pi^{T-1})_H + \epsilon/4 \end{aligned} \tag{17}$$

since agents make κ-accurate value estimations, $\kappa = \epsilon/8$. Since at step $t = T$, the algorithm converges as per Corollary 1, we must have $\Delta_i^T = \hat{V}_i(\hat{\pi}_i^T, \pi_{-i}^{T-1})_H - \hat{V}_i(\pi^{T-1})_H \leq \epsilon/2$ for all $i \in \mathcal{N}$. This implies

$$V_i(\hat{\pi}_i^T, \pi_{-i}^{T-1}) - V_i(\pi^{T-1}) \leq 3\epsilon/4 \tag{18}$$

for all $i \in \mathcal{N}$, from (17). Conditioning on $\mathcal{E}_{\approx}$, we get

$$\max_{\pi \in \Pi_i^C(\pi_{-i}^{T-1})} V_i(\pi, \pi_{-i}^{T-1}) - V_i(\hat{\pi}_i^T, \pi_{-i}^{T-1}) \leq \epsilon/4 \tag{19}$$

for all $i \in \mathcal{N}$. Thus,

$$\begin{aligned} \max_{\pi \in \Pi_i^C(\pi_{-i}^{T-1})} V_i(\pi, \pi_{-i}^{T-1}) &\leq \epsilon/4 + V_i(\hat{\pi}_i^T, \pi_{-i}^{T-1}) \\ &\leq \epsilon/4 + 3\epsilon/4 + V_i(\pi^{T-1}) = \epsilon + V_i(\pi^{T-1}), \end{aligned} \tag{20}$$

where we use (18) to obtain the second inequality. Thus π^{T-1} is an ϵ-NE policy. Since at time $t = T$, the policy is not updated, $\pi^T = \pi^{T-1}$. ∎

A. Proof of Main Result

Proof: Define the events $\mathcal{E}_1 := \{\forall i \in \mathcal{N}, \forall t \in [T], |\hat{V}_i(\hat{\pi}_i^t, \pi_{-i}^{t-1})_H - V_i(\hat{\pi}_i^t, \pi_{-i}^{t-1})| \leq \epsilon/8\}$ and $\mathcal{E}_2 := \{\forall i \in \mathcal{N}, \forall t \in [T], |\hat{V}_i(\pi^{t-1})_H - V_i(\pi^{t-1})| \leq \epsilon/8\}$ in which $\hat{\pi}_i^t$ is a solution to $\mathcal{G}_{-i}$ obtained using the PAC learner L_i and $T \leq \frac{4n\alpha'}{\epsilon - 4\alpha}$. Given the conditions on ϵ, M and T, in the following we show that

$$\mathbb{P}[\mathcal{E}_{\approx} \cap \mathcal{E}_1 \cap \mathcal{E}_2] \geq 1 - \delta. \tag{21}$$

where $\mathcal{E}_{\approx}$ is as given in Definition 7. Once we establish the inequality (21), the convergence in T steps follows from Lemma 3.

Claim [$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] \geq 1 - \delta/2$]: For some policy $\pi \in \Pi$ and $\kappa \geq 0$, consider the event,

$$\xi_\kappa := |\hat{V}_j(\pi)_H - V_j(\pi)| \geq \kappa, \quad \forall j \in \mathcal{N}. \tag{22}$$

See that $\hat{V}_j(\pi)_H$ and $\hat{V}_j(\pi)$ (the non-truncated version) are related as $\hat{V}_j(\pi)_H - \hat{V}_j(\pi) \leq \frac{1}{M} M \sum_{t=H+1}^{\infty} \gamma^t = \frac{\gamma^{H+1}}{1-\gamma}$, given normalized rewards and recalling $\hat{V}_j(\pi)$ definition in (3). Thus, we have

$$|\hat{V}_j(\pi)_H - \hat{V}_j(\pi)| \leq \frac{\gamma^{H+1}}{1-\gamma}. \tag{23}$$

See that event ξ_κ implies the event $\chi_\kappa := |\hat{V}_j(\pi) - V_j(\pi)| \geq \kappa - \frac{\gamma^{H+1}}{1-\gamma}, \forall j \in \mathcal{N}$ using (23). That is, $\mathbb{P}\big[\chi_\kappa\big] \geq \mathbb{P}\big[\xi_\kappa\big]$.

Consider the estimator of the value function in the l-th episode $\tilde{V}_j(\pi)^{(l)}$ for some $j \in \mathcal{N}$ and $l \in [M]$, for which we have $\mathbb{E}[\tilde{V}_j(\pi)^{(l)}] = V_j(\pi)$ by definition of the (deterministic) value function in (2). Also, $\tilde{V}_j(\pi)^{(l)}$ for $l \in [M]$ are independent random variables, bounded between $-\frac{1}{1-\gamma}$ and $\frac{1}{1-\gamma}$. Using Hoeffding's inequality, we get

$$\mathbb{P}\big[|\sum_{l=1}^{M} \tilde{V}_j(\pi)^{(l)} - M V_j(\pi)| \geq M\kappa'\big] \leq 2e^{\frac{-2M^2\kappa'^2}{\sum_{l=1}^{M}(2(1-\gamma)^{-1})^2}} \tag{24}$$

where $\kappa' := \kappa - \frac{\gamma^{H+1}}{1-\gamma}$. Thus, plugging in $\kappa = \epsilon/8$ and using $\hat{V}_j(\pi) = \frac{1}{M}\sum_{l=1}^{M} \tilde{V}_j(\pi)^{(l)}$, after some arithmetic, we get

$$\mathbb{P}[\xi_{\epsilon/8}] \leq \mathbb{P}[\chi_{\epsilon/8}] \leq 2e^{-\frac{M}{2}(\frac{\epsilon(1-\gamma)}{8} - \gamma^{H+1})^2} = 2e^{-\frac{M}{128}\epsilon'^2} \tag{25}$$

where we use the assumption $\epsilon > 8\frac{\gamma^{H+1}}{1-\gamma}$, $\mathbb{P}\big[\chi_\kappa\big] \geq \mathbb{P}\big[\xi_\kappa\big]$, and $\epsilon' = \epsilon(1-\gamma) - 8\gamma^{H+1}$. Now, using the union bound and then (25), we obtain the following bound

$$\begin{aligned}\mathbb{P}[\mathcal{E}_1^C \cup \mathcal{E}_2^C] &\leq \mathbb{P}[\mathcal{E}_1^C] + \mathbb{P}[\mathcal{E}_2^C] \\ &\leq \sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{P}\big[|\hat{V}_i(\hat{\pi}_i^t, \pi_{-i}^{t-1})_H - V_i(\hat{\pi}_i^t, \pi_{-i}^{t-1})| \geq \epsilon/8\big] \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{n} \mathbb{P}\big[|\hat{V}_i(\pi^{t-1})_H - V_i(\pi^{t-1})| \geq \epsilon/8\big] \\ &\leq 4nTe^{-\frac{M}{128}\epsilon'^2}. \end{aligned} \tag{26}$$

Using $T \leq \frac{4n\alpha'}{\epsilon - 4\alpha}$ and $M = \lceil \frac{128}{\epsilon'^2} log \frac{32n^2\alpha'}{\delta(\epsilon-4\alpha)} \rceil \geq \frac{128}{\epsilon'^2} log \frac{32n^2\alpha'}{\delta(\epsilon-4\alpha)}$, we get that $\mathbb{P}[\mathcal{E}_1^C \cup \mathcal{E}_2^C] \leq \delta/2$, which further implies that $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] \geq 1 - \delta/2$. Thus with probability at least $1 - \delta/2$ agents make $\epsilon/8$-accurate value estimations during the course of Algorithm 1. At this point, also note that each value estimation consumes MH samples, thus the total number of samples used by the algorithm for value estimation is bounded by $2nTMH \leq \frac{8n^2\alpha' MH}{\epsilon - 4\alpha}$ using the upper bound for T.

Claim [$\mathbb{P}[\mathcal{E}_{\approx}] \geq 1 - \delta/2$]: Using the union bound, we get

$$\mathbb{P}[\mathcal{E}_{\approx}^C] \leq \sum_{i=1}^{n}\sum_{t=1}^{T} \mathbb{P}[\hat{\pi}_i^t \text{ is worse than } \epsilon/4 \text{ optimal.}] \tag{27}$$

because $\mathcal{E}_{\approx}^C = \bigcup_{i,t}\{\hat{\pi}_i^t$ is worse than $\epsilon/4$ optimal.$\}$ Since agents have access to the PAC-learner for solving $\mathcal{G}_{-i}$, agents are guaranteed to get an arbitrary accurate solution ($\hat{\epsilon} > 0$) with arbitrary confidence ($\hat{\delta} \in (0, 1)$), when the adequate number of samples have been used. Plugging in $\hat{\epsilon} = \epsilon/4$ and $\hat{\delta} = \delta'$, agents are guaranteed to get an $\epsilon/4$-optimal solution to $\mathcal{G}_{-i}$ with probability at least $1 - \delta'$ with a sample complexity bounded by $SC_{i,t} := poly(|S|, |\mathcal{A}_i|, \frac{1}{\zeta}, \frac{4}{\epsilon}, \frac{1}{\delta'}, \frac{1}{1-\gamma}, r_{\max})$. This implies that $\mathbb{P}[\mathcal{E}_{\approx}^C] \leq nT\delta'$, with agents consuming $\sum_{i,t} SC_{i,t}$ samples overall. Plugging in $\delta' = \frac{2\delta(\epsilon-4\alpha)}{n^2\alpha'}$, $T = \frac{4n\alpha'}{\epsilon-4\alpha}$, after some arithmetic, we have $\mathbb{P}[\mathcal{E}_{\approx}^C] \leq nT\delta' = \delta/2$ or $\mathbb{P}[\mathcal{E}_{\approx}] \geq 1 - \delta/2$.

Since $\mathcal{E}_{\approx}$ is independent of $\mathcal{E}_1 \cap \mathcal{E}_2$, we have that the probability in (21) is bounded by $(1 - \delta/2)^2$ from which we have the inequality in (21) since $(1 - \delta/2)^2 \geq 1 - \delta$.

Given (21), using lemmas 1, 2 and 3, we guarantee that Algorithm 1 converges in at most T steps to an ϵ-NE policy. Noting that the sum of a finite number of polynomials is a polynomial, we derive the sample complexity in (10) by summing the sample complexity requirement of the PAC solvers with the samples needed for the value estimations obtained by multiplying M value with T bound. ■

V. Conclusion

In this paper, we showed that when the constrained Markov game has an α-potential for finite α, and agents have an efficient PAC learner for the constrained MDP subproblem, the sequential best-response algorithm provably converges to a stationary ϵ-approximate NE policy profile with probability at least $1 - \delta$ in finite time. The sample complexity grows as a polynomial of the reciprocal of specific accuracy and confidence parameters and the reciprocal of the problem dependent Slater's constants.

References

[1] S. Le Cleac'h, M. Schwager, and Z. Manchester, "LUCIDGames: Online unscented inverse dynamic games for adaptive trajectory prediction and planning," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5485–5492, 2021. [Online]. Available: https://arxiv.org/abs/2011.08152

[2] M. Wang, Z. Wang, J. Talbot, J. C. Gerdes, and M. Schwager, "Game-Theoretic planning for self-driving cars in multivehicle competitive scenarios," *IEEE Transactions on Robotics*, 2021.

[3] M. Liu, I. Kolmanovsky, H. E. Tseng, S. Huang, D. Filev, and A. Girard, "Potential game based decision-making frameworks for autonomous driving," *arXiv preprint arXiv:2201.06157*, 2022.

[4] K. Zhang, A. Koppel, H. Zhu, and T. Basar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *SIAM Journal on Control and Optimization*, vol. 58, no. 6, pp. 3586–3612, 2020.

[5] S. Cayci, N. He, and R. Srikant, "Linear convergence of entropy-regularized natural policy gradient with linear function approximation," *arXiv preprint arXiv:2106.04096*, 2021.

[6] D. Ding, C.-Y. Wei, K. Zhang, and M. Jovanovic, "Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence," in *International Conference on Machine Learning*. PMLR, 2022, pp. 5166–5220.

[7] C. Daskalakis, D. J. Foster, and N. Golowich, "Independent policy gradient methods for competitive reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5527–5540, 2020.

[8] M. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar, "Decentralized Q-learning in zero-sum Markov games," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 320–18 334, 2021.

[9] A. S. Nowak, "Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space," *Journal of Optimization Theory and Applications*, vol. 45, pp. 591–602, 1985.

[10] E. Altman and A. Shwartz, "Constrained Markov games: Nash equilibria," in *Advances in Dynamic Games and Applications*. Springer, 2000, pp. 213–221.

[11] F. Dufour and T. Prieto-Rumeau, "Stationary Markov Nash equilibria for nonzero-sum constrained ARAT Markov games," *SIAM Journal on Control and Optimization*, vol. 60, no. 2, pp. 945–967, 2022.

[12] A. Jaśkiewicz and A. S. Nowak, "On approximate and weak correlated equilibria in constrained discounted stochastic games," *Applied Mathematics & Optimization*, vol. 87, no. 2, p. 23, 2023.

[13] X. Guo, X. Li, C. Maheshwari, S. Sastry, and M. Wu, "Markov α-potential games: Equilibrium approximation and regret analysis," *arXiv preprint arXiv:2305.12553*, 2023.

[14] X. Guo, X. Li, and Y. Zhang, "An α-potential game framework for n-player games," *arXiv preprint arXiv:2403.16962*, 2024.

[15] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras, "Global convergence of multi-agent policy gradient in Markov potential games," *arXiv preprint arXiv:2106.01969*, 2021.

[16] S. Vaswani, L. Yang, and C. Szepesvári, "Near-optimal sample complexity bounds for constrained MDPs," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3110–3122, 2022.

[17] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal, "Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 3682–3689.

[18] H. Wei, X. Liu, and L. Ying, "A provably-efficient model-free algorithm for constrained Markov decision processes," *arXiv preprint arXiv:2106.01577*, 2021.

[19] P. Alatur, G. Ramponi, N. He, and A. Krause, "Provably learning Nash policies in constrained Markov potential games," *arXiv preprint arXiv:2306.07749*, 2023.

[20] Z. Song, S. Mei, and Y. Bai, "When can we learn general-sum Markov games with a large number of players sample-efficiently?" *arXiv preprint arXiv:2110.04184*, 2021.