Learning Nash Equilibrium of Markov Potential Games with a Shared Constraint via Primal-Dual Optimization

Songtao Feng[1], **Michael Dorothy**[2], **Jie Fu**[1]
[1]University of Florida, FL, USA
[2]DEVCOM Army Research Laboratory, MD, USA
sfeng1@ufl.edu, michael.r.dorothy.civ@army.mil, fujie@ufl.edu

Abstract

The problem of constrained Markov game has recently attracted interests in the study of multi-agent reinforcement learning (MARL). The existing literature has focused on safe MARL problems where safety constraints are imposed for each agent individually. In this work, we consider Markov potential game (MPG) with a shared constraint, where the cost function with respect to the constraint depends on states and joint actions of all agents. We adopt a primal-dual framework to tackle the problem and establish the Slater condition to ensure the strong duality. Moreover, we propose our primal-dual learning algorithm for learning approximate Nash equilibrium in MPG with shared constraint. Thanks to the novel design of the dual update, we provide asymptotic convergence on the weighted output policy. Specifically, we prove that both the value function gap and the constraint violation of the output policy converge at the rate $\mathcal{O}(\epsilon+1/\sqrt{T})$, where ϵ is the accuracy level of the primal update, and T is the number of iterations. We further show that the weighted output policy outperforms the existing uniformly chosen policy.

Introduction

The problem of multi-agent reinforcement learning (MARL) focuses on analyzing the strategic interactions among multiple players in a shared dynamic environment, where reward and state transitions jointly depend on all players' actions. The goal of MARL is to find Nash equilibrium (NE), under which no player can benefit from a unilateral deviation. It has broad applications in practical scenarios, including autonomous driving (Shalev-Shwartz, Shammah, and Shashua 2016), traffic signal control (K.J., A.N, and Bhatnagar 2014), and market pricing (Kononen and Oja 2004).

Although MARL has been extensively studied, its applicability is limited because in many real-life applications, agents are subject to certain constraints, such as anti-jamming system where jamming signal is subject to average power constraint (Hanawal, Abdel-Rahman, and Krunz 2016). In those applications, agents aim to maximize their rewards while ensuring their joint policy satisfies the constraints, which motivates the formulation of constrained Markov games (Altman and Shwartz 2000). In the existing constrained Markov game literature, most works considered individual cost functions for each agent, which depends on the state and the action associated with that agent (instead of the joint actions). There is limited work on Markov games with shared constraints (Jordan, Barakat, and He 2024) where cost functions of the constraints depend on the joint state and joint action space.

In general, finding NE in MARL turns out to be PPAD-complete (Deng et al. 2022; Jin, Muthukumar, and Sidford 2022), and an alternative goal of finding correlated equilibrium (CE) or coarse correlated equilibrium (CCE) has been widely adopted (Liu et al. 2021; Mao and Başar 2022; Song, Mei, and Bai 2022; Jin et al. 2022; Mao et al. 2022; Daskalakis, Golowich, and Zhang 2023; Wang et al. 2023). Finding NE for a special class of Markov games, called Markov potential games (MPGs), has been studied in Macua, Zazo, and Zazo (2018); Leonardos et al. (2022); Zhang, Ren, and Li (2022); Ding et al. (2022); Mao et al. (2022); Zhang et al. (2024); Maheshwari et al. (2023); Narasimha et al. (2022); Perolat et al. (2017); Guo et al. (2024); Zhou et al. (2023). In MPGs, it is assumed that there exists an underlying potential function that can measure the change of the expected reward of any player caused by deviating from his own policy. If such a potential function is publicly available, then an NE policy can be simply obtained by finding the global optimum of the potential function. Even if the potential function is unknown, such structural assumption on the Markov game model can be beneficial in the algorithm design. Moreover, the MPGs cover a wide range of mixed cooperative and competitive Markov games, including identical interest Markov games and congestion games.

The concept of constrained MPGs has been recently introduced in Alatur et al. (2024), and they proposed a coordinate accent algorithm for finding an NE. Later Jordan, Barakat, and He (2024) proposed an independent proximal-policy algorithm to relax the coordination requirement between agents. A more natural duality approach has been shown to fail when the strong duality does not hold (Alatur et al. 2024; Jordan, Barakat, and He 2024). In this work, we consider the problem of finding NE in MPGs with a shared constraint, and re-investigate the primal-dual approach. Our contributions are summarized below.

First, we investigate the Lagrangian function in MPGs with a shared constraint, and introduce the Slater-type condition to establish the strong duality. Moreover, a sample-

efficient algorithm for checking Slater condition is proposed based on a reduction of the original MPG with a shared constraint to an unconstrained two-player zero-sum Markov game. At the same time, the algorithm also outputs the safe margin defined in the Slater condition, which is of great importance in the analysis of our subsequent primal-dual based MARL algorithm.

Second, we propose our primal-dual algorithm for finding NE in MPGs with a shared constraint. The primal update calls for an oracle for learning an approximate ϵ-NE in an unconstrained MPG, which can be solved efficiently by existing algorithms including independent and decentralized learning in MPGs (Maheshwari et al. 2023). For the dual update, we propose two variants of dual projected descent, which can lead to weighted mixed output policy.

Third, we show that both the value function gap and the constraint violation of our output policies under two different dual update rules converge at the rate of $\widetilde{\mathcal{O}}(\epsilon + 1/\sqrt{T})$, where ϵ is the accuracy in the primal update and T is the number of iterations. We observe that the weighted mixed output policy achieves a better performance under compared to the state-of-the-art uniformly chosen policy.

Related Work

Markov potential games. Macua, Zazo, and Zazo (2018) introduced MPGs as an extension of normal form potential games (Monderer and Shapley 1996) and state-based potential games (Marden 2012). Leonardos et al. (2022) proposed independent stochastic policy gradient methods and Zhang, Ren, and Li (2022) proposed a model-based algorithm, and they both achieve $\mathcal{O}(\epsilon^{-6})$ to reach the ϵ-approximate NE. Maheshwari et al. (2023) proposed two timescale algorithm for independent and decentralized MPGs where players do not have knowledge of model and cannot coordinate. Narasimha et al. (2022) provided structural assumptions for verifying whether a Markov game is an MPG and proposed several algorithms for solving MPGs. When coordination between players is available, Song, Mei, and Bai (2022) proposed a coordinate ascent algorithm for achieving $\tilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity. Guo et al. (2024) introduced a new class of α-MPGs generalizing the MPGs by allowing its potential function to deviate at most α. Zhou et al. (2023) recently introduced a class of networked MPGs and proposed localized actor-critic algorithm with linear function approximation.

Constrained Markov games and constrained MPGs. There has been a thrust in MARL with safety constraints in practice (ElSayed-Aly et al. 2021; Gu et al. 2023). Constrained Markov games were first introduced in Altman and Shwartz (2000), and they provide sufficient conditions for the existence of NE. Chen, Ma, and Zhou (2022) introduced the notion of CE in constrained Markov games and proposed a primal-dual algorithm for finding CE. Ding et al. (2023) developed an upper confidence learning algorithm and established the regret guarantees for two-player zero-sum constrained Markov games. Constrained MPGs were first studied in Alatur et al. (2024), and they proposed a coordinate ascent algorithm for finding NE. Inspired by the recent progress in nonconvex optimization under nonconvex constraints, Jordan, Barakat, and He (2024) removed the requirement of coordination and proposed an independent learning algorithm for finding NE in constrained MPGs. Our work is closely related to Jordan, Barakat, and He (2024) and we re-investigate the primal-dual approach for constraint MPGs by introducing the Slater condition assumption and an algorithm for checking such condition. Based on the Slater condition, we further propose the primal-dual algorithm for finding NE in constrained MPGs.

Preliminaries

Notations. Denote $[N]$ as the set of integers ranging from 1 to N. We also define function $(a)_+ = \max\{0, a\}$. For $i \in [N]$, we define $-i = [N] \setminus \{i\}$.

MPGs with a shared constraint. Consider an N-agent MPG with a shared constraint $\mathcal{M} = (N, \mathcal{S}, \mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_N, P, \{r_i\}_{i=1}^N, (c, \alpha), \gamma, \rho)$, where N is the number of agents, $\mathcal{S}$ is the (joint) state space, $\mathcal{A}_i$ is the action space for agent i, $P(s'|s, a)$ is the transition probability from state s to state s' upon taking a joint action $a \in \mathcal{A}$, $r_i : \mathcal{S} \times \mathcal{A} \to \mathbf{R}$ is the reward function for agent i, $c : \mathcal{S} \times \mathcal{A} \to \mathbf{R}$ is the shared cost function and α is the threshold, γ is the discount factor, and ρ is the initial state distribution. We assume that the reward and cost functions are bounded, i.e., $0 \le r_i(s, a) \le R$, $0 \le c(s, a) \le C$ for any valid i, s, a.

Policy and value function. Each agent i decides its own policy $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ and the set of policies for agent i is denoted as Π_i. A joint policy π is a collection of individual agents' policies, $\pi = \{\pi_1, \ldots, \pi_n\}$.

Given a joint policy π of N agent and a state s, the value function of i-th agent is defined as: For all $s \in \mathcal{S}$,

$$V_{r_i}^{\pi}(s) = \mathbb{E}_{(s_t, a_t) \sim (P, \pi)} \left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \middle| s_0 = s \right],$$

where (P, π) specifies the dynamics of the induced stochastic process $\{(s_t, a_t), t \in \mathbb{N}\}$. Similarly, given a joint policy π of N agents and a state-action pair (s, a), the Q-value function of i-th agent is defined as

$$Q_{r_i}^{\pi}(s, a) = \mathbb{E}_{(s_t, a_t) \sim (P, \pi)} \left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \middle| (s_0, a_0) = (s, a) \right].$$

With a little abuse of notation, we use $V_{r_i}(\pi)$ to represent the corresponding value function when the initial state follows from distribution ρ. That is,

$$V_{r_i}(\pi) = \mathbb{E}_{s \sim \rho} \left[V_{r_i}^{\pi}(s) \right].$$

The value functions $V_c(\cdot), Q_c(\cdot)$ given the cost $c(\cdot)$ and a joint policy π are defined analogously. Define **permissible policy set** as $\Pi \subset \otimes_i \Pi_i$, where for any permissible policy $\pi \in \Pi$, it holds that $V_c(\pi) \le \alpha$. For ease of exposition, let $R_{\max} \ge \{\max_{\pi, i} V_{r_i}(\pi)\}$, $C_{\max} \ge \{\max_{\pi} V_c(\pi), \alpha\}$ be upper bounds on the total discounted reward and cost, respectively. Proper choices can be $R_{\max} = \frac{1}{1-\gamma} R$ and $C_{\max} = \max\{\alpha, \frac{1}{1-\gamma} C\}$.

For an MPG, there exists an underlying potential function $\phi : \Pi \to \mathbb{R}$ such that $\forall \pi_i, \pi_i' \in \Pi_i, \forall \pi_{-i} \in \Pi \backslash \Pi_i, \forall i \in [N]$,

$$V_{r_i}(\pi_i, \pi_{-i}) - V_{r_i}(\pi_i', \pi_{-i}) = \phi(\pi_i, \pi_{-i}) - \phi(\pi_i', \pi_{-i}).$$

We remark that the underlying potential function is not revealed in the learning process.

Properties of MPGs with a shared constraint. It is well-known that for unconstrained Markov potential games, the maximizer of the potential function has to be an NE policy. It remains valid for Markov potential games with shared constraints, as summarized below.

Lemma 1. Let ϕ be a potential function for an MPG with shared constraints, and $\pi^* \in \arg\max_{\pi\in\Pi} \phi(\pi)$, where Π is the permissible policy set. Then π^* has to be an NE policy for such MPG with shared constraints problem.

Proof. The proof is by contradiction, which is essentially the same as that in MPG. Suppose π^* is not an NE policy. Then, there exists an agent i's policy π_i such that $V_{r_i}(\pi_i, \pi^*_{-i}) > V_{r_i}(\pi^*_i, \pi^*_{-i})$. Therefore, we have

$$\phi(\pi_i, \pi^*_{-i}) - \phi(\pi^*) = V_{r_i}(\pi_i, \pi^*_{-i}) - V_{r_i}(\pi^*_i, \pi^*_{-i}) > 0,$$

which contradicts to the assumption that $\pi^*_i \in \arg\max_{\pi\in\Pi} \phi(\pi)$. □

Nash equilibrium and Learning objective. Recall the definition of permissible policy set Π under which any permissible policy $\pi \in \Pi$ satisfies the safety constraint $V_c(\pi) \le \alpha$. For constrained MPGs, a permissible joint policy π is an Nash equilibrium (NE) if it satisfies the following two properties: First, a joint policy π is a product policy where each agent takes independent action at each decision state; Second, for any agent $i \in [N]$ and any policy π'_i of i-th agent, it holds that $V_{r_i}(\pi) \ge V_{r_i}(\pi'_i, \pi_{-i})$ where π_{-i} denotes the joint policy of all agents' policies excluding agent i. In other words, for an NE policy $\pi \in \Pi$, no agent can improve its value function without violating the shared constraint by deviating from its own policy.

For ease of exposition, we use $\pi^\dagger_i(\pi_{-i})$ to denote the best response of the i-th agent when other agents' joint policy is π_{-i} under constrained MPGs. Here policy $(\pi^\dagger_i(\pi_{-i}), \pi_{-i})$ still satisfies the safety constraint. A joint permissible policy $\pi = (\pi_1, \ldots, \pi_N) \in \Pi$ is an ϵ-approximate NE policy if for all $i \in [N]$

$$V_{r_i}(\pi^\dagger(\pi_{-i}), \pi_{-i}) - V_{r_i}(\pi) \le \epsilon.$$

Our goal is to design sample efficient algorithm for learning ϵ-approximate Nash equilibrium (NE) of the Markov potential games with a shared constraint.

Strong Duality for Markov Potential Games with a Shared Constraint

In this section, we establish the strong duality for MPGs with a shared constraint by first introducing the Slater-type condition in this context, and then proposing an algorithm for checking if such a condition is met for a given MPG with a shared constraint.

Since a policy maximizes the potential function is an NE policy (Lemma 1), we begin with our original optimization problem

$$\max_{\pi} \phi(\pi), \quad \text{s.t.} \quad V_c(\pi) \le \alpha.$$

Consider the equivalent Lagrangian primal problem

$$\max_{\pi} \min_{\lambda\ge 0} L(\pi, \lambda),$$

where the Lagrangian function $L(\pi, \lambda)$ is defines as

$$L(\pi, \lambda) = \phi(\pi) + \lambda(\alpha - V_c(\pi)).$$

The Lagrangian dual problem is defined as

$$\min_{\lambda\ge 0} \max_{\pi\in\Pi} L(\pi, \lambda). \tag{1}$$

By the minimax inequality, weak duality can thus be established, which claims that the optimal value of the primal problems upper bounds the optimal value of the dual problems.

We first show that the inner optimization problem of the dual problem (1) is equivalent to optimizing the potential function of an unconstrained MPG with certain modified rewards.

Lemma 2. Given $\lambda \ge 0$, define $\widetilde{\phi}(\pi) = \phi(\pi) + \lambda(\alpha - V_c(\pi))$. Then the inner optimization problem of (1) is equivalent to finding a policy $\widehat{\pi} \in \arg\max_\pi \widetilde{\phi}(\pi)$ under the unconstrained MPGs with potential function $\widetilde{\phi}(\cdot)$ and modified reward functions: for any $i \in [N]$, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\widetilde{r}_i(s, a) = r_i(s, a) - \lambda c(s, a)$.

Proof. Fix dual variable λ. For any policy $\pi = (\pi_i, \pi_{-i})$ and $\widehat{\pi}_i$, The difference between $\widetilde{\phi}(\widehat{\pi}_i, \pi_{-i})$ and $\widetilde{\phi}(\pi)$ is

$$\begin{aligned}
&\widetilde{\phi}(\widehat{\pi}_i, \pi_{-i}) - \widetilde{\phi}(\pi) \\
&= \phi(\widehat{\pi}_i, \pi_{-i}) - \phi(\pi) - \lambda(V_c(\widehat{\pi}_i, \pi_{-i}) - V_c(\pi)) \\
&= V_{r_i}(\widehat{\pi}_i, \pi_{-i}) - V_{r_i}(\pi) - \lambda(V_c(\widehat{\pi}_i, \pi_{-i}) - V_c(\pi)) \\
&= V_{r_i - \lambda c}(\pi_i, \pi_{-i}) - V_{r_i - \lambda c}(\pi),
\end{aligned}$$

which is exactly $V_{\widetilde{r}_i}(\pi_i, \pi_{-i}) - V_{\widetilde{r}_i}(\pi)$. □

We will make use of this property in our primal-dual algorithm, as elaborated in the next section.

Next, we introduce the strong Slater condition, which is the key to establish the strong duality result.

Assumption 1 (Slater Condition)**.** For any agent i and any joint policy $\pi = (\pi_i, \pi_{-i})$, there exist a stochastic policy $\widetilde{\pi}_i$ and positive real number u such that $V_c(\widetilde{\pi}_i, \pi_{-i}) \le \alpha - u$.

The Slater condition assumes that any agent i can always find a strictly feasible policy given a fixed policy π_{-i} for all other agents. With the above Slater condition, the strong duality can thus be established.

Theorem 1 (Strong Duality)**.** Let Assumption 1 hold, then the strong duality holds for Markov potential games with a shared constraint.

Next, we first propose Algorithm 1 for checking Slater condition in a given MPG with a shared constraint. Recall product policy set $\otimes_{i\in[N]}\Pi_i$, and we define $\Pi_{-i} = \otimes_{i\in[N]\setminus\{i\}}\Pi_i$. Assume we have access to an oracle for computing the NE policy $(\pi_i, \pi_{-i}) \in (\Pi_i, \Pi_{-i})$, which can be solved efficiently (Feng et al. 2024). We remark that finding NE in the two-player zero-sum Markov games can be

Algorithm 1: Algorithm for Checking Slater Condition

1: Input: shared constraint threshold α, accuracy $\epsilon' > 0$.
2: **for** $i \leftarrow 1, \ldots, N$ **do**
3: Find an ϵ'-approximate NE policy $\pi^i = (\pi_i^i, \pi_{-i}^i)$ under the unconstrained zero-sum Markov game

$$V_c(\pi_i^i, \pi_{-i}^+(\pi_i^i)) - V_c(\pi_i^+(\pi_{-i}^i), \pi_{-i}^i) < \epsilon'.$$

4: $\widetilde{u}_i = \alpha - V_c(\pi_i^i, \pi_{-i}^i) - \epsilon'$.
5: **end for**
6: Return: $\widetilde{u} := \min_{i \in [N]} \widetilde{u}_i$.

solved efficiently despite such algorithms do not exist under multi-player general-sum Markov games.

Algorithm description. For clarity, we use $\pi_i^+(\pi_{-i})$ to denote the best response of agent i under unconstained MPGs. For each agent i, we view all other agents $-i$ as a combined single agent, and consider the following unconstrained two-player zero-sum Markov games: The combined single agent $-i$ serves as the max-player, and plays against the min-player agent i. That is, the agent i aims to minimize V_c while combined agents $-i$ aim to maximize V_c. The reward specified in the algorithm is the joint cost function c of the original MPGs with a shared constraint. The newly defined unconstrained two-player zero-sum Markov game focuses on finding the i-th agent "duality gap" with respect to the original cost function (rather than the original reward function). The next lemma establishes the relationship between Algorithm 1 and the Slater condition (Assumption 1).

Lemma 3. For any policy $\pi_{-i} \in \Pi_{-i}$, agent i's policy π_i^i satisfies the inequality $V_c(\pi_i^i, \pi_{-i}) \leq \alpha$ if $\widetilde{u}_i > 0$. Further, the gap $\alpha - V_c(\pi_i^i, \pi_{-i})$ is upper bounded by $\widetilde{u}_i$.

Proof. Since $\pi^i = (\pi_i^i, \pi_{-i}^i)$ is an ϵ'-approximate NE of the unconstrained zero-sum Markov game specified in line 3 and the combined single-player $-i$ plays as the max-player, $V_c(\pi_i^i, \pi_{-i}) \leq V_c(\pi_i^i, \pi_{-i}^i) + \epsilon'$ for any joint policy π_{-i}. If $\widetilde{u}_i > 0$, it holds that

$$V_c(\pi_i^i, \pi_{-i}) \leq V_c(\pi_i^i, \pi_{-i}^i) + \epsilon' \leq \alpha,$$

where the last inequality follows from line 4 and $\widetilde{u}_i > 0$.

The gap $\alpha - V_c(\pi_i^i, \pi_{-i})$ can be bounded as follows

$$\alpha - V_c(\pi_i^i, \pi_{-i}) \geq \alpha - V_c(\pi_i^i, \pi_{-i}^i) - \epsilon' = \widetilde{u}_i,$$

where the first inequality holds since π^i is an ϵ'-approximate NE policy, and the last equality follows from the definition of $\widetilde{u}_i$ in line 4. □

Thanks to Lemma 3, if all those gaps $\widetilde{u}_i$ are positive (line 6), Slater condition (Assumption 1) holds and $\widetilde{u}$ serves as the duality gap. If $\widetilde{u} \leq 0$, then we say that the original problem *may* not satisfy the Slater condition (Assumption 1). We remark that obtaining a negative $\widetilde{u}$ does not necessarily mean that the Slater condition fails, since raising the accuracy level (selecting a smaller ϵ') could result in a positive $\widetilde{u}_i$.

Algorithm 2: Primal-dual Algorithm for finding NE in MPGs with a shared constraint

1: Initialization: $\lambda_t = 0$, $\rho_t = \log(t+1)$, $\eta = \eta_t = 1/\sqrt{T}$, $\xi = 2HR_{\max}/u$ where u is the duality gap.
2: **for** $t = 0, \ldots, T-1$ **do**
3: primal update: solve the unconstrained MPGs with reward $r_i - \lambda_t c$ to obtain an approximate NE π_t

$$\max_{\pi_i'} V_{r_i - \lambda c}(\pi_i', \pi_{-i,t}) - V_{r_i - \lambda c}(\pi_t) \leq \epsilon, \quad \forall i \quad (2)$$

4: Dual update: update λ_{t+1} via projected dual descent

$$\lambda_{t+1} = \mathrm{Proj}_{[0,\xi]}\left(\left(\rho_t \lambda_t - \eta \nabla_\lambda L(\pi_t, \lambda_t)\right)/\rho_{t+1}\right) \quad (3)$$

$$\text{or} \quad \lambda_{t+1} = \mathrm{Proj}_{[0,\xi]}\left(\lambda_t - \eta_t \nabla_\lambda L(\pi_t, \lambda_t)\right) \quad (4)$$

5: **end for**
6: Output: $\widetilde{\pi}_{t'}$ where t' is sampled from $\{0, 1, \ldots, T-1\}$ with weight $\{\rho_0, \rho_1, \ldots, \rho_{T-1}\}$ for (3) and $\{\eta_0, \eta_1, \ldots, \eta_{T-1}\}$ for (4).

Primal-dual Based Algorithm

In this section, we propose a primal-dual algorithm for learning ϵ-approximate NE in MPG with a shared constraint.

Our algorithm builds upon the primal-dual algorithm for constrained Markov games (Chen, Ma, and Zhou 2022) and features a special design of the dual variable update. Specifically, we introduce two variants of projected dual descent, and the resulting output policies are weighted mixed policies, which outperforms the uniform policy in certain circumstances.

Primal Update: Oracle for finding NE in unconstrained MPGs. For the primal update (line 3), instead of conducting projected primal ascent step to update primal variable, we seek to find the optimal primal variable π_t given fixed dual variable λ_t. Thanks to Lemma 2, such step is essentially equivalent to finding ϵ-approximate NE policy for certain unconstrained MPG (line 3). We employ an oracle (any existing algorithm) for finding an ϵ-NE policy in the unconstrained MPG (e.g. Maheshwari et al. (2023), Song, Mei, and Bai (2022)) with reward $r_i - \lambda_t c$.

Projected dual descent. Dual variable λ_t is updated via projected dual descent with fixed primal variable π_t. In dual update rule (3) and (4), $\nabla_\lambda L(\pi_t, \lambda_t) = \alpha - V_c(\pi)$. Note that the projection set $[0, \xi]$ must contain the optimal dual variable. Instead of the vanilla dual update $\lambda_{t+1} = \mathrm{Proj}_{[0,\xi]}\left(\lambda_t - \eta \nabla_\lambda L(\pi_t, \lambda_t)\right)$ where η is the parameter to be optimized, we consider two different dual update rules (3) and (4). Note first that both can be reduced to the vanilla dual update with appropriate choices of parameters ρ_t, η and η_t. The benefits of our new update rules lies in the versatility. By proper choices of parameters ρ_t, η and η_t, the resulting mixed output policy weighs more on the recent iterations while their mixed output policy uniformly sample from all iterations.

Compare with Chen, Ma, and Zhou (2022). Our algorithm is inspired by the primal-dual algorithm therein with the following differences. First, we consider the MPGs with

a shared constraint while they study the Markov games. Despite their model is more general, they focus on finding an approximate CE rather than NE (as efficient algorithms for finding NE does not exist). Thanks to the special structural assumption on the MPGs, we are able to find an approximate NE. Second, Chen, Ma, and Zhou (2022) identified the Slater-type condition in constrained Markov game, while in our work, besides identifying the Slater-type condition in constrained Markov potential games, we also propose an efficient algorithm for checking Slater condition. Third, we provide two variants of dual projected descent update rules, and the resulting mixed output policy weighs more on the recent iterations while their mixed output policy uniformly sample from all iterations. Our convergence results suggest weighted policy can outperform the uniformly chosen policy and can be potentially used to improve existing result.

Theoretical Guarantees

In this section, we provide the theoretical guarantees for our proposed primal-dual algorithm and their proof sketches.

Define value function gap of i-th agent as

$$D_i(\pi) = \max_{\text{permissible } \widetilde{\pi}_i} V_{r_i}(\widetilde{\pi}_i, \pi_{-i}) - V_{r_i}(\pi).$$

Here permissible policy set contains all policies that satisfy the shared constraint.

The performance of our algorithm is evaluated by both the value function gap and the constraint violation, and the theoretical guarantee is summarized as follows.

Theorem 2. Let Assumption 1 holds. Run algorithm 2 for T iterations with dual update rule (3), the output policy $\widetilde{\pi}_t$ is sampled from iterations $[1, 2, \ldots, T-1]$ with increasing weight $[\rho_0, \rho_1, \ldots, \rho_{T-1}]$ and its value function gap and the constraint violation satisfy

$$\mathbb{E}[D_i(\widetilde{\pi}_t)] \le \epsilon + 4C_{\max}^2 \cdot \frac{\eta T}{\sum_{t=0}^{T-1} \rho_t}, \quad \forall i \in [N],$$

$$\mathbb{E}[(V_c(\widetilde{\pi}_t) - \alpha)_+] \le \frac{1}{\left(\sum_{t=0}^{T-1} \rho_t\right)} \cdot \Big[\frac{2\epsilon}{\xi} \sum_{t=0}^{T-1} \rho_t + \frac{\xi}{\eta}\left(\sum_{t=0}^{T} \|\rho_t - \rho_{t+1}\|^2 + \|\rho_0\|^2\right) + \frac{4T\eta C_{\max}^2}{\xi} + 2C_{\max}(\rho_T - \rho_0) + \frac{4\xi}{\eta} \sum_{t=0}^{T-1} \rho_t(\rho_{t+1} - \rho_t)\Big].$$

Similarly, Run algorithm 2 for T iterations with dual update rule (4), the output policy $\widetilde{\pi}_t$ is sampled from iterations $[1, 2, \ldots, T-1]$ with weight $[\eta_0, \eta_1, \ldots, \eta_{T-1}]$ and its value function gap and the constraint violation satisfy

$$\mathbb{E}[D(\widetilde{\pi}_T)] = \epsilon + 2C_{\max}^2 \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t},$$

$$\mathbb{E}[(V_c(\widetilde{\pi}_t) - \alpha)_+] \le \frac{2\epsilon}{\xi} + \frac{1}{\sum_{t=0}^{T-1} \eta_t}\left(\xi + \frac{4C_{\max}^2}{\xi} \sum_{t=0}^{T-1} \eta_t^2\right).$$

The above theorem establishes the performance guarantee of Algorithm 2 for both value function gap and the constraint violation associated with the output policy $\widetilde{\pi}_t$. Note that the dual update reduces to standard projected dual descent when $\rho_t = 1$ under (3) or $\eta_t = \eta$ under (4), and the corresponding value gap and the constraint violation is of order $\widetilde{O}(\epsilon + 1/\sqrt{T})$. In the following, we show that by properly selecting parameters in the update rules (3) and (4), the overall value function gap and the constraint violation remain the same.

Theorem 3. Under the same assumption as in Theorem 2, if we select $\rho_t = \log(t+1)$ and $\eta = 1/\sqrt{T}$ in (3) which leads to a weighted policy, the value gap and the constraint violation are of order

$$\mathbb{E}[D_i(\widetilde{\pi}_t)] \le \mathcal{O}\left(\epsilon + \frac{C_{\max}^2}{\sqrt{T}\log T}\right),$$

$$\mathbb{E}[(V_c(\widetilde{\pi}_t) - \alpha)_+] \le \mathcal{O}\left(\frac{\epsilon}{\xi} + \frac{\xi + C_{\max}^2/\xi}{\sqrt{T}\log T} + \frac{C_{\max}}{T}\right).$$

Similarly, if we select $\eta_t = 1/\sqrt{T}$ in (4) which leads to a uniformly chosen policy, the value gap and the constraint violation are of order

$$\mathbb{E}[D_i(\widetilde{\pi}_t)] \le \mathcal{O}\left(\epsilon + \frac{C_{\max}^2}{\sqrt{T}}\right),$$

$$\mathbb{E}[(V_c(\widetilde{\pi}_t) - \alpha)_+] \le \mathcal{O}\left(\frac{\epsilon}{\xi} + \frac{\xi}{\sqrt{T}} + \frac{C_{\max}^2}{\xi\sqrt{T}}\right).$$

Benefit of weighted policy. Note that both the value function gap and the constraint violation under the weighted policy have faster convergence rate compared to those under the uniformly chosen policy. Further, both of our dual update rule (3) and (4) may render better weighted policies by properly selecting parameters η_t, ρ_t, and we leave it as future works.

Compare with Jordan, Barakat, and He (2024). If we choose Nash-CA (Song, Mei, and Bai 2022) as our primal update oracle, then the total sample complexity of our algorithm is $O(\epsilon^{-5})$ sample complexity, which coincides with the best existing algorithm proposed in Jordan, Barakat, and He (2024). Our algorithm follows the primal-dual approach while their policy update rule guarantees that the sampling policies are always feasible. Moreover, their algorithm require the initial policy to be feasible while our algorithm does not need such assumption. Further, the lower bound for Markov games is $\Omega(\epsilon^2)$, which also serves the lower bound for the constrained Markov potential games. Both algorithms suffer $O(\epsilon^{-2})$ gap, and it is open whether this gap can be narrowed.

Proof Sketch of Theorem 2

We provide the proof sketch for the uniformly chosen policy under the dual update rule (3) and the proof for the weighted policy under the dual update rule (4) is similar. We focus on bounding the value function gap in step I-II, and the performance of the constraint violation is analyzed in step III-V.

Step I: First, we bound the value difference between policy $(\pi_{i,t}^*, \pi_{-i,t})$ and policy π_t, where $\pi_{i,t}^* = \pi_i^\dagger(\pi_{-i,t})$ is player i's best response policy when other agents choose policy $\pi_{-i,t}$ under the shared constraint.

Lemma 4. Let π_t be the approximate NE policy of the primal update in the t-th iteration and $\pi^*_{i,t} = \pi^\dagger_i(\pi_{-i,t})$. Then, we have

$$\sum_{t=0}^{T-1} \rho_t \lambda_t (V_c(\pi^*_{i,t}, \pi_{-i,t}) - V_c(\pi_t)) \le 2\eta T C^2_{\max}. \quad (5)$$

Proof. Note that

$$\begin{aligned}
0 &\ge -\|\rho_T \lambda_T\|^2 \\
&\overset{(i)}{=} \sum_{t=0}^{T-1} \left(\|\rho_t \lambda_t\|^2 - \|\rho_{t+1}\lambda_{t+1}\|^2 \right) \\
&\overset{(ii)}{\ge} \sum_{t=0}^{T-1} \left(\|\rho_t \lambda_t\|^2 - \|\rho_t \lambda_t - \eta(\alpha - V_c(\pi_t))\|^2 \right) \\
&\overset{(iii)}{\ge} -\eta^2 \sum_{t=0}^{T-1} (\|\alpha\| + \|V_c(\pi_t)\|)^2 + 2\eta \sum_{t=0}^{T-1} \rho_t \lambda_t (\alpha - V_c(\pi_t)) \\
&\overset{(iv)}{\ge} -4\eta^2 T C^2_{\max} + 2\eta \sum_{t=0}^{T-1} \rho_t \lambda_t (\alpha - V_c(\pi_t)) \\
&\overset{(v)}{\ge} -4\eta^2 T C^2_{\max} + 2\eta \sum_{t=0}^{T-1} \rho_t \lambda_t (V_c(\pi^*_{i,t}, \pi_{-i,t}) - V_c(\pi_t)),
\end{aligned}$$

where (i) uses the fact that $\lambda_0 = 0$, (ii) follows from the fact that $0 \in [0, \xi]$ and the nonexpansiveness of projection, (iii) follows from triangle inequality, (iv) uses the bound $C_{\max} \ge \max\{\alpha, \max_\pi V_c(\pi)\}$, and (v) is based on the fact that $(\pi^*_{i,t}, \pi_{-i,t})$ is a feasible policy. Rearranging the above inequality gives the desired inequality. □

Step II: We establish the value function gap below.

Lemma 5. Let $D_i(\pi_t)$ be the value function gap under t-th iteration policy π_t. Then, it holds that

$$\sum_{t=0}^{T-1} \rho_t D_i(\pi_t) \le \epsilon \sum_{t=0}^{T-1} \rho_t + 2\eta T C^2_{\max}.$$

Proof. Note that

$$\begin{aligned}
0 &\le \sum_{t=0}^{T-1} \rho_t \left(\max_{\widetilde{\pi}_{i,t}} L((\widetilde{\pi}_{i,t}, \pi_{-i,t}), \lambda_t) - L((\pi^*_{i,t}, \pi_{-i,t}), \lambda_t) \right) \\
&\le \sum_{t=0}^{T-1} \rho_t \left(\max_{\widetilde{\pi}_{i,t}} V_{r_i - \lambda_t c}((\widetilde{\pi}_{i,t}, \pi_{-i,t}), \lambda_t) \right. \\
&\qquad \left. - V_{r_i - \lambda_t c}((\pi^*_{i,t}, \pi_{-i,t}), \lambda_t) \right) \\
&\overset{(i)}{\le} \sum_{t=0}^{T-1} \rho_t \left(\epsilon + V_{r_i - \lambda_t c}(\pi_t) - V_{r_i - \lambda_t c}((\pi^*_{i,t}, \pi_{-i,t}), \lambda_t) \right) \\
&\le \sum_{t=0}^{T-1} \rho_t \left(\epsilon - \left(V_{r_i}(\pi^*_{i,t}, \pi_{-i,t}) - V_{r_i}(\pi_t) \right) \right) \\
&\qquad + \sum_{t=0}^{T-1} \rho_t \lambda_t \left(V_c(\pi^*_{i,t}, \pi_{-i,t}) - V_c(\pi_t) \right) \quad (6) \\
&\overset{(ii)}{\le} \sum_{t=0}^{T-1} \rho_t \left(\epsilon - D_i(\pi_t) \right) + 2\eta T C^2_{\max}
\end{aligned}$$

where (i) follows from the primal update rule and ϵ is the accuracy level in the primal update, (ii) follows from the definition of duality gap $D_i(\cdot)$ and the bound on the duality gap (see Lemma 4). Rearranging the above inequality gives the result. □

We remark that $\pi^*_{i,t}$ is the best response of agent i under the shared constraint while $\widetilde{\pi}_{i,t}$ is the optimizer that maximizes the Lagrangian function. In general, $(\widetilde{\pi}_{i,t}, \pi_{-i,t})$ is not necessarily permissible.

If we define the mixed output policy $\widetilde{\pi}_t$ such that $\mathbb{P}[\widetilde{\pi}_t = \pi_t] = \rho_t / \sum_{t=0}^{T-1} \rho_t$, we immediately get the value function gap guarantee:

$$\mathbb{E}[D_i(\widetilde{\pi}_t)] \le \epsilon + 4C^2_{\max} \cdot \frac{\eta T}{\sum_{t=0}^{T-1} \rho_t}.$$

Note that, in the uniform sampling $\rho_t = 1$ for all t, $\mathbb{E}[D_i(\widetilde{\pi}_t)] \le \epsilon + 4C^2_{\max} \cdot \eta$. The proposed weighted sampling provides a tighter bound on the expectation of the gap given a finite horizon T.

Step III: For any given $\widetilde{\lambda}$, we first show an inequality regarding $\sum_{t=0}^{T-1} \rho_t (\lambda_t - \widetilde{\lambda})(\alpha - V_c(\pi_t))$.

Lemma 6. Suppose $\rho_{t+1} \ge \rho_t > 0$ for all t. For any $\widetilde{\lambda} \in [0, \xi]$, it holds that

$$\eta \sum_{t=0}^{T-1} \rho_t (\lambda_t - \widetilde{\lambda})(\alpha - V_c(\pi_t)) \quad (7)$$

$$\le \frac{1}{2} \left(\sum_{t=0}^{T} \|\rho_t - \rho_{t+1}\|^2 + \|\rho_0\|^2 \right) \|\widetilde{\lambda}\|^2 + 2T\eta^2 C^2_{\max} \quad (8)$$

$$+ \eta \xi C_{\max} (\rho_T - \rho_0) + 2\xi^2 \sum_{t=0}^{T-1} \rho_t (\rho_{t+1} - \rho_t). \quad (9)$$

Proof. Given $\widetilde{\lambda} \in [0, \xi]$, it holds that

$$\begin{aligned}
&\|\rho_{t+1}\lambda_{t+1} - \rho_{t+1}\widetilde{\lambda}\|^2 \\
&\overset{(i)}{\le} \|\rho_t \lambda_t - \eta(\alpha - V_c(\pi_t)) - \rho_{t+1}\widetilde{\lambda}\|^2 \\
&= \|\rho_t(\lambda_t - \widetilde{\lambda}) + (\rho_t - \rho_{t+1})\widetilde{\lambda} - \eta(\alpha - V_c(\pi_t))\|^2 \\
&\le \|\rho_t \lambda_t - \rho_t \widetilde{\lambda}\|^2 + \|(\rho_t - \rho_{t+1})\widetilde{\lambda}\|^2 + \|\eta(\alpha - V_c(\pi_t))\|^2 \\
&- 2\eta\rho_t(\lambda_t - \widetilde{\lambda})(\alpha - V_c(\pi_t)) - 2\eta(\rho_t - \rho_{t+1})\widetilde{\lambda}(\alpha - V_c(\pi_t)) \\
&\quad + 2\rho_t(\rho_t - \rho_{t+1})\widetilde{\lambda}(\lambda_t - \widetilde{\lambda}) \\
&\overset{(ii)}{\le} \|\rho_t \lambda_t - \rho_t \widetilde{\lambda}\|^2 + \|(\rho_t - \rho_{t+1})\widetilde{\lambda}\|^2 + 4\eta^2 C^2_{\max} \\
&\quad - 2\eta\rho_t(\lambda_t - \widetilde{\lambda})(\alpha - V_c(\pi_t)) + 4\eta\xi C_{\max}(\rho_{t+1} - \rho_t) \\
&\quad + 4\xi^2 \rho_t(\rho_{t+1} - \rho_t),
\end{aligned}$$

where (i) follows from the nonexpansiveness of projection, and (ii) uses triangle inequality and the fact that $\rho_{t+1} > \rho_t > 0$ for all valid t. Summing both sides over $t = 0, 1, \ldots, T-1$ gives

$$0 \le \|\rho_T \lambda_T - \rho_T \widetilde{\lambda}\|^2$$

$$\leq \left(\sum_{t=0}^{T}\|\rho_t-\rho_{t+1}\|^2+\|\rho_0\|^2\right)\|\widetilde{\lambda}\|^2+4T\eta^2C_{\max}^2$$
$$-2\eta\sum_{t=0}^{T-1}\rho_t(\lambda_t-\widetilde{\lambda})(\alpha-V_c(\pi_t))$$
$$+2\eta\xi C_{\max}(\rho_T-\rho_0)+4\xi^2\sum_{t=0}^{T-1}\rho_t(\rho_{t+1}-\rho_t).$$

The proof is complete by rearranging the above terms. □

Step IV: By the strong duality and the property of MPGs, we can show that the difference between $V_{r_i}(\pi_{i,t}^*,\pi_{-i,t})$ and $V_{r_i}(\pi_t)-\widetilde{\lambda}_t^*(V_c(\pi_t)$ upper bounds by $\widetilde{\lambda}_t^*(V_c(\pi_t)-\alpha)_+$, where $\widetilde{\lambda}_t^*$ will be specified in the next lemma.

Lemma 7. Define

$$\widetilde{\lambda}_t^*=\arg\min_{\widetilde{\lambda}_t>0}\max_{\widetilde{\pi}_{i,t}}L((\widetilde{\pi}_{i,t},\pi_{-i,t}),\widetilde{\lambda}_t),$$

then it holds that

$$V_{r_i}(\pi_{i,t}^*,\pi_{-i,t})\geq V_{r_i}(\pi_t)-\widetilde{\lambda}_t^*(V_c(\pi_t)-\alpha)_+. \quad (10)$$

Proof. For ease of exposition, we denote $\Pi_{i,t}(\pi)=\{\pi_i:V_c(\pi_i,\pi_{-i})\leq\max\{\alpha,V_c(\pi)\}$. Note that

$$\begin{aligned}
&\phi(\pi_{i,t}^*,\pi_{-i,t})=\max_{\widetilde{\pi}_{i,t}}\min_{\widetilde{\lambda}_t>0}L((\widetilde{\pi}_{i,t},\pi_{-i,t}),\widetilde{\lambda}_t)\\
&\overset{(i)}{=}\min_{\widetilde{\lambda}_t>0}\max_{\widetilde{\pi}_{i,t}}L((\widetilde{\pi}_{i,t},\pi_{-i,t}),\widetilde{\lambda}_t)\\
&\overset{(ii)}{\geq}\max_{\widetilde{\pi}_{i,t}\in\Pi_{i,t}(\pi_t)}L((\widetilde{\pi}_{i,t},\pi_{-i,t}),\widetilde{\lambda}_t^*)\\
&=\max_{\widetilde{\pi}_{i,t}\in\Pi_{i,t}(\pi_t)}\phi(\widetilde{\pi}_{i,t},\pi_{-i,t})+\widetilde{\lambda}_t^*\left(\alpha-V_c(\widetilde{\pi}_{i,t},\pi_{-i,t})\right)\\
&\geq\max_{\widetilde{\pi}_{i,t}\in\Pi_{i,t}(\pi_t)}\phi(\widetilde{\pi}_{i,t},\pi_{-i,t})+\widetilde{\lambda}_t^*\min\{0,\alpha-V_c(\pi_t)\}\\
&\overset{(iii)}{\geq}\phi(\pi_t)-\widetilde{\lambda}_t^*\left(V_c(\pi_t)-\alpha\right)_+,
\end{aligned}$$

where (i) follows from the strong duality, (ii) uses the definition of $\widetilde{\lambda}_t^*$ and $\Pi_{i,t}(\pi_t)$ is a subset of the feasible set of agent i when other agents follows $\pi_{-i,t}$, (iii) follows from the fact that $\pi_{i,t}\in\Pi_{i,t}(\pi_t)$.

Now we exploit the property of MPGs, i.e. $V_{r_i}(\pi_i,\pi_{-i})-V_{r_i}(\pi_i',\pi_{-i})=\phi(\pi_i,\pi_{-i})-\phi(\pi_i',\pi_{-i})$, we can obtain the inequality $V_{r_i}(\pi_{i,t}^*,\pi_{-i,t})\geq V_{r_i}(\pi_t)-\widetilde{\lambda}_t^*(V_c(\pi_t)-\alpha)_+$. □

Step V: We are ready to establish the constraint violation guarantee, as summarized in the following lemma.

Lemma 8. Suppose $\rho_{t+1}\geq\rho_t>0$ for all t. It holds that

$$\begin{aligned}
&\sum_{t=0}^{T-1}\rho_t(V_c(\pi_t)-\alpha)_+\\
&\leq\frac{2\epsilon}{\xi}\sum_{t=0}^{T-1}\rho_t+\frac{\xi}{\eta}\left(\sum_{t=0}^{T}\|\rho_t-\rho_{t+1}\|^2+\|\rho_0\|^2\right)+\frac{4T\eta C_{\max}^2}{\xi}\\
&\quad+2C_{\max}(\rho_T-\rho_0)+\frac{4\xi}{\eta}\sum_{t=0}^{T-1}\rho_t(\rho_{t+1}-\rho_t).
\end{aligned}$$

Proof. Combining (9) and (10) in Lemma 6-7 gives

$$\begin{aligned}
&\eta\sum_{t=0}^{T-1}\rho_t\left(\widetilde{\lambda}(V_c(\pi_t)-\alpha)-\widetilde{\lambda}_t^*(V_c(\pi_t)-\alpha)_+\right)\\
&\leq\eta\epsilon\sum_{t=0}^{T-1}\rho_t+\frac{1}{2}\left(\sum_{t=0}^{T}\|\rho_t-\rho_{t+1}\|^2+\|\rho_0\|^2\right)\|\widetilde{\lambda}\|^2\\
&\quad+\eta\xi C_{\max}(\rho_T-\rho_0)+2\xi^2\sum_{t=0}^{T-1}\rho_t(\rho_{t+1}-\rho_{t+1})\\
&\quad+2T\eta^2C_{\max}^2. \quad (11)
\end{aligned}$$

Similar to Lemma E.1 in Chen, Ma, and Zhou (2022), we can show that the optimal dual variable $\lambda_t^*\leq HR_{\max}/u$ where u is the duality gap in Slater condition (Assumption 1). Recall $\xi=2HR_{\max}/u$ and $\lambda_t^*\in[0,\xi/2]$. Selecting $\widetilde{\lambda}=\xi\mathbb{1}\{V_c(\pi_t)\leq\alpha\}$ gives

$$\widetilde{\lambda}(V_c(\pi_t)-\alpha)-\widetilde{\lambda}_t^*(V_c(\pi_t)-\alpha)_+\geq\frac{\xi}{2}(V_c(\pi_t)-\alpha)_+. \quad (12)$$

Combining (11) and (12) gives the desired result. □

Conclusion

In this work, we investigated the primal-dual approach for learning NE in MPGs with a shared constraint. We first introduced the Slater condition assumption for MPGs with a shared constraint that leads to the strong duality, and proposed an algorithm for checking the Slater condition. Based on the result, we developed a primal-dual algorithm that learns an ϵ-approximate NE at the rate of $\mathcal{O}(\epsilon+1/\sqrt{T})$. Different from the existing literature, our primal-dual algorithm features two variants of the standard dual updating rules, which can lead to weighted output policies. We showed that the weighted output policy can achieve better performance in certain regimes compared to the uniform output policy.

Appendix A: Proof of Theorem 1

The proof is similar to that in Chen, Ma, and Zhou (2022), and we provide the proof sketch for completeness.

Let policy-induced measure $p_\pi(s_{1:h},a_{1:h'})$ be the distribution of $(s_{1:h},a_{1:h'})$ induced by joint policy π. By Appendix A in Chen, Ma, and Zhou (2022), the value function and the Lagrangian function can be written as a linear function of p_π, i.e., $V_{r_i}(\pi)=\tilde{V}_{r_i}(p_\pi)=\sum_{s_{1:H},a_{1:H}}p_\pi(s_{1:H},a_{1:H})\sum_{h=0}^{H}\gamma^h r_i(s_h,a_h)$. Note that for any reference policy $\pi^{\text{ref}}=(\pi_1^{\text{ref}},\ldots,\pi_N^{\text{ref}})$, policy $\pi=(\pi_1,\ldots,\pi_N)$ can be decomposed as

$$\pi=\pi-(\pi_1^{\text{ref}},\pi_{-1})+(\pi_1^{\text{ref}},\pi_{-1})+\cdots+\pi^{\text{ref}}-\pi^{\text{ref}},$$

where adjacent terms are the same except for only one agent's policy. By the definition of MPG, $\phi(\pi)=\phi(\pi^{\text{ref}})+V(\pi)-V(\pi^{\text{ref}})$ is therefore a linear function in p_π since $V(\pi)$ is linear in p_π.

Let $\tilde{L}(p_\pi,\lambda)=L(\pi,\lambda)$. Based on the minimax theorem (Lemma 9.2 of Altman (2021)), it suffices to prove the following properties.

1) $\tilde{L}(p_\pi, \cdot)$ is convex and lower semi-continuous, and $\tilde{L}(\cdot, \lambda)$ is concave.
2) Domain of dual variable is convex.
3) Domain of primal variable p_π is convex and compact. The Slater-type condition assumption guarantees there always exists a feasible p_π, and such domain exists.

The last two properties hold while the first property follows from the definition of $\tilde{L}(\cdot, \cdot)$ and the linear property with respect to p_π of the Lagrangian function, and the proof is complete.

Appendix B: Proof of Theorem 3

For ease of exposition, we write $y_t \lesssim x_t$ if there exists some absolute constant C such that $y_t < Cx_t$ for all $t \in \mathbb{N}$, and we we write $y_t \sim x_t$ if there exists some absolute constant C_1, C_2 such that $C_1 x_t < y_t < C_2 x_t$ for all $t \in \mathbb{N}$.

The following lemma is useful for proving Theorem 3.

Lemma 9. The following inequalities hold:

1) $\sum_{t=1}^{T} \log t \sim T \log T$.
2) $\sum_{t=1}^{T} \log^2(1 + 1/t) \leq \sum_{t=1}^{T} 1/t^2 < \infty$.
3) $\sum_{t=1}^{T} \log t \log(1 + 1/t) \lesssim \sum_{t=1}^{T} \log t$.

Under weighed policy, $\rho_t = \log(t + 1)$ and $\eta = 1/\sqrt{T}$ in update rule (3). By Theorem 2 and Lemma 9 1), the value function gap of agent i is

$$\mathbb{E}[D_i(\widetilde{\pi}_t)] \lesssim \epsilon + C_{\max}^2 \cdot \frac{1/\sqrt{T} \cdot T}{T \log T} = \epsilon + \frac{C_{\max}^2}{\sqrt{T} \log T},$$

By Theorem 2 and Lemma 9 1-3), the constraint violation is

$$\begin{aligned}
\mathbb{E}[(V_c(\widetilde{\pi}_t) - \alpha)_+] &\lesssim \frac{\epsilon}{\xi} + \frac{1}{T \log T} \frac{\xi}{1/\sqrt{T}} \\
&+ \frac{1}{T \log T} \left[\frac{C_{\max}^2 T/\sqrt{T}}{\xi} + C_{\max} \log T + \frac{\xi}{1/\sqrt{T}} \right] \\
&= \frac{\epsilon}{\xi} + \frac{\xi}{\sqrt{T} \log T} + \frac{C_{\max}^2/\xi}{\sqrt{T} \log T} + \frac{C_{\max}}{T}.
\end{aligned}$$

Under uniformly chosen policy, the upper bounds for the value function gap and constraint violation under uniformly chosen policy can be easily obtained by Theorem 2, i.e.,

$$\mathbb{E}[D(\widetilde{\pi}_T)] \lesssim \epsilon + \frac{C_{\max}^2 T \cdot (1/\sqrt{T})^2}{T \cdot (1/\sqrt{T})} = \epsilon + \frac{C_{\max}^2}{\sqrt{T}},$$

$$\begin{aligned}
\mathbb{E}[(V_c(\widetilde{\pi}_t) - \alpha)_+] &\lesssim \frac{\epsilon}{\xi} + \frac{\xi}{T(1/\sqrt{T})} + \frac{C_{\max^2}}{\xi} \frac{T(1/\sqrt{T})^2}{T(1/\sqrt{T})} \\
&= \frac{\epsilon}{\xi} + \frac{\xi}{\sqrt{T}} + \frac{C_{\max}^2}{\xi\sqrt{T}}.
\end{aligned}$$

Appendix C: Sample Complexity

In our algorithm, any existing algorithms for finding ϵ-approximate NE in unconstrained MPGs can be served as the oracle for the primal update. In the following, we choose Nash-CA algorithm developed in Song, Mei, and Bai (2022), which achieves $\tilde{O}(\epsilon^{-3})$ sample complexity.

In the following, we aim to provide sample complexity of finding a policy π such that both the value function gap and the constraint violation are bounded by 2ϵ. We will focus on the sample complexity of the uniformly chosen policy, and the sample complexity of the weighted policy can be obtained similarly.

For the value function gap, setting $\mathbb{E}[D(\widetilde{\pi}_T)] \lesssim \epsilon + \frac{C_{\max}^2}{\sqrt{T}} \leq 2\epsilon$ gives $T \geq \tilde{\mathcal{O}}(\epsilon^2)$. Similarly, $T \geq \tilde{\mathcal{O}}(\epsilon^2)$ also guarantees that the constraint violation is bounded by $O(\epsilon)$. Therefore, the sample complexity of finding (2ϵ)-approximate NA is $T \cdot \tilde{O}(\epsilon^{-3}) = \tilde{O}(\epsilon^{-5})$. By rescaling, we conclude that our algorithm achieves $\tilde{O}(\epsilon^{-5})$ sample complexity for finding ϵ-approximate NE policy.

Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-22-2-0233. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

References

Alatur, P.; Ramponi, G.; He, N.; and Krause, A. 2024. Provably Learning Nash Policies in Constrained Markov Potential Games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*.

Altman, E. 2021. *Constrained Markov decision processes*. Routledge.

Altman, E.; and Shwartz, A. 2000. Constrained Markov Games: Nash Equilibria. In *Advances in Dynamic Games and Applications*. Birkhäuser Boston.

Chen, Z.; Ma, S.; and Zhou, Y. 2022. Finding Correlated Equilibrium of Constrained Markov Game: A Primal-Dual Approach. In *Advances in Neural Information Processing Systems*, volume 35.

Daskalakis, C.; Golowich, N.; and Zhang, K. 2023. The Complexity of Markov Equilibrium in Stochastic Games. In *Proceedings of Thirty Sixth Conference on Learning Theory*.

Deng, X.; Li, N.; Mguni, D.; Wang, J.; and Yang, Y. 2022. On the complexity of computing Markov perfect equilibrium in general-sum stochastic games. *National Science Review*, 10(1): nwac256.

Ding, D.; Wei, C.-Y.; Zhang, K.; and Jovanovic, M. 2022. Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence. In *Proceedings of the 39th International Conference on Machine Learning*.

Ding, D.; Wei, X.; Yang, Z.; Wang, Z.; and Jovanovic, M. 2023. Provably Efficient Generalized Lagrangian Policy Optimization for Safe Multi-Agent Reinforcement Learning. In *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*.

ElSayed-Aly, I.; Bharadwaj, S.; Amato, C.; Ehlers, R.; Topcu, U.; and Feng, L. 2021. Safe Multi-Agent Reinforcement Learning via Shielding. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems*.

Feng, S.; Yin, M.; Wang, Y.; Yang, J.; and Liang, Y. 2024. Improving Sample Efficiency of Model-Free Algorithms for Zero-Sum Markov Games. In *Forty-first International Conference on Machine Learning*.

Gu, S.; Grudzien Kuba, J.; Chen, Y.; Du, Y.; Yang, L.; Knoll, A.; and Yang, Y. 2023. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319: 103905.

Guo, X.; Li, X.; Maheshwari, C.; Sastry, S.; and Wu, M. 2024. Markov α-Potential Games. arXiv:2305.12553.

Hanawal, M. K.; Abdel-Rahman, M. J.; and Krunz, M. 2016. Joint Adaptation of Frequency Hopping and Transmission Rate for Anti-Jamming Wireless Systems. *IEEE Transactions on Mobile Computing*, 15(9): 2247–2259.

Jin, C.; Liu, Q.; Wang, Y.; and Yu, T. 2022. V-Learning – A Simple, Efficient, Decentralized Algorithm for Multiagent RL. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*.

Jin, Y.; Muthukumar, V.; and Sidford, A. 2022. The Complexity of Infinite-Horizon General-Sum Stochastic Games. In *Information Technology Convergence and Services*.

Jordan, P.; Barakat, A.; and He, N. 2024. Independent Learning in Constrained Markov Potential Games. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*.

K.J., P.; A.N, H. K.; and Bhatnagar, S. 2014. Multi-agent reinforcement learning for traffic signal control. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*.

Kononen, V.; and Oja, E. 2004. Asymmetric multiagent reinforcement learning in pricing applications. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*.

Leonardos, S.; Overman, W.; Panageas, I.; and Piliouras, G. 2022. Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*.

Liu, Q.; Yu, T.; Bai, Y.; and Jin, C. 2021. A Sharp Analysis of Model-based Reinforcement Learning with Self-Play. In *Proceedings of the 38th International Conference on Machine Learning*.

Macua, S. V.; Zazo, J.; and Zazo, S. 2018. Learning Parametric Closed-Loop Policies for Markov Potential Games. In *International Conference on Learning Representations*.

Maheshwari, C.; Wu, M.; Pai, D.; and Sastry, S. 2023. Independent and Decentralized Learning in Markov Potential Games. arXiv:2205.14590.

Mao, W.; and Başar, T. 2022. Provably Efficient Reinforcement Learning in Decentralized General-Sum Markov Games. *Dynamic Games and Applications*.

Mao, W.; Yang, L.; Zhang, K.; and Basar, T. 2022. On Improving Model-Free Algorithms for Decentralized Multi-Agent Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*.

Marden, J. R. 2012. State based potential games. *Automatica*, 48(12): 3075–3088.

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

Narasimha, D.; Lee, K.; Kalathil, D.; and Shakkottai, S. 2022. Multi-Agent Learning via Markov Potential Games in Marketplaces for Distributed Energy Resources. In *2022 IEEE 61st Conference on Decision and Control (CDC)*.

Perolat, J.; Strub, F.; Piot, B.; and Pietquin, O. 2017. Learning Nash Equilibrium for General-Sum Markov Games from Batch Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Shalev-Shwartz, S.; Shammah, S.; and Shashua, A. 2016. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. arXiv:1610.03295.

Song, Z.; Mei, S.; and Bai, Y. 2022. When Can We Learn General-Sum Markov Games with a Large Number of Players Sample-Efficiently? In *International Conference on Learning Representations*.

Wang, Y.; Liu, Q.; Bai, Y.; and Jin, C. 2023. Breaking the Curse of Multiagency: Provably Efficient Decentralized Multi-Agent RL with Function Approximation. In *Proceedings of Thirty Sixth Conference on Learning Theory*.

Zhang, R.; Mei, J.; Dai, B.; Schuurmans, D.; and Li, N. 2024. On the global convergence rates of decentralized softmax gradient play in Markov potential games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Zhang, R. C.; Ren, Z.; and Li, N. 2022. Gradient Play in Stochastic Games: Stationary Points and Local Geometry. *IFAC-PapersOnLine*, 55(30): 73–78. 25th International Symposium on Mathematical Theory of Networks and Systems MTNS 2022.

Zhou, Z.; Chen, Z.; Lin, Y.; and Wierman, A. 2023. Convergence rates for localized actor-critic in networked Markov potential games. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*.