Finding Correlated Equilibrium of Constrained Markov Game: A Primal-Dual Approach

Ziyi Chen, Shaocong Ma, Yi Zhou
Department of Electrical and Computer Engineering
University of Utah
Salt lake city, UT 84112
{u1276972,s.ma,yi.zhou}@utah.edu

Abstract

Constrained Markov game is a fundamental problem that covers many applications, where multiple agents compete with each other under behavioral constraints. The existing literature has proved the existence of Nash equilibrium for constrained Markov games, which turns out to be PPAD-complete and cannot be computed in polynomial time. In this work, we propose a surrogate notion of correlated equilibrium (CE) for constrained Markov games that can be computed in polynomial time, and study its fundamental properties. We show that the modification structure of CE of constrained Markov games is fundamentally different from that of unconstrained Markov games. Moreover, we prove that the corresponding Lagrangian function has zero duality gap. Based on these results, we develop the first primal-dual algorithm that provably converges to CE of constrained Markov games. In particular, we prove that both the duality gap and the constraint violation of the output policy converge at the rate $\mathcal{O}(\frac{1}{\sqrt{T}})$. Moreover, when adopting the V-learning algorithm as the subroutine in the primal update, our algorithm achieves an approximate CE with ϵ duality gap with the sample complexity $\mathcal{O}(H^9SA^2\epsilon^{-4})$.

1 Introduction

Markov game [43] is a fundamental problem in game theory where multiple agents compete with each other by interacting with a dynamic environment. It has broad applications in diverse fields, including board games [44], inverse reinforcement learning [61], market pricing [34], etc.

In the existing literature, the goal of Markov game is often formulated as achieving Nash equilibrium (NE) – a product policy under which no agent can benefit via deviating from its own policy alone. Although NE has been shown to exist for general Markov games [22], computing NE turns out to be a PPAD-complete problem that cannot be solved in polynomial time [17, 32], except for some special Markov games with zero-sum structure [11, 16, 21, 63, 64] or potential structure [35, 58]. This further motivates researchers to consider other surrogate notions of equilibrium that are slightly weaker than NE but are computationally tractable, and correlated equilibrium (CE) is such a classic and popular surrogate notion [40] (see Section 2 for the formal definition) for which many convergent and efficient algorithms have been developed very recently [31, 37, 46].

While Markov game has been well studied, its formulation has limited applicability to real-life applications where the behavior of agents is usually subject to certain constraints. For example, in an uplink time division multiple access (TDMA) cognitive radio network, each user's total latency cannot exceed a threshold on average [62]. In an anti-jamming system, the jammer aims to jam ligitimate signal transmitter under average power constraint [27]. To overcome this limitation, constrained Markov game has been introduced where agents compete with each other under local behavioral constraints encoded by state value functions [4]. However, most existing works on

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

constrained Markov game only studied Nash equilibrium (NE) that is PPAD-complete to compute [3–6, 20, 24, 27–30, 45, 48, 51, 52, 55, 59, 60, 62], and a tractable notion of correlated equilibrium (CE) has not been formally defined for constrained Markov games. To the best of our knowledge, [26] is the only work that explored CE in constrained Markov games. However, they define CE based on the unconstrained Lagrangian function associated with the constrained Markov game, which does not correspond to an equilibrium of the original constrained Markov game. Therefore, **the goal of this work** is to formally define a tractable notion of CE for constrained Markov games, study its fundamental properties and develop provably convergent and computation-efficient algorithms for achieving such an equilibrium, as we further elaborate below.

To achieve this goal, we need to explore and address several fundamental problems. First, while one can generalize the definition of CE of Markov game to the constrained case, it may possess fundamentally different structures due to the presence of constraints. Specifically, as we introduce later in Definition 2.1, *modification* (of actions) is a key structure of CE. In standard Markov games (without constraints), CE is equivalent under both stochastic and deterministic modifications. However, it is unclear if such equivalence still holds in constrained Markov games, which critically affects the formal definition of CE as well as the algorithm design. Second, a classic and powerful tool to handle constrained problems is the strong duality of the associated Lagrangian, which has been established for NE of constrained Markov games [30, 62]. Thus, we are inspired to explore if strong duality holds for CE of constrained Markov games, which is a challenging problem due to the complex modification structure of CE and the nonconvex Lagrangian. Lastly, once we have a comprehensive understanding of the above two problems, we can hope for a provably convergent and efficient primal-dual algorithm for finding CE of constrained Markov games. In particular, it is much desired to analyze the non-asymptotic convergence rate and sample complexity of the proposed algorithm, as the existing algorithms for constrained Markov games either lack convergence guarantee [30] or only have asymptotic convergence guarantee [26, 62].

1.1 Our Contributions

We define a notion of CE for constrained Markov games as a generalization of the CE for Markov games. It is defined as an equilibrium at which no feasible stochastic modification of the joint policy from any agent alone can improve the associated value function (see Definition 3.2). Such a generalized CE turns out to possess a structure of *modification* that is fundamentally different from the CE of Markov games. Specifically, we prove in Theorem 1 that our proposed CE for constrained Markov games is strictly stronger with stochastic modification than with deterministic modification. As a comparison, the CE of Markov games is equivalent with both types of modifications. This fact motivates defining CE with stochastic modification for constrained Markov games, which further lays the foundation for developing the strong duality result and the primal-dual algorithm later.

We study the Lagrangian function associated with the proposed CE for constrained Markov games and prove that it has zero duality gap (See Theorem 3). Such a result is the key for developing a convergent primal-dual algorithm. Compared with the strong duality established in the existing works [2, 30, 42], the proof of our result for constrained Markov games requires new technical developments. To elaborate, the existing works yield Lagrangian functions that are defined over the entire unconstrained policy space. As a comparison, our Lagrangian function is defined over the set of stochastically modified policies, which corresponds to a subset of the policy space over which the Lagrangian function is nonconvex. To address this challenge, we transform the Lagrangian function from the policy space to the corresponding space of probability measures on episodes induced by the stochastically modified policies. We show that the transformed Lagrangian function is linear and the transformed space is convex and compact, which further implies the desired strong duality.

Then, we propose the first primal-dual algorithm that provably converges to CE of constrained Markov games. In particular, the primal update of our algorithm requires to compute an approximate CE of an unconstrained Markov game associated with the Lagrangian function, which can be efficiently solved by many existing algorithms, e.g., V-learning [31, 46] and Nash Value Iteration [37]. We prove that both the duality gap and the constraint violation of the output policy produced by our algorithm converge at the rate $\mathcal{O}(\frac{1}{\sqrt{T}})$, where T denotes the number of iterations. This rate is comparable to that of the primal-dual algorithm for single-agent constrained reinforcement learning [18, 19]. Moreover, when adopting the V-learning algorithm as the subroutine in the primal update, our algorithm achieves an approximate CE with ϵ duality gap and constrained violation with the sample

complexity $\mathcal{O}(H^9SA^2\epsilon^{-4})$, where H denotes the episode length, S, A denote the cardinality of the state space and action space, respectively.

1.2 Other Related Work

Markov game: [53] introduced coarse correlated equilibrium (CCE) for Markov game that is weaker than CE. Many algorithms such as V-learning [31, 39, 46] and Nash Value Iteration [37] can achieve an approximate CCE with finite time convergence guarantee. Various other Markov game settings have also been studied. For example, [14, 25] studied mean-field Markov game with a large number of agents. In a Stackelberg game [8, 50], agents are partitioned into leaders and followers. The followers select the best response to the leaders' actions, and the leaders' goal is to achieve Stackelberg equilibrium which corresponds to the leaders' optimal strategy. Strategic game can be seen as a special case of Markov game without state and transition [41]. Extensive-form game [10, 23, 33, 65] can be seen as a special case of Markov game where any state can only be reached from a unique state at the previous step due to the tree structure of the game.

Constrained reinforcement learning: Single-agent constrained reinforcement learning (RL) is a single-agent RL problem with behavioral constraints. It is usually formulated as constrained Markov decision process [2], which can be seen as a special case of constrained Markov game with only one agent. Various primal-dual algorithms have been proposed for single-agent constrained RL [1, 2, 18, 19, 36, 47, 49, 56, 57]. Other types of algorithms include Lyapunov-based algorithms that use Lyapunov functions [12, 13], interior point methods that use logarithmic barrier functions [38], policy network with a special layer to encode constraints [15], and primal algorithms that update the policy via alternatively maximizing the objective function and minimizing the violated constraints [54].

2 Preliminaries of Markov Game

We consider M agents playing a Markov game over a finite time horizon of episode length H. Specifically, at every time step $h = 1, ..., H$, the agents observe a global state $s_h \in \mathcal{S}$ of the environment, where $\mathcal{S}$ denotes the state space. Then, the agents take a joint action $a_h = [a_h^{(1)}, \ldots, a_h^{(M)}]$ following a joint stochastic policy $\pi_h(\cdot|s_{1:h}, a_{1:(h-1)})$, which corresponds to a distribution on the joint action space $\mathcal{A} := \prod_{m=1}^{M} \mathcal{A}^{(m)}$ and depends on the past states $s_{1:h} := \{s_t\}_{t=1}^{h}$ and past actions $a_{1:(h-1)} := \{a_t\}_{t=1}^{h-1}$. After that, the global state of the environment transfers to a new state s_{h+1} following the state transition kernel $\mathcal{P}(\cdot|s_h, a_h)$, and each agent m receives a local reward $r_{0,h}^{(m)}(s_h, a_h)$ from the environment. We note that here the joint policy π_h can be a non-product (i.e., cannot be factorized into a product of independent policies associated with the agents) and non-Markov policy.

In the above Markov game, each agent m collects its own rewards over the episodes. Denote ρ as the initial state distribution, we can define the following value function associated with agent m.

$$V_0^{(m)}(\pi) := \mathbb{E}_\pi\Big[\sum_{h=1}^{H} r_{0,h}^{(m)}(s_h, a_h)\Big| s_1 \sim \rho\Big], \tag{1}$$

which corresponds to the expected cumulative reward received by agent m under joint policy $\pi := \{\pi_h\}_{h=1}^{H}$ and initial state s. The goal of agent m is to optimize its own policy $\pi^{(m)} := \{\pi_h^{(m)}\}_{h=1}^{H}$ in order to maximize its associated value function $V_0^{(m)}(\pi)$. However, since every agent's value function is also affected by the other agents' policies and actions, the agents must compete with each other to gain more rewards until they reach a certain equilibrium. Here, we introduce two popular equilibrium notions that will be discussed throughout the paper.

Nash Equilibrium (NE). A joint policy π is called an NE if the following two conditions are met: (i) for any time step h, the joint policy π_h is a product policy that can be factorized into a product of *independent* policies, i.e., $\pi_h = \pi_h^{(1)} \times \ldots \times \pi_h^{(M)}$ so that the agents take independent actions; (ii) For any agent m with any associated policy $\widetilde{\pi}^{(m)}$, it holds that $V_0^{(m)}(\pi) \geq V_0^{(m)}(\widetilde{\pi}^{(m)} \times \pi^{(\setminus m)})$. Here, $\pi^{(\setminus m)}$ denotes the joint policy of all the other agents excluding the agent m, and '$\times$' means that $\widetilde{\pi}^{(m)}$ is independent from $\pi^{(\setminus m)}$.

In other words, at NE, no agent can improve its value function by deviating from its associated policy. In the existing literature, it has been shown that computing NE is a PPAD-complete problem [17, 32],

for which it is not possible to develop polynomial-time algorithms. This observation has motivated researchers to propose a surrogate correlated equilibrium (CE) notion [40]. Before introducing the formal definition of CE, we first define the following stochastic modification operator.

Definition 2.1 (Stochastic modification)**.** *For any time step h, denote $a_h^{(m)}$ as agent m's action induced by joint policy π_h. Given the past states and actions $s_{1:h}, a_{1:(h-1)}$, a stochastic modification $\phi_h^{(m)}$ (associated with agent m) is a distribution that randomly maps $a_h^{(m)}$ to another action $\widetilde{a}_h^{(m)}$, i.e., $\widetilde{a}_h^{(m)} \sim \phi_h^{(m)}(\cdot|s_{1:h}, a_{1:h-1}, a_h^{(m)})$. Moreover, we denote $\phi_h^{(m)} \circ \pi_h$ as the joint policy modified by $\phi_h^{(m)}$, i.e., π_h first generates a joint action $a_h := [a_h^{(m)}, a_h^{(\setminus m)}]$, and then $\phi_h^{(m)}$ modifies $a_h^{(m)}$ to another $\widetilde{a}_h^{(m)}$ at random. We also denote $\phi^{(m)} := \{\phi_h^{(m)}\}_{h=1}^H$, $\phi^{(m)} \circ \pi := \{\phi_h^{(m)} \circ \pi_h\}_{h=1}^H$.*

Correlated Equilibrium (CE). A joint policy π is called a CE if for any agent m and any stochastic modification $\phi^{(m)}$, it holds that $V_0^{(m)}(\pi) \geq V_0^{(m)}(\phi^{(m)} \circ \pi)$.

Intuitively, at CE, no agent can improve its value function by modifying its own action induced by the joint CE policy. Compare to NE policies, CE policies do not require joint independence among all the agents. In fact, it has been shown that any NE policy is guaranteed to be a CE policy [31, 37, 46], and hence CE is a weaker equilibrium notion than NE. Moreover, CE can be reformulated as linear programming and therefore is computationally tractable.

3 Correlated Equilibrium of Constrained Markov Game

In this section, we introduce constrained Markov game – a generalization of the standard Markov game with constraints on agents' behavior. We then propose and study a new notion of correlated equilibrium for constrained Markov games.

3.1 Constrained Markov Game

In many real-life Markov games, the behavior of the agents is usually restricted by certain constraints, e.g., constraints on total time delay, average power, etc. This motivates us to consider a constrained version of Markov game [2]. Constrained Markov game follows the same state/action transition as that of the standard Markov game introduced in Section 2. The key difference is that, after the state transition at each time step h in a constrained Markov game, every agent m receives d_m+1 number of rewards $\{r_{j,h}^{(m)}(s_h, a_h)\}_{j=0}^{d_m}$. Here, the accumulation of the reward $r_{0,h}^{(m)}$ corresponds to the target that agent m aims to maximize. The other rewards $\{r_{j,h}^{(m)}\}_{j=1}^{d_m}$ are introduced to encode the behavioral constraints of agent m. In this paper, we consider the type of covering constraints on the accumulation of the rewards $\{r_{j,h}^{(m)}(s_h, a_h)\}_{j=1}^{d_m}$ over the episode, which impose constraints on the long-term behavior of agents. More specifically, define the following value functions associated with the reward functions of agent m:

$$V_j^{(m)}(\pi) := \mathbb{E}_\pi\Big[\sum_{h=1}^H r_{j,h}^{(m)}(s_h, a_h)\Big| s_1 \sim \rho\Big], \quad j = 0, \ldots, d_m. \tag{2}$$

Every agent m aims to optimize its own policy $\pi^{(m)}$ via the following constrained Markov game.

$$\text{(Constrained Markov Game):} \quad \max_{\pi^{(m)}} V_0^{(m)}(\pi), \ \text{ s.t. } V_j^{(m)}(\pi) \geq c_j^{(m)}, \ j = 1, \ldots, d_m. \tag{3}$$

where $\{c_j^{(m)}\}_{j=1}^{d_m} \in \mathbb{R}$ denote the thresholds of the constraints $V_j^{(m)}(\pi) \geq c_j^{(m)}$. Intuitively, the constraints on the value functions implicitly impose constraints on the joint policy π, and all the agents aim to maximize their own target value functions and achieve a certain equilibrium under their local constraints.

3.2 Nash Equilibrium and Correlated Equilibrium

For constrained Markov games, we can define the following notion of Nash equilibrium (NE) that is similar to the standard Nash equilibrium but takes into account the constraints [3, 27, 45, 51, 55, 60].

Definition 3.1 (NE of constrained Markov game)**.** *For a constrained Markov game, a joint policy π is called an NE if the following conditions are met: (i) for any time step h, the joint policy π_h is a product policy that can be factorized into a product of independent policies; (ii) π is feasible, i.e., it satisfies the constraints of all the agents; (iii) For any agent m with any associated policy $\widetilde{\pi}^{(m)}$ such that $\widetilde{\pi}^{(m)} \times \pi^{(\setminus m)}$ is feasible, it holds that $V_0^{(m)}(\pi) \geq V_0^{(m)}(\widetilde{\pi}^{(m)} \times \pi^{(\setminus m)})$.*

Compared with the NE of Markov game defined in Section 2, NE of constrained Markov game further requires the joint policies π and $\widetilde{\pi}^{(m)} \times \pi^{(\setminus m)}$ to be feasible. It has been shown that such an NE exists for general constrained Markov games [4]. Moreover, since constrained Markov game generalizes the Markov game, computing NE of constrained Markov game is also a PPAD-complete problem. Therefore, we are motivated to propose the following surrogate notion of correlated equilibrium (CE).

Definition 3.2 (CE of constrained Markov game)**.** *For a constrained Markov game, a joint policy π is called a CE if the following conditions hold: (i) π is feasible, i.e., it satisfies the constraints of all the agents; (ii) for any agent m and any stochastic modification $\phi^{(m)}$ such that $\phi^{(m)} \circ \pi$ is feasible, it holds that $V_0^{(m)}(\pi) \geq V_0^{(m)}(\phi^{(m)} \circ \pi)$.*

In particular, if π is a product policy, it can be seen that $\phi^{(m)} \circ \pi$ is also a product policy and then CE would reduce to NE. Therefore, the above definition of CE generalizes NE of constrained Markov games. We note that [26] also defined CE for constrained Markov games, but in a very different aspect. Specifically, they only defined CE based on the **unconstrained** Markov game associated with a Lagrange function (i.e., with surrogate reward defined in eq. (8)), which does not necessarily require π and $\phi^{(m)} \circ \pi$ to be feasible but our Definition 3.2 requires. Therefore, the existing CE proposed in [26] essentially follows the CE of **unconstrained** Markov games, whose definition is stated after our Definition 2.1. Moreover, [26] defined CE based on the state-action value function and consider stationary Markov policies in the discounted infinite horizon setting, whereas we define CE based on the state value function and consider non-Markov policies in the finite horizon setting.

It may seem that the CE in Definition 3.2 is a straightforward generalization of that of the standard Markov game. However, it turns out to possess a fundamentally different structure of the modification operator at the equilibrium point due to the presence of constraints, as elaborated in the following fundamental result. Throughout the paper, we call a stochastic modification operator deterministic if its pdf is supported on a single action.

Theorem 1 (Modification of CE)**.**

1. *For unconstrained Markov games. If a policy π satisfies that $V_0^{(m)}(\pi) \geq V_0^{(m)}(\phi^{(m)} \circ \pi)$ for any agent m and any deterministic modification $\phi^{(m)}$, then π must be a CE.*
2. *For constrained Markov games. Even if a feasible policy π satisfies that $V_0^{(m)}(\pi) \geq V_0^{(m)}(\phi^{(m)} \circ \pi)$ for any agent m and any feasible deterministic modification $\phi^{(m)}$, π may not be a CE.*

Theorem 1 shows that the structure of modification at the CE of Markov games is different from that of constrained Markov games. To elaborate item 1, recall that a CE policy π of a Markov game is defined to satisfy $V^{(m)}(\pi) \geq V^{(m)}(\phi^{(m)} \circ \pi)$ for any stochastic modification $\phi^{(m)}$ (see Section 2). Since stochastic modifications include deterministic modifications, item 1 proves that CE of Markov games can always be achieved by a deterministic modification. Therefore, it suffices to define CE of Markov games using deterministic modifications, as adopted by many existing works [31, 37, 46]. As a comparison, item 2 shows that for constrained Markov games, the CE defined using deterministic modification is in general weaker than the CE defined using stochastic modification (see Definition 3.2), which is the focus of this paper. More importantly, as we show later in Theorem 2, defining CE of constrained Markov games using stochastic modification allows us to establish a strong duality result, which is the key to develop convergent primal-dual algorithms.

The proof of Theorem 1 is non-trivial and we provide a proof sketch here. To prove item 1, our strategy is to construct a simple deterministic modification $\phi^{(m)}$ (see Appendix C for the construction) for any optimal stochastic modification $\widetilde{\phi}^{(m)}$ and show that they achieve the same optimal value function, i.e., $V_0^{(m)}(\phi^{(m)} \circ \pi) = V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi)$. To show this, we transform the value functions from the policy space π to the space of probability measure p_π of episodes induced by the associated policy, i.e., $V_0^{(m)}(\pi) = \widetilde{V}_0^{(m)}(p_\pi)$ (see eq. (22)), so that the transformed value function $\widetilde{V}_0^{(m)}$ is linear in p_π. Then we consider the extrapolated probability measure $p_\lambda := \lambda p_{\phi^{(m)} \circ \pi} + (1-\lambda) p_{\widetilde{\phi}^{(m)} \circ \pi}$ and show that there exists $\lambda < 0$ such that p_λ is a measure induced by a modified policy $\phi_\lambda^{(m)} \circ \pi$ for

certain proper stochastic modification $\phi_\lambda^{(m)}$. This together with the linearity of $\widetilde{V}_0^{(m)}$ proves item 1. To prove item 2, we construct a counterexample of constrained Markov game such that its CE under stochastic modification is shown to be strictly stronger than its CE under deterministic modification.

4 A Primal-Dual Algorithm for Constrained Markov Game

In this section, we develop a primal-dual algorithm for finding correlated equilibrium (CE) of general constrained Markov games. Throughout this section, to simplify the notation, we denote $V^{(m)}(\pi) := [V_1^{(m)}(\pi), ..., V_{d_m}^{(m)}(\pi)]$ as the collection of all the value functions in the constraints, and denote $c^{(m)} := [c_1^{(m)}, ..., c_{d_m}^{(m)}]$ as the collection of all the thresholds of the constraints.

4.1 Strong Duality for Correlated Equilibrium

Consider any feasible policy π of a constrained Markov game, i.e., $V^{(m)}(\pi) \geq c^{(m)}$ holds entrywise. The CE of constrained Markov game is related to the following constrained optimization problem.

$$\text{For } m = 1, ..., M, \quad \max_{\text{stochastic } \phi^{(m)}} V_0^{(m)}(\phi^{(m)} \circ \pi) \quad \text{s.t.} \quad V^{(m)}(\phi^{(m)} \circ \pi) \geq c^{(m)}, \tag{4}$$

where '$\geq$' applies entrywise and the maximization is taken over all stochastic modifications. To further explain, define the following duality gap for every agent m.

$$D^{(m)}(\pi) := \max_{\text{feasible } \phi^{(m)}} V_0^{(m)}(\phi^{(m)} \circ \pi) - V_0^{(m)}(\pi), \tag{5}$$

where a modification $\phi^{(m)}$ is called feasible if $\phi^{(m)} \circ \pi$ satisfies the constraint in eq. (4). Clearly, we have $D^{(m)}(\pi) \geq 0$ as the identity modification (i.e., $\phi^{(m)} \circ \pi = \pi$) is feasible. Moreover, it can be seen that π is a CE of constrained Markov game if and only if $\max_m D^{(m)}(\pi) = 0$.

Since the problem in eq. (4) takes a constrained form, it is natural to consider its equivalent Lagrangian formulation, which is written below.

$$\max_{\phi^{(m)}} \min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)}) := V_0^{(m)}(\phi^{(m)} \circ \pi) + \lambda^{(m)\top}\big(V^{(m)}(\phi^{(m)} \circ \pi) - c^{(m)}\big), \tag{6}$$

where $\lambda^{(m)} \in \mathbb{R}_+^{d_m}$ is the dual variable. In order to develop convergent primal-dual algorithms for solving the constrained problem, a fundamental problem that needs to be addressed is the duality gap of the above Lagrangian function. In conventional constrained convex optimization [9], strong duality can be achieved under a Slater's condition, which requires the existence of a strictly feasible point under the constraints. Recently, this condition has also been introduced to establish strong duality for single-agent constrained reinforcement learning [18, 19, 42]. For constrained Markov game, we introduce the following Slater-type condition to study its duality gap.

Assumption 1 (Slater's condition)**.** *For any agent m and any joint policy π, there exists a stochastic modification $\phi^{(m)}$ such that $V^{(m)}(\phi^{(m)} \circ \pi) - c^{(m)} \geq \xi^{(m)}$ for some strictly positive $\xi^{(m)}$.*

The above Slater's condition assumes that agents can always find a strictly feasible stochastic modification, regardless of the joint policy. Such a condition is closely related to the existing Slater-type condition for constrained Markov games [3, 4, 45], which replaces the modification operator $\phi^{(m)}$ by an independent local policy $\widetilde{\pi}^{(m)}$ and is used to study NE. With the above Slater's condition, we are able to establish the following strong duality result for constrained Markov game.

Theorem 2 (Strong duality)**.** *Let Assumption 1 hold. Then, strong duality holds for constrained Markov games, i.e., for any joint policy π and any agent m,*

$$\max_{\phi^{(m)}} \min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)}) = \min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} \max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)}). \tag{7}$$

To the best of our knowledge, such a strong duality result has not been established before for CE of constrained Markov games, and later we show that this is essential for developing a primal-dual algorithm for finding CE of constrained Markov games. In particular, compared to the strong duality results of single-agent constrained RL [2, 42] and those for finding Nash equilibrium (NE) of

constrained Markov game [30], the proof of Theorem 1 for finding CE of constrained Markov game is substantially different and more challenging, as we elaborate below.

Technical novelty. Note that the constrained settings considered in the aforementioned existing works [2, 30, 42] lead to Lagrangian functions defined over the entire unconstrained policy space, which is clearly convex and helps establish the strong duality result. As a comparison, in order to find CE, our Lagrangian function $L^{(m)}$ is a nonlinear function in $\phi^{(m)}$ and is defined over the set of modified policies $\{\phi^{(m)} \circ \pi : \phi^{(m)} \text{ is stochastic modification}\}$ for a fixed π. Such a set is a subset of the entire policy space and is not easy to analyze directly. To overcome this challenge, we first rewrite $L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)})$ as $\widetilde{L}^{(m)}(p_{\phi^{(m)}\circ\pi}, \lambda^{(m)})$ (see eq. (23)), which is a function of $p_{\phi^{(m)}\circ\pi}$ – the probability measure of the episode $(s_{1:H}, a_{1:H})$ induced by the modified policy $\phi^{(m)} \circ \pi$.[1] Since we showed that $\widetilde{L}^{(m)}$ is linear in both of its arguments, strong duality would follow if we can prove that the domain of the first argument $\mathcal{X} := \{p_{\phi^{(m)}\circ\pi} : \phi^{(m)} \text{ is stochastic modification}\}$ is a convex and compact set, and the proof of this result requires substantial effort. More specifically, to prove convexity we need to show that the convex combination $\lambda p_{\phi^{(m)}\circ\pi} + (1-\lambda) p_{\widetilde{\phi}^{(m)}\circ\pi}$ for any $\lambda \in [0, 1]$ and any modifications $\phi^{(m)}, \widetilde{\phi}^{(m)}$ is a valid probability measure of the episode induced by a certain modified policy, and to prove compactness we need to show the closedness of any converging sequence $\{p_{\phi^{(m)}_{[k]}\circ\pi}\}_k$. To prove these claims, we need to study the fundamental properties of the measure p_π and the modified measure $p_{\phi^{(m)}\circ\pi}$ in Lemma A.1 and Lemma B.1.

4.2 A Primal-Dual Algorithm for Finding Correlated Equilibrium

Based on the strong duality result in Theorem 2, finding CE of constrained Markov game is equivalent to solving the minimax optimization problem in eq. (7) for all agents m, which can be effectively solved by a primal-dual algorithm that alternatively updates the primal variable $\phi^{(m)}$ and the dual variable $\lambda^{(m)}$. While $\lambda^{(m)}$ can be updated using the standard projected dual descent update, the update of $\phi^{(m)}$ requires further exploration as we elaborate below.

Specifically, with a fixed $\lambda^{(m)}$, the primal update needs to solve the optimization problem $\max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)})$. Denote $r_h^{(m)}(s, a) := [r_{1,h}^{(m)}(s, a), \ldots, r_{d_m,h}^{(m)}(s, a)]^\top$ as the collection of the rewards associated with the constraints, and define the following weighted surrogate reward $R_{\lambda,h}^{(m)}$ and its associated value function $V_\lambda^{(m)}$.

$$R_{\lambda,h}^{(m)}(s_h, a_h) := r_{0,h}^{(m)}(s_h, a_h) + \lambda^{(m)\top} r_h^{(m)}(s_h, a_h), \tag{8}$$

$$V_\lambda^{(m)}(\pi) := \mathbb{E}_\pi\Big[\sum_{h=1}^{H} R_{\lambda,h}^{(m)}(s_h, a_h)\Big| s_1 \sim \rho\Big]. \tag{9}$$

Then, the optimization problem of the primal update can be rewritten as follows.

$$\max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)}) = \max_{\phi^{(m)}} \Big\{V_\lambda^{(m)}(\phi^{(m)} \circ \pi)\Big\} - \lambda^{(m)\top} c^{(m)}. \tag{10}$$

Interestingly, the problem $\max_{\phi^{(m)}} V_\lambda^{(m)}(\phi^{(m)} \circ \pi)$ can be understood as finding CE of an unconstrained Markov game with associated reward $R_{\lambda,h}^{(m)}$ and value function $V_\lambda^{(m)}$. Therefore, to perform the primal update, we can apply any black-box algorithm for finding CE of unconstrained Markov games to update the policy. In particular, some existing popular algorithms include the V-learning algorithm [31, 46] and the Nash Value Iteration algorithm [37], both of which can output a policy that achieves ϵ-duality gap (defined in eq. (11)) with provable finite-time complexities.

We summarize this primal-dual algorithm in Algorithm 1, which alternatively updates the policy π_t via finding an approximated CE of an unconstrained Markov game (with fixed $\lambda_t^{(m)}$), and updates $\lambda_t^{(m)}$ via projected dual descent (with fixed π_t). In particular, the projection set $\Lambda^{(m)} \subset \mathbb{R}_+^{d_m}$ must contain the optimal dual variable and also be bounded to ensure a bounded surrogate reward $R_{\lambda,h}^{(m)}$. We give a specific choice of $\Lambda^{(m)}$ in the next convergence analysis Section 5.

[1] [30, 42] use occupation measure induced by policy instead of the joint probability measure p_π, as they consider stationary Markov policies that select action a_t based on the current state s_t.

Algorithm 1 Primal-Dual Algorithm for Finding CE of Constrained Markov Game

Input: $\epsilon, \eta > 0; \Lambda^{(m)} \subset \mathbb{R}_+^{d_m}$ for all $m = 1, ..., M$.

Initialize: $\lambda_0^{(m)} = 0$ for all $m = 1, ..., M$.

for *iterations* $t = 0, 1, \ldots, T-1$ **do**

- Let $\lambda^{(m)} \leftarrow \lambda_t^{(m)}$. Solve unconstrained Markov game with rewards $R_{\lambda,h}^{(m)}$ (defined in eq. (8)) and obtain an approximate CE policy π_t that satisfies

$$\max_{\phi^{(m)}} V_\lambda^{(m)}(\phi^{(m)} \circ \pi_t) - V_\lambda^{(m)}(\pi_t) \leq \epsilon, \quad \forall m = 1, ..., M. \tag{11}$$

- Update $\lambda_t^{(m)}$ via the following projected dual descent step

$$\lambda_{t+1}^{(m)} = \text{proj}_{\Lambda^{(m)}}\Big(\lambda_t^{(m)} - \eta\big(V^{(m)}(\pi_t) - c^{(m)}\big)\Big), \quad \forall m = 1, ..., M. \tag{12}$$

end

Output: $\pi_{\widetilde{t}}$ where $\widetilde{t}$ is sampled from $\{0, 1, \ldots, T-1\}$ uniformly at random.

Comparison to the existing art. Some existing works have developed other primal-dual algorithms with different primal policy improvement updates for constrained Markov games. For example, value iteration (Algorithm 11.1 of [62]), simultaneous perturbation stochastic approximation (Algorithm 11.2 of [62]) and policy gradient [30] are used as the primal update to find NE of constrained Markov games, and Q-learning [26] is used as the primal update to find CE of the Lagrangian function associated with the constrained Markov game. However, none of these works have established non-asymptotic convergence rate and sample complexity for the proposed algorithms. As a comparison, our Algorithm 1 can use any black-box primal update that can compute an approximate CE of unconstrained Markov games, and enjoys a provable non-asymptotic convergence rate and low sample complexity as we show in the next section.

5 Convergence Analysis and Sample Complexity

In this section, we analyze the non-asymptotic convergence rate of the primal-dual Algorithm 1. Then, we adopt the V-learning algorithm [31] as the primal update of Algorithm 1, and derive the overall sample complexity required by Algorithm 1 to achieve an approximate CE with ϵ duality gap.

We adopt the following assumption of bounded reward that is commonly adopted in the analysis of constrained RL [18, 19, 42], Markov game [31, 37, 46] and constrained Markov game [26].

Assumption 2 (Bounded reward)**.** *For all agents $m = 1, ..., M$, there exists constant $r_{j,\max}^{(m)} > 0$ such that $0 \leq r_{j,h}^{(m)} \leq r_{j,\max}^{(m)}$ for all $1 \leq h \leq H$ and $0 \leq j \leq d_m$.*

Under Assumptions 1 and 2, we choose the following projection set $\Lambda^{(m)}$ for Algorithm 1. In particular, we prove in Lemma E.1 that such a set contains the optimal dual variable.

$$\Lambda^{(m)} := \left[0, \frac{2Hr_{0,\max}^{(m)}}{\xi_1^{(m)}}\right] \times \left[0, \frac{2Hr_{0,\max}^{(m)}}{\xi_2^{(m)}}\right] \times \ldots \times \left[0, \frac{2Hr_{0,\max}^{(m)}}{\xi_{d_m}^{(m)}}\right] \subset \mathbb{R}_+^{d_m}. \tag{13}$$

Theorem 3 (Convergence rate)**.** *Let Assumptions 1 and 2 hold. Run Algorithm 1 for T iterations with stepsize $\eta = \frac{1}{\sqrt{T}}$ and projection set $\Lambda^{(m)}$ defined in eq.* (13)*. Then, the output policy $\pi_{\widetilde{t}}$ achieves the following duality gap (defined in eq.* (5)*).*

$$\mathbb{E}\big[D^{(m)}(\pi_{\widetilde{t}})\big] \leq \frac{2(HR_{\max}^{(m)})^2}{\sqrt{T}} + \epsilon, \quad \forall m = 1, ..., M, \tag{14}$$

where $R_{\max}^{(m)} := \sqrt{\sum_{j=1}^{d_m}(r_{j,\max}^{(m)})^2}$. Moreover, define the constraint violation of the output policy $\pi_{\widetilde{t}}$ as $W^{(m)}(\pi_{\widetilde{t}}) := \sum_{j=1}^{d_m}(\xi_j^{(m)})^{-1}\big(c_j^{(m)} - V_j^{(m)}(\pi_{\widetilde{t}})\big)_+$, where $(a)_+ := \max\{0, a\}$. We obtain that

$$\mathbb{E}\big[W^{(m)}(\pi_{\widetilde{t}})\big] \leq \frac{2HR_{\max}^{(m)}}{\sqrt{T}}\sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2} + \frac{2H(R_{\max}^{(m)})^2}{r_{0,\max}^{(m)}\sqrt{T}} + \frac{\epsilon}{Hr_{0,\max}^{(m)}}, \quad \forall m = 1, ..., M. \tag{15}$$

The above theorem shows that both the duality gap and the constraint violation associated with the output policy converge at the rate $\mathcal{O}(\frac{1}{\sqrt{T}} + \epsilon)$, where ϵ is the target duality gap of the unconstrained Markov game in the primal update. In particular, a larger slackness $\{\xi_j^{(m)}\}_{j,m}$ implies a faster convergence of the constraint violation. These convergence rates are comparable to those of primal-dual algorithms for single-agent constrained RL [18, 19, 36], and the analysis is a generalization of the single-agent case. Specifically, the updated policy π_t in the primal-dual algorithm for single-agent constrained RL is an approximate maximizer of the value function under the constraints, whereas the π_t in our case corresponds to an approximate CE at which all agents can benefit little (up to ϵ) by modifying their own actions alone. Particularly for each agent, the duality gap convergence criterion in eq. (11) can be understood as a convergence criterion of single-agent constrained RL, and there the proof logic partially follows the analysis of single-agent case [19].

As our Algorithm 1 can adopt any existing algorithm for unconstrained Markov games in the primal update, we consider adopting the V-learning algorithm developed in [31]. This is a decentralized algorithm that incrementally updates the state value functions, which are further used to update the agents' policies via adversarial bandit algorithms. In particular, thanks to the use of state value function, it requires much less memory than other algorithms based on state-action value functions [26]. In Theorem 7 of [31], it is proved that the V-learning algorithm achieves an approximate CE with ϵ duality gap for any unconstrained Markov game with the state-of-the-art sample complexity $\widetilde{\mathcal{O}}(H^5SA^2\epsilon^{-2})$, where $\widetilde{\mathcal{O}}$ hides the logarithm factors and $S = |\mathcal{S}|$, $A = \max_m |\mathcal{A}^{(m)}|$ correspond to the cardinality of the state space and action space, respectively. We then obtain the following sample complexity result.

Corollary 5.1 (Sample complexity)**.** *Let Theorem 3 hold and assume $r_{0,\max}^{(m)} \geq \frac{1}{H}$. Apply the V-learning algorithm [31] to the primal update of Algorithm 1 and choose $T = \max_m 4\epsilon^{-2}(HR_{\max}^{(m)})^2\big(\sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2} + HR_{\max}^{(m)}\big)^2$. Then, the output policy $\pi_{\tilde{t}}$ achieves $\max\big(\mathbb{E}[D^{(m)}(\pi_{\tilde{t}})], \mathbb{E}[W^{(m)}(\pi_{\tilde{t}})]\big) \leq 2\epsilon$ with the sample complexity $\widetilde{\mathcal{O}}(H^9SA^2\epsilon^{-4})$.*

To the best of our knowledge, this is the first finite-time sample complexity result establsihed for constrained Markov games. Next, we compare our sample complexity with that of two popular RL algorithms. The first one is the V-learning algorithm [31] developed for finding CE of unconstrained Markov games, and it achieves the state-of-the-art sample complexity $\widetilde{\mathcal{O}}(H^5SA^2\epsilon^{-2})$. The information theoretical lower bound is $\Omega(H^3SA\epsilon^{-2})$ [31]. As a comparison, the sample complexity of our primal-dual V-learning algorithm is higher since we require additional projected dual descent steps (12) to account for the constraints. The other one is the basic primal-dual algorithm for solving single-agent constrained RL problems, which uses natural policy gradient in the primal update and achieves the state-of-the-art sample complexity $\mathcal{O}(\epsilon^{-4})$ [7]. Such a complexity has the same dependence on ϵ as the sample complexity of our algorithm.

6 Conclusion

In this work, we propose a surrogate notion of correlated equilibrium (CE) for constrained Markov games and show that it possesses a fundamentally different modification structure from that of CE of Markov game. Moreover, we prove that the corresponding Lagrangian function has zero duality gap. Based on this result, we develop the first primal-dual algorithm that provably converges to CE of constrained Markov games with the convergence rate $\mathcal{O}(\frac{1}{\sqrt{T}})$. Moreover, when adopting the V-learning algorithm as the subroutine in the primal update, our algorithm achieves an approximate CE with ϵ duality gap with the sample complexity $\mathcal{O}(H^9SA^2\epsilon^{-4})$. **Limitations:** In this work, we only develop the basic primal-dual algorithm for constrained Markov games. An interesting future direction is to develop other advanced algorithms such as Nesterov's momentum-accelerated primal-dual algorithms. **Negative social impacts:** This is a fundamental theoretical study and does not have any potential negative social impact.

Acknowledgment

The work of Z. Chen, S. Ma and Y. Zhou was supported in part by U.S. National Science Foundation under the Grants CCF-2106216 and DMS-2134223.

References

[1] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *Proc. International conference on machine learning (ICML)*, pages 22–31, 2017.

[2] E. Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

[3] E. Altman, K. Avrachenkov, N. Bonneau, M. Debbah, R. El-Azouzi, and D. S. Menasche. Constrained cost-coupled stochastic games with independent state processes. *Operations Research Letters*, 36(2):160–164, 2008.

[4] E. Altman and A. Shwartz. Constrained markov games: Nash equilibria. In *Advances in dynamic games and applications*, pages 213–221. 2000.

[5] E. Altman and E. Solan. Constrained games: The impact of the attitude to adversary's constraints. *IEEE Transactions on Automatic Control*, 54(10):2435–2440, 2009.

[6] J. Alvarez-Mena and O. Hernández-Lerma. Existence of nash equilibria for constrained stochastic games. *Mathematical Methods of Operations Research*, 63(2):261–285, 2006.

[7] Q. Bai, A. S. Bedi, and V. Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. *arXiv preprint arXiv:2206.05850*, 2022.

[8] Y. Bai, C. Jin, H. Wang, and C. Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. *Proc. Advances in Neural Information Processing Systems (Neurips)*, 34, 2021.

[9] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[10] A. Celli, A. Marchesi, G. Farina, and N. Gatti. No-regret learning dynamics for extensive-form correlated equilibrium. *Proc. Advances in Neural Information Processing Systems (Neurips)*, 33:7722–7732, 2020.

[11] Z. Chen, S. Ma, and Y. Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum markov game. In *Proc. International Conference on Learning Representations (ICLR)*, 2022.

[12] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Proc. Advances in neural information processing systems (Neurips)*, volume 31, 2018.

[13] Y. Chow, O. Nachum, A. Faust, E. Duenez-Guzman, and M. Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *ArXiv:1901.10031*, 2019.

[14] K. Cui and H. Koeppl. Learning graphon mean field games and approximate nash equilibria. *ArXiv:2112.01280*, 2021.

[15] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa. Safe exploration in continuous action spaces. *ArXiv:1801.08757*, 2018.

[16] C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (Neurips)*, volume 33, 2020.

[17] X. Deng, Y. Li, D. H. Mguni, J. Wang, and Y. Yang. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *ArXiv:2109.01795*, 2021.

[18] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3304–3312, 2021.

[19] D. Ding, K. Zhang, T. Basar, and M. R. Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *Proc. International Conference on Neural Information Processing Systems (Neurips)*, volume 2020, 2020.

[20] F. Dufour and T. Prieto-Rumeau. Stationary markov nash equilibria for nonzero-sum constrained arat markov games. *ArXiv:2109.13003*, 2021.

[21] J. Fan, Z. Wang, Y. Xie, and Z. Yang. A theoretical analysis of deep q-learning. In *Proc. Learning for Dynamics and Control (L4DC)*, pages 486–489, 2020.

[22] J. Filar and K. Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[23] A. Gilpin and T. Sandholm. Finding equilibria in large sequential games of imperfect information. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 160–169, 2006.

[24] E. Gómez-Ramırez, K. Najim, and A. Poznyak. Saddle-point calculation for constrained finite markov chains. *Journal of Economic Dynamics and Control*, 27(10):1833–1853, 2003.

[25] X. Guo, A. Hu, R. Xu, and J. Zhang. Learning mean-field games. *Proc. Advances in Neural Information Processing Systems (Neurips)*, 32:4966–4976, 2019.

[26] V. Hakami and M. Dehghan. Learning stationary correlated equilibria in constrained general-sum stochastic games. *IEEE Transactions on Cybernetics*, 46(7):1640–1654, 2015.

[27] M. K. Hanawal, M. J. Abdel-Rahman, and M. Krunz. Joint adaptation of frequency hopping and transmission rate for anti-jamming wireless systems. *IEEE Transactions on Mobile Computing*, 15(9):2247–2259, 2015.

[28] A. Jaśkiewicz and A. S. Nowak. Constrained discounted stochastic games. *ArXiv:2112.07960*, 2021.

[29] A. Jaśkiewicz and A. S. Nowak. On approximate and weak correlated equilibria in constrained discounted stochastic games. *ArXiv:2201.01065*, 2022.

[30] X. Jiang, S. Chen, J. Yang, H. Hu, and Z. Zhang. Finding the equilibrium for continuous constrained markov games under the average criteria. *IEEE Transactions on Automatic Control*, 65(12):5399–5406, 2020.

[31] C. Jin, Q. Liu, Y. Wang, and T. Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *ArXiv:2110.14555*, 2021.

[32] Y. Jin, V. Muthukumar, and A. Sidford. The complexity of infinite-horizon general-sum stochastic games. *ArXiv:2204.04186*, 2022.

[33] D. Koller and N. Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and economic behavior*, 4(4):528–552, 1992.

[34] V. Kononen and E. Oja. Asymmetric multiagent reinforcement learning in pricing applications. In *Proc. IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 2, pages 1097–1102, 2004.

[35] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *ArXiv:2106.01969*, 2021.

[36] T. Li, Z. Guan, S. Zou, T. Xu, Y. Liang, and G. Lan. Faster algorithm and sharper analysis for constrained markov decision process. *ArXiv:2110.10351*, 2021.

[37] Q. Liu, T. Yu, Y. Bai, and C. Jin. A sharp analysis of model-based reinforcement learning with self-play. In *Proc. International Conference on Machine Learning (ICML)*, pages 7001–7010, 2021.

[38] Y. Liu, J. Ding, and X. Liu. Ipo: Interior-point policy optimization under constraints. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 4940–4947, 2020.

[39] W. Mao and T. Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

[40] H. Moulin and J.-P. Vial. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3):201–221, 1978.

[41] M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT press, 1994.

[42] S. Paternain, L. F. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained reinforcement learning has zero duality gap. In *Proc. International Conference on Neural Information Processing Systems (Neurips)*, pages 7555–7565, 2019.

[43] L. S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[44] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[45] V. V. Singh and N. Hemachandra. A mathematical programming based characterization of nash equilibria of some constrained stochastic games. *ArXiv:1206.1672*, 2012.

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

[47] A. Stooke, J. Achiam, and P. Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *Proc. International Conference on Machine Learning (ICML)*, pages 9133–9143, 2020.

[48] H. Tembine, E. Altman, R. ElAzouzi, and Y. Hayel. Stochastic population games with individual independent states and coupled constraints. In *Proc. the 3rd International Conference on Performance Evaluation Methodologies and Tools*, pages 1–10, 2008.

[49] C. Tessler, D. J. Mankowitz, and S. Mannor. Reward constrained policy optimization. In *Proc. International Conference on Learning Representations (ICML)*, 2018.

[50] D. Vasal. Stochastic stackelberg games. *ArXiv:2005.01997*, 2020.

[51] Q. Wei. Constrained expected average stochastic games for continuous-time jump processes. *Applied Mathematics & Optimization*, 83(3):1277–1309, 2021.

[52] Q. Wei and X. Chen. Constrained stochastic games with the average payoff criteria. *Operations Research Letters*, 43(1):83–88, 2015.

[53] Q. Xie, Y. Chen, Z. Wang, and Z. Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Proc. Conference on learning theory (COLT)*, pages 3674–3682, 2020.

[54] T. Xu, Y. Liang, and G. Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *Proc. International Conference on Machine Learning (ICML)*, pages 11480–11491, 2021.

[55] V. G. Yaji and S. Bhatnagar. Necessary and sufficient conditions for optimality in constrained general sum stochastic games. *Systems & Control Letters*, 85:8–15, 2015.

[56] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge. Projection-based constrained policy optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[57] M. Yu, Z. Yang, M. Kolar, and Z. Wang. Convergent policy optimization for safe reinforcement learning. *Proc. Advances in Neural Information Processing Systems (Neurips)*, 32, 2019.

[58] R. Zhang, Z. Ren, and N. Li. Gradient play in multi-agent markov stochastic games: Stationary points and convergence. *ArXiv:2106.00198*, 2021.

[59] W. Zhang. Continuous-time constrained stochastic games under the discounted cost criteria. *Applied Mathematics & Optimization*, 77(2):275–296, 2018.

[60] W. Zhang, B. Wang, and D. Chen. Continuous-time constrained stochastic games with average criteria. *Operations Research Letters*, 46(1):109–115, 2018.

[61] X. Zhang, K. Zhang, E. Miehling, and T. Basar. Non-cooperative inverse reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (Neurips)*, volume 32, pages 9487–9497, 2019.

[62] Y. Zhang and M. Guizani. *Game theory for wireless communications and networking*. CRC press, 2011.

[63] Y. Zhao, Y. Tian, J. D. Lee, and S. S. Du. Provably efficient policy gradient methods for two-player zero-sum markov games. *ArXiv:2102.08903*, 2021.

[64] Y. Zhu and D. Zhao. Online minimax q network learning for two-player zero-sum markov games. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[65] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the "Conclusion" section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the "Conclusion" section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Policy-Induced Joint Measure **14**

B Properties of Modification **15**

C Proof of Theorem 1 **17**
C.1 Proof of item 1 for unconstrained Markov game 17
C.2 Proof of item 2 for constrained Markov game 18

D Proof of Theorem 2 **19**

E The Range of the Optimal Dual Variable **20**

F Proof of Theorem 3 **20**

G Proof of Corollary 5.1 **23**

A Policy-Induced Joint Measure

We introduce policy-induced joint measure as a useful tool to help prove Theorems 1 and 2. Specifically, for any h, h', we define $p_\pi(s_{1:h}, a_{1:h'})$ as the joint distribution of $(s_{1:h}, a_{1:h'})$ induced by policy π. Based on the underlying Markov decision process, it can be easily verified that p_π takes the following forms.

$$p_\pi(s_1) = \rho(s_1), \tag{16}$$

$$p_\pi(s_{1:h}, a_{1:h-1}) = p_\pi(s_{1:h-1}, a_{1:h-1})\mathcal{P}(s_h|s_{h-1}, a_{h-1}), \tag{17}$$

$$p_\pi(s_{1:h}, a_{1:h}) = \rho(s_1)\pi_1(a_1|s_1)\prod_{h'=2}^{h}\mathcal{P}(s_{h'}|s_{h'-1}, a_{h'-1})\pi_{h'}(a_{h'}|s_{1:h'}, a_{1:h'-1}). \tag{18}$$

Eqs. (17) and (18) further imply that

$$p_\pi(s_{1:h}, a_{1:h}) = p_\pi(s_{1:h}, a_{1:h-1})\pi(a_h|s_{1:h}, a_{1:h-1}). \tag{19}$$

Conversely, given a proper joint probability measure p, we can infer its corresponding inducing policy π as follows.

Lemma A.1. *Consider any joint probability measure p that satisfies eqs.* (16) *and* (17) *(replace all p_π with p). Then, the following policy π induces p.*

$$\pi_h(a_h|s_{1:h}, a_{1:h-1}) = \begin{cases} \dfrac{p(s_{1:h}, a_{1:h})}{p(s_{1:h}, a_{1:h-1})}, & \textit{if } p(s_{1:h}, a_{1:h-1}) > 0 \\ \textit{arbitrary distribution}, & \textit{if } p(s_{1:h}, a_{1:h-1}) = 0 \end{cases}. \tag{20}$$

Proof. It suffices to prove that $p_\pi(s_{1:h}, a_{1:h}) = p(s_{1:h}, a_{1:h})$ for any $s_{1:h}, a_{1:h}$, i.e., p is exactly the joint measure p_π induced by π. We consider the following two cases.

Case 1: $p(s_{1:h}, a_{1:h-1}) > 0$. In this case, we must have that $p(s_{1:h'}, a_{1:h'-1}) > 0$ for any $1 \le h' \le h$. Therefore, by eq. (20) we have $\pi_{h'}(a_{h'}|s_{1:h'}, a_{1:h'-1}) = \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'}, a_{1:h'-1})}$ for any $1 \le h' \le h$. Substitute this policy π into eq. (18) and note that $\rho(s_1) = p(s_1)$, we obtain that

$$p_\pi(s_{1:h}, a_{1:h}) = p(s_1)\frac{p(s_1, a_1)}{p(s_1)}\prod_{h'=2}^{h}\mathcal{P}(s_{h'}|s_{h'-1}, a_{h'-1})\frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'}, a_{1:h'-1})}$$

$$\overset{(i)}{=} p(s_1, a_1) \prod_{h'=2}^{h} \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'-1}, a_{1:h'-1})} = p(s_{1:h}, a_{1:h}),$$

where (i) follows from eq. (17) (replace all p_π with p).

Case 2: $p(s_{1:h}, a_{1:h-1}) = 0$. In this case, we have that $p(s_{1:h}, a_{1:h}) = 0$. Hence, it suffices to prove that $p_\pi(s_{1:h}, a_{1:h}) = 0$ as well. We further consider the following two subcases.

(Case 2.1) If $p(s_1) = \rho(s_1) = 0$, then $p_\pi(s_{1:h}, a_{1:h}) = 0$ by substituting $\rho(s_1) = 0$ into eq. (18).

(Case 2.2) If $p(s_1) = \rho(s_1) > 0$ and because Case 2 assumes that $p(s_{1:h}, a_{1:h-1}) = 0$, then there must exist $1 \le h' \le h-1$ such that $p(s_{1:h'}, a_{1:h'-1}) > 0$ and $p(s_{1:h'+1}, a_{1:h'}) = 0$. On the other hand, note that eq. (18) implies that $p_\pi(s_{1:h}, a_{1:h})$ contains the following multiplicative factor

$$\begin{aligned} &\pi_{h'}(a_{h'}|s_{1:h'}, a_{1:h'-1})\mathcal{P}(s_{h'+1}|s_{h'}, a_{h'}) \\ &= \mathcal{P}(s_{h'+1}|s_{h'}, a_{h'}) \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'}, a_{1:h'-1})} \overset{(i)}{=} \frac{p(s_{1:h'+1}, a_{1:h'})}{p(s_{1:h'}, a_{1:h'-1})} = 0 \end{aligned} \tag{21}$$

where (i) uses eq. (17). Thus we conclude that $p_\pi(s_{1:h}, a_{1:h}) = 0$. □

With the policy-induced joint measure p_π, we can rewrite the value function $V_j^{(m)}(\pi) := \mathbb{E}_\pi\big[\sum_{h=1}^{H} r_{j,h}^{(m)} \big| s_1 \sim \rho\big]$ and the Lagrangian function eq. (6) as follows.

$$V_j^{(m)}(\pi) = \widetilde{V}_j^{(m)}(p_\pi) := \sum_{s_{1:H}, a_{1:H}} p_\pi(s_{1:H}, a_{1:H}) \sum_{h=1}^{H} r_{j,h}^{(m)}(s_h, a_h), \tag{22}$$

$$L^{(m)}(\pi, \lambda^{(m)}) = \widetilde{L}^{(m)}(p_\pi, \lambda^{(m)}) := \widetilde{V}_0^{(m)}(p_\pi) + \sum_{j=1}^{d_m} \lambda_j \big(\widetilde{V}_j^{(m)}(p_\pi) - c_j^{(m)}\big). \tag{23}$$

Thus, both the value function $V_j^{(m)}(\pi)$ and the Lagrangian function $L^{(m)}(\pi, \lambda^{(m)})$ can be rewritten as linear functions of p_π. Such a linear form helps simplify the problem and prove the key Theorems 1 and 2.

B Properties of Modification

In this section, we present some useful properties of the modification operator. Recall that for any policy π and any modification operator $\phi^{(m)}$, the modified policy $\phi_h^{(m)} \circ \pi_h$ at time step h is defined as follows: we first generate joint action $a_h = [a_h^{(m)}, a_h^{(\setminus m)}] \sim \pi_h(\cdot|s_{1:h}, a_{1:h-1})$. Then, $\phi_h^{(m)}$ randomly modifies $a_h^{(m)}$ to $\widetilde{a}_h^{(m)} \sim \phi_h^{(m)}(\cdot|s_{1:h}, a_{1:h-1}, a_h^{(m)})$. To summarize, the modified policy $\phi_h^{(m)} \circ \pi_h$ takes the following form.

$$\begin{aligned} &(\phi_h^{(m)} \circ \pi_h)\big([\widetilde{a}_h^{(m)}, a_h^{(\setminus m)}] | s_{1:h}, a_{1:h-1}\big) \\ &\qquad = \sum_{a_h^{(m)}} \phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \pi_h(a_h|s_{1:h}, a_{1:h-1}). \end{aligned} \tag{24}$$

Next, let $p_{\phi^{(m)} \circ \pi}$ and $p_{\widetilde{\phi}^{(m)} \circ \pi}$ be the joint measures induced by the modified policies $\phi^{(m)} \circ \pi$ and $\widetilde{\phi}^{(m)} \circ \pi$, respectively, for any $\phi^{(m)}, \widetilde{\phi}^{(m)}$. In the proof of Theorems 1 and 2, we introduce the following linear combination of these two joint measures.

$$p_\lambda := \lambda p_{\phi^{(m)} \circ \pi} + (1-\lambda) p_{\widetilde{\phi}^{(m)} \circ \pi}, \lambda \in \mathbb{R}. \tag{25}$$

We note that $\sum_{s_{1:h}, a_{1:h}} p_\lambda(s_{1:h}, a_{1:h}) = 1$ for any $\lambda \in \mathbb{R}$, so p_λ is also a proper probability measure if λ is selected such that $p_\lambda(s_{1:h}, a_{1:h}) \ge 0$ for any $s_{1:h}, a_{1:h}$. In this case, since the joint measures $p_{\phi^{(m)} \circ \pi}$ and $p_{\widetilde{\phi}^{(m)} \circ \pi}$ satisfy eqs. (16) and (17) by definition, it is easy to verify that p_λ also satisfies eqs. (16) and (17) and hence is a proper joint measure. Therefore, by Lemma A.1 we can find its inducing policy π_λ using eq. (20). Next, we show that such an inducing policy π_λ can actually be viewed as the policy π modified by a certain stochastic modification.

Lemma B.1. *Regarding the p_λ defined in eq.* (25)*, if $\lambda \in \mathbb{R}$ is selected such that the following two conditions hold for any $s_{1:h}, a_{1:h-1}, a_h^{(m)}, \widetilde{a}_h^{(m)}$:*

$$p_\lambda(s_{1:h}, a_{1:h-1}) \geq 0 \tag{26}$$

$$\begin{aligned}&\lambda p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})\\&\quad + (1-\lambda)p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\widetilde{\phi}_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \geq 0,\end{aligned} \tag{27}$$

then its inducing policy (defined by eq. (20)*) can be written as $\pi_\lambda = \phi_\lambda^{(m)} \circ \pi$, where the stochastic modification $\phi_\lambda^{(m)}$ takes the following form.*

$$\phi_{\lambda,h}^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) = \begin{cases}\frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})}\big[\lambda p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \\ \qquad +(1-\lambda)p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\widetilde{\phi}_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})\big], & \text{if } p_\lambda(s_{1:h}, a_{1:h-1}) > 0 \\ \\ \textit{Arbitrary distribution}, & \text{if } p_\lambda(s_{1:h}, a_{1:h-1}) = 0\end{cases} \tag{28}$$

Proof. We first show that $\phi_{\lambda,h}^{(m)}(\cdot|s_{1:h}, a_{1:h-1}, a_h^{(m)}))$ is a proper stochastic modification. By eq. (28), we only need to consider the case $p_\lambda(s_{1:h}, a_{1:h-1}) > 0$. In this case, based on the condition in eq. (27), we conclude that $\phi_{\lambda,h}^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \geq 0$. In addition,

$$\begin{aligned}&\sum_{\widetilde{a}_h^{(m)}\in\mathcal{A}^{(m)}} \phi_{\lambda,h}^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})\\&= \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})}\Big(\lambda p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1}) \sum_{\widetilde{a}_h^{(m)}\in\mathcal{A}^{(m)}} \phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})\\&\quad + (1-\lambda)p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1}) \sum_{\widetilde{a}_h^{(m)}\in\mathcal{A}^{(m)}} \widetilde{\phi}_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})\Big)\\&= \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})}\big(\lambda p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1}) + (1-\lambda)p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\big) = 1.\end{aligned}$$

Therefore, ϕ_λ is a proper stochastic modification.

Next, we prove that the policy that induces p_λ takes the form $\pi_\lambda = \phi_\lambda^{(m)} \circ \pi$. We consider two cases.

Case 1: $p_\lambda(s_{1:h}, a_{1:h-1}) > 0$. In this case, we obtain that

$$\begin{aligned}&\pi_{\lambda,h}(a_h|s_{1:h}, a_{1:h-1})\\&\overset{(i)}{=} \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})}\big(\lambda p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h}) + (1-\lambda)p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h})\big)\\&\overset{(ii)}{=} \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})}\big(\lambda p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})(\phi_h^{(m)}\circ\pi_h)(a_h|s_{1:h}, a_{1:h-1})\\&\quad + (1-\lambda)p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})(\widetilde{\phi}_h^{(m)}\circ\pi_h)(a_h|s_{1:h}, a_{1:h-1})\big)\\&\overset{(iii)}{=} \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})}\sum_{\widetilde{a}_h^{(m)}}\big(\lambda p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\pi_h^{(m)}([\widetilde{a}_h^{(m)}, a_h^{(\backslash m)}]|s_{1:h}, a_{1:h-1})\\&\quad \phi_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widetilde{a}_h^{(m)}) + (1-\lambda)p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\\&\quad \pi_h^{(m)}([\widetilde{a}_h^{(m)}, a_h^{(\backslash m)}]|s_{1:h}, a_{1:h-1})\widetilde{\phi}_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widetilde{a}_h^{(m)})\big)\\&\overset{(iv)}{=} \sum_{\widetilde{a}_h^{(m)}}\pi_h^{(m)}([\widetilde{a}_h^{(m)}, a_h^{(\backslash m)}]|s_{1:h}, a_{1:h-1})\phi_{\lambda,h}^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widetilde{a}_h^{(m)})\end{aligned}$$

where (i) uses eqs. (20) and (25), (ii) uses eq. (19), (iii) uses eq. (24), and (iv) uses eq. (28). This proves the claim due to eq. (24).

Case 2: $p_\lambda(s_{1:h}, a_{1:h-1}) = 0$. In this case, both $\phi^{(m)}_{\lambda,h}$ and π_λ can be arbitrarily defined. Hence, we can simply define π_λ by eq. (24). □

C Proof of Theorem 1

C.1 Proof of item 1 for unconstrained Markov game

Throughout the proof, for any policy π, we denote $\widetilde{\phi}^{(m)}$ as the *optimal* stochastic modification associated with π, i.e., $V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi)$ achieves the maximum value over all stochastic modifications. In order for π to be a CE, it must satisfy $V_0^{(m)}(\pi) \geq V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi)$.

If for any optimal stochastic modification $\widetilde{\phi}^{(m)}$ associated with π, we can construct a corresponding deterministic modification $\phi^{(m)}$ such that $V_0^{(m)}(\phi^{(m)} \circ \pi) = V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi)$, then the condition of item 1 guarantees that π is a CE and then item 1 is proved.

Next, for any policy π and any associated optimal stochastic modification $\widetilde{\phi}^{(m)}$, we construct a deterministic modification $\phi^{(m)}$ as follows: for any $s_{1:h}, a_{1:h-1}, a_h^{(m)}$, select an arbitrary $\widetilde{a}_h^{(m)}$ such that $\widetilde{\phi}^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) > 0$ (this always exists) and then simply define $\phi^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) = 1$, and 0 otherwise. It suffices to prove that $V_0^{(m)}(\phi^{(m)} \circ \pi) = V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi)$ for any π satisfying the condition of item 1.

To proceed, we claim that one can find $\lambda < 0$ such that the joint measure $p_\lambda := \lambda p_{\phi^{(m)} \circ \pi} + (1 - \lambda) p_{\widetilde{\phi}^{(m)} \circ \pi}$ satisfies eqs. (26) and (27). We will prove the validity of this claim later. Suppose this claim holds. Then based on Lemma B.1, the inducing policy of p_λ takes the form $\pi_\lambda = \phi_\lambda^{(m)} \circ \pi$, where $\phi_\lambda^{(m)}$ is defined by eq. (28). Then, we obtain that

$$
\begin{aligned}
V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi) &\overset{(i)}{\geq} V_0^{(m)}(\phi_\lambda^{(m)} \circ \pi) \overset{(ii)}{=} \widetilde{V}_0^{(m)}(p_\lambda) \\
&= \widetilde{V}_0^{(m)}\big(\lambda p_{\phi^{(m)} \circ \pi} + (1-\lambda) p_{\widetilde{\phi}^{(m)} \circ \pi}\big) \\
&\overset{(iii)}{=} \lambda \widetilde{V}_0^{(m)}(p_{\phi^{(m)} \circ \pi}) + (1-\lambda) \widetilde{V}_0^{(m)}(p_{\widetilde{\phi}^{(m)} \circ \pi}) \\
&\overset{(iv)}{=} \lambda V_0^{(m)}(\phi^{(m)} \circ \pi) + (1-\lambda) V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi),
\end{aligned}
$$

where (i) uses the optimality of $\widetilde{\phi}^{(m)}$, (ii)-(iv) use the linear form of $\widetilde{V}_0^{(m)}(p_\pi)$ defined in eq. (22). The above inequality along with $\lambda < 0$ implies that $V_0^{(m)}(\phi^{(m)} \circ \pi) \geq V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi)$. On the other hand, $V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi) \geq V_0^{(m)}(\phi^{(m)} \circ \pi)$ based on the optimality of the stochastic modification $\widetilde{\phi}^{(m)}$. Hence, $V_0^{(m)}(\phi^{(m)} \circ \pi) = V_0^{(m)}(\widetilde{\phi}^{(m)} \circ \pi)$ as desired. All left is to find $\lambda < 0$ such that the joint measure p_λ satisfies eqs. (26) and (27). We prove them as follows.

Proof of eq. (26): Recall that $p_\lambda = \lambda p_{\phi^{(m)} \circ \pi} + (1-\lambda) p_{\widetilde{\phi}^{(m)} \circ \pi}$ and $\lambda < 0$. If $p_{\phi^{(m)} \circ \pi}(s_{1:h}, a_{1:h-1}) = 0$, it is clear that eq. (26) holds. So we just need to consider the other case where $p_{\phi^{(m)} \circ \pi}(s_{1:h}, a_{1:h-1}) > 0$. In this case and by eq. (18), we must have that $\rho(s_1), (\phi_1^{(m)} \circ \pi_1)(a_1|s_1), \mathcal{P}(s_h|s_{h-1}, a_{h-1}), (\phi_h^{(m)} \circ \pi_h)(a_h|s_{1:h}, a_{1:h-1}) > 0$ for any $h = 2, \ldots, H$. Then, eq. (24) implies that for any $h = 2, \ldots, H$,

$$
\begin{aligned}
0 &< (\phi_h^{(m)} \circ \pi_h)\big(a_h|s_{1:h}, a_{1:h-1}\big) \\
&= \sum_{\widetilde{a}_h^{(m)}} \phi_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widetilde{a}_h^{(m)}) \pi_h\big(\big[\widetilde{a}_h^{(m)}, a_h^{(\backslash m)}\big]|s_{1:h}, a_{1:h-1}\big).
\end{aligned}
$$

Hence, there must exist $\widehat{a}_h^{(m)}$ such that $\phi_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widehat{a}_h^{(m)}) \pi_h\big(\big[\widehat{a}_h^{(m)}, a_h^{(\backslash m)}\big]|s_{1:h}, a_{1:h-1}\big) > 0$. As $\phi_h^{(m)}$ is the deterministic modification constructed at the beginning of this proof, we must have

$\phi_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widehat{a}_h^{(m)}) = 1$ and therefore the corresponding stochastic modification satisfies $\widetilde{\phi}_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widehat{a}_h^{(m)}) > 0$. Then, eq. (24) implies that

$$\begin{aligned}
&(\widetilde{\phi}_h^{(m)} \circ \pi_h)\big(a_h|s_{1:h}, a_{1:h-1}\big) \\
&= \sum_{\widetilde{a}_h^{(m)}} \widetilde{\phi}_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widetilde{a}_h^{(m)})\pi_h\big(\big[\widetilde{a}_h^{(m)}, a_h^{(\backslash m)}\big]|s_{1:h}, a_{1:h-1}\big) \\
&\geq \widetilde{\phi}_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \widehat{a}_h^{(m)})\pi_h\big(\big[\widehat{a}_h^{(m)}, a_h^{(\backslash m)}\big]|s_{1:h}, a_{1:h-1}\big) > 0.
\end{aligned}$$

Similarly, we can prove that $(\widetilde{\phi}_1^{(m)} \circ \pi_1)(a_1|s_1) > 0$ from $(\phi_1^{(m)} \circ \pi_1)(a_1|s_1) > 0$. Therefore, based on eq. (18), it is proved that whenever $p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1}) > 0$, we have

$$\begin{aligned}
&p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h}) \\
&= \rho(s_1)(\widetilde{\phi}_1^{(m)} \circ \pi_1)(a_1|s_1) \prod_{h'=2}^{h} \mathcal{P}(s_{h'}|s_{h'-1}, a_{h'-1})(\widetilde{\phi}_{h'}^{(m)} \circ \pi_{h'})(a_{h'}|s_{1:h'}, a_{1:h'-1}) > 0. \quad (29)
\end{aligned}$$

Therefore, eq. (26) holds for

$$0 > \lambda \geq -\frac{p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h})}{p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h}) - p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h})} := w(s_{1:h}, a_{1:h}),$$

for any $s_{1:h}, a_{1:h}$ whenever $p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h}) > p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h})$, which implies that $p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h}) > 0$ and therefore $p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h}) > 0$ based on eq. (29). Thus, we conclude that $w(s_{1:h}, a_{1:h}) < 0$. Consider the finite (and possibly empty) set $A_1 := \{w(s_{1:h}, a_{1:h}) : 1 \leq h \leq H, p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h}) > p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h})\}$. If it is non-empty, eq. (26) holds for all $0 > \lambda \geq \max A_1$ for constant $\max A_1 < 0$; Otherwise, eq. (26) holds for all $\lambda < 0$.

Proof of eq. (27): If $p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) = 0$ and $\lambda < 0$, then eq. (27) holds. Consider the other case $p_{\phi^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1})\phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) > 0$ and $\lambda < 0$. In this case, we have $p_{\widetilde{\phi}^{(m)}\circ\pi}(s_{1:h}, a_{1:h-1}) > 0$ as proved in the proof of eq. (26), and we also have $\phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) = 1 > 0$ and thus $\widetilde{\phi}_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) > 0$ based on the construction of $\phi_h^{(m)}$. Following the same proof logic as that of eq. (26), we can find a constant $A_2 < 0$ such that eq. (27) holds for all $0 > \lambda \geq A_2$.

In summary, we have proved that there exists $\lambda < 0$ that guarantees both eqs. (26) and (27).

C.2 Proof of item 2 for constrained Markov game

Here we construct a counter example to prove item 2. Consider a constrained Markov game with only one state $\mathcal{S} = \{s\}$, two agents with action spaces $\mathcal{A}_1 = \mathcal{A}_2 = \{0, 1\}$ and horizon $H = 1$. For simplicity, we drop the time step index $h = 1$ and state s in all notations throughout this example. Specifically, we denote $\pi(a^{(1)}, a^{(2)})$, $\pi^{(m)}(a^{(m)})$, $\phi^{(m)}(\widetilde{a}^{(m)}|a^{(m)})$, $m = 1, 2$, as the joint policy, marginal policy and stochastic modification, respectively.

For both agents $m = 1, 2$, we define rewards $r_0^{(m)} = a^{(m)}$, $r_1^{(m)} = a^{(0)} + a^{(1)}$, $r_2^{(m)} = 2 - a^{(0)} - a^{(1)}$ and constraint thresholds $c_1^{(m)} = c_2^{(m)} = 0.6$. Therefore, $V_0^{(m)}(\pi) = \mathbb{E}_\pi r_0^{(m)} = \pi^{(m)}(1)$, $V_1^{(m)}(\pi) = \mathbb{E}_\pi r_1^{(m)} = \pi^{(1)}(1) + \pi^{(2)}(1)$ and $V_2^{(m)}(\pi) = \mathbb{E}_\pi r_2^{(m)} = 2 - \pi^{(1)}(1) - \pi^{(2)}(1) = \pi^{(1)}(0) + \pi^{(2)}(0)$. Therefore, for both agents $m = 1, 2$, their value function constraints $V_1^{(m)}(\pi) \geq 0.6, V_2^{(m)}(\pi) \geq 0.6$ are equivalent to the following condition

$$0.6 \leq \pi^{(1)}(1) + \pi^{(2)}(1) \leq 1.4. \quad (30)$$

Now consider a uniform policy $\overline{\pi}$ where $\overline{\pi}(a^{(1)}, a^{(2)}) = 0.25$ for all $a^{(1)}, a^{(2)} \in \{0, 1\}$. This is a product policy which generates independent uniformly distributed actions $a^{(1)}, a^{(2)}$ with $\overline{\pi}^{(1)}(1) = \overline{\pi}^{(2)}(1) = 0.5$ that satisfy the constraints in eq. (30). Note that $\mathcal{A}^{(1)}$ only includes two actions. Hence, the set of all possible deterministic modifications $\phi^{(1)}$ includes the following three cases.

(i) $\phi^{(1)} \circ \overline{\pi} = \overline{\pi}$: either $\phi^{(1)}$ modifies any $a^{(1)}$ to $a^{(1)}$ or modifies any $a^{(1)}$ to $1 - a^{(1)}$;
(ii) $\phi^{(1)} \circ \overline{\pi} = \pi'$ that always generates $a^{(1)} = 0$ and generates $a^{(2)}$ uniformly at random: $\phi^{(1)}$ modifies any $a^{(1)}$ to 0;
(iii) $\phi^{(1)} \circ \overline{\pi} = \pi''$ that always generates $a^{(1)} = 1$ and generates $a^{(2)}$ uniformly at random: $\phi^{(1)}$ modifies any $a^{(1)}$ to 1.

However, π' and π'' do not satisfy the constraint (30) since $\pi'^{(1)}(1) + \pi'^{(2)}(1) = 0.5$ and $\pi''^{(1)}(1) + \pi''^{(2)}(1) = 1.5$. Hence, the only feasible deterministic modifications $\phi^{(1)}$ are the two ones in (i) with $\phi^{(1)} \circ \overline{\pi} = \overline{\pi}$, which implies that $V_0^{(1)}(\phi^{(1)} \circ \overline{\pi}) = V_0^{(1)}(\overline{\pi}) = \overline{\pi}^{(1)}(1) = 0.5$. Therefore, such a $\overline{\pi}$ satisfies the assumption of item 2.

Now consider a stochastic modification $\phi^{(1)}$ defined by $\phi^{(1)}(1|a_1) = 0.9$ and $\phi^{(1)}(0|a_1) = 0.1$ for $a_1 \in \{0, 1\}$. Then $\phi^{(1)} \circ \overline{\pi}$ independently generates Bernoulli distributed actions $a^{(1)} \sim \text{Bern}(0.9)$ and $a^{(2)} \sim \text{Bern}(0.5)$. Hence, $(\phi^{(1)} \circ \overline{\pi})^{(1)}(1) + (\phi^{(1)} \circ \overline{\pi})^{(2)}(1) = 1.4$, which means $\phi^{(1)}$ is feasible based on eq. (30). In addition, $V_0^{(1)}(\phi^{(1)} \circ \overline{\pi}) = (\phi^{(1)} \circ \overline{\pi})^{(1)}(1) = 0.9$, which is strictly larger than $V_0^{(1)}(\overline{\pi}) = 0.5$. Therefore, $\overline{\pi}$ is not a CE as defined in Definition 3.2.

D Proof of Theorem 2

For any policy π and its associated joint measure p_π, recall the following equivalent Lagrangian functions defined in eq. (23).

$$L^{(m)}(\pi, \lambda^{(m)}) = \widetilde{L}^{(m)}(p_\pi, \lambda^{(m)}).$$

Then, the desired strong duality result shown in eq. (7) is equivalent to the following equation.

$$\max_{p \in \mathcal{X}} \min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} \widetilde{L}^{(m)}(p, \lambda^{(m)}) = \min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} \max_{p \in \mathcal{X}} \widetilde{L}^{(m)}(p, \lambda^{(m)}),$$

where the set $\mathcal{X} := \{p_{\phi^{(m)} \circ \pi} : \phi^{(m)} \text{ is a stochastic modification}\}$ is defined for the fixed π. The nice property of the Lagrangian function $\widetilde{L}^{(m)}(p, \lambda^{(m)})$ is that it is a linear function in p, which has an advantage toward establishing strong duality.

Based on the minimax theorem (Lemma 9.2 of [2]), it suffices to prove the following properties:

(I). $\widetilde{L}^{(m)}(p, \cdot)$ is convex and lower semi-continuous, and $\widetilde{L}^{(m)}(\cdot, p)$ is concave. These properties directly follow from the definition of $\widetilde{L}$ in eq. (23).

(II). $\mathbb{R}_+^{d_m}$ is a convex set, which holds obviously.

(III). $\mathcal{X}$ is a convex set, which follows from Lemma B.1 since eqs. (26) and (27) always hold for $\lambda \in [0, 1]$.

(IV). $\mathcal{X}$ is a compact set.

Hence, it remains to prove (IV).

As the state space $\mathcal{S}$, action apace $\mathcal{A}$ and the horizon H are finite, we can represent p_π as a vector with entries $p_\pi(s_{1:H}, a_{1:H})$ for every $s_{1:H}, a_{1:H} \in \mathcal{S}^H \times \mathcal{A}^H$. Hence, the set $\mathcal{X} \subset [0, 1]^{(|\mathcal{S}||\mathcal{A}|)^H}$ is bounded. Then, it suffices to prove that $\mathcal{X}$ is a closed set, i.e., $p \in \mathcal{X}$ if $p_{\phi_{[k]}^{(m)} \circ \pi}(s_{1:H}, a_{1:H}) \xrightarrow{k} p(s_{1:H}, a_{1:H}), \forall s_{1:H}, a_{1:H}$ for some $p_{\phi_{[k]}^{(m)} \circ \pi} \in \mathcal{X}$ (Note that the notation $\phi_{[k]}^{(m)}$ indexed by k differs from $\phi_h^{(m)}$ where h denotes time step).

Similar to $\mathcal{X}$, any stochastic modification $\phi^{(m)}$ can also be seen as a bounded finite-dimensional vector with entries $\phi^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \in [0, 1]$. Hence, $\{\phi_{[k]}^{(m)} : k \in \mathbb{N}^+\}$ has a convergent subsequence $\{\phi_{[k_i]}^{(m)} : i \in \mathbb{N}^+\}$ such that $\phi_{[k_i]}^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \xrightarrow{i} \phi^*(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})$ for any $s_{1:h}, a_{1:h-1}, a_h^{(m)}, \widetilde{a}_h^{(m)}$, which implies that $\phi^*(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \geq 0$ and $\sum_{\widetilde{a}_h^{(m)}} \phi^*(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) = 1$. Therefore, ϕ^* is a proper stochastic modification.

Then based on eq. (24), it holds for any $s_{1:h}, a_{1:h}$ that

$$
\begin{aligned}
&(\phi^{(m)}_{[k_i],h} \circ \pi_h)\big(a_h|s_{1:h}, a_{1:h-1}\big) \\
&= \sum_{\widetilde{a}^{(m)}_h} \phi^{(m)}_{[k_i],h}(a^{(m)}_h|s_{1:h}, a_{1:h-1}, \widetilde{a}^{(m)}_h)\pi_h\big(\big[\widetilde{a}^{(m)}_h, a^{(\backslash m)}_h\big]|s_{1:h}, a_{1:h-1}\big) \\
&\xrightarrow{i} \sum_{\widetilde{a}^{(m)}_h} \phi^{*}_h(a^{(m)}_h|s_{1:h}, a_{1:h-1}, \widetilde{a}^{(m)}_h)\pi_h\big(\big[\widetilde{a}^{(m)}_h, a^{(\backslash m)}_h\big]|s_{1:h}, a_{1:h-1}\big) \\
&= (\phi^*_h \circ \pi_h)(a_h|s_{1:h}, a_{1:h-1}). \qquad (31)
\end{aligned}
$$

On one hand, the above inequality and eq. (18) imply that for any $s_{1:h}, a_{1:h}$, $p_{\phi^{(m)}_{[k_i]}\circ\pi}(s_{1:h}, a_{1:h}) \xrightarrow{i} p_{\phi^*\circ\pi}(s_{1:h}, a_{1:h})$. On the other hand, $p_{\phi^{(m)}_{[k_i]}\circ\pi}(s_{1:h}, a_{1:h}) \xrightarrow{i} p(s_{1:h}, a_{1:h})$. Therefore, $p = p_{\phi^*\circ\pi}$ for ϕ^* being a stochastic modification, and thus $p \in \mathcal{X}$.

E The Range of the Optimal Dual Variable

Before proving Theorem 3 and Corollary 5.1 on the non-asymptotic convergence of Algorithm 1, we first consider the optimal dual variable $\lambda^{(m)}_*$ of the minimax optimization problem in eq. (7) and derive its range below, which is important for the selection of the projection set $\Lambda^{(m)}$ in Algorithm 1.

Lemma E.1. *The optimal dual variable $\lambda^{(m)}_*$ satisfies the following range.*

$$
\lambda^{(m)}_{*,j} \leq \frac{Hr^{(m)}_{0,\max}}{\xi^{(m)}_j}, j = 1, \ldots, d_m. \qquad (32)
$$

Proof. Given π, denote $\phi^{(m)}_*$ as the optimal solution to the constrained optimization problem in eq. (4) and denote $\widetilde{\phi}^{(m)}$ as the stochastic modification that satisfies Assumption 1, i.e., $V^{(m)}(\widetilde{\phi}^{(m)} \circ \pi) - c^{(m)} \geq \xi^{(m)}$. Then we have

$$
\begin{aligned}
Hr^{(m)}_{0,\max} &\overset{(i)}{\geq} V^{(m)}_0\big(\phi^{(m)}_* \circ \pi\big) \\
&\overset{(ii)}{=} \max_{\phi^{(m)}} L^{(m)}\big(\phi^{(m)} \circ \pi, \lambda^{(m)}_*\big) \\
&\geq L^{(m)}\big(\widetilde{\phi}^{(m)} \circ \pi, \lambda^{(m)}_*\big) \\
&= V^{(m)}_0(\widetilde{\phi}^{(m)} \circ \pi) + \sum_{j=1}^{d_m} \lambda^{(m)}_{*,j}\big(V^{(m)}_j(\widetilde{\phi}^{(m)} \circ \pi) - c^{(m)}_j\big) \\
&\overset{(iii)}{\geq} \sum_{j=1}^{d_m} \lambda^{(m)}_{*,j}\xi^{(m)}_j,
\end{aligned}
$$

where (i) and (iii) use $V^{(m)}_0(\pi) \in [0, Hr^{(m)}_{0,\max}], \forall j = 0, 1, \ldots, d_m$ which is directly implied by Assumption 2, (ii) uses Theorem 2 which implies the equivalence between the constrained optimization problem in eq. (4) and the minimax optimization problem in eq. (7), and (iii) also uses $\lambda^{(m)}_{*,j} \geq 0$ and $V^{(m)}_j(\pi) - c^{(m)}_j \geq \xi^{(m)}_j$. Since $\xi^{(m)}_j > 0$, the above inequality implies eq. (32) □

F Proof of Theorem 3

Assumption 2 and the value functions defined in eq. (4) imply that for any $m = 1, \ldots, M$, $j = 0, 1, \ldots, d_m$ and joint policy π, we have

$$
0 \leq V^{(m)}_j(\pi) = \mathbb{E}_\pi\Big[\sum_{h=1}^{H} r^{(m)}_{j,h}(s_h, a_h)\Big| s_1 \sim \rho\Big] \leq Hr^{(m)}_{j,\max}. \qquad (33)
$$

Hence, for any $m = 1, \dots, M$ and joint policy π

$$\|V^{(m)}(\pi)\| = \sqrt{\sum_{j=1}^{d_m} V_j^{(m)}(\pi)^2} \le H\sqrt{\sum_{j=1}^{d_m} (r_{j,\max}^{(m)})^2} = HR_{\max}^{(m)} \tag{34}$$

Furthermore, Assumption 1 implies that there is a joint policy π' such that $0 \le c^{(m)} \le V^{(m)}(\pi')$, so

$$\|c^{(m)}\| \le \|V^{(m)}(\pi')\| \le HR_{\max}^{(m)}. \tag{35}$$

Then,

$$\begin{aligned}
0 &\le \|\lambda_T^{(m)}\|^2 \\
&\overset{(i)}{=} \sum_{t=0}^{T-1} \big(\|\lambda_{t+1}^{(m)}\|^2 - \|\lambda_t^{(m)}\|^2\big) \\
&\overset{(ii)}{\le} \sum_{t=0}^{T-1} \big(\big\|\lambda_t^{(m)} - \eta\big(V^{(m)}(\pi_t) - c^{(m)}\big)\big\|^2 - \|\lambda_t^{(m)}\|^2\big) \\
&\overset{(iii)}{\le} 2\eta \sum_{t=0}^{T-1} \lambda_t^{(m)\top}\big(c^{(m)} - V^{(m)}(\pi_t)\big) + \eta^2 \sum_{t=0}^{T-1} \big(\big\|V^{(m)}(\pi_t)\big\| + \|c^{(m)}\|\big)^2 \\
&\overset{(iv)}{\le} 2\eta \sum_{t=0}^{T-1} \lambda_t^{(m)\top}\big(V^{(m)}(\phi_{t*}^{(m)} \circ \pi_t) - V^{(m)}(\pi_t)\big) + 4T(\eta HR_{\max}^{(m)})^2,
\end{aligned}$$

where (i) uses the initialization $\lambda_0^{(m)} = 0$, (ii) uses eq. (12) and $0 \in \Lambda^{(m)}$, (iii) uses triangular inequality, and (iv) uses eqs. (34) and (35) and the constraint that $V^{(m)}(\phi_{t*}^{(m)} \circ \pi) \ge c^{(m)}$ satisfied by the optimal modification $\phi_{t*}^{(m)}$ of the constrained optimization problem in eq. (4) for $\pi = \pi_t$. Rearranging the above inequality yields that

$$\sum_{t=0}^{T-1} \lambda_t^{(m)\top}\big(V^{(m)}(\pi_t) - V^{(m)}(\phi_t^{(m)} \circ \pi_t)\big) \le 2\eta T(HR_{\max}^{(m)})^2. \tag{36}$$

Note that

$$\begin{aligned}
0 &\le \sum_{t=0}^{T-1} \Big(\max_{\phi^{(m)}} L^{(m)}\big(\phi^{(m)} \circ \pi_t, \lambda_t^{(m)}\big) - L^{(m)}\big(\phi_{t*}^{(m)} \circ \pi_t, \lambda_t^{(m)}\big)\Big) \\
&\overset{(i)}{=} \sum_{t=0}^{T-1} \Big(\max_{\phi^{(m)}} V_{\lambda_t}^{(m)}\big(\phi^{(m)} \circ \pi_t\big) - V_{\lambda_t}^{(m)}\big(\phi_{t*}^{(m)} \circ \pi_t\big)\Big) \\
&\overset{(ii)}{\le} \sum_{t=0}^{T-1} \Big(\epsilon + V_{\lambda_t}^{(m)}\big(\pi_t\big) - V_{\lambda_t}^{(m)}\big(\phi_{t*}^{(m)} \circ \pi_t\big)\Big) \\
&\overset{(iii)}{=} \sum_{t=0}^{T-1} \Big(\epsilon + V_0^{(m)}(\pi_t) - V_0^{(m)}(\phi_{t*}^{(m)} \circ \pi_t) + \lambda_t^{(m)\top}\big(V^{(m)}(\pi_t) - V^{(m)}(\phi_{t*}^{(m)} \circ \pi_t)\big)\Big) \qquad (37) \\
&\overset{(iv)}{\le} \sum_{t=0}^{T-1} \big(\epsilon - D^{(m)}(\pi_t)\big) + 2\eta T(HR_{\max}^{(m)})^2,
\end{aligned}$$

where (i) uses the rewritten Lagrangian function $L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)}) = V_\lambda^{(m)}(\phi^{(m)} \circ \pi) - \lambda^{(m)\top} c^{(m)}$, (ii) uses eq. (11), (iii) uses $V_\lambda^{(m)}(\pi) = V_0^{(m)}(\pi) + \lambda^{(m)\top} V^{(m)}(\pi), \forall \pi$ implies by eqs. (2) and (9), and (iv) uses eqs. (5) and (36). Rearranging the above inequality yields that

$$\mathbb{E}_{\tilde{t}}\big[D^{(m)}(\pi_t)\big] = \frac{1}{T}\sum_{t=0}^{T-1} D^{(m)}(\pi_t)\big) \le 2\eta(HR_{\max}^{(m)})^2 + \epsilon,$$

which proves the duality gap in eq. (14) by substituting $\eta = \frac{1}{\sqrt{T}}$.

Next, we prove the constraint violation in eq. (15).

For any $\lambda^{(m)} \in \Lambda^{(m)}$, it holds that

$$\begin{aligned}
&\|\lambda_{t+1}^{(m)} - \lambda^{(m)}\|^2 \\
&\overset{(i)}{\leq} \left\|\lambda_t^{(m)} - \eta\big(V^{(m)}(\pi_t) - c^{(m)}\big) - \lambda^{(m)}\right\|^2 \\
&\overset{(ii)}{\leq} \|\lambda_t^{(m)} - \lambda^{(m)}\|^2 - 2\eta(\lambda_t^{(m)} - \lambda^{(m)})^\top \big(V^{(m)}(\pi_t) - c^{(m)}\big) + \eta^2\big(\|V^{(m)}(\pi_t)\| + \|c^{(m)}\|\big)^2 \\
&\overset{(iii)}{\leq} \|\lambda_t^{(m)} - \lambda^{(m)}\|^2 - 2\eta(\lambda_t^{(m)} - \lambda^{(m)})^\top \big(V^{(m)}(\pi_t) - c^{(m)}\big) + 4(\eta H R_{\max}^{(m)})^2
\end{aligned}$$

where (i) uses eq. (12) and $\lambda^{(m)} \in \Lambda^{(m)}$, (ii) uses triangular inequality, (iii) uses eqs. (34) and (35). Telescoping the above inequality over $t = 0, 1, \ldots, T-1$ and using $\lambda_0^{(m)} = 0$ yields that

$$\eta \sum_{t=0}^{T-1} (\lambda_t^{(m)} - \lambda^{(m)})^\top \big(V^{(m)}(\pi_t) - c^{(m)}\big) \leq \frac{1}{2}\left\|\lambda^{(m)}\right\|^2 + 2T(\eta H R_{\max}^{(m)})^2. \tag{38}$$

Since $V^{(m)}(\phi_{t*}^{(m)} \circ \pi_t) \geq c^{(m)}$ and $\lambda_t^{(m)} \in \mathbb{R}_+^{d_m}$, eq. (37) implies that

$$\eta \sum_{t=0}^{T-1} \lambda_t^{(m)\top} \big(c^{(m)} - V^{(m)}(\pi_t)\big) \leq \eta \sum_{t=0}^{T-1} \big(\epsilon + V_0^{(m)}(\pi_t) - V_0^{(m)}(\phi_{t*}^{(m)} \circ \pi)\big) \tag{39}$$

Summing up eqs. (38) and (39) yields that

$$\begin{aligned}
\eta \sum_{t=0}^{T-1} \lambda^{(m)\top} \big(c^{(m)} - V^{(m)}(\pi_t)\big) \leq &\eta \sum_{t=0}^{T-1} \big(\epsilon + V_0^{(m)}(\pi_t) - V_0^{(m)}(\phi_{t*}^{(m)} \circ \pi)\big) \\
&+ \frac{1}{2}\left\|\lambda^{(m)}\right\|^2 + 2T(\eta H R_{\max}^{(m)})^2.
\end{aligned} \tag{40}$$

Denote $\Phi_t^{(m)} := \big\{\phi^{(m)} : V^{(m)}(\phi^{(m)} \circ \pi_t) \geq \min\big(c^{(m)}, V^{(m)}(\pi_t)\big)\big\}$, which is a non-empty set that includes identity modification $\phi^{(m)}$ such that $I^{(m)} \circ \pi_t = \pi_t$. Hence,

$$\begin{aligned}
V_0^{(m)}(\phi_{t*}^{(m)} \circ \pi_t) &= \max_{\phi^{(m)}} \min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} L^{(m)}(\phi^{(m)} \circ \pi_t, \lambda^{(m)}) \\
&\overset{(i)}{=} \min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} \max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi_t, \lambda^{(m)}) \\
&\overset{(ii)}{\geq} \max_{\phi^{(m)} \in \Phi_t^{(m)}} L^{(m)}\big(\phi^{(m)} \circ \pi_t, \lambda_{t*}^{(m)}\big) \\
&= \max_{\phi^{(m)} \in \Phi_t^{(m)}} \big(V_0^{(m)}\big(\phi^{(m)} \circ \pi_t\big) + (\lambda_{t*}^{(m)})^\top \big[V^{(m)}\big(\phi^{(m)} \circ \pi_t\big) - c^{(m)}\big]\big) \\
&\overset{(iii)}{\geq} \max_{\phi^{(m)} \in \Phi_t^{(m)}} V_0^{(m)}\big(\phi^{(m)} \circ \pi_t\big) + (\lambda_{t*}^{(m)})^\top \min\big(0, V^{(m)}(\pi_t) - c^{(m)}\big) \\
&\overset{(iv)}{\geq} V_0^{(m)}(\pi_t) - (\lambda_{t*}^{(m)})^\top \big(c^{(m)} - V^{(m)}(\pi_t)\big)_+
\end{aligned}$$

where (i) uses Theorem 2, (ii) uses the fact that $\Phi_t^{(m)}$ is only a subset of stochastic modifications and denotes that $\lambda_{t*}^{(m)} = \arg\min_{\lambda^{(m)} \in \mathbb{R}_+^{d_m}} \max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi_t, \lambda^{(m)})$, (iii) uses $\lambda_{t*}^{(m)} \in \mathbb{R}_+^{d_m}$ and the definition of $\Phi_t^{(m)}$, and (iv) uses the fact that the identity modification $\phi^{(m)} \in \Phi_t^{(m)}$. Substituting the above inequality into eq. (40) and rearranging it, we obtain that

$$\eta \sum_{t=0}^{T-1} \Big(\lambda^{(m)\top} \big(c^{(m)} - V^{(m)}(\pi_t)\big) - (\lambda_{t*}^{(m)})^\top \big(c^{(m)} - V^{(m)}(\pi_t)\big)_+\Big)$$

$$\leq \frac{1}{2}\|\lambda^{(m)}\|^2 + 2T(\eta H R_{\max}^{(m)})^2 + \eta T\epsilon. \tag{41}$$

Using eq. (32) and selecting $\lambda_j^{(m)} = \frac{2Hr_{0,\max}^{(m)}}{\xi_j^{(m)}}\mathbb{1}\{V_j^{(m)}(\pi_t) \leq c_j^{(m)}\}$ (this satisfies $\lambda^{(m)} \in \Lambda^{(m)}$), we obtain that

$$\begin{aligned}
&\lambda^{(m)\top}\big(c^{(m)} - V^{(m)}(\pi_t)\big) - (\lambda_{t*}^{(m)})^\top\big(c^{(m)} - V^{(m)}(\pi_t)\big)_+ \\
&\geq \sum_{j=1}^{d_m} \frac{Hr_{0,\max}^{(m)}}{\xi_j^{(m)}}\big(c_j^{(m)} - V_j^{(m)}(\pi_t)\big)_+,
\end{aligned}$$

where the last inequality uses eq. (32). Substituting the above inequality into eq. (41) yields that

$$\begin{aligned}
&\eta H r_{0,\max}^{(m)} \sum_{t=0}^{T-1}\sum_{j=1}^{d_m}(\xi_j^{(m)})^{-1}\big(c_j^{(m)} - V_j^{(m)}(\pi_t)\big)_+ \\
&\leq \frac{1}{2}\|\lambda^{(m)}\|^2 + 2T(\eta H R_{\max}^{(m)})^2 + \eta T\epsilon \\
&\overset{(i)}{\leq} 2(Hr_{0,\max}^{(m)})^2\sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2} + 2T(\eta H R_{\max}^{(m)})^2 + \eta T\epsilon,
\end{aligned}$$

where (i) uses $\|\lambda^{(m)}\| \leq 2Hr_{0,\max}^{(m)}\sqrt{\sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2}}$ for our choice $\lambda_j^{(m)} = \frac{2Hr_{0,\max}^{(m)}}{\xi_j^{(m)}}$ $\mathbb{1}\{V_j^{(m)}(\pi_t) \leq c_j^{(m)}\}$. Dividing both sides of the above inequality by $\eta T H r_{0,\max}^{(m)}$ and substituting $\eta = \frac{1}{\sqrt{T}}$, we prove the constraint violation in eq. (15).

G Proof of Corollary 5.1

The surrogate rewards defined in eq. (8) has the following bound

$$\begin{aligned}
0 &\leq R_{\lambda_t,h}^{(m)}(s_h, a_h) = r_{0,h}^{(m)}(s_h, a_h) + \lambda_t^{(m)\top} r_h^{(m)}(s_h, a_h) \\
&\leq r_{0,h}^{(m)}(s_h, a_h) + \|\lambda_t^{(m)}\|\|r_h^{(m)}(s_h, a_h)\| \\
&\overset{(i)}{\leq} r_{0,\max}^{(m)} + 2Hr_{0,\max}^{(m)} R_{\max}^{(m)}\sqrt{\sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2}} := \widetilde{R}_{\max}^{(m)}
\end{aligned} \tag{42}$$

where (i) uses Assumption 2 and $\lambda_{t,j}^{(m)} \in \big[0, \frac{2Hr_{0,\max}^{(m)}}{\xi_j^{(m)}}\big]$ (since $\lambda_t^{(m)} \in \Lambda^{(m)}$ based on eq. (12)). Note that the V-learning in [31] assumes the rewards to range in $[0, 1]$. To adjust to this assumption, we apply V-learning to the scaled rewards $\frac{1}{\widetilde{R}_{\max}^{(m)}} R_{\lambda_t,h}^{(m)}(s_h, a_h) \in [0, 1]$ with corresponding value function $\frac{1}{\widetilde{R}_{\max}^{(m)}} V_{\lambda_t}^{(m)}$. Then based on Theorem 7 of [31], it takes $\widetilde{\mathcal{O}}(H^5SA^2(\epsilon/\widetilde{R}_{\max}^{(m)})^{-2}) = \widetilde{\mathcal{O}}(H^5SA^2\epsilon^{-2})$ samples to reach the $\epsilon/\widetilde{R}_{\max}^{(m)}$-CE of this scaled Markov game with probability at least $1 - \delta/T$ for any $\delta \in (0, 1)$ (we replace δ with δ/T which only changes the hidden logarithm factor in $\widetilde{O}$), that is,

$$\max_{\phi^{(m)}} \frac{1}{\widetilde{R}_{\max}^{(m)}} V_\lambda^{(m)}(\phi^{(m)} \circ \pi_t) - \frac{1}{\widetilde{R}_{\max}^{(m)}} V_\lambda^{(m)}(\pi_t) \leq \frac{\epsilon}{\widetilde{R}_{\max}^{(m)}},$$

which is equivalent to eq. (11). Applying union bound over the T iterations yields that eq. (11) holds for all iterations $t = 0, 1, \ldots, T-1$ with probability at least $1 - \delta$. In that case, the convergence rates in eqs. (14) and (15) hold. Substituting $T = \max_m 4\epsilon^{-2}(HR_{\max}^{(m)})^2\big(\sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2} + HR_{\max}^{(m)}\big)^2$ and $r_{0,\max}^{(m)} \geq \frac{1}{H}$ into these convergence rates yields that

$$\mathbb{E}_{\widetilde{t}}\big(D^{(m)}(\pi_{\widetilde{t}})\big) \leq \frac{2(HR_{\max}^{(m)})^2}{\sqrt{T}} + \epsilon \leq 2\epsilon,$$

$$\mathbb{E}_{\widetilde{t}}\big(W^{(m)}(\pi_{\widetilde{t}})\big) \leq \frac{2HR_{\max}^{(m)}}{\sqrt{T}} \sum_{j=1}^{d_m} (\xi_j^{(m)})^{-2} + \frac{2H(R_{\max}^{(m)})^2}{r_{0,\max}^{(m)}\sqrt{T}} + \frac{\epsilon}{Hr_{0,\max}^{(m)}}$$

$$\leq \frac{1}{\sqrt{T}} \Big(2HR_{\max}^{(m)} \sum_{j=1}^{d_m} (\xi_j^{(m)})^{-2} + 2(HR_{\max}^{(m)})^2 \Big) + \epsilon \leq 2\epsilon.$$

The above two inequalities prove that $\max\big(\mathbb{E}_{\widetilde{t}} D^{(m)}(\pi_{\widetilde{t}}), \mathbb{E}_{\widetilde{t}} W^{(m)}(\pi_{\widetilde{t}})\big) \leq 2\epsilon$.

Since each of the $T = \mathcal{O}(H^4\epsilon^{-2})$ iterations takes $\widetilde{\mathcal{O}}(H^5SA^2\epsilon^{-2})$ samples, the required sample complexity is $T\widetilde{\mathcal{O}}(H^5SA^2\epsilon^{-2}) = \widetilde{\mathcal{O}}(H^9SA^2\epsilon^{-4})$.