Proceedings of Machine Learning Research vol 195:1–55, 2023 36th Annual Conference on Learning Theory

The Complexity of Markov Equilibrium in Stochastic Games

Constantinos Daskalakis COSTIS@MIT.EDU
MIT, Cambridge, MA, 02139

Noah Golowich NZG@MIT.EDU
MIT, Cambridge, MA, 02139

Kaiqing Zhang KAIQING@UMD.EDU
University of Maryland, College Park, MD, 20740

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We show that computing approximate stationary Markov coarse correlated equilibria (CCE) in general-sum stochastic games is PPAD-hard, even when there are two players, the game is turn-based, the discount factor is an absolute constant, and the approximation is an absolute constant. Our intractability results stand in sharp contrast to the results in normal-form games, where exact CCEs are efficiently computable. A fortiori, our results imply that, in the setting of multi-agent reinforcement learning (MARL), it is computationally hard to learn stationary Markov CCE policies in stochastic games, even when the interaction is two-player and turn-based, and both the discount factor and the desired approximation of the learned policies is an absolute constant. In turn, these results stand in sharp contrast to single-agent reinforcement learning (RL) where near-optimal stationary Markov policies can be computationally efficiently learned. Complementing our intractability results for *stationary* Markov CCEs, we provide a decentralized algorithm (assuming shared randomness among players) for learning a *nonstationary* Markov CCE policy with polynomial time and sample complexity in all problem parameters. Previous work for learning Markov CCE policies all required exponential time and sample complexity in the number of players. In the balance, our work advocates for the use of nonstationary Markov CCE policies as a computationally and statistically tractable solution concept in MARL, advancing an important and outstanding frontier in machine learning.

1. Introduction

Learning in multi-agent, dynamic environments lies at the heart of many important advances and outstanding challenges in artificial intelligence, from playing Go (Silver et al., 2016) and Poker (Brown and Sandholm, 2019) to improving algorithms for multi-robot interaction and autonomous driving (Shalev-Shwartz et al., 2016) and evaluating the outcomes of economic policies (Zheng et al., 2020). A prominent and general learning framework capturing these and other important applications is that of *multi-agent reinforcement learning (MARL)*, which generalizes its single-agent analogue, reinforcement learning (RL) (Busoniu et al., 2008; Zhang et al., 2021). In the same way that RL is mathematically grounded on the model of Markov Decision Processes (MDPs), MARL is grounded on the model of *Stochastic Games (SGs)*, the multi-agent analog of MDPs introduced in the seminal work of Shapley (Shapley, 1953). In contrast to RL, however, where a range of algorithms for learning optimal policies are known, it has remained challenging to pin down what types of policies are efficiently learnable in MARL, unless the setting has a very special structure, as we discuss below. The goal of this work is to shed light on this central challenge, by showing

that (i) a prominent type of policy, namely *stationary Markov coarse correlated equilibrium (CCE)*, is intractable, even when the MARL setting is fully known and relatively simple, and (ii) another prominent type of policy, namely *nonstationary Markov CCE*, is efficiently learnable via distributed learning dynamics, even when the environment is unknown.

To place our results in their equilibrium complexity and machine learning context, recall that in a stochastic game, several agents interact in an environment over multiple steps: at each step, each agent takes an action, and then the environment transitions to a new state and each agent receives a reward. The rewards and transitions depend on both the current state and the profile of actions chosen by the players at the current step. In contrast to the study of MDPs, where a standard goal is to learn a near-*optimal* policy, a standard goal in the study of SGs is for the agents to learn a near-*equilibrium* policy by interacting. Since their introduction by Shapley, SGs have received extensive study in game theory (Neyman and Sorin, 2003; Solan and Vieille, 2015), machine learning (Littman, 1994; Hu and Wellman, 2003; Busoniu et al., 2008; Zhang et al., 2021), and various other fields, due to their broad applicability.

When there is a single state and the interaction lasts for a single step, SGs degenerate to *normal-form* games. In this case, our understanding of equilibrium existence, computational complexity, and learnability is quite advanced. If the game is two-player and zero-sum, Nash equilibria are identical to minimax equilibria (von Neumann, 1928), which can be computed efficiently using linear programming (Dantzig, 1951), and a large number of (decentralized) learning algorithms have been discovered which converge to minimax equilibrium when employed by the agents to iteratively update their strategies, even when the game is a priori unknown to them; see e.g., Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012) for overviews. Beyond two-player zero-sum games, it is known that computing a Nash equilibrium is intractable in general (Daskalakis et al., 2009; Chen et al., 2009), but (coarse) correlated equilibria can be computed efficiently using linear programming, or decentralized learning (Cesa-Bianchi and Lugosi, 2006; Bubeck and Cesa-Bianchi, 2012).

When there are more states and steps, questions of equilibrium existence, computation, and learning become much more intricate, occupying many works in the literature; see e.g., Solan and Vieille (2015); Zhang et al. (2021). Indeed, there are various versions of (coarse) correlated equilibrium, and it is often unclear which we should search for. In particular, when the players interact over multiple steps, strategic behavior might be *history-dependent*, giving rise to notions of equilibrium that are also history-dependent and, thus, extremely complex. Circumventing this complexity, a compelling type of strategic behavior, introduced by Shapley and studied in much of the game theory and machine learning literature, is *Markovian*, i.e., strategic behavior wherein the actions chosen by the players at every step of the game depend on the *current state* (and potentially the step count), but not the history of states visited and actions played so far. Indeed, under broad and natural conditions, e.g., future payoff discounting, there exist Markov Nash equilibria that are also *stationary*, i.e., the actions played at every state are also step-count independent; see e.g., Shapley (1953); Takahashi (1962); Fink (1964); Solan and Vieille (2015).

On the computation and learning front, most of the progress has been on efficient computation and learning of (approximate) Nash equilibria in *two-player zero-sum* stochastic games; see e.g., Brafman and Tennenholtz (2002); Wei et al. (2017); Xie et al. (2020); Zhang et al. (2020); Sidford et al. (2020); Bai and Jin (2020); Daskalakis et al. (2020); Bai et al. (2020); Liu et al. (2021); Jin et al. (2021). Indeed, some of these works provide time- and sample-efficient learning algorithms for computing Nash equilibria that are also Markovian. Beyond the two-player zero-sum

case, however, our understanding is lagging. On the one hand, Nash equilibria are computationally intractable (namely, PPAD-hard), as SGs are more expressive than normal-form games. On the other hand, the complexity and learnability of (coarse) correlated equilibria are not well-understood in SGs. It is easy to see that approximate *nonstationary* Markov (coarse) correlated equilibria can be computed efficiently via *backward induction*, but the complexity of *stationary* Markov (coarse) correlated equilibria remained unknown prior to this work. At the same time, a flurry of recent work has provided *learning* algorithms for nonstationary (coarse) correlated equilibria in finite-horizon episodic SGs (Liu et al., 2021; Mao and Başar, 2021; Song et al., 2021; Jin et al., 2021). However, each of these algorithms suffers from one of two shortfalls: either they cannot output Markov equilibria (Mao and Başar, 2021; Song et al., 2021; Jin et al., 2021), or they require exponentially many samples in the number of agents (Liu et al., 2021), i.e., suffering from the so-called "curse of multi-agents;" see Jin et al. (2021) for additional explanation. We defer a more detailed literature review to Appendix A.

Overview of results. In this work, we settle the complexity of computing stationary Markov (coarse) correlated equilibria, showing that they are intractable; we then complement these results with time- and sample-efficient decentralized learning algorithms for computing nonstationary Markov coarse correlated equilibria. In particular, we show the following results (which are summarized and compared to existing results in Table 1):

- In Theorems 3 and 4, we establish intractability of computing approximate stationary Markov coarse correlated equilibria (CCE) in 2-player, discounted general-sum stochastic games, even when both the approximation and the discount factors are absolute constants. In particular, a notion of stationary Markov CCE called *perfect CCE* (Definition 1) are PPAD-hard to approximate up to a constant, and a relaxed notion (*stationary CCE*) are PPAD-hard to approximate up to a constant assuming the "PCP for PPAD" conjecture (Conjecture 10).

- To circumvent the above intractability results, we then consider the computation of *Markov nonstationary CCE*, a relaxation of stationary CCE. While it is trivial to *compute* an approximate Markov nonstationary CCE using backward induction, the learning problem, in which the SG is *unknown* and agents must employ exploratory policies to learn its transitions and rewards, is more challenging. In Theorem 5, we establish the first guarantee for learning a Markov nonstationary CCE which has sample and computational cost polynomial in all parameters, including the number of agents. In particular, our algorithm (`SPoCMAR`, Algorithm 1) avoids the curse of multi-agents suffered by prior work Liu et al. (2021), which required sample complexity exponential in the number of agents for computing Markov nonstationary CCE. We also show that `SPoCMAR` can be implemented in a decentralized manner (Section 4.4), assuming that agents have access to shared common randomness.

Contemporaneous & subsequent work. Since the initial release of this paper, there have been several works that prove related results. On the lower bound front, the most closely related work is the concurrent work of Jin et al. (2022), which shows a similar (but weaker) result to Theorem 3, where the number of players is equal to the number of states. As we will discuss in Appendix B.5, it requires nontrivial ideas to reduce the number of players in the hardness result down to 2. Moreover, no decentralized learning algorithms were investigated therein.

Subsequent to our work, several papers studied decentralized algorithms for learning CCE in general-sum Markov games, though generally under significantly stronger assumptions than our

Table 1: Complexity results of finding approximate CCE in general-sum stochastic games. The two rows correspond to computational and sample complexities, respectively. `Polynomial` means computational and sample costs with polynomial dependence on all problem parameters (namely the number of states, actions, players, and the inverse approximation), and `Exponential` means computational and sample costs which are exponential in the number of players.

	Markovian		**Non-Markovian**
	Stationary	**Nonstationary**	
Computation	PPAD-hard **(Theorems 3, 4)**	`Polynomial` (Folklore, via backward induction)	`Polynomial` (Song et al., 2021; Mao and Başar, 2021; Jin et al., 2021)
Learning	PPAD-hard **(Theorems 3, 4)**	`Exponential` (Liu et al., 2021); `Polynomial` **(Theorem 5)**	

own. Li et al. (2022) find a decentralized algorithm for learning a Markov nonstationary CCE in the setting of a generative model (which is easier than our setting due to the ability to query the transition distribution at any state-action pair). Erez et al. (2022) study decentralized no-regret algorithms in Markov games; while their no-regret algorithms do succeed at finding approximate CCE through standard reductions, their results only hold under the strong assumption that all policies visit each state with a lower-bounded probability, which makes the need for exploration less of a challenge compared to our general setting. Furthermore, their rates are worse than ours in their dependence on the approximation, namely their sample complexity scales as ϵ^{-9} vs the ϵ^{-3} of our Theorem 5. Finally, several recent papers Giannou et al. (2022); Ding et al. (2022); Fox et al. (2022); Cen et al. (2022); Zhang et al. (2022); Leonardos et al. (2022) have studied the convergence of *policy gradient* methods to equilibria in stochastic games; while these algorithms are decentralized and do avoid the curse of multi-agents, they all either are restricted to the full-information setting, in which an oracle can return exact policy gradients, or make strong assumptions on the game dynamics which avoid the need for exploration. Additionally, most of them only apply to special classes of games, such as Markov potential games (Ding et al., 2022; Fox et al., 2022; Cen et al., 2022; Zhang et al., 2022; Leonardos et al., 2022). Indeed, understanding how to *explore* in general settings while at the same time avoid the curse of multi-agents is a key technical obstacle we face.

Discussion & future work. Our work provides significant progress in understanding the computational and statistical complexities of equilibrium computation in stochastic games. Theorems 3 and 4 indicate that if we wish to avoid computational intractability, then the "correct" notion of Markov CCE is the nonstationary variant. Algorithm 1 (Theorem 5) then provides the first guarantee for learning such a Markov nonstationary CCE which has sample (and computational) cost polynomial in all parameters and which does not make simplifying assumptions that ease the challenge of exploration. We remark, however, that the sample complexity of Algorithm 1 is suboptimal in its dependence on H, S, and $1/\epsilon$. Improving the dependence on ϵ (from $1/\epsilon^3$ to $1/\epsilon^2$) seems to be a particularly interesting direction for future work, as it would likely require new exploration techniques: as discussed further in Section 4.1, Algorithm 1 uses `Rmax`-type exploration bonuses (Brafman and Tennenholtz, 2002; Jin et al., 2020), which seem insufficient to get tight rates. It would be interesting to see if upper confidence bound-type bonuses (e.g., as in V-learning (Song et al., 2021; Mao and Başar, 2021; Jin et al., 2021)) can achieve the optimal dependence on ϵ.

Notation. We use $\perp$ to denote a null element. For $m \in \mathbb{N}$, $[m]$ denotes the set $\{1, \cdots, m\}$. For a finite set $\mathcal{T}$, $\Delta(\mathcal{T})$ denotes the space of distributions over $\mathcal{T}$, $|\mathcal{T}|$ denotes the cardinality of the set. Let $\sqcup$ denote the disjoint union of sets. For $x \in \mathbb{R}$, $\mathrm{sign}(x) \in \{\pm 1\}$ denotes the sign of x.

2. Problem Formulation & Preliminaries

2.1. Stochastic games

We begin with some background regarding the terminology and equilibrium concepts in general-sum stochastic games. Formally, for some $m \in \mathbb{N}$, an infinite-horizon discounted m-player[1] stochastic game $\mathbb{G}$ is defined to be a tuple $(\mathcal{S}, (\mathcal{A}_i)_{i\in[m]}, \mathbb{P}, (r_i)_{i\in[m]}, \gamma, \mu)$, where:

- $\mathcal{S}$ denotes the (finite) *state space*, and we denote $S = |\mathcal{S}|$.
- $\mathcal{A}_i$ denotes the (finite) action space of each player $i \in [m]$, and we denote $A_i = |\mathcal{A}_i|$. We will write $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ to denote the *joint action space*, and, for $i \in [m]$, $\mathcal{A}_{-i} := \prod_{i' \neq i} \mathcal{A}_{i'}$.
- $\gamma \in [0, 1)$ denotes the *discount factor*.
- $\mu \in \Delta(\mathcal{S})$ denotes the distribution over initial states.
- $r_i : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ denotes the *reward function* for player i.
- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the *transition kernel*: $\mathbb{P}(\cdot|s, \boldsymbol{a}) \in \Delta(\mathcal{S})$ denotes the distribution over the next state if joint action profile is played at a state.

We denote joint action profiles $\boldsymbol{a} \in \mathcal{A}$ with boldface; to denote the action of some agent $i \in [m]$ when the joint action profile is $\boldsymbol{a}$, we write $a_i \in \mathcal{A}_i$. Similarly, we denote a joint reward profile as $\boldsymbol{r} \in \mathbb{R}^m$, with $r_i \in \mathbb{R}$ denoting the reward to agent i.

Policies: Stationary and nonstationary. We primarily consider two types of policies in this paper, namely *stationary* Markov policies and *nonstationary* Markov policies: a *stationary Markov policy* for some player i is a mapping $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$, and a *nonstationary Markov policy* for player i is a sequence of maps $\pi_{i,1}, \pi_{i,2}, \ldots : \mathcal{S} \to \Delta(\mathcal{A}_i)$, which we denote by $\pi_i = (\pi_{i,1}, \pi_{i,2}, \ldots)$. A stationary Markov policy π_i maps each state s to a distribution over actions $\pi_i(s) \in \Delta(\mathcal{A}_i)$ for player i; in the nonstationary case, the distribution over actions taken, $\pi_{i,h}(s)$, depends also on the current step h. Furthermore, we will often write $\pi_i(a_i|s)$ to denote the probability of taking a_i under the distribution $\pi_i(s)$. We denote the set of all stationary Markov policies of player i by $\Delta(\mathcal{A}_i)^{\mathcal{S}}$, and the set of all nonstationary Markov policies of player i by $\Delta(\mathcal{A}_i)^{\mathbb{N}\times\mathcal{S}}$.[2]

Joint Markov policies are defined analogously to policies for individual players, except they prescribe a distribution over *joint* actions at each state: in particular a joint stationary Markov policy is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, and a joint nonstationary Markov policy with horizon H is a sequence $\pi = (\pi_1, \pi_2, \ldots)$, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. With slight abuse of terminology, we will drop

1. Hereafter, we use "player" and "agent" interchangeably.
2. Notice that it takes infinite space to specify a general nonstationary policy: to obtain efficient algorithms which output nonstationary policies, we fully specify the policy for some number H of steps and then specify a fixed policy (e.g., playing a uniform action) for all remaining steps. As long as $H \gg \frac{1}{1-\gamma}$, any suboptimality of the policy played at steps $h > H$ incurs only a small approximation error.

"Markov" and "joint" from our terminology when discussing policies if the context is clear. We say that the stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a *product policy* if there are policies $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ so that $\pi(s) = \pi_1(s) \times \cdots \times \pi_m(s)$ for all $s \in \mathcal{S}$. A nonstationary policy is a product policy if each of its constituent policies $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ is a product policy. Given a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and a player $i \in [m]$, let $\pi_{-i} : \mathcal{S} \to \Delta(\mathcal{A}_{-i})$ denote the joint policy which at each state s outputs the marginal distribution of $\pi(s)$ over $\mathcal{A}_{-i}$. For a joint nonstationary policy $\pi \in \Delta(\mathcal{A})^{\mathbb{N} \times \mathcal{S}}$, we write $\pi_{-i,h} := (\pi_h)_{-i} : \mathcal{S} \to \Delta(\mathcal{A}_{-i})$, and define π_{-i} to be the sequence $(\pi_{-i,1}, \pi_{-i,2}, \ldots)$.

Value functions. Consider first a joint stationary policy π. The evolution of the stochastic game $\mathbb{G}$ proceeds as follows: the system starts at some $s_1 \in \mathcal{S}$, drawn according to μ, and at each step $h \geq 1$, all players observe s_h, draw a joint action $\boldsymbol{a}_h \sim \pi(s_h)$, and then the system transitions to some $s_{h+1} \sim \mathbb{P}(\cdot | s_h, \boldsymbol{a}_h)$. We call the tuple $(s_1, \boldsymbol{a}_1, s_2, \boldsymbol{a}_2, \ldots)$ a *trajectory*, and will write $(s_1, \boldsymbol{a}_1, s_2, \boldsymbol{a}_2, \ldots) \sim (\mathbb{G}, \pi)$ to denote a trajectory drawn in this manner. For any agent $i \in [m]$, their value function $V_i^\pi : \mathcal{S} \to [-1, 1]$ is defined as the expected γ-discounted cumulative reward that player i receives if the game starts at state $s_1 = s$ and the players act according to π:

$$V_i^\pi(s) := (1-\gamma) \cdot \mathbb{E}_{(s_1, \boldsymbol{a}_1, s_2, \boldsymbol{a}_2, \ldots) \sim (\mathbb{G}, \pi)} \left[\sum_{h=1}^{\infty} \gamma^{h-1} \cdot r_i(s_h, \boldsymbol{a}_h) \,\middle|\, s_1 = s \right].$$

Furthermore, set $V_i^\pi(\mu) := \mathbb{E}_{s \sim \mu}[V_i^\pi(s)]$. The value function is defined similarly for a joint *nonstationary* policy $\pi \in \Delta(\mathcal{A})^{\mathbb{N} \times \mathcal{S}}$, except that, due to nonstationarity, it is useful to define separate value functions at each step $h \geq 1$: thus, we write, for $h \geq 1$,

$$V_{i,h}^\pi(s) = (1-\gamma) \cdot \mathbb{E}_{(s_h, \boldsymbol{a}_h, s_{h+1}, \boldsymbol{a}_{h+1}, \ldots) \sim (\mathbb{G}, \pi)} \left[\sum_{h'=h}^{\infty} \gamma^{h'-h} \cdot r_i(s_{h'}, \boldsymbol{a}_{h'}) \,\middle|\, s_h = s \right], \tag{1}$$

and for simplicity write $V_i^\pi(s) = V_{i,1}^\pi(s)$. In the expectation in (1), for $h' \geq h$ the action $\boldsymbol{a}_{h'}$ is drawn from $\pi_{h'}(s_{h'})$. Similarly to above, we define $V_{i,h}^\pi(\mu) := \mathbb{E}_{s \sim \mu}\left[V_{i,h}^\pi(s)\right]$.

2.2. Equilibrium notions

To define the equilibrium notions we work with, we begin by introducing best-response policies.

Best-response policies. For any $i \in [m]$ and for stationary Markov policies $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$, $\pi_{-i} : \mathcal{S} \to \Delta(\mathcal{A}_{-i})$, we let $\pi_i \times \pi_{-i}$ refer to the policy which at each state s, samples an action profile according to the product distribution $\pi_i(s) \times \pi_{-i}(s)$. Fix any $i \in [m]$, and consider any joint stationary policy $\pi_{-i} : \mathcal{S} \to \Delta(\mathcal{A}_{-i})$ of all players except player i. There is a stationary policy of the ith player, $\pi_i^\dagger(\pi_{-i}) : \mathcal{S} \to \Delta(\mathcal{A}_i)$, so that $V_i^{\pi_i^\dagger(\pi_{-i}) \times \pi_{-i}}(s) = \sup_{\pi_i' : \mathcal{S} \to \Delta(\mathcal{A}_i)} V_i^{\pi_i' \times \pi_{-i}}(s)$ for all $s \in \mathcal{S}$. The policy $\pi_i^\dagger(\pi_{-i})$ is called the *best-response policy* of player i, and we will write $V_i^{\dagger, \pi_{-i}}(s) := V_i^{\pi_i^\dagger(\pi_{-i}) \times \pi_{-i}}(s)$, and $V_i^{\dagger, \pi_{-i}}(\mu) := \mathbb{E}_{s \sim \mu}\left[V_i^{\dagger, \pi_{-i}}(s)\right]$.[3]

3. It is well-known that when π_{-i} is Markov (as is assumed here), the best response amongst all *history-dependent* policies is Markovian (and is in fact deterministic), as it reduces to a single-agent Markov decision process problem; thus it is without loss of generality to constrain ourselves to Markov policies π_i' above; an analogous fact also holds for nonstationary policies.

Best-response policies for nonstationary policies are defined similarly: for a nonstationary policy $\pi_{-i} \in \Delta(\mathcal{A}_{-i})^{\mathbb{N}\times\mathcal{S}}$, there is a nonstationary *best-response* policy of the ith player $\pi_i^\dagger(\pi_{-i}) \in \Delta(\mathcal{A}_i)^{\mathbb{N}\times\mathcal{S}}$ so that for all $(h,s) \in \mathbb{N}\times\mathcal{S}$, $V_{i,h}^{\pi_i^\dagger(\pi_{-i}),\pi_{-i}}(s) = \sup_{\pi_i' \in \Delta(\mathcal{A}_i)^{\mathbb{N}\times S}} V_{i,h}^{\pi_i'\times\pi_{-i}}(s)$. As above we write $V_{i,h}^{\dagger,\pi_{-i}}(s) := V_{i,h}^{\pi_i^\dagger(\pi_{-i})\times\pi_{-i}}(s)$ and $V_{i,h}^{\dagger,\pi_{-i}}(\mu) := \mathbb{E}_{s\sim\mu}\left[V_{i,h}^{\dagger,\pi_{-i}}(s)\right]$.

Coarse correlated equilibrium. We first define approximate Markov CCE in stochastic games.

Definition 1 (Coarse correlated equilibrium) *For $\epsilon > 0$:*

- *A stationary policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$ is an* ϵ-approximate stationary Markov coarse correlated equilibrium *(abbreviated* ϵ-stationary CCE*) if* $\max_{i\in[m]}\left\{V_i^{\dagger,\pi_{-i}}(\mu) - V_i^{\pi}(\mu)\right\} \le \epsilon$.
- *A nonstationary policy $\pi \in \Delta(\mathcal{A})^{\mathbb{N}\times\mathcal{S}}$ is an* ϵ-approximate nonstationary Markov coarse correlated equilibrium *(abbreviated* ϵ-nonstationary CCE*) if* $\max_{i\in[m]}\left\{V_i^{\dagger,\pi_{-i}}(\mu) - V_i^{\pi}(\mu)\right\} \le \epsilon$.
- *A stationary policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$ is an* ϵ-approximate perfect Markov coarse correlated equilibrium *(abbreviated* ϵ-perfect CCE*) if it holds that* $\max_{i\in[m],s\in\mathcal{S}}\left\{V_i^{\dagger,\pi_{-i}}(s) - V_i^{\pi}(s)\right\} \le \epsilon$.

It is also possible to define ϵ-perfect nonstationary CCE in a natural way, but we will not need to do so (as such equilibria are easily seen to be computationally feasible to compute, yet also impossible to learn in the model of PAC learning of stochastic games we consider, see Section 2.3). When stationarity (or lack thereof) of π is clear from context, we will drop the words "stationary" and "nonstationary" from the above definitions; furthermore, we will drop the word "Markov" when referring to the above definitions since all equilibria we consider are Markovian.

Nash equilibrium. We next define approximate Markov-Nash equilibria in stochastic games.

Definition 2 (Nash equilibrium) *For $\epsilon > 0$, the notions:*

- ϵ-approximate stationary Markov Nash equilibrium *(abbreviated* ϵ-stationary NE*)*
- ϵ-approximate nonstationary Markov Nash equilibrium *(abbreviated* ϵ-nonstationary NE*)*
- ϵ-approximate perfect Markov Nash equilibrium *(abbreviated* ϵ-perfect NE*)*

are defined to be ϵ-stationary CCE, ϵ-nonstationary CCE, and ϵ-perfect CCE, respectively, which are also product policies.

In the literature, perfect NE is also referred to as *Markov perfect equilibrium* (Maskin and Tirole, 1988) for stochastic games. It is known that perfect a Markov NE always *exists* for discounted SGs (Shapley, 1953; Fink, 1964), thus so do the stationary and nonstationary NE, and the corresponding CCE counterparts.

2.3. The PAC-RL model for stochastic games

Our results on *learning SGs* (Section 4) operate in the *probably approximately correct (PAC) learning model* of RL, which is standard in the literature (Kakade, 2003; Azar et al., 2017). In particular, at the onset of the algorithm, the agents have no information about the transitions $\mathbb{P}$, the reward functions r_i, or the initial state distribution μ; only the parameters S, γ are known to all agents, and each agent i knows A_i. The agents' only access to the SG is through the ability to repeatedly choose some joint (perhaps nonstationary) policy π and then sample a trajectory $(s_1, \boldsymbol{a}_1, \boldsymbol{r}_1, s_2, \boldsymbol{a}_2, \boldsymbol{r}_2, \ldots) \sim (\mathbb{G}, \pi)$. In the *centralized setting* (studied, for instance, in Liu et al. (2021)), all agents may communicate with a central coordinator who may choose π and observe the entire trajectory. Our algorithm may in fact be implemented in the stricter *decentralized setting* with public randomness, which is discussed in Section 4.4. Note that in either case, the agents need to efficiently *explore* the environment, as the trajectory data they access might not visit all the state-action pairs with large rewards often enough, usually leading to a poor sample complexity.

In order to have computationally efficient algorithms, each trajectory must be truncated at some step $H \in \mathbb{N}$, after which another trajectory is started anew. In the infinite-horizon discounted setting, we, therefore, assume that agents can choose to stop playing the SG at some point: in particular for a desired error parameter $\epsilon > 0$, all agents will truncate after $H := \frac{\log 1/\epsilon}{1-\gamma}$ steps, incurring only ϵ loss from steps $h > H$.

2.4. Turn-based stochastic games

A stochastic game $\mathbb{G}$ is called a *turn-based stochastic game* if, at each state $s \in \mathcal{S}$, there is a single player $i \in [m]$ (called the *controller* of state s, and denoted $i = \mathtt{cr}(s)$) whose action at s entirely determines the reward and the transition to the next state. Formally, for all $j \in [m]$ there is some function $r'_j : \mathcal{S} \times (\mathcal{A}_1 \sqcup \cdots \sqcup \mathcal{A}_m) \to [-1, 1]$ and some transition kernel $\mathbb{P}' : \mathcal{S} \times (\mathcal{A}_1 \sqcup \cdots \sqcup \mathcal{A}_m) \to \Delta(\mathcal{S})$ so that $r_j(s, \boldsymbol{a}) = r'_j(s, a_{\mathtt{cr}(s)})$ for all $s \in \mathcal{S}$, $j \in [m]$, $\boldsymbol{a} = (a_1, \ldots, a_m) \in \mathcal{A}$, and so that $\mathbb{P}(\cdot|s, \boldsymbol{a}) = \mathbb{P}'(\cdot|s, a_{\mathtt{cr}(s)})$ for all $s \in \mathcal{S}$, $\boldsymbol{a} \in \mathcal{A}$. It is evident that in turn-based stochastic games, the notions of ϵ-CCE and ϵ-NE are equivalent, both for stationary and nonstationary policies (and the same holds for the perfect versions of the equilibria), since in such games we may restrict to product policies without loss of generality.

2.5. PPAD and the generalized circuit problem

The problems of computing equilibria of the types defined in Definitions 1 and 2 are instances of *total search problems* (Megiddo and Papadimitriou, 1991). In particular, they lie in the class TFNP, which is the class of binary relations $\mathcal{P} \subset \{0,1\}^\star \times \{0,1\}^\star$ so that for all $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^\star$, there is a polynomial-time algorithm that can determine whether $\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})$ holds, and so that for all $\boldsymbol{x} \in \{0,1\}^\star$, there is some $\boldsymbol{y} \in \{0,1\}^\star$ with $|\boldsymbol{y}| \leq \mathrm{poly}(|\boldsymbol{x}|)$ so that $\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})$ holds. Approximate equilibrium computation in stochastic games is seen to be in TFNP as follows: $\boldsymbol{x}$ represents the description of the stochastic game and the approximation requirement, $\boldsymbol{y}$ represents a proposed approximate equilibrium policy, and $\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})$ holds if $\boldsymbol{y}$ is an approximate equilibrium of $\boldsymbol{x}$. For all notions of equilibria we have defined, an equilibrium always exists (Fink, 1964; Solan and Vieille, 2015) and it may thus be easily seen that there exists an approximate one that has polynomial bit description in the description of the game and the approximation requirement. Moreover, it may be efficiently checked whether a proposed policy $\boldsymbol{y}$ is indeed an approximate equilibrium for the game represented by $\boldsymbol{x}$.

The class PPAD (Papadimitriou, 1994) is defined as the class of all problems in TFNP which have a polynomial-time reduction to the End-of-the-Line (EOTL) problem; we refer the reader to Papadimitriou (1994); Daskalakis et al. (2009) for a description of EOTL, as we do not need to directly use its definition. To establish our hardness results (i.e. PPAD-completeness for computing approximate stationary CCE), we instead use the fact, proven in Rubinstein (2016), that the ϵ-GCircuit problem (Definition 8) is PPAD-complete for some absolute constant ϵ (Theorem 9). Additional preliminaries regarding PPAD and ϵ-GCircuit are presented in Appendix B.1.

3. PPAD-hardness for *Stationary* Equilibria

We next state our main lower bounds, which establish hardness for finding stationary CCE in infinite-horizon discounted SGs. To do so, we first prove hardness for finding stationary NE in the special case of turn-based discounted SGs, and then note that in such games, stationary Nash equilibria and stationary coarse correlated equilibria coincide (so do the perfect versions).

Theorem 3 (PPAD-hardness for perfect equilibria) *There is a constant $\epsilon > 0$ so that the problem of computing ϵ-perfect NE in 2-player, $1/2$-discounted turn-based stochastic games is PPAD-hard. Thus, computing ϵ-perfect CCE in 2-player, $1/2$-discounted stochastic games is PPAD-hard.*

We next turn to showing intractability results for the weaker, "non-perfect" notions of equilibria, namely ϵ-stationary NE in turn-based stochastic games, and more generally, ϵ-stationary CCE in stochastic games. The motivation for pursuing these results is two-fold: First, they are standard equilibrium concepts and thus a natural target where to extend our intractability results. Second, if the initial state distribution μ is not sufficiently *exploratory*, it is impossible, in the PAC-RL model, to *learn* notions of equilibrium which are perfect because these require a condition to hold for *each state*. Accordingly, in the learning setting we consider in Section 4, our algorithms only learn the (non-perfect) notions of ϵ-nonstationary NE in turn-based games and ϵ-nonstationary CCE in general stochastic games. By establishing intractability results for their stationary counterparts, we are able to argue that our learning results cannot be extended to stationary non-perfect equilibria.

Theorem 4 (PPAD-hardness for non-perfect equilibria) *There are constants $\epsilon, c > 0$ so that:*

- *The problem of computing c/S-stationary NE in 2-player, 1/2-discounted turn-based stochastic games is PPAD-hard; and thus so is the problem of computing c/S-stationary CCE in 2-player, 1/2-discounted stochastic games.*
- *Under the "PCP for PPAD conjecture" (Conjecture 10), the problem of computing ϵ-stationary NE in 2-player, 1/2-discounted turn-based stochastic games is PPAD-hard; and thus so is the problem of computing ϵ-stationary CCE in 2-player, 1/2-discounted stochastic games.*

3.1. Proof overview for Theorems 3 and 4

The proofs of Theorems 3 and 4 proceed by reducing the (ϵ, δ)-GCircuit problem, introduced in Appendix B.1, to the problem of finding approximate stationary Nash equilibria in 2-player general-sum turn-based stochastic games. We overview here the proof of Theorem 3, which uses PPAD-hardness of the $(\epsilon, 0)$-GCircuit problem (Theorem 9); the proof of Theorem 4 is similar, except that the second part of the theorem relies on the PPAD-hardness of the (ϵ, δ)-GCircuit problem for some constants $\epsilon, \delta > 0$, which is not yet known but is rather the content of Conjecture 10.

Given an instance $\mathcal{C}$ of the $(\epsilon, 0)$-GCircuit problem, we wish to construct a 2-player, 1/2-discounted turn-based stochastic game $\mathbb{G}$ so that given an ϵ'-perfect NE of $\mathbb{G}$, for some $\epsilon' > 0$, we can compute an assignment of values to the nodes of the generalized circuit $\mathcal{C}$ which ϵ-satisfies all gates. To do so, we construct $\mathbb{G}$ by creating a number of "gadgets" (Appendix B.4), each of which implements one gate in the generalized circuit instance $\mathcal{C}$. Each such gadget consists of a constant number of states of $\mathbb{G}$, each of which is controlled by a single player who can take one of two actions, say $\{0, 1\}$, at the state. A stationary policy of $\mathbb{G}$ then is equivalent to a mapping $\pi : \mathcal{S} \to [0, 1]$, where $\mathcal{S}$ is the state space of $\mathbb{G}$ and, for $s \in \mathcal{S}$, $\pi(s)$ denotes the probability that the agent controlling s chooses action 1 at state s. We then define transitions and rewards for the states in each gadget in a way that forces any ϵ'-perfect NE π of $\mathbb{G}$ to have the property that the restriction of π to the gadget ϵ-approximately satisfies the gate corresponding to the gadget. By defining transitions between the gadgets in a way that mirrors the structure of the circuit $\mathcal{C}$, we achieve the desired reduction. The gadgets we use are somewhat reminiscent of the gadgets used in Daskalakis et al. (2009) to show PPAD-hardness of computing Nash equilibria in graphical games.

The reduction as described above suffices to show PPAD-hardness of computing ϵ'-perfect NE of $|\mathcal{S}|$-player stochastic turn-based games, namely games in which a different player chooses an action at each state. In this case, because of the discounting, each player essentially only strategizes about their short-term rather than long-term reward, and the equilibrium constraints remain essentially local and faithful to the intended gadget functionality. To establish hardness for *2-player* games, care must be taken to ensure that for each player, the rewards assigned to them from different gadgets do not conflict with each other. In fact, as we describe in Appendix B.5, conflicts could arise for the gadgets that we use. To overcome this issue, we show (in Lemma 25) how to map the given generalized circuit instance $\mathcal{C}$ to an equivalent one, $\mathcal{C}'$, which has the property that conflicts as described above *cannot* arise. In particular, we introduce the notion of *valid colorings* (Definition 24) to establish a formal condition on the circuit instance $\mathcal{C}'$ which guarantees that there will be no conflicts. We then show, using our game gadgets, how to map the instance $\mathcal{C}'$ (equipped with a valid coloring) to a 2-player stochastic game $\mathbb{G}$ whose ϵ'-perfect NE yields an assignment of $\mathcal{C}$ that ϵ-satisfies all gates.

The above description omits some details; for instance, rather than reducing directly to the problem of ϵ'-perfect NE in stochastic games, we instead reduce ϵ-GCircuit to the problem of computing *perfect well-supported Nash equilibria in stage games* (Definition 15), which in turn reduces to perfect NE (Lemma 14). The full proofs for our lower bounds are in Appendix B.

4. A Decentralized MARL Algorithm

Given the intractability results discussed in the previous section, it is natural to relax the notion of (Markov) *stationary* equilibria. A very natural relaxation, and indeed one considered in a number of recent works, is to drop the requirement of *stationarity* of the equilibrium policy. While *computing* approximate nonstationary (coarse) correlated equilibria in general-sum discounted stochastic games is straightforward via backward induction, the *learning* problem, in which the stochastic game is *unknown* and the players must employ *exploratory* policies to learn an equilibrium, is significantly less trivial. All prior work (on finite-horizon episodic SGs) for learning nonstationary equilibria either requires a number of samples exponential in the number of players (Bai and Jin, 2020; Liu et al., 2021), or else does not compute Markov policies (Mao and Başar, 2021; Jin et al., 2021; Song et al., 2021). In this section, we present Theorem 5, which establishes a decentralized

learning algorithm that learns a Markov nonstationary equilibrium in time polynomial in the number of players.

Reduction to the finite-horizon case. To learn ϵ-nonstationary CCE in infinite-horizon discounted games, we use a standard reduction to computing ϵ-nonstationary CCE in *finite-horizon undiscounted* games, which is the setting studied by most of the aforementioned work (Mao and Başar, 2021; Jin et al., 2021; Song et al., 2021; Liu et al., 2021). A finite-horizon stochastic game $\mathbb{G} = (\mathcal{S}, (\mathcal{A}_i)_{i\in[m]}, \mathbb{P}, (r_i)_{i\in[m]}, H, \mu)$ is defined identically to the infinite-horizon case (Section 2.1), except that the discount factor γ is replaced by an integer $H \in \mathbb{N}$, denoting the *horizon*; as such, the total reward is no longer discounted, but is summed from steps $h = 1$ to H (see Appendix C.3 for further details).

4.1. The SPoCMAR algorithm

We next introduce our main algorithm, called SPoCMAR (**S**tage-based **Po**licy **C**over for **M**ulti-**A**gent Learning with **R**max), presented in full in Algorithm 1 (see Appendix C). The SPoCMAR algorithm combines multiple tools from the literature in order to learn a Markov equilibrium while breaking the curse of multi-agents: it uses an adversarial bandit routine at each state (see Appendix C.2), similar to the recent works of Mao and Başar (2021); Song et al. (2021); Jin et al. (2021), optimistic rewards inspired by those in the Rmax algorithm (Brafman and Tennenholtz, 2002) to induce exploration, as well as a *policy cover* (see, e.g., Agarwal et al. (2020); Foster et al. (2021); Jin et al. (2020)) to ensure that exploratory policies learned in the past are not forgotten.

High-level overview & challenges. At a high level, these ingredients are combined in the following manner. Suppose first that we had a collection of policies that explored the entire state space (namely, a policy cover); then we could learn an approximate equilibrium in a backward-inductive manner, as follows. For each $h \in \{H, H-1, \ldots, 1\}$, and for each state at step h, play some element of the policy cover which reaches that state, and then choose an action according to a bandit no-regret algorithm at that state, with rewards for the bandit learner given by an estimate of the value function for the approximate CCE that we have already learned at steps $h+1$ through H. After sufficiently many rounds, the bandit no-regret learners ensure that there is no useful deviation at all states at each step h, which suffices to show that no player can usefully deviate to any fixed policy, thus establishing that we have found a Markov nonstationary CCE.

A key technical challenge is how to cope with the fact that we do not have a policy cover that explores the entire state space; in fact, such a "complete" cover may not exist, since some states may not be reachable under *any* policy. Additionally, it may be the case that there are some reachable states but that it takes exponentially (in the number of agents) many samples to reach. For instance, suppose there is a fixed starting state s_1, and some state $s_2^\star$ at step 2 is only reachable if the agents play according to a fixed joint action $\boldsymbol{a}_1^\star \in \mathcal{A}$ at s_1. Via standard multi-armed bandit lower bounds (Lattimore and Szepesvári, 2020), it is straightforward to show that finding a policy which reaches $s_2^\star$ with nontrivial probability requires $\Omega(|\mathcal{A}|) = \Omega\left(\prod_{i=1}^m A_i\right) \geq \Omega(\exp(m))$ many samples.

We overcome both of the aforementioned challenges by adding an exploration bonus to each state which has not been sufficiently explored in the past (similar to the Rmax algorithm (Brafman and Tennenholtz, 2002)). The use of such bonuses allows the algorithm SPoCMAR to reach any state that any agent can reach as a result of deviating from the approximate CCE being learned. Roughly speaking, this follows since, in the presence of such a deviation, the exploration bonuses

will propagate, via Bellman updates, to some profitable action that some player's bandit learner at some state can take. In such a case, the bandit learner will eventually take that action over time. Hence, by taking note of which states are visited by each policy played in the course of SPoCMAR, we can build up a sufficiently good policy cover over time. Since the order in which we find new states may not necessarily be decreasing in h (e.g., we may first visit a new state at step H, then step $H/2$, then at step $H-1$), the algorithm needs to operate in multiple *stages*, computing new value function estimates for each stage.

Detailed description. In more detail, the algorithm proceeds as follows: it takes as input the parameters $m, \mathcal{S}, \mathcal{A}, H$ of the finite-horizon stochastic game, as well as additional parameters $K, N_{\text{visit}} \in \mathbb{N}$, $p \in (0,1)$ whose interpretations will be explained below. SPoCMAR's computation proceeds in a number of *stages* q (step 4). At a high level, the algorithm will attempt to find, for each pair (h,s), a policy $\pi_{h,s}^{\text{cover}}$ which visits (h,s) with nontrivial probability (namely, at least p; see step 31). For each stage q, the algorithm loops over $h = H, H-1, \ldots, 1$ (step 7), and for each such value of h, the algorithm first re-initializes all adversarial bandit instances according to a bandit algorithm satisfying the guarantee of Theorem 28 (step 9), and then loops over all policies in a *policy cover* set Π_h^q (step 10), which is the set of current non-null cover policies $\pi_{h,s}^{\text{cover}}$. To deal with the case that Π_h^q is empty (e.g., if $q=1$), in step 10 the algorithm also loops over the policy $\pi^{\mathcal{U}}$ which prescribes all agents to choose their action uniformly at random at each state.

For each $\pi \in \Pi_h^q \cup \pi^{\mathcal{U}}$, the algorithm executes a policy $\overline{\pi}$ (step 13), which is identical to π except that at step h each agent plays according to her adversarial bandit instance. The policy $\overline{\pi}$ is executed for a total of K *episodes* (step 11). The algorithm then uses the trajectory data drawn from $\overline{\pi}$ at each episode to update the adversarial bandit instances at step h (steps 15 to 19). Using the data collected from all K episodes for each cover policy in $\Pi_h^q \cup \pi^{\mathcal{U}}$, the algorithm then computes a function $\overline{V}_{i,h}^q : \mathcal{S} \to \mathbb{R}$, representing a value function estimate for a coarse correlated equilibrium, in steps 22 through 25. Crucially, the estimates $\overline{V}_{i,h}^q$ depend on $\overline{V}_{i,h+1}^q$, necessitating the backward loop over h in step 7.

After this backward loop over h has been completed, the algorithm constructs a policy $\widetilde{\pi}^q$ (step 27) representing an estimate for an approximate CCE given the data collected at stage q. By drawing a total of N_{visit} additional trajectories from $\widetilde{\pi}^q$, in step 28 it uses the sub-procedure EstVisitation (Algorithm 2) to estimate the state visitation probabilities for $\widetilde{\pi}^q$. If $\widetilde{\pi}^q$ does not visit any new pairs (h,s) with significant probability, SPoCMAR terminates, outputting $\widetilde{\pi}^q$; otherwise, it sets $\pi_{h,s}^{\text{cover}} \leftarrow \widetilde{\pi}^q$ for newly visited pairs (h,s), and then proceeds to the following stage.

4.2. Guarantee for SPoCMAR

In Theorem 5 we state the main guarantee for SPoCMAR, which shows that for finite-horizon general-sum stochastic games, SPoCMAR achieves sample and computational complexities polynomial in all relevant parameters, including the number of players.

Theorem 5 *Fix any $\epsilon, \delta > 0$. For appropriate settings of the parameters N_{visit}, K, p (specified in Appendix C.4), SPoCMAR outputs an ϵ-nonstationary Markov CCE with probability at least $1-\delta$ after sampling at most $O\left(\frac{H^{10}S^3\iota^2 \max_{i\in[m]} A_i}{\epsilon^3}\right)$ trajectories, where $\iota = \log\left(\frac{SH \max_i A_i}{\epsilon\delta}\right)$. The computational complexity of SPoCMAR is polynomial in $H, S, \max_i A_i, 1/\epsilon, \log 1/\delta$.*

Combining SPoCMAR with the reduction from infinite-horizon discounted games to finite-horizon (undiscounted) games described in Appendix C.3, we obtain the following as an immediate corollary of Theorem 5:

Corollary 6 *There is a polynomial-time algorithm which learns a ϵ-nonstationary Markov CCE in γ-discounted general-sum stochastic games using $\widetilde{O}\left(\frac{S^3 \max_{i\in[m]} A_i}{(1-\gamma)^{10}\epsilon^3}\right)$ trajectories.*

The proof of Theorem 5 can be found in Appendix C, with an overview provided next. We will also explain next how SPoCMAR (Algorithm 1) can be implemented in a *decentralized* manner (with access to shared randomness).

4.3. Proof overview for Theorem 5

We now overview the proof of Theorem 5. Let $\widehat{\mathcal{V}}$ denote the value of $\mathcal{V}$ at the termination of SPoCMAR, and $\widehat{q}$ denote the value of the final stage of SPoCMAR. The main tool in the proof is to construct an intermediate game, denoted $\mathbb{G}_{\widehat{\mathcal{V}}}$ (Appendix C.5): we will first show that the output policy of SPoCMAR is an ϵ-CCE with respect to the game $\mathbb{G}_{\widehat{\mathcal{V}}}$, and then, using the termination criterion of SPoCMAR, we will show that this implies that $\widehat{\pi}$ is an ϵ-CCE with respect to the true game $\mathbb{G}$.

The game $\mathbb{G}_{\widehat{\mathcal{V}}}$ is constructed in a similar way as an intermediate MDP used in the analysis of the Rmax algorithm (Brafman and Tennenholtz, 2002; Jin et al., 2020). For tuples $(h,s) \notin \widehat{\mathcal{V}}$, $\mathbb{G}_{\widehat{\mathcal{V}}}$ transitions, at (h,s), to a special sink state at which all agents receive reward 1 (the maximum possible reward) at all future steps; for all $(h,s) \in \widehat{\mathcal{V}}$, the rewards and transitions of $\mathbb{G}_{\widehat{\mathcal{V}}}$ at (h,s) are identical to those of $\mathbb{G}$. By ensuring that the parameter K passed to SPoCMAR is sufficiently large, we may guarantee that, during stage $\widehat{q}$, SPoCMAR visits all $(h,s) \in \widehat{\mathcal{V}}$ sufficiently many times to compute accurate estimates of $V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)$ for such $(h,s) \in \widehat{\mathcal{V}}$. Since, for all $(h,s) \notin \widehat{\mathcal{V}}$, we have $V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s) = H+1-h$, it is possible to show (Lemma 32) that $\left|\overline{V}_{i,h}^{\widehat{q}}(s) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)\right|$ is small for *all* $(h,s) \in [H] \times \mathcal{S}$ and all $i \in [m]$.

Using the no-regret property of the adversarial bandit instances used by each player for each (h,s), we then obtain (Lemmas 33 and 34) that $\widehat{\pi}$ is an ϵ-CCE of $\mathbb{G}_{\widehat{\mathcal{V}}}$. To derive such a guarantee for the true SG $\mathbb{G}$, we use two facts: first, by the optimistic nature of the reward of $\mathbb{G}_{\widehat{\mathcal{V}}}$, the value function of $\mathbb{G}_{\widehat{\mathcal{V}}}$ is always an *upper bound* on the value function of $\mathbb{G}$, and second, by the termination criterion of SPoCMAR, the probability that a trajectory $(s_1, s_2, \ldots, s_H) \sim (\mathbb{G}, \widehat{\pi})$ visits any state $(h, s_h) \notin \widehat{\mathcal{V}}$ is small (Lemma 35). These arguments are worked out in detail in Lemma 36.

Reduction from infinite-horizon to finite-horizon. Finally, to derive Corollary 6, we remark that there is a simple reduction from episodic learning of an infinite-horizon discounted game with discount factor γ to episodic learning of a finite-horizon game with horizon $H := \frac{\log 1/\epsilon}{1-\gamma}$ (see Appendix C.3). Owing to the fact that $\gamma^H \leq \epsilon$, this reduction preserves nonstationary equilibria up to an additive approximation of ϵ.

4.4. Implementing SPoCMAR in a decentralized manner

So far, we have described SPoCMAR as a centralized algorithm, declining to make distinctions between the computations performed by each agent. We now proceed to explain how SPoCMAR can be implemented in a *decentralized* way, namely in the following setting:

1. All agents know $m, \mathcal{S}, \mathcal{A}, H$, as well as ϵ, δ, so that they may each compute the additional parameters K, N_{visit}, p passed to `SPoCMAR`.

2. For each trajectory of $\mathbb{G}$ sampled from a policy π, each agent $i \in [m]$ sees only the states, their actions, and their rewards.

3. The agents may access a common string of uniformly random bits during the course of the algorithm, but no communication between agents is allowed.

4. The agents are required to be able to sample from the output policy $\widehat{\pi}$, again using only common randomness (and no communication).

The only existing decentralized learning algorithm for multi-player general-sum stochastic games, V-learning (Song et al., 2021; Mao and Başar, 2021; Jin et al., 2021), shares all requirements above except item 3. In particular, while V-learning requires public random bits to sample a trajectory from its output CCE policy, such bits are not used in the process of *learning* the policy.

To implement `SPoCMAR` in a decentralized manner, we first describe how agents can sample trajectories from $\overline{\pi}$ (defined in step 13) without communicating: note that the first $h-1$ steps of $\overline{\pi}$ are given by $(\widetilde{\pi}^q_1, \ldots, \widetilde{\pi}^q_{h-1})$, for some stage q: furthermore, $\widetilde{\pi}^q_{h'}(\cdot|s)$ is a uniform mixture over some number $J^q_{h',s}$ of joint action profiles (step 24; we denote the parameter $J_{h',s}$ at stage q by $J^q_{h',s}$). Thus, if each agent stores its action taken in each of the $J^q_{h',s}$ such steps for all s, h', q, the agents may draw an action sampled from $\widetilde{\pi}^q_{h'}$ by using the public randomness to sample a uniformly random element of $[J^q_{h',s}]$. Noting that $J^q_{h',s} \le K(S+1)$ for all h', s, we see that the total number of common random bits needed to execute $\overline{\pi}$ is $O(H^3S^2K\log(SK))$.

It is straightforward that the bandit updates in steps 15 through 19 as well as the computation of $\overline{V}^q_{i,h}$ in steps 22 through 25 may be implemented in a decentralized way (in particular, each agent i only computes its own value estimate $\overline{V}^q_{i,h}$). Finally, the procedure `EstVisitation` allows each agent $i \in [m]$ to compute their own estimates of $\widehat{d}^q_h$, for all h, q, which all coincide since the states drawn from each trajectory are common knowledge. In order to play the policy $\widetilde{\pi}^q$ passed to `EstVisitation`, the same strategy as described above may be used, which requires a total of $O(H^2S\log(SK))$ bits of common randomness over all stages $q \ge 1$.

In sum, executing `SPoCMAR` in a decentralized way requires $O(H^3S^2K\log(SK))$ bits of common randomness. For the K described in Appendix C.4, this leads to $\widetilde{O}\left(\frac{H^7S^3\max_{i\in[m]}A_i}{\epsilon^3}\right)$ bits.

Acknowledgement

The authors would like to thank the anonymous reviewers for the helpful comments. C.D. and N.G. were supported by NSF Awards CCF-1901292, DMS-2022448 and DMS2134108, a Simons Investigator Award, the Simons Collaboration on the Theory of Algorithmic Fairness, and a DSTA grant. N.G. was also supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship. K.Z. was supported by a DSTA grant and Simons-Berkeley Research Fellowship. This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing.

References

Alekh Agarwal, Mikael Henaff, Sham M. Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Robert J Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18, 1987.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

Yakov Babichenko, Christos Papadimitriou, and Aviad Rubinstein. Can almost everybody be almost happy? PCP for PPAD and the inapproximability of Nash. *arXiv preprint arXiv:1504.02411*, 2015.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Clément L. Canonne. A short note on learning discrete distributions. 2020.

Shicong Cen, Fan Chen, and Yuejie Chi. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization, 2022.

Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Krishnendu Chatterjee, Rupak Majumdar, and Marcin Jurdziński. On Nash equilibria in stochastic games. In *International Workshop on Computer Science Logic*, pages 26–40. Springer, 2004.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Computing Nash equilibria: Approximation and smoothed complexity. In *In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009.

Anne Condon. On algorithms for simple stochastic games. In *Advances in Computational Complexity Theory*, pages 51–72, 1990.

Anne Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, 1992.

Vincent Conitzer and Tuomas Sandholm. New complexity results about Nash equilibria. *Games and Economic Behavior (GEB)*, 63(2):621–641, 2008.

Qiwen Cui and Simon S Du. When is offline two-player zero-sum Markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022.

George B. Dantzig. A proof of the equivalence of the programming problem and the game problem. *Koopmans, T. C., editor(s), Activity Analysis of Production and Allocation*, 1951.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Constantinos Daskalakis, Dylan Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Xiaotie Deng, Yuhao Li, David Henry Mguni, Jun Wang, and Yaodong Yang. On the complexity of computing Markov perfect equilibrium in general-sum stochastic games. *arXiv preprint arXiv:2109.01795*, 2021.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5166–5220. PMLR, 17–23 Jul 2022.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum markov games, 2022.

Kousha Etessami and Mihalis Yannakakis. On the complexity of Nash equilibria and other fixed points. *SIAM Journal on Computing (JoC)*, 39(6):2531–2597, 2010.

John Fearnley, Spencer Gordon, Ruta Mehta, and Rahul Savani. Unique end of potential line. *CoRR*, abs/1811.03841, 2018.

Arlington M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, series A-I (mathematics)*, 28(1):89–93, 1964.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *CoRR*, abs/2112.13487, 2021.

Roy Fox, Stephen M. Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4414–4425. PMLR, 28–30 Mar 2022.

Angeliki Giannou, Kyriakos Lotidis, Panayotis Mertikopoulos, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general stochastic games, 2022.

Itzhak Gilboa and Eitan Zemel. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior (GEB)*, 1(1):80–93, 1989.

Amy Greenwald, Keith Hall, Roberto Serrano, et al. Correlated Q-learning. In *International Conference on Machine Learning*, volume 3, pages 242–249, 2003.

Junling Hu and Michael P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4(Nov):1039–1069, 2003.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 4870–4879. PMLR, 2020.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–A simple, efficient, decentralized algorithm for multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021.

Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games, 2022.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pages 267–274, 2002.

Sham M Kakade. On the sample complexity of reinforcement learning, 2003.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent rl in markov games with a generative model, 2022.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Michael L Littman. Friend-or-foe Q-learning in general-sum games. In *International Conference on Machine Learning (ICML)*, volume 1, pages 322–328, 2001.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum Markov games. *arXiv preprint arXiv:2110.05682*, 2021.

Eric Maskin and Jean Tirole. A theory of dynamic oligopoly, I: Overview and quantity competition with large fixed costs. *Econometrica: Journal of the Econometric Society*, pages 549–569, 1988.

Nimrod Megiddo and Christos H. Papadimitriou. On total functions, existence theorems and computational complexity. *Theor. Comput. Sci.*, 81(2):317–324, apr 1991. ISSN 0304-3975.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3168–3176, 2015.

Abraham Neyman and Sylvain (editors) Sorin. *Stochastic games and applications*, volume 570. Springer Science & Business Media, 2003.

Christos H Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and system Sciences*, 48(3):498–532, 1994.

Christos H Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)*, 55(3):1–29, 2008.

Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 258–265. IEEE, 2016.

Aviad Rubinstein. Inapproximability of Nash equilibrium. *SIAM Journal on Computing*, 47(3): 917–959, 2018.

Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *CoRR*, abs/1610.03295, 2016.

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

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Eilon Solan and Nicolas Vieille. Stochastic games. *Proceedings of the National Academy of Sciences (PNAS)*, 112(45):13743–13746, 2015.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

Masayuki Takahashi. Stochastic games with infinitely many strategies. *Journal of Science of the Hiroshima University, Series A-I (mathematics)*, 26(2):123–134, 1962.

John von Neumann. Zur Theorie der Gesellschaftsspiele. In *Math. Ann.*, pages 295–320, 1928.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online Reinforcement Learning in Stochastic Games. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Proceedings of the the 33rd Annual Conference on Learning Theory (COLT)*, 2020.

Kaiqing Zhang, Sham M. Kakade, Tamer Basar, and Lin F. Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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

Runyu Zhang, Qinghua Liu, Huan Wang, Caiming Xiong, Na Li, and Yu Bai. Policy optimization for markov games: Unified framework and faster convergence, 2022.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C. Parkes, and Richard Socher. The ai economist: Improving equality and productivity with ai-driven tax policies, 2020.

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. *arXiv preprint arXiv:2202.07511*, 2022.

Martin Zinkevich, Amy Greenwald, and Michael Littman. Cyclic equilibria in Markov games. *Advances in Neural Information Processing Systems (NeurIPS)*, 18, 2005.

Uri Zwick and Mike Paterson. The complexity of mean payoff games on graphs. *Theoretical Computer Science*, 158(1-2):343–359, 1996.

Appendix A. Related Work

In this section, we summarize the most related work in the literature.

Equilibrium computation complexity in games. The computational complexity of finding equilibria has been extensively studied in normal-form games. It is known that for two-player zero-sum normal-form games, Nash equilibria can be computed efficiently using either linear programming (Dantzig, 1951) or decentralized no-regret learning algorithms (Cesa-Bianchi and Lugosi, 2006). For general-sum normal-form games, however, computing a Nash equilibrium is known to be PPAD-complete (Daskalakis et al., 2009; Chen et al., 2006; Rubinstein, 2016), even for the two-player case. In contrast, (coarse) correlated equilibria (Aumann, 1987) can be found via either linear programming (Gilboa and Zemel, 1989; Papadimitriou and Roughgarden, 2008) or no-regret learning (Cesa-Bianchi and Lugosi, 2006) efficiently, even in this general-sum setting. The study of the complexity of equilibria computation in stochastic games has been comparatively scarce. Since stochastic games generalize normal-form games, the complexity of computing Markov perfect Nash equilibrium in general-sum SGs is thus at least PPAD-hard. Very recently, Deng et al. (2021) confirmed that computing Markov perfect NE is PPAD-complete (meaning that the problem of computing perfect CCE is also in PPAD). Other computational complexity results for stochastic games include the following: determining whether a pure-strategy NE exists in an SG is PSPACE-hard (Conitzer and Sandholm, 2008); determining if there exists a memoryless ϵ-NE in reachability SGs is NP-hard (Chatterjee et al., 2004); in *simple stochastic games* (Condon, 1990), a special class of zero-sum SGs introduced in (Shapley, 1953), deciding which player has the greater chance of winning is in NP $\cap$ co-NP (Condon, 1992; Zwick and Paterson, 1996), and computing an equilibrium is in UEOPL, a subclass of CLS, which is, in turn, a subclass of PPAD (see Etessami and Yannakakis (2010); Fearnley et al. (2018)).

Multi-agent RL in stochastic games. Stochastic games (Shapley, 1953) have served as the foundational framework of multi-agent reinforcement learning since Littman (1994). There is a rich literature on multi-agent RL in two-player zero-sum SGs, including the early studies of Littman (1994); Brafman and Tennenholtz (2002) as well as more recent ones with finite-sample complexity guarantees (Wei et al., 2017; Xie et al., 2020; Zhang et al., 2020; Sidford et al., 2020; Bai and Jin, 2020; Daskalakis et al., 2020; Bai et al., 2020; Liu et al., 2021; Cui and Du, 2022; Zhong et al., 2022). On the general-sum front, Q-learning based algorithms, e.g., Nash Q-learning (Hu and Wellman, 2003) and Friend-or-Foe Q-learning (Littman, 2001), have been shown to converge to the Nash equilibrium asymptotically under certain restrictive assumptions. Another variant, Correlated Q-learning (Greenwald et al., 2003), which also aims to find a correlated equilibrium in a similar spirit to the present work, was shown to converge empirically in several SGs. Related to our findings, Zinkevich et al. (2005) demonstrated that *value-based* RL methods cannot find *stationary* equilibria in arbitrary general-sum SGs, and advocated instead for an alternative *nonstationary* equilibrium concept – cyclic equilibria. Finally, decentralized multi-agent RL has attracted increased attention recently (Daskalakis et al., 2020; Sayin et al., 2021; Jin et al., 2021; Song et al., 2021; Mao and Başar, 2021), due to the fact that decentralized algorithms are more natural, require fewer assumptions, and typically avoid exponential dependence on the number of agents. Most relevant to our paper are the works of Jin et al. (2021); Song et al. (2021); Mao and Başar (2021), which developed the *V-learning* algorithm for learning nonstationary (C)CE in general-sum SGs. These algorithms have tighter sample complexity than ours, but the output policies are not Markovian.

Appendix B. Proofs for Section 3

In this section, we prove Theorems 3 and 4. To this end, we show how to reduce solving an instance of ϵ-GCircuit (or (ϵ, δ)-GCircuit; defined in Section B.1 below) to computing the appropriate notion of equilibria in a 2-player stochastic game. The proof is organized into the following parts:

- In Section B.1 we introduce some additional preliminaries regarding PPAD-hardness and (ϵ, δ)-GCircuit, and in Section B.2 we introduce some additional preliminaries regarding stochastic games.
- In Section B.3, we introduce a notion of Nash equilibrium in turn-based games, namely *well-supported Nash equilibrium in stage games (WSNE-SG)*. We show (roughly speaking) that computing approximate WSNE-SG reduces to computing approximate NE in turn-based games, i.e., it suffices to show PPAD-hardness for computing approximate WSNE-SG.
- In Section B.4, we show how each of the gates in Definition 8 can be implemented via a *gadget* with a constant number of states, transitions, and rewards in a turn-based stochastic game; these gadgets are similar in nature to those introduced by Daskalakis et al. (2009) in showing that computing Nash equilibria in graphical games is PPAD-complete.
- In Section B.5, we show how to combine the gadgets from Section B.4 to construct a turn-based stochastic game whose approximate WSNE-SG correspond to approximate assignments to a given GCircuit instance.

Unless otherwise stated, the policies considered in this section are Markov stationary policies.

B.1. Additional preliminaries for PPAD

For some $\epsilon > 0$ and reals x, y, we use $x = y \pm \epsilon$ to denote $x \in [y - \epsilon, y + \epsilon]$ throughout this section.[4]

Definition 7 (Generalized circuit) *A generalized circuit $\mathcal{C} = (V, \mathcal{G})$ is a finite set of* nodes V *and* gates $\mathcal{G}$. *Each gate in $\mathcal{G}$ is characterized as $G(\ell|v_1, v_2|v)$, where $G \in \{G_{\leftarrow}, G_{\times,+}, G_{<}\}$ denotes a gate type, $\ell \in \mathbb{R}^\star$ is a vector of real parameters (perhaps of length 0), $v_1, v_2 \in V \cup \{\perp\}$ denote the gate's* input nodes*, and $v \in V$ denotes the gate's* output node. *The collection of gates $\mathcal{G}$ satisfies the following property: for every two gates $G(\ell|v_1, v_2|v)$ and $G'(\ell'|v_1', v_2'|v')$, it holds that $v \neq v'$ (i.e., each gate computes a distinct, and thus well-defined, output node).*

For the purposes of proving hardness results, it is without loss of generality to assume that each node $v \in V$ is the output node of some gate in $\mathcal{G}$: for each node which is not the output node of some gate, we can add a gate specifying that the node is equal to $1 \pm \epsilon$. The resulting circuit is still a valid instance of the generalized circuit problem, and it has a solution by Brouwer's fixed point theorem. Any such solution is certainly a valid solution to the original generalized circuit instance.

Definition 8 ((ϵ, δ)-GCircuit) *Fix $\epsilon, \delta \in (0, 1)$. Given a generalized circuit $\mathcal{C} = (V, \mathcal{G})$, an assignment $\pi : V \to [0, 1]$ is said to* ϵ-approximately satisfy *some gate $G \in \mathcal{G}$, if the following holds:*

4. We remark that in Rubinstein (2016), $x \pm \epsilon$ was used to denote the fact that $x \in (y - \epsilon, y + \epsilon)$; this difference does not materially change any of the hardness results from Rubinstein (2016), since ϵ may be scaled down by any constant factor.

- *If for some constant $\zeta \in \{0, 1\}$, the gate G is of the form $G_{\leftarrow}(\zeta||v)$, then we have $\pi(v) = \zeta$;*
- *If for some constants $\xi, \zeta \in [-1, 1]$, the gate G is of the form $G_{\times,+}(\xi, \zeta|v_1, v_2|v_3)$, then we have*
$$\pi(v_3) = \max\{\min\{\xi \cdot \pi(v_1) + \zeta \cdot \pi(v_2), 1\}, 0\} \pm \epsilon;$$
- *If the gate G is of the form $G_{<}(|v_1, v_2|v_3)$, then we have $\pi(v_3) = \begin{cases} 1 \pm \epsilon, & \pi(v_1) \le \pi(v_2) - \epsilon \\ 0 \pm \epsilon, & \pi(v_1) \ge \pi(v_2) + \epsilon. \end{cases}$*

We stress that in the gates $G_{\times,+}(\xi, \zeta|v_1, v_2|v_3)$ and in $G_{<}(|v_1, v_2|v_3)$, we allow for $v_1 = v_2$. The problem (ϵ, δ)-`GCircuit` is the following: Given a generalized circuit $\mathcal{C} = (V, \mathcal{G})$, find an assignment $\pi : V \to [0, 1]$ (represented in binary) which ϵ-approximately satisfies all but a δ-fraction of the gates in $\mathcal{G}$. We then define the ϵ-`GCircuit` problem to be the $(\epsilon, 0)$-`GCircuit` problem.

The following theorem will be crucial for our hardness results.

Theorem 9 ((Rubinstein, 2018)) *There is a constant $\epsilon > 0$ so that ϵ-`GCircuit` is PPAD-complete.*

The problem of (ϵ, δ)-`GCircuit` for positive δ is not (yet) known to be PPAD-hard, but it has been conjectured to be so:

Conjecture 10 (PCP for PPAD conjecture (Babichenko et al., 2015)) *There are constants $\epsilon, \delta > 0$ so that `EOTL` has a polynomial-time reduction to (ϵ, δ)-`GCircuit`.*

We remark that Conjecture 10 is slightly weaker (i.e., more plausible) than (Babichenko et al., 2015, Conjecture 2), which states that the reduction is *quasilinear*.

Further, we remark that the definition of ϵ-`GCircuit` in Rubinstein (2018) uses some additional gates; it is straightforward to see that these gates may be implemented using the gates in Definition 8, meaning that the ϵ-`GCircuit` problem with the set of gates listed above is still PPAD-complete for constant ϵ (and Conjecture 10 is implied by (Babichenko et al., 2015, Conjecture 2)). For completeness, we have presented the details of this reduction in Appendix D.

B.2. Additional preliminaries for stochastic games

In this section, we introduce some additional definitions and lemmas which will be helpful in our proofs. Consider an infinite-horizon discounted stochastic game $\mathbb{G}$ and a stationary policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$. Then one can define the state-action value function $Q_i^{\pi} : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ under policy π as

$$Q_i^{\pi}(s, \boldsymbol{a}) := (1 - \gamma) \cdot \mathbb{E}_{(s_1, \boldsymbol{a}_1, s_2, \boldsymbol{a}_2, \ldots) \sim (\mathbb{G}, \pi)} \left[\sum_{h=1}^{\infty} \gamma^{h-1} r_i(s_h, \boldsymbol{a}_h) \,\middle|\, s_1 = s, \boldsymbol{a}_1 = \boldsymbol{a} \right],$$

which corresponds to the γ-discounted cumulative reward starting from $(s, \boldsymbol{a})$.

The *state-visitation distribution* under the policy π given that the initial state is s is defined as follows:

$$d_s^{\pi}(s') := (1 - \gamma) \cdot \sum_{h=1}^{\infty} \gamma^{h-1} \cdot \mathbb{P}_{s_h \sim (\mathbb{G}, \pi)} \left[s_h = s' | s_1 = s \right].$$

Note that d_s^{π} is a valid distribution over $\mathcal{S}$, i.e., $d_s^{\pi}(s') \ge 0$ for all s' and $\sum_{s' \in \mathcal{S}} d_s^{\pi}(s') = 1$.

Lemma 11 (Performance difference lemma (Kakade and Langford, 2002)) *Consider any two stationary policies $\pi : \mathcal{S} \to \Delta(\mathcal{A}),\ \pi' : \mathcal{S} \to \Delta(\mathcal{A})$. Then for all $s_1 \in \mathcal{S}$ and $i \in [m]$,*

$$V_i^{\pi}(s_1) - V_i^{\pi'}(s_1) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d_{s_1}^{\pi}} \mathbb{E}_{\boldsymbol{a} \sim \pi(s)}[Q_i^{\pi'}(s, \boldsymbol{a}) - V_i^{\pi'}(s)].$$

The following lemma is standard but we give a proof in Section F.1 for completeness:

Lemma 12 *For policies $\pi, \pi' \in \Delta(\mathcal{A})^{\mathcal{S}}$, it holds that, for all states $s \in \mathcal{S}$, joint actions $\boldsymbol{a} \in \mathcal{A}$, and agents $i \in [m]$,*

$$\left|V_i^{\pi}(s) - V_i^{\pi'}(s)\right| \leq \frac{1}{1-\gamma} \cdot \max_{s' \in \mathcal{S}} \|\pi(\cdot|s') - \pi'(\cdot|s')\|_1$$
$$\left|Q_i^{\pi}(s, \boldsymbol{a}) - Q_i^{\pi'}(s, \boldsymbol{a})\right| \leq \frac{\gamma}{1-\gamma} \cdot \max_{s' \in \mathcal{S}} \|\pi(\cdot|s') - \pi'(\cdot|s')\|_1.$$

B.3. Reductions between notions of equilibria

We begin by introducing a variant of Nash equilibrium which holds with respect to each stage game:

Definition 13 *We say that a product Markov policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is an* ϵ-perfect NE in stage games *(abbreviated* ϵ-PNE-SG*) if for all states $s \in \mathcal{S}$ and all agents $i \in [m]$,*

$$\max_{a_i \in \mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}(s)} \left[Q_i^{\pi}(s, (a_i, \boldsymbol{a}_{-i}))\right] - V_i^{\pi}(s) \leq \epsilon. \tag{2}$$

We also say that π is an ϵ-NE in stage games *(abbreviated* ϵ-NE-SG*) if for all agents $i \in [m]$,*

$$\mathbb{E}_{s \sim \mu}\left[\max_{a_i \in \mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}(s)} \left[Q_i^{\pi}(s, (a_i, \boldsymbol{a}_{-i}))\right] - V_i^{\pi}(s)\right] \leq \epsilon. \tag{3}$$

The below lemma reduces the problem of computing an ϵ-(P)NE-SG to computing an ϵ-(perfect) NE.

Lemma 14 *Consider a stationary product policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$. Then:*

- *If π is an ϵ-perfect NE, then it is an ϵ-PNE-SG.*
- *If π is an ϵ-NE, then it is an ϵ-NE-SG.*

The proof of Lemma 14 may be found in Section F.2. Next, we introduce a well-supported variant of the stage-game Nash equilibrium of Definition 13.

Definition 15 *Consider a product Markov policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$. For each $i \in [m]$ and $s \in \mathcal{S}$, define*

$$\epsilon_{i,s} := \max_{a_i' \in \mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}(s)} \left[Q_i^{\pi}(s, (a_i', \boldsymbol{a}_{-i}))\right] - \min_{a_i \in \mathcal{A}_i : \pi_i(a_i|s) > 0} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}(s)} \left[Q_i^{\pi}(s, (a_i, \boldsymbol{a}_{-i}))\right]. \tag{4}$$

We say that:

- π *is an* ϵ-perfect well-supported Nash equilibrium in stage games *(abbreviated* ϵ*-PWSNE-SG) if*

$$\max_{i\in[m],s\in\mathcal{S}} \epsilon_{i,s} \leq \epsilon;$$

- π *is an* ϵ-well-supported Nash equilibrium in stage games *(abreviated* ϵ*-WSNE-SG) if*

$$\max_{i\in[m]} \mathbb{E}_{s\sim\mu}[\epsilon_{i,s}] \leq \epsilon.$$

Lemma 16 reduces the problem of computing a well-supported Nash equilibrium in stage games to that of computing a Nash equilibrium in stage games.

Lemma 16 *Suppose the stochastic game* $\mathbb{G}$ *satisfies the property that at each state, all but* p *players have trivial action space (i.e., equal to a singleton). Given a product stationary policy* $\pi : \mathcal{S} \to \Delta(\mathcal{A})$*, then:*

- *If* π *is an* ϵ*-PNE-SG, we can construct in polynomial time a policy* $\pi' : \mathcal{S} \to \Delta(\mathcal{A})$ *which is a* $6 \cdot \sqrt{\frac{p\epsilon}{1-\gamma}}$*-PWSNE-SG.*

- *If* π *is an* ϵ*-NE-SG, we can construct in polynomial time a policy* $\pi' : \mathcal{S} \to \Delta(\mathcal{A})$ *which is a* $6 \cdot \sqrt{\frac{p\epsilon}{1-\gamma}}$*-WSNE-SG.*

The proof of Lemma 16 may be found in Section F.2. Combining the results presented in this section, we have the following:

Lemma 17 *Consider a turn-based stochastic game* $\mathbb{G}$*. Given a product stationary policy* $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$*, the following statements hold:*

- *If* π *is an* ϵ*-perfect NE, then we can construct in polynomial time a policy* $\pi' \in \Delta(\mathcal{A})^{\mathcal{S}}$ *which is a* $6 \cdot \sqrt{\frac{\epsilon}{1-\gamma}}$*-PWSNE-SG.*

- *If* π *is an* ϵ*-NE, then we can construct in polynomial time a policy* $\pi' \in \Delta(\mathcal{A})^{\mathcal{S}}$ *which is a* $6 \cdot \sqrt{\frac{\epsilon}{1-\gamma}}$*-WSNE-SG.*

Proof The lemma is an immediate consequence of Lemmas 14 and 16, noting that since $\mathbb{G}$ is turn-based we may take $p = 1$ in Lemma 16. ■

B.4. Implementing the gates with "stochastic game gadgets"

In this section, we introduce several "game gadgets" which show how to implement each of the arithmetic gates of Definition 8 using a constant number of states in a turn-based stochastic game.

Notation for turn-based games. Throughout this section, we will consider an m-player turn-based stochastic game $\mathbb{G} = (\mathcal{S}, (\mathcal{A}_i)_{i\in[m]}, \mathbb{P}, (r_i)_{i\in[m]}, \gamma, \mu)$ (Though we will eventually take $m = 2$, we will introduce the gadgets in this section for games with an arbitrary number of players). In our construction, we will have that for each player $i \in [m]$, $\mathcal{A}_i = \{0, 1\}$. Since, at each state $s \in \mathcal{S}$, there is a single agent (namely, $\texttt{cr}(s)$) whose action affects the reward and transition at that state, the value functions induced by a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ depend only on the values of $\pi_{\texttt{cr}(s)}(s)$, for each $s \in \mathcal{S}$. Thus, we may represent such a policy as a mapping from $\mathcal{S}$ to $[0, 1] = \Delta(\{0, 1\})$; with a slight abuse of notation, we denote this mapping also as $\pi : \mathcal{S} \to [0, 1]$. In particular, $\pi(s)$ is to be interpreted as the probability that agent $\texttt{cr}(s)$ plays the action $1 \in \mathcal{A}_{\texttt{cr}(s)} = \{0, 1\}$ at state s.

We will furthermore work with games $\mathbb{G}$ that have a designated sink state $s^0_{\text{sink}} \in \mathcal{S}$, so that $\mathbb{P}(s^0_{\text{sink}}|s^0_{\text{sink}}, \boldsymbol{a}) = 1$ for all $\boldsymbol{a} \in \mathcal{A}$, and $r_i(s^0_{\text{sink}}, \boldsymbol{a}) = 0$ for all $i \in [m], \boldsymbol{a} \in \mathcal{A}$. We allow $\texttt{cr}(s^0_{\text{sink}})$ to be arbitrary; the value of $\texttt{cr}(s^0_{\text{sink}})$ will have no relevance to any of our results. In particular, whenever the system reaches state s^0_{sink}, it stays there for all future steps and all agents accumulate 0 additional reward.

Unimprovable states. Given a turn-based game $\mathbb{G}$ and a policy $\pi : \mathcal{S} \to [0, 1]$, note that, for each $s \in \mathcal{S}$ and $\boldsymbol{a} = (a_1, \ldots, a_m) \in \mathcal{A}$, $Q^\pi_{\texttt{cr}(s)}(s, \boldsymbol{a})$ depends only on $a_{\texttt{cr}(s)} \in \mathcal{A}_{\texttt{cr}(s)}$. Thus, to simplify notation, we will use $Q^\pi_{\texttt{cr}(s)}(s, a_{\texttt{cr}(s)}) := Q^\pi_{\texttt{cr}(s)}(s, \boldsymbol{a})$. We say that a state $s \in \mathcal{S}$ is *ϵ-unimprovable under π* if

$$\max_{a'\in\{0,1\}} Q^\pi_{\texttt{cr}(s)}(s, a') - \min_{a\in\{0,1\}:\pi(s)\neq 1-a} Q^\pi_{\texttt{cr}(s)}(s, a) \leq \epsilon. \tag{5}$$

Note that the set of actions $a \in \{0, 1\}$ so that $\pi(s) \neq 1 - a$ is exactly the set of actions on which $\pi(s)$ puts positive probability. Hence the quantity in (5) is identical to the quantity $\epsilon_{i,s}$ defined more generally in (4). Thus, if π is an ϵ-PWSNE-SG (Definition 15), it holds that all states are ϵ-unimprovable. Furthermore, if π is an ϵ-WSNE-SG, then by Markov's inequality, for all $k \geq 1$, a fraction $1 - 1/k$ of states are $\epsilon \cdot k$-unimprovable.

Implementing the $G_{\times,+}$ gate. We first define a gadget that implements the $G_{\times,+}(\xi, \zeta|v_1, v_2|v_3)$ gate: in particular, the gadget will consist of states v_1, v_2, v_3 of the stochastic game $\mathbb{G}$ together with a helper state w of $\mathbb{G}$. We will choose the transitions and rewards of $\mathbb{G}$ so that, roughly speaking, for any equilibrium policy $\pi : \mathcal{S} \to [0, 1]$, $\pi(v_3)$ is close to $\max\{\min\{\xi \cdot \pi(v_1) + \zeta \cdot \pi(v_2), 1\}, 0\}$.

Definition 18 *Consider any $\alpha, \psi, \beta \in \mathbb{R}$, each with absolute value at most $1 - \gamma$. We say that a $G_{\times,+}(\frac{\alpha}{2\beta}, \frac{\psi}{2\beta}|v_1, v_2|v_3)$ gate* embeds in a stochastic game $\mathbb{G}$ via the states (v_1, v_2, v_3, w) and the constants (α, ψ, β), *for states $v_1, v_2, v_3, w \in \mathcal{S}$ if the following holds:*

1. *The transitions out of v_3 and w satisfy the following*
   - $\mathbb{P}(v_1|w, 0) = \min\{\frac{1}{2}, \frac{|\alpha|}{2|\beta|}\}$, $\mathbb{P}(v_2|w, 0) = \min\{\frac{1}{2}, \frac{|\psi|}{2|\beta|}\}$, $\mathbb{P}(s^0_{\text{sink}}|w, 0) = 1-\mathbb{P}(v_1|w, 0)-\mathbb{P}(v_2|w, 0)$, *and* $\mathbb{P}(v_3|w, 1) = 1$*;*
   - $\mathbb{P}(w|v_3, 0) = 1$, *and* $\mathbb{P}(s^0_{\text{sink}}|v_3, 1) = 1$.
2. *The rewards to the players controlling v_3, w at states v_1, v_2, v_3, w satisfy the following:*
   - $r_{\texttt{cr}(w)}(v_1, 1) = \frac{\alpha\cdot\max\{1,\frac{|\beta|}{|\alpha|}\}}{1-\gamma}$, $r_{\texttt{cr}(w)}(v_2, 1) = \frac{\psi\cdot\max\{1,\frac{|\beta|}{|\psi|}\}}{1-\gamma}$, *and* $r_{\texttt{cr}(w)}(v_3, 1) = \frac{\beta}{1-\gamma}$*;*

- $r_{\texttt{cr}(v_3)}(w,1) = \frac{\beta}{1-\gamma}$ *and* $r_{\texttt{cr}(v_3)}(w,0) = -\frac{\beta}{1-\gamma}$*;*
- *For all* $a \in \{0,1\}$*, it holds that*

$$r_{\texttt{cr}(w)}(v_1,0) = r_{\texttt{cr}(w)}(v_2,0) = r_{\texttt{cr}(w)}(v_3,0) = r_{\texttt{cr}(w)}(w,a) = r_{\texttt{cr}(v_3)}(v_3,a) = 0.$$

Recall from Definition 8 that we allow for $v_1 = v_2$ *in the gate* $G_{\times,+}(\xi,\zeta|v_1,v_2|v_3)$*. If this is the case, we require that* $\alpha = \psi$ *and instead require that* $\mathbb{P}(v_1|w,0) = \mathbb{P}(v_2|w,0) = 2 \cdot \min\left\{\frac{1}{2}, \frac{|\alpha|}{2|\beta|}\right\}$ *above.*

In the context of Definition 18, we will at times refer to w as the *helper node* for the embedded gate.

Lemma 19 *Suppose* $G_{\times,+}(\frac{\alpha}{2\beta}, \frac{\psi}{2\beta}|v_1,v_2|v_3)$ *embeds in a stochastic game* $\mathbb{G}$ *via the tuple* (v_1,v_2,v_3,w) *and the vector* (α,ψ,β)*, and consider any* $\epsilon,\epsilon' \in (0,1)$*. Suppose that* $\gamma|\beta|\epsilon - 2\gamma^2 > \epsilon'$*. Then for any policy* $\pi : \mathcal{S} \to [0,1]$ *for which* v_3, w *are each* ϵ'*-unimprovable under* π*, it holds that*

$$\pi(v_3) = \max\left\{\min\left\{\frac{\alpha}{2\beta}\cdot\pi(v_1) + \frac{\psi}{2\beta}\cdot\pi(v_2), 1\right\}, 0\right\} \pm \epsilon.$$

Proof Since $\gamma|\beta|\epsilon - 2\gamma^2 > \epsilon'$ and $\epsilon < 1$, we have that $\gamma|\beta| - \gamma^2 > \epsilon' > 0$ and $\beta \neq 0$.

Consider any policy $\pi : \mathcal{S} \to [0,1]$. First note that, since $r_{\texttt{cr}(w)}(v_1,0) = r_{\texttt{cr}(w)}(v_2,0) = 0$, it holds that

$$V^{\pi}_{\texttt{cr}(w)}(v_1) = (1-\gamma)\cdot\pi(v_1)\cdot r_{\texttt{cr}(w)}(v_1,1) \pm \gamma = \alpha \cdot \max\left\{1, \frac{|\beta|}{|\alpha|}\right\}\cdot\pi(v_1) \pm \gamma \tag{6}$$

$$V^{\pi}_{\texttt{cr}(w)}(v_2) = (1-\gamma)\cdot\pi(v_2)\cdot r_{\texttt{cr}(w)}(v_2,1) \pm \gamma = \psi \cdot \max\left\{1, \frac{|\beta|}{|\psi|}\right\}\cdot\pi(v_2) \pm \gamma. \tag{7}$$

In particular, we have used that the total contribution of rewards of $\texttt{cr}(w)$ to $V^{\pi}_{\texttt{cr}(w)}(v_1)$ (respectively, to $V^{\pi}_{\texttt{cr}(w)}(v_2)$) at states at least 1 step out from v_1 (respectively, v_2) is at most $(1-\gamma)\cdot(\gamma + \gamma^2 + \cdots) = \gamma$.

We first compute $Q^{\pi}_{\texttt{cr}(w)}(w,b)$ for $b \in \{0,1\}$, as follows. Using that $r_{\texttt{cr}(w)}(v_3,0) = r_{\texttt{cr}(w)}(w,0) = r_{\texttt{cr}(w)}(w,1) = 0$, we have:

- $Q^{\pi}_{\texttt{cr}(w)}(w,1) = \gamma\cdot\pi(v_3)\cdot\beta \pm \gamma^2$.
- $Q^{\pi}_{\texttt{cr}(w)}(w,0) = \frac{1}{2}\gamma\alpha\cdot\pi(v_1) + \frac{1}{2}\gamma\psi\cdot\pi(v_2) \pm \gamma^2$, which may be seen as follows:

$$\begin{aligned}
Q^{\pi}_{\texttt{cr}(w)}(w,0) =& \gamma\cdot\min\left\{\frac{1}{2}, \frac{|\alpha|}{2|\beta|}\right\}\cdot V^{\pi}_{\texttt{cr}(w)}(v_1) + \gamma\cdot\min\left\{\frac{1}{2}, \frac{|\psi|}{2|\beta|}\right\}\cdot V^{\pi}_{\texttt{cr}(w)}(v_2) \\
=& \gamma\cdot\min\left\{\frac{1}{2}, \frac{|\alpha|}{2|\beta|}\right\}\cdot\left(\alpha\cdot\max\left\{1,\frac{|\beta|}{|\alpha|}\right\}\cdot\pi(v_1) \pm \gamma\right) \\
&+ \gamma\cdot\min\left\{\frac{1}{2}, \frac{|\psi|}{2|\beta|}\right\}\cdot\left(\psi\cdot\max\left\{1,\frac{|\beta|}{|\psi|}\right\}\cdot\pi(v_2) \pm \gamma\right) \\
=& \frac{1}{2}\gamma\alpha\cdot\pi(v_1) + \frac{1}{2}\gamma\psi\cdot\pi(v_2) \pm \gamma^2.
\end{aligned}$$

For computation of $Q^{\pi}_{\mathtt{cr}(w)}(w,0)$ above, we have used (6) and (7), as well as the fact that $\min\left\{\frac{1}{2}, \frac{|\alpha|}{2|\beta|}\right\} \cdot \max\left\{1, \frac{|\beta|}{|\alpha|}\right\} = \frac{1}{2}$ (and an analogous equality clearly holds with ψ replacing α).

We next compute $Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,b)$ for $b \in \{0,1\}$ in the particular case where $\pi(w) \in \{0,1\}$, using that $r_{\mathtt{cr}(v_3)}(v_3,a) = 0$ for each $a \in \{0,1\}$:

- If $\pi(w) = 1$, then:
  - $Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,0) = \gamma \cdot \beta \pm \gamma^2$;
  - $Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,1) = 0$.
- If $\pi(w) = 0$, then:
  - $Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,0) = -\gamma \cdot \beta \pm \gamma^2$;
  - $Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,1) = 0$.

We consider two cases, depending on the sign of β.

Case 1: $\beta > 0$**.** If $\pi(v_3) > \max\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 0\} + \epsilon$, then $\beta \cdot \pi(v_3) > \frac{\alpha}{2} \cdot \pi(v_1) + \frac{\psi}{2} \cdot \pi(v_2) + \beta\epsilon$, and so we have

$$Q^{\pi}_{\mathtt{cr}(w)}(w,1) - Q^{\pi}_{\mathtt{cr}(w)}(w,0) \geq \gamma \cdot \left(\beta \cdot \pi(v_3) - \frac{1}{2}\alpha \cdot \pi(v_1) - \frac{1}{2}\psi \cdot \pi(v_2)\right) - 2\gamma^2 > \gamma\beta \cdot \epsilon - 2\gamma^2 > \epsilon',$$

which implies, since w is ϵ'-unimprovable under π, that $\pi(w) = 1$. But then

$$Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,0) - Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,1) \geq \gamma\beta - \gamma^2 > \epsilon',$$

which implies that, since v_3 is ϵ'-unimprovable under π, $\pi(v_3) = 0$. But we have assumed above that $\pi(v_3) > \max\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 0\} + \epsilon > 0$, which is a contradiction.

Next suppose that $\pi(v_3) < \min\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 1\} - \epsilon$, which implies that $\beta \cdot \pi(v_3) \leq \frac{\alpha}{2} \cdot \pi(v_1) + \frac{\psi}{2} \cdot \pi(v_2) - \beta\epsilon$. Then

$$Q^{\pi}_{\mathtt{cr}(w)}(w,0) - Q^{\pi}_{\mathtt{cr}(w)}(w,1) \geq \gamma \cdot \left(\frac{1}{2}\alpha \cdot \pi(v_1) + \frac{1}{2}\psi \cdot \pi(v_2) - \beta \cdot \pi(v_3)\right) - 2\gamma^2 \geq \gamma\beta \cdot \epsilon - 2\gamma^2 > \epsilon',$$

which implies that, since w is ϵ'-unimprovable under π, $\pi(w) = 0$. But then

$$Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,1) - Q^{\pi}_{\mathtt{cr}(v_3)}(v_3,0) \geq \gamma\beta - \gamma^2 > \epsilon',$$

which implies that, since v_3 is ϵ'-unimprovable under π, $\pi(v_3) = 1 > 1 - \epsilon$, a contradiction to $\pi(v_3) < \min\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 1\} - \epsilon$.

Case 2: $\beta < 0$**.** Roughly speaking, this case is similar to Case 1, except that some inequalities are reversed. We work out the details for completeness. If $\pi(v_3) > \max\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 0\} + \epsilon$, then $\beta \cdot \pi(v_3) < \frac{\alpha}{2} \cdot \pi(v_1) + \frac{\psi}{2} \cdot \pi(v_2) + \beta\epsilon$, and so we have

$$Q^{\pi}_{\mathtt{cr}(w)}(w,0) - Q^{\pi}_{\mathtt{cr}(w)}(w,1) \geq \gamma \cdot \left(\frac{1}{2}\alpha \cdot \pi(v_1) + \frac{1}{2}\psi \cdot \pi(v_2) - \beta \cdot \pi(v_3)\right) - 2\gamma^2 \geq -\gamma\beta \cdot \epsilon - 2\gamma^2 > \epsilon',$$

which implies that, since w is ϵ'-unimprovable under π, then $\pi(w) = 0$. But then

$$Q^{\pi}_{\mathtt{cr}(v_3)}(v_3, 0) - Q^{\pi}_{\mathtt{cr}(v_3)}(v_3, 1) \geq -\gamma\beta - \gamma^2 > \epsilon',$$

which implies that, since v_3 is ϵ'-unimprovable under π, $\pi(v_3) = 0 < \epsilon$, a contradiction to $\pi(v_3) > \max\{\frac{\alpha}{2\beta} + \frac{\psi}{2\beta} \cdot \pi(v_2), 0\} + \epsilon$, which we assumed above.

Next suppose that $\pi(v_3) < \min\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 1\} - \epsilon$, which implies that $\beta \cdot \pi(v_3) \geq \frac{\alpha}{2} \cdot \pi(v_1) + \frac{\psi}{2} \cdot \pi(v_2) - \beta\epsilon$. Then

$$Q^{\pi}_{\mathtt{cr}(w)}(w, 1) - Q^{\pi}_{\mathtt{cr}(w)}(w, 0) \geq \gamma \cdot \left(\beta \cdot \pi(v_3) - \frac{1}{2}\alpha \cdot \pi(v_1) - \frac{1}{2}\psi \cdot \pi(v_2)\right) - 2\gamma^2 > -\gamma\beta \cdot \epsilon - 2\gamma^2 > \epsilon',$$

which implies that, since w is ϵ'-unimprovable under π, $\pi(w) = 1$. But then

$$Q^{\pi}_{\mathtt{cr}(v_3)}(v_3, 1) - Q^{\pi}_{\mathtt{cr}(v_3)}(v_3, 0) \geq -\gamma\beta - \gamma^2 > \epsilon',$$

which implies that, since v_3 is ϵ'-unimprovable under π, $\pi(v_3) = 1 > 1 - \epsilon$, a contradiction to $\pi(v_3) < \min\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 1\} - \epsilon$.

In all possible cases, we have established that $\pi(v_3) \geq \min\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 1\} - \epsilon$ and $\pi(v_3) \leq \max\{\frac{\alpha}{2\beta} \cdot \pi(v_1) + \frac{\psi}{2\beta} \cdot \pi(v_2), 0\} + \epsilon$, which establishes the statement of the lemma. ■

Implementing the $G_{\leftarrow}$ gate. Next, we define a gadget that implements the $G_{\leftarrow}(b||v)$ gate, for a constant $b \in \{0, 1\}$.

Definition 20 *For $b \in \{0, 1\}$, we say that a $G_{\leftarrow}(b||v)$ gate* embeds in a stochastic game $\mathbb{G}$ via the state $v \in \mathcal{S}$ and the constant *b, if the following holds:*

1. *The transitions out of v satisfy $\mathbb{P}(s^0_{\text{sink}}|v, 0) = \mathbb{P}(s^0_{\text{sink}}|v, 1) = 1$;*
2. *The rewards to player $\mathtt{cr}(v)$ at the state v satisfy $r_{\mathtt{cr}(v)}(v, 1) = b$ and $r_{\mathtt{cr}(v)}(v, 0) = 1 - b$.*

Lemma 21 *Suppose that $G_{\leftarrow}(b||v)$ embeds in a stochastic game $\mathbb{G}$ via the state v and the constant $b \in \{0, 1\}$, and consider a policy $\pi : \mathcal{S} \to [0, 1]$. Then if the state v is ϵ-unimprovable under π with $\epsilon < (1 - \gamma)/2$, it holds that $\pi(v) = b$.*

Proof For $b \in \{0, 1\}$, it is clear that $Q^{\pi}_{\mathtt{cr}(v)}(v, 1) = (1 - \gamma) \cdot b$ and $Q^{\pi}_{\mathtt{cr}(v)}(v, 0) = (1 - \gamma) \cdot (1 - b)$. Thus, since v is ϵ-unimprovable under π and $\epsilon < (1 - \gamma)/2$, we must have that $\pi(v) = b$. ■

Implementing the $G_{<}$ gate. Finally, we define a gadget which implements the $G_{<}(|v_1, v_2|v_3)$ gate, for states v_1, v_2, v_3 of the stochastic game $\mathbb{G}$.

Definition 22 *Consider any $\beta \in \mathbb{R}$ with absolute value at most $1 - \gamma$. We say that a $G_{<}(|v_1, v_2|v_3)$ gate* embeds in a stochastic game $\mathbb{G}$ via the states (v_1, v_2, v_3, w) and the constant β, *for states $v_1, v_2, v_3, w \in \mathcal{S}$ if the following holds:*

1. *The transitions out of v_3 and w satisfy the following:*

- $\mathbb{P}(v_1|w,0) = 1$ *and* $\mathbb{P}(v_2|w,1) = 1$*;*
- $\mathbb{P}(w|v_3,1) = 1$ *and* $\mathbb{P}(s^0_{\text{sink}}|v_3,0) = 1$.

2. *The reward of the players controlling* w, v_3 *at states* v_1, v_2, w *satisfy the following:*
   - $r_{\texttt{cr}(w)}(v_1,1) = r_{\texttt{cr}(w)}(v_2,1) = \frac{\beta}{1-\gamma}$*;*
   - $r_{\texttt{cr}(w)}(v_1,0) = r_{\texttt{cr}(w)}(v_2,0) = 0$*;*
   - *For each* $a \in \{0,1\}$, $r_{\texttt{cr}(w)}(w,a) = r_{\texttt{cr}(v_3)}(v_3,a) = 0$*;*
   - $r_{\texttt{cr}(v_3)}(w,1) = \frac{\beta}{1-\gamma}$ *and* $r_{\texttt{cr}(v_3)}(w,0) = -\frac{\beta}{1-\gamma}$.

In the context of Definition 22, we will at times refer to w as the *helper node* for the gate.

Lemma 23 *Suppose that* $G_<(|v_1,v_2|v_3)$ *embeds in a stochastic game* $\mathbb{G}$ *via the states* (v_1,v_2,v_3,w) *and the constant* β*, and consider any* $\epsilon, \epsilon' \in (0,1)$*. Suppose that* $\gamma|\beta|\epsilon - 2\gamma^2 > \epsilon'$*. Then for any policy* $\pi : \mathcal{S} \to [0,1]$ *so that* v_3, w *are* ϵ'*-unimprovable under* π*, it holds that*

$$\pi(v_3) = \begin{cases} 1 \pm \epsilon & : \pi(v_1) \leq \pi(v_2) - \epsilon \\ 0 \pm \epsilon & : \pi(v_1) \geq \pi(v_2) + \epsilon. \end{cases}$$

The proof of Lemma 23 uses similar ideas to that of Lemma 19 and may be found in Section F.3.

B.5. Gluing gadgets via valid colorings

Next, we discuss how to combine the gadgets introduced in the previous section into a 2-player turn-based stochastic game such that an approximate Nash equilibrium yields an approximate assignment to a given instance of the generalized circuit problem.

Thought experiment & challenges. If we were willing to allow each player to control a different node (so that the number of players would be polynomial in the input length), then this procedure would be quite straightforward. Since we aim to show hardness for *2-player* games, however, we have to be more careful, since some of the constraints induced by the embedding of gates in Definitions 18-22 may conflict with each other when multiple states are controlled by a single player.

For instance, Definition 18 requires that for the embedded gate $G = G_{\times,+}(\frac{\alpha}{2\beta}, \frac{\psi}{2\beta}|v_1,v_2|v_3)$ with helper node w, we must have $r_{\texttt{cr}(w)}(v_1,1) = \frac{\alpha \cdot \max\left\{1, \frac{|\beta|}{|\alpha|}\right\}}{1-\gamma}$. Now suppose we were to attempt to embed gate $G' = G_{\times,+}(\frac{\alpha'}{2\beta'}, \frac{\psi'}{2\beta'}|v_1',v_2'|v_3')$ with some helper node w'. Suppose further that the output node v_3' of G' equals v_1 (which corresponds to the output of G' feeding into the gate G). Then the constraints of Definition 18 for this gate would require that $r_{\texttt{cr}(w')}(v_3',1) = r_{\texttt{cr}(w')}(v_1,1) = \frac{\beta'}{1-\gamma}$. It is possible that $\frac{\beta'}{1-\gamma} \neq \frac{\alpha \cdot \max\left\{1, \frac{|\beta|}{|\alpha|}\right\}}{1-\gamma}$, which implies that $\texttt{cr}(w) \neq \texttt{cr}(w')$. It is a straightforward consequence of Definition 18 that we must also have that $\texttt{cr}(w') \neq \texttt{cr}(v_3') = \texttt{cr}(v_1)$. Similar constraints may arise involving $\texttt{cr}(v_2)$ and $\texttt{cr}(v_3)$, and it is evident that the task of assigning a controller to each node becomes quite nontrivial. If we assign all non-helper nodes (denoted by v, v_1, v_2, v_3 in Definitions 18, 20, 22) to a single player, it is possible to show that, assuming the given generalized circuit instance has fan-out 2 (which is without loss of generality by (Rubinstein, 2018)), by greedily assigning each of the helper nodes to one of 4 players (for a total of 5 players), we may satisfy all constraints of the embedded gadgets.

To obtain hardness for 2-player (as opposed to 5-player) games, we instead take a different approach. As discussed above, we will assign all non-helper states to a single player, which we call $\mathtt{V}$, and all helper states (denoted by w in Definitions 18 and 22) to a second player, which we call $\mathtt{W}$. As mentioned above, this will cause conflicts; however, it is straightforward to check that the only type of conflict that arises is that for states w, w' controlled by $\mathtt{W}$, there is some state v controlled by player $\mathtt{V}$ so that the constraints on the embedded gates require that $r_{\mathtt{cr}(w)}(v,1) = c$ and $r_{\mathtt{cr}(w')}(v,1) = c'$ for some $c \neq c'$. To avoid this type of conflict, we will show how to convert a given generalized circuit instance into an equivalent instance for which such conflicts cannot arise. In particular, we introduce a notion of *valid coloring* of the nodes of a generalized circuit, so that a circuit equipped with a valid coloring has the property that no conflicts of the above type can arise when embedding the circuit into a 2-player turn-based stochastic game.

Valid colorings. Given a generalized circuit $\mathcal{C} = (V, \mathcal{G})$, we will say that the assignment of a real number to each node, denoted by $\phi : V \to \mathbb{R}$, is a *coloring* of V. Below we define the notion of *valid coloring*, which requires, loosely speaking, that the colorings of nodes are consistent with the rewards given to the $\mathtt{cr}(w)$ player in each of the gate gadgets defined in the previous section:

Definition 24 *Given a coloring $\phi : V \to \mathbb{R}$, we say that ϕ is* valid *if the following holds:*

1. *For each gate $G_<(|v_1, v_2|v_3)$, it holds that $\phi(v_1) = \phi(v_2)$;*
2. *For each gate $G_{\times,+}(\xi, \zeta|v_1, v_2|v_3)$, it holds that*

$$\frac{\phi(v_1)}{\phi(v_3)} = \begin{cases} 2 \cdot \xi & : |\xi| \geq 1/2 \\ \text{sign}(\xi) & : |\xi| < 1/2 \end{cases}, \text{ and } \frac{\phi(v_2)}{\phi(v_3)} = \begin{cases} 2 \cdot \zeta & : |\zeta| \geq 1/2 \\ \text{sign}(\zeta) & : |\zeta| < 1/2. \end{cases}$$

We say that a gate G is valid *if, in the case that it is one of the above types of gates, the respective condition above is met (If G is not one of the above types of gates, i.e., the $G_{\leftarrow}$ gate, then it is automatically defined to be valid).*

To understand Definition 24, $\phi(v)$ may be interpreted as the reward being given to the $\mathtt{W}$ player in the hard instance of stochastic games we are constructing: indeed, when reducing the ϵ-$\mathtt{GCircuit}$ problem to finding approximate equilibria in stochastic games, in Lemma 27, the reward to $\mathtt{W}$ at a state v is given by the scaling $\frac{\phi(v)}{1-\gamma}$. Then the constraints of Definition 24 are defined so as to ensure that there will be no conflicts in terms of the rewards given to the $\mathtt{W} = \mathtt{cr}(w)$ player in Definition 22 for the gate $G_<(|v_1, v_2|v_3)$ and in Definition 18 for the gate $G_{\times,+}(\xi, \zeta|v_1, v_2|v_3)$. In particular, Definition 22 requires that if $G_<(|v_1, v_2|v_3)$ embeds in $\mathbb{G}$ via the states (v_1, v_2, v_3, w), then $r_{\mathtt{cr}(w)}(v_1, 1) = r_{\mathtt{cr}(w)}(v_2, 1) = \frac{\beta}{1-\gamma}$, which corresponds to the first item in Definition 24. Similarly, the second item of Definition 24 corresponds to the constraints in item 2 of Definition 18 on $r_{\mathtt{cr}(w)}(v_1, 1), r_{\mathtt{cr}(w)}(v_2, 1), r_{\mathtt{cr}(w)}(v_3, 1)$.

We define the *range* of ϕ to be the set $\{\phi(v) \ : \ v \in V\} \subset \mathbb{R}$. The below lemma shows, loosely speaking, how to convert a circuit with some coloring ϕ to an equivalent circuit with a valid coloring. For simplicity, we use the following terminology: given a generalized circuit $\mathcal{C} = (V, \mathcal{G})$, we say that an assignment $\pi : V \to [0, 1]$ is an (ϵ, δ)-*assignment* of $\mathcal{C}$ if at least a $1 - \delta$ fraction of the gates are ϵ-approximately satisfied by π (see Definition 8). We say that π is an ϵ-*assignment* if it is an $(\epsilon, 0)$-assignment.

Lemma 25 *There is an absolute constant $C_0 > 0$ so that the following holds. Let $\mathcal{C} = (V, \mathcal{G})$ be a generalized circuit and $\epsilon > 0$. Then one can construct, in polynomial time, a circuit $\mathcal{C}' = (V', \mathcal{G}')$ together with a* valid *coloring $\phi : V' \to \mathbb{R}$, so that:*

1. *$V \subset V'$;*
2. *The range of ϕ is contained in $[1/4, 1/2] \cup [-1/2, -1/4]$;*
3. *For any $\delta \geq 0$, given an (ϵ, δ)-approximate assignment $\pi : V' \to [0, 1]$ of $\mathcal{C}'$, the restriction of π to V constitutes a $(133\sqrt{\epsilon}, C_0\delta/\sqrt{\epsilon})$-approximate assignment of $\mathcal{C}$.*

We first sketch the proof of Lemma 25. For an appropriate choice of V' which contains V as a subset, we will define $\phi : V' \to \mathbb{R}$ so that $\phi(v) = 1/4$ for all $v \in V$. Notice that we would be able to choose $V' = V, \mathcal{G}' = \mathcal{G}$ and would immediately have a valid coloring if it were not for gates of the type $G_{\times,+}(\xi, \zeta | v_1, v_2 | v_3)$, which can require that different nodes have different colors under ϕ (i.e., when $\xi \neq 1/2$ or $\zeta \neq 1/2$).

To circumvent this obstacle, for each gate of the form $G_{\times,+}(\xi, \zeta | v_1, v_2 | v_3)$, we introduce a sequence of gates and nodes connecting each of v_1 (respectively, v_2) to some (new) node v_1' (respectively, v_2'), which approximately implements the identity map. Importantly, this sequence of gates and nodes will have the property that there is a cut (i.e., a separating set) consisting solely of gates of the type $G_<(|u_1, u_2|u_3)$, which place no restriction on $\phi(u_3)$ for a valid coloring ϕ. We will be able to use this cut to ensure that the value of ϕ at vertices on one side of the cut differs from the value of ϕ at vertices on the other side of the cut, thus ensuring that $\phi(v_1')$ (respectively, $\phi(v_2')$) can differ from $\phi(v_1)$ (respectively, $\phi(v_2)$), while maintaining validity.

Proof [Proof of Lemma 25] We follow the outline sketched above. In particular, we build up the circuit $(V', \mathcal{G}')$ according to the following procedure. We initialize $V' = V$ and set $\phi(v) = 1/4$ for all $v \in V$. Moreover, for all gates apart from those of the type $G_{\times,+}$, we add the same gate (with the same input and output nodes) to $\mathcal{G}'$. It is immediate that any such gate satisfies the validity constraint in Definition 24 under the coloring ϕ (if applicable). Now define the function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 2x & : |x| \geq 1/2 \\ \text{sign}(x) & : |x| < 1/2 \end{cases}.$$

Consider each gate of the form $G_{\times,+}(\xi, \zeta | v_1, v_2 | v_3)$ in turn. For each such gate, we perform the following steps: We add nodes v_1', v_2' to V', and add the gate $G_{\times,+}(\xi, \zeta | v_1', v_2' | v_3)$ to $\mathcal{G}'$. Furthermore, we set $\phi(v_1') := f(\xi) \cdot \phi(v_3) = f(\xi) \cdot \frac{1}{4}$ and $\phi(v_2') := f(\zeta) \cdot \phi(v_3) = f(\zeta) \cdot \frac{1}{4}$, thus ensuring that the validity constraint for $G_{\times,+}(\xi, \zeta | v_1', v_2' | v_3)$ is satisfied. Furthermore, since $|f(x)| \geq 1$ for all $x \in \mathbb{R}$, we have that $|\phi(v_1')| \geq 1/4$ and $|\phi(v_2')| \geq 1/4$. Moreover, since $|\zeta| \leq 1$ and $|\xi| \leq 1$ (by Definition 8), it holds that $|\phi(v_1')| \leq 1/2$ and $|\phi(v_2')| \leq 1/2$.

In Claim 26 below, we add a sequence of gates and nodes to $\mathcal{G}', V'$, respectively (together with respective colors ensuring validity), that lie between v_1 and v_1', which ensure that in any ϵ-approximate assignment π, $|\pi(v_1) - \pi(v_1')| \leq O(\sqrt{\epsilon})$. By symmetry, the same construction can be implemented for the nodes v_2, v_2'; as we discuss following the proof of Claim 26 the result of Lemma 25 will follow in a straightforward manner.

Claim 26 *Consider two nodes $a, a' \in V'$ so that $\phi(a), \phi(a')$ are defined. Then it is possible to add a set $\widetilde{V}$ of $O(1/\sqrt{\epsilon})$ nodes to V' and a set $\widetilde{\mathcal{G}}$ of $O(1/\sqrt{\epsilon})$ gates to $\mathcal{G}'$ so that the gates in $\widetilde{\mathcal{G}}$ have all their input and output nodes in $\widetilde{V} \cup \{a, a'\}$, and the following holds:*

1. *ϕ may be extended to a mapping on $\widetilde{V}$ so that for all $v \in \widetilde{V}$, $\phi(v) \in \{\phi(a), \phi(a')\}$. Furthermore, the resulting ϕ is so that all gates in $\widetilde{\mathcal{G}}$ are valid;*
2. *Given any assignment π of the circuit (including that of the nodes in $\widetilde{V}$), if all gates in $\widetilde{\mathcal{G}}$ are satisfied under π, then $|\pi(a) - \pi(a')| \le 66 \cdot \sqrt{\epsilon}$.*

Proof The construction we introduce mirrors that of Algorithms 6 and 7 of (Rubinstein, 2018), with a few differences. Choose $\epsilon' \ge \sqrt{\epsilon}$ as small as possible so that $4/\epsilon'$ is a power of 2 (so that $\epsilon' \le 2\sqrt{\epsilon}$). Initialize $\widetilde{V}, \widetilde{\mathcal{G}}$ to be empty sets. We now introduce the following nodes and gates, which are added to $\widetilde{V}$ and $\widetilde{\mathcal{G}}$, respectively:

1. Add a gate $G_{\leftarrow}(1||\sigma)$ to $\widetilde{\mathcal{G}}$, whose output node σ is added to $\widetilde{V}$. Define $\phi(\sigma) := \phi(a)$.
2. For each $k \in [4/\epsilon']$:
   (a) Add a gate $G_{\times,+}(k\epsilon'/8, k\epsilon'/8|\sigma, \sigma|\sigma_k)$ to $\widetilde{\mathcal{G}}$, whose output node σ_k is added to $\widetilde{V}$. Define $\phi(\sigma_k) := \phi(\sigma)$. *(Validity of this gate is ensured since $0 < k\epsilon'/8 \le 1/2$ and we have $\phi(\sigma_k) = \phi(\sigma)$.)*
   (b) Add a gate $G_{<}(|\sigma_k, a|b_k)$ to $\widetilde{\mathcal{G}}$, whose output node b_k is added to $\widetilde{V}$. Define $\phi(b_k) := \phi(a')$. *(Validity of this gate is ensured since $\phi(\sigma_k) = \phi(\sigma) = \phi(a)$; importantly, we are allowed to set $\phi(b_k)$ to something which does* not *equal $\phi(a)$.)*
3. For each $j \in [\log_2(4/\epsilon')]$:
   (a) For each $k \in [(4/\epsilon')/2^j]$:
      i. If $j = 1$, add a gate $G_{\times,+}(1/2, 1/2|b_{2k-1}, b_{2k}|d_{1,k})$ to $\widetilde{\mathcal{G}}$, whose output node $d_{1,k}$ is added to $\widetilde{V}$. Define $\phi(d_{1,k}) := \phi(a')$. *(Validity of this gate is ensured since $\phi(b_{2k-1}) = \phi(b_{2k}) = \phi(d_{1,k}) = \phi(a')$.)*
      ii. If $j > 1$, add a gate $G_{\times,+}(1/2, 1/2|d_{j-1,2k-1}, d_{j-1,2k}|d_{j,k})$ to $\widetilde{\mathcal{G}}$, whose output node $d_{j,k}$ is added to $\widetilde{V}$. Define $\phi(d_{j,k}) := \phi(a')$. *(Validity of this gate is ensured since $\phi(d_{j-1,2k-1}) = \phi(d_{j-1,2k}) = \phi(d_{j,k}) = \phi(a')$.)*
4. Add a gate $G_{\times,+}(1/2, 1/2|d_{\log_2(4/\epsilon'),1}, d_{\log_2(4/\epsilon'),1}|a')$ to $\widetilde{\mathcal{G}}$. *(Validity of this gate is ensured since $\phi(d_{\log_2(4/\epsilon'),1}) = \phi(a')$.)*

It is straightforward to see that $|\widetilde{V}|, |\widetilde{\mathcal{G}}|$ are bounded above by $O(1/\epsilon') = O(1/\sqrt{\epsilon})$ at the end of the above procedure.

It is clear that at all nodes v added to $\widetilde{V}$ in the above construction, we have $\phi(v) \in \{\phi(a), \phi(a')\}$, thus verifying the first item in the claim's statement. To see the second item, let π denote an assignment for the generalized circuit $(V', \mathcal{G}')$, after $\widetilde{V}$ has been added to V' and $\widetilde{\mathcal{G}}$ has been added to $\mathcal{G}'$, and suppose that π ϵ-approximately satisfies all gates in $\widetilde{\mathcal{G}}$. By definition of the gate $G_{\times,+}$, we must have that for each $k \in [4/\epsilon']$, $\pi(\sigma_k) = k\epsilon'/4 \pm \epsilon$. Thus, the number of integers $k \in [4/\epsilon']$ so that $\pi(b_k) \ge 1 - \epsilon$ lies in the range $\left[\frac{4}{\epsilon'} \cdot (\pi(a) - 3\epsilon), \frac{4}{\epsilon'} \cdot (\pi(a) + 3\epsilon)\right]$, and the number of integers $k \in [4/\epsilon']$ so that $\pi(b_k) \le \epsilon$ lies in the range $\left[\frac{4}{\epsilon'} \cdot (1 - \pi(a) - 3\epsilon), \frac{4}{\epsilon'} \cdot (1 - \pi(a) + 3\epsilon)\right]$.

It follows that

$$\sum_{k=1}^{4/\epsilon'} \pi(b_k) = \frac{4}{\epsilon'} \cdot \pi(a) \pm \left(\epsilon \cdot \frac{4}{\epsilon'} + \frac{4}{\epsilon'} \cdot 6\epsilon\right) = \frac{4}{\epsilon'} \cdot \pi(a) \pm 28\epsilon'.$$

By the definition of the gate $G_{\times,+}$ and the triangle inequality, it holds that $\pi(d_{\log_2(4/\epsilon'),1}) = \frac{\epsilon'}{4}\sum_{k=1}^{4/\epsilon'}\pi(b_k) \pm \log_2(4/\epsilon')\cdot\epsilon = \frac{\epsilon'}{4}\sum_{k=1}^{4/\epsilon'}\pi(b_k) \pm 4\epsilon'$, since $\log_2(4/\epsilon') \le 4/\epsilon'$. Since also $\pi(a') = \pi(d_{\log_2(4/\epsilon'),1}) \pm \epsilon$, we conclude that

$$\pi(a') = \frac{\epsilon'}{4}\sum_{k=1}^{4/\epsilon'}\pi(b_k) \pm (4\epsilon' + \epsilon) = \pi(a) \pm (4\epsilon' + \epsilon + 28(\epsilon')^2) = \pi(a) \pm 33\epsilon' = \pi(a) \pm 66\sqrt{\epsilon},$$

where the last step uses that $\epsilon \le \epsilon' \le 2\sqrt{\epsilon}$. ■

Given Claim 26, we complete the proof of Lemma 25. For the gate $G_{\times,+}(\xi,\zeta|v_1,v_2|v_3)$ (as was introduced above), we apply Claim 26 once with $a = v_1, a' = v_1'$, adding sets $\widetilde{V}_1, \widetilde{\mathcal{G}}_1$ to $V', \mathcal{G}'$, respectively, and once with $a = v_2, a' = v_2'$, adding sets $\widetilde{V}_2, \widetilde{\mathcal{G}}_2$ to $V', \mathcal{G}'$, respectively. After this procedure, it still holds that for all $v \in V'$, $|\phi(v)| \in [1/4, 1/2]$ by item 1 of Claim 26. Furthermore, item 2 of Claim 26 gives that in any assignment $\pi : V' \to [0,1]$ for which all gates in $\widetilde{\mathcal{G}}_1 \cup \widetilde{\mathcal{G}}_2 \cup \{G_{\times,+}(\xi,\zeta|v_1',v_2'|v_3)\}$ are ϵ-approximately satisfied,

$$\begin{aligned}\pi(v_3) &= \max\left\{\min\left\{\xi\cdot\pi(v_1') + \zeta\cdot\pi(v_2'), 1\right\}, 0\right\} \pm \epsilon \\ &= \max\left\{\min\left\{\xi\cdot(\pi(v_1) \pm 66\sqrt{\epsilon}) + \zeta\cdot(\pi(v_2) \pm 66\sqrt{\epsilon}), 1\right\}, 0\right\} \pm \epsilon \\ &= \max\left\{\min\left\{\xi\cdot\pi(v_1) + \zeta\cdot\pi(v_2), 1\right\}, 0\right\} \pm 133\sqrt{\epsilon},\end{aligned}$$

where the final step uses that $|\xi|, |\zeta| \le 1$. Note that $|\widetilde{\mathcal{G}}_1 \cup \widetilde{\mathcal{G}}_2| \le O(1/\sqrt{\epsilon})$ by Claim 26. Thus, after applying Claim 26 for each gate $G_{\times,+}$ in the original circuit $\mathcal{C}$, we note that for an (ϵ,δ)-approximate assignment $\pi : V' \to [0,1]$ of $\mathcal{C}'$, it must hold that, for some constant $C > 1$, for at least a fraction $1 - C\delta/\sqrt{\epsilon}$ fraction of the gates G of the original circuit $\mathcal{C}$, the gate G is $133\sqrt{\epsilon}$-satisfied. (This holds because at least a $1 - C\delta/\sqrt{\epsilon}$ fraction of the gates in $\mathcal{C}$ are either not of the type $G_{\times,+}$ or have all of the $O(1/\sqrt{\epsilon})$ supplementary gates added in course of Claim 26 ϵ-satisfied by π.) This completes the proof of the lemma. ■

Next we show that the problem of finding an approximate assignment of a circuit which has a valid coloring can be reduced to the problem of finding an approximate (P)WSNE-SG of an infinite-horizon discounted stochastic game.

Lemma 27 *Fix any $\epsilon \in (0, \frac{1}{12})$ and $\delta \in (0,1)$. Set $\gamma = \epsilon^2$, $\epsilon' = \epsilon^4$, and $\epsilon'' = \epsilon'\cdot\delta$. Then the following statements hold.*

- *The problem of finding an ϵ-assignment to a generalized circuit instance equipped with a valid coloring with range contained in $[-1/2,-1/4]\cup[1/4,1/2]$ has a polynomial-time reduction to the problem of computing an ϵ'-PWSNE-SG in 2-player turn-based γ-discounted stochastic games.*

- *The problem of finding an $(\epsilon, 3\delta)$-assignment to a generalized circuit instance equipped with a valid coloring with range contained in $[-1/2,-1/4]\cup[1/4,1/2]$ has a polynomial-time reduction to the problem of computing a ϵ''-WSNE-SG in 2-player turn-based γ-discounted stochastic games.*

Proof Let $\mathcal{C} = (V, \mathcal{G})$ be a generalized circuit together with some valid coloring $\phi : V \to \mathbb{R}$, and $\epsilon \in (0, 1)$. We construct a γ-discounted 2-player turn-based stochastic game $\mathbb{G}$, as follows: the two players are denoted by $\mathtt{W}$ and $\mathtt{V}$, the action spaces of each player of $\mathbb{G}$ satisfiy $\mathcal{A}_{\mathtt{V}} = \mathcal{A}_{\mathtt{W}} = \{0, 1\}$, and the state space $\mathcal{S}$ satisfies:

$$\mathcal{S} = V \sqcup W \sqcup s^0_{\text{sink}},$$

where s^0_{sink} is a special sink state which transitions to itself indefinitely and at which all players receive 0 reward, and W is in bijection with $\mathcal{G}$, consisting of a designated node w_G for each gate $G \in \mathcal{G}$. The ownership of the states is as follows: for all $v \in V$, we have $\mathtt{cr}(v) = \mathtt{V}$, for all $w \in W$, we have $\mathtt{cr}(w) = \mathtt{W}$. Finally, we arbitrarily set $\mathtt{cr}(s^0_{\text{sink}}) = \mathtt{V}$.

For each gate of the form $G(\ell|v_1, v_2|v_3)$, we will ensure that G embeds in $\mathbb{G}$ via the tuple (v_1, v_2, v_3, w_G) or the tuple (v_1, v_2, v_3) (depending on the type of gate G), and via an appropriate vector of constants (if applicable, again depending on the type of gate G). To do so, we construct the transitions and rewards of $\mathbb{G}$ as follows: intially set the reward at each state (for all agents and actions) to be 0, and define the transitions so that each state transitions to s^0_{sink} under any action. We will then make several modifications to the rewards and transitions: First, for each state $v \in V$, define

$$r_{\mathtt{W}}(v, 1) = \frac{\phi(v)}{1 - \gamma}. \tag{8}$$

Next, for each state $v \in V$ which is the outgoing node of a gate of the form $G_{\leftarrow}(b||v)$, set

$$r_{\mathtt{V}}(v, b) = b, \qquad r_{\mathtt{V}}(v, 0) = 1 - b.$$

Next, for each gate $G \in \mathcal{G}$, we make the following modifications to $\mathbb{G}$'s transitions and rewards, depending on the type of gate G:

1. If G is of the form $G_{\times,+}(\xi, \zeta|v_1, v_2|v_3)$, then define

$$\alpha = \begin{cases} \phi(v_1) & : |\xi| \geq 1/2 \\ \phi(v_1) \cdot 2|\xi| & : |\xi| < 1/2 \end{cases}, \qquad \psi = \begin{cases} \phi(v_2) & : |\zeta| \geq 1/2 \\ \phi(v_2) \cdot 2|\zeta| & : |\zeta| < 1/2, \end{cases}, \qquad \beta = \phi(v_3),$$

   and modify the outgoing transitions from the states v_3, w_G and the rewards at state v_3 to satisfy the requirements of Definition 18 with the above values of α, ψ, β. The above definitions and the validity of ϕ ensure that $\frac{\alpha}{2\beta} = \xi$ and $\frac{\psi}{2\beta} = \zeta$, as in Definition 18. Furthermore, we do not have to modify $r_{\mathtt{W}}(v_1, 1)$, $r_{\mathtt{W}}(v_2, 1)$, or $r_{\mathtt{W}}(v_3, 1)$ since they are set to $\frac{\phi(v_1)}{1-\gamma}$, $\frac{\phi(v_2)}{1-\gamma}$, and $\frac{\phi(v_3)}{1-\gamma}$, respectively, in (8), and it holds by validity of ϕ and the definitions of α, β, ψ above that

$$\alpha \cdot \max\left\{1, \frac{|\beta|}{|\alpha|}\right\} = \phi(v_1), \quad \psi \cdot \max\left\{1, \frac{|\beta|}{|\psi|}\right\} = \phi(v_2), \quad \beta = \phi(v_3). \tag{9}$$

2. If G is of the form $G_{\leftarrow}(1||v)$, then modify the outgoing transitions from the state v to satisfy the requirements of Definition 20.

3. If G is of the form $G_{<}(|v_1, v_2|v_3)$, then modify the outgoing transitions from the states v_3, w_G and the rewards at state w_G to satisfy the requirements of Definition 22, with $\beta = \phi(v_1)$ (We do not have to modify $r_{\mathtt{W}}(v_1, 1)$ or $r_{\mathtt{W}}(v_2, 1)$ since both are set to $\frac{\beta}{1-\gamma} = \frac{\phi(v_1)}{1-\gamma} = \frac{\phi(v_2)}{1-\gamma}$, by validity of ϕ and (8)).

Note that for each gate G, whose output node is denoted by v, in the above procedure we have modified only the outgoing transitions at v and w_G, and the two players' rewards at state w_G. Since each node is the output node of a unique gate, this process ensures that neither the transitions nor reward at any state are modified twice in the below procedure, thus ensuring that the embedding requirements for each gate are still satisfied at the end of the above procedure.

Since $\max_{v \in V} |\phi(v)| \leq 1/2$, all nonzero rewards assigned to state-action pairs in the above procedure are bounded in magnitude by $\frac{\max_{v \in V} |\phi(v)|}{1-\gamma}$, and $\gamma < 1/2$, it holds that all rewards of $\mathbb{G}$ have absolute value at most 1.

Now we verify the condition in Lemma 19. Let us write $\beta_0 := \min_{v \in V} |\phi(v)| \geq 1/4$. We claim that $\gamma \cdot |\beta_0| \cdot \epsilon - 2\gamma^2 > \epsilon'$. This holds since $\gamma\epsilon/4 - 2\gamma^2 > \epsilon'$, which is guaranteed by our choice of $\gamma = \epsilon^2$ and $\epsilon' = \epsilon^4$ and since $\epsilon^3/4 - 2\epsilon^4 > \epsilon^4$, which holds as long as $\epsilon < 1/12$, which was assumed in the lemma statement.

Consider a policy π of $\mathbb{G}$, is represented by a function $\pi : \mathcal{S} \to [0, 1]$. If π is an ϵ'-PWSNE-SG of $\mathbb{G}$, then by Definition 15 and (5), all states s of $\mathbb{G}$ are ϵ'-unimprovable under π. By Lemmas 19, 21, and 23, since $\gamma \cdot |\beta_0| \cdot \epsilon - 2\gamma^2 > \epsilon'$, in any ϵ'-PWSNE-SG π of $\mathbb{G}$, it holds that each gate is ϵ-approximately satisfied by the restriction of π to V. Thus, the restriction of an ϵ'-PWSNE-SG π to the nodes $V \subset \mathcal{S}$ furnishes an ϵ-approximate assignment to the ϵ-`GCircuit` instance $\mathcal{C}$, as desired.

Next, suppose that π is an ϵ''-WSNE-SG of $\mathbb{G}$. Then by Definition 15, (5), and Markov's inequality, a fraction $1 - \delta$ of states of $\mathbb{G}$ are $\epsilon''/\delta = \epsilon'$-unimprovable. Since each node $v \in V$ is the output node of some (unique) gate in $\mathcal{G}$, a fraction $1 - 3\delta$ of gates G in $\mathcal{G}$ have the following property:

- If $G = G_{\times,+}(\xi, \zeta | v_1, v_2 | v_3)$, then v_3, w_G are ϵ'-unimprovable.
- If $G = G_{\leftarrow}(1||v)$, then v is ϵ'-unimprovable.
- If $G = G_{<}(|v_1, v_2|v_3)$, then v_3, w_G are ϵ'-unimprovable.

By Lemmas 19, 21, and 23, it follows that a fraction $1 - 3\delta$ of the gates of $\mathcal{G}$ are ϵ-approximately satisfied by the restriction of π to V. ■

Finally, we are ready to prove Theorems 3 and 4.

Proof [Proof of Theorem 3] By Theorem 9, it suffices to show that there is a constant $c > 0$ so that for all $\epsilon_0 < 1/12$, the ϵ_0-`GCircuit` problem has a polynomial-time reduction to the problem of computing $c \cdot \epsilon_0^{16}$-perfect NE in $1/2$-discounted 2-player stochastic games.

Fix any $\epsilon \in (0, 1/12)$, and write $\gamma = \epsilon^2$. Consider an instance $\mathcal{C} = (V, \mathcal{G})$ of ϵ-`GCircuit`. We construct a circuit $\mathcal{C}' = (V', \mathcal{G}')$, together with a valid coloring $\phi : V' \to \mathbb{R}$, as guaranteed in the statement of Lemma 25 (where we take $\delta = 0$). By Lemma 27, we may further construct in polynomial time, given $\mathcal{C}'$ together with ϕ, a γ-discounted 2-player turn-based stochastic game $\mathbb{G}$ so that given an ϵ^4-PWSNE-SG of $\mathbb{G}$, we may compute an ϵ-approximate assignment to $\mathcal{C}'$, which, by Lemma 25, yields a $133\sqrt{\epsilon}$-approximate assignment of $\mathcal{C}$.

By Lemma 17, the problem of computing an ϵ^4-PWSNE-SG of $\mathbb{G}$ reduces to the problem of computing a $\frac{\epsilon^8}{144}$-perfect NE of $\mathbb{G}$. Finally, noting that the game $\mathbb{G}$ is γ-discounted and we wish to reduce to the problem of computing equilibria in $1/2$-discounted games, we argue as follows: given the game $\mathbb{G}$, we may construct a $1/2$-discounted stochastic game $\mathbb{G}'$ whose states, actions,

and rewards are identical to that of $\mathbb{G}$, and whose transitions $\mathbb{P}'(\cdot|s,a)$ are determined from the transitions $\mathbb{P}(\cdot|s,a)$ of $\mathbb{G}$ as follows:

$$\mathbb{P}'(s'|s,a) = \begin{cases} \frac{\gamma}{1/2} \cdot \mathbb{P}(s'|s,a) & : s' \neq s^0_{\text{sink}} \\ \frac{\gamma}{1/2} \cdot \mathbb{P}(s'|s,a) + (1-2\gamma) & : s' = s^0_{\text{sink}}. \end{cases} \tag{10}$$

Let $V_i^{\mathbb{G},\pi}$ denote the value function of $\mathbb{G}$ and $V_i^{\mathbb{G}',\pi}$ denote the value function of $\mathbb{G}'$. It is clear that for all π and all $i \in [m]$, $V_i^{\mathbb{G},\pi}(s^0_{\text{sink}}) = V_i^{\mathbb{G}',\pi}(s^0_{\text{sink}}) = 0$. It is now straightforward to see that for any joint stationary policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$, we have, for all $s \in \mathcal{S}$ and $i \in [m]$,

$$\begin{aligned} V_i^{\mathbb{G}',\pi}(s) =& \mathbb{E}_{\boldsymbol{a}\sim\pi(s)}\left[r(s,\boldsymbol{a}) + \frac{1}{2}\cdot\sum_{s'\in\mathcal{S}}\frac{\gamma}{1/2}\cdot V_i^{\mathbb{G}',\pi}(s')\right] \\ V_i^{\mathbb{G},\pi}(s) =& \mathbb{E}_{\boldsymbol{a}\sim\pi(s)}\left[r(s,\boldsymbol{a}) + \gamma\cdot\sum_{s'\in\mathcal{S}} V_i^{\mathbb{G},\pi}(s')\right], \end{aligned}$$

which immediately implies that $V_i^{\mathbb{G},\pi} \equiv V_i^{\mathbb{G}',\pi}$. Hence the $\frac{\epsilon^8}{144}$-perfect Nash equilibria of $\mathbb{G}$ and $\mathbb{G}'$ coincide. Choosing $\epsilon_0 = \sqrt{\epsilon}$ shows that the ϵ_0-GCircuit problem reduces to the problem of finding $c \cdot \epsilon_0^{16}$-perfect NE in 1/2-discounted 2-player stochastic games, as desired.

Finally, to show PPAD-hardness of computing approximate stationary CCE in 2-player 1/2-discounted stochastic games, we note that in any turn-based stochastic game, any stationary policy π is equivalent to some product policy π' (in the sense that $V_i^{\pi} \equiv V_i^{\pi'}$ for all i): in particular, π' is the policy where at each state s, all players except $\texttt{cr}(s)$ take some fixed action in their action set and $\texttt{cr}(s)$ plays according to their marginal in $\pi(s)$. Thus, for any $\epsilon > 0$, an ϵ-perfect CCE may be converted into an ϵ-perfect NE in polynomial time. ■

On larger discount factors. We remark that it is evident from the above proof that the constant $1/2$ for the discount factor can be replaced by any constant $\gamma' \in (1/2, 1)$. In particular, we would simply modify the transition probabilities of the game $\mathbb{G}'$ in (10) by replacing the constant $1/2$ with γ'.

Proof [Proof of Theorem 4] The first part of the theorem is an immediate consequence of Theorem 3, as we proceed to explain. Consider a turn-based stochastic game $\mathbb{G}$, and $\epsilon > 0$. Note that an ϵ/S-stationary NE π of $\mathbb{G}$ must satisfy $\max_{i\in[m]} \mathbb{E}_{s\sim\mu}\left[V_i^{\dagger,\pi_{-i}}(s) - V_i^{\pi}(s)\right] \leq \epsilon/S$. Since π is a product policy, we have that $V_i^{\dagger,\pi_{-i}}(s) - V_i^{\pi}(s) \geq 0$ for all $i \in [m]$. Thus, for all $i \in [m], s \in \mathcal{S}$, we have $V_i^{\dagger,\pi_{-i}}(s) - V_i^{\pi}(s) \leq \epsilon$, i.e., π is an ϵ-perfect NE of $\mathbb{G}$, which is PPAD-hard to compute by Theorem 3.

We proceed to prove the second part of Theorem 3. Since we assume Conjecture 10, it suffices to show that there is a constant $c > 0$ so that for all $\epsilon_0 < 1/12$ and $\delta_0 < 1$, the (ϵ_0, δ_0)-GCircuit problem has a polynomial-time reduction to the problem of computing $c \cdot \epsilon_0^{18}\delta_0^2$-stationary NE in 1/2-discounted 2-player stochastic games.

To do so, fix $\epsilon \in (0, 1/12), \delta \in (0,1)$ and write $\gamma = \epsilon^2$. Consider an instance $\mathcal{C} = (V, \mathcal{G})$ of the (ϵ,δ)-GCircuit problem. We construct a circuit $\mathcal{C}'$, together with a valid coloring $\phi : V' \to \mathbb{R}$, as guaranteed in the statement of Lemma 25: in particular, given an (ϵ,δ)-approximate assignment π :

$V' \to [0,1]$ of $\mathcal{C}'$, the restriction of π to V constitutes a $(133\sqrt{\epsilon}, C_0\delta/\sqrt{\epsilon})$-approximate assignment of $\mathcal{C}$ (for some constant $C_0 > 1$).

By Lemma 27, we may further construct in polynomial time, given $\mathcal{C}'$ together with ϕ, a γ-discounted 2-player turn-based stochastic game $\mathbb{G}$ so that, given a $\epsilon^4\delta/3$-WSNE-SG of $\mathbb{G}$, we may compute an (ϵ, δ)-assignment to $\mathcal{C}'$, which thus yields a $(133\sqrt{\epsilon}, C_0\delta/\sqrt{\epsilon})$-approximate assignment of $\mathcal{C}$. By Lemma 17, the problem of computing an $\epsilon^4\delta/3$-WSNE-SG of $\mathbb{G}$ reduces to computing an $\frac{\epsilon^8\delta^2}{144\cdot 9}$-stationary NE of $\mathbb{G}$. The same construction as in the proof of Theorem 3 allows us to reduce further to the problem of computing an $\frac{\epsilon^8\delta^2}{144\cdot 9}$-stationary NE of a 2-player 1/2-discounted game $\mathbb{G}'$. Choosing $\epsilon_0 = \sqrt{\epsilon}$ and $\delta_0 = \delta/\sqrt{\epsilon}$, we have shown that the (ϵ_0, δ_0)-GCircuit problem reduces to computing a $c\epsilon_0^{18}\delta_0^2$-stationary NE in 1/2-discounted 2-player stochastic games, for some constant $c > 0$. ■

Appendix C. Proofs for Section 4

In this section, we prove Theorem 5, which gives a PAC-RL guarantee for SPoCMAR (Algorithm 1). First, in Section C.1, we give an overview of the proof of Theorem 5. Then, in Section 4.4, we explain how Algorithm 1 can be implemented in a decentralized manner with access to shared randomness. Section C.2 reviews some preliminaries regarding adversarial bandit regret bounds. In Section C.3 we proceed to review some basic preliminaries for finite-horizon stochastic games (closely mirroring the analogous definitions in Section 2). In Section C.4 we introduce some parameters and concentration inequalities used in the proof. In Section C.5, we introduce an intermediate stochastic game that is used in the analysis, which is reminiscient of the analysis of Rmax (Brafman and Tennenholtz, 2002; Jin et al., 2020). In Section C.6, we complete the proof of Theorem 5.

C.1. Proof overview for Theorem 5

We now overview the proof of Theorem 5. Let $\widehat{\mathcal{V}}$ denote the value of $\mathcal{V}$ at termination of SPoCMAR and $\widehat{q}$ denote the value of the final stage of SPoCMAR. The main tool in the proof is to construct an intermediate game, denoted $\mathbb{G}_{\widehat{\mathcal{V}}}$ (Section C.5): we will first show that the output policy of SPoCMAR is an ϵ-CCE with respect to the game $\mathbb{G}_{\widehat{\mathcal{V}}}$, and then, using the termination criterion of SPoCMAR, we will show that this implies that $\widehat{\pi}$ is an ϵ-CCE with respect to the true game $\mathbb{G}$.

The game $\mathbb{G}_{\widehat{\mathcal{V}}}$ is constructed in a similar way to an intermediate MDP used in the analysis of the Rmax algorithm (Brafman and Tennenholtz, 2002; Jin et al., 2020). For tuples $(h,s) \notin \widehat{\mathcal{V}}$, $\mathbb{G}_{\widehat{\mathcal{V}}}$ transitions, at (h,s), to a special sink state at which all agents receive reward 1 (the maximum possible reward) at all future steps; for all $(h,s) \in \widehat{\mathcal{V}}$, the rewards and transitions of $\mathbb{G}_{\widehat{\mathcal{V}}}$ at (h,s) are identical to those of $\mathbb{G}$. By ensuring that the parameter K passed to SPoCMAR is sufficiently large, we may guarantee that, during stage $\widehat{q}$, SPoCMAR visits all $(h,s) \in \widehat{\mathcal{V}}$ sufficiently many times to compute accurate estimates of $V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)$ for such $(h,s) \in \widehat{\mathcal{V}}$. Since, for all $(h,s) \notin \widehat{\mathcal{V}}$, we have $V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s) = H + 1 - h$, it is possible to show (Lemma 32) that $\left|\overline{V}_{i,h}^{\widehat{q}}(s) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)\right|$ is small for *all* $(h,s) \in [H] \times \mathcal{S}$ and all $i \in [m]$.

Using the no-regret property of the adversarial bandit instances used by each player for each (h,s), we then obtain (Lemmas 33 and 34) that $\widehat{\pi}$ is an ϵ-CCE of $\mathbb{G}_{\widehat{\mathcal{V}}}$. To derive such a guarantee for the true SG $\mathbb{G}$, we use two facts: first, by the optimistic nature of the rewards of $\mathbb{G}_{\widehat{\mathcal{V}}}$ the value function of $\mathbb{G}_{\widehat{\mathcal{V}}}$ is always an *upper bound* on the value function of $\mathbb{G}$, and second, by the termination

Algorithm 1 SPoCMAR (**S**tage-based **Po**licy **C**over for **M**ulti-**A**gent Learning with **R**max)

procedure SPoCMAR($m, \mathcal{S}, \mathcal{A}, H, K, N_{\text{visit}}, p$)
  Set $\mathcal{V} = \emptyset$. ($\mathcal{V}$ *denotes the set of "well-visited" states, updated at each stage.*)
  For each $h \in [H]$, $s \in \mathcal{S}$, set $\pi^{\text{cover}}_{h,s} = \perp$. ($\pi^{\text{cover}}_{h,s}$ *will be set to a joint policy in* $\Delta(\mathcal{A})^{[H]\times\mathcal{S}}$.)
  for $q \geq 1$ and while $\tau = 0$ **do**
    Set $\tau = 1$ (τ *is a bit indicating whether we should terminate at the current stage*).
    Set $\Pi^q_h := \{\pi^{\text{cover}}_{h,s} : s \in \mathcal{S}\}$ for each $h \in [H]$. (*Note that* $|\Pi^q_h| \leq S$ *for each* h.)
    for $h = H, H-1, \ldots, 1$ **do**
      Set $k = 0$, and $\overline{V}^q_{i,H+1}(s) = 0$ for all $s \in \mathcal{S}$ and $i \in [m]$.
      Each player i initializes an adversarial bandit instance at each state $s \in \mathcal{S}$ for the step h, according to some algorithm satisfying the guarantee of Theorem 28.
      for each $\pi \in \Pi^q_h \cup \pi^{\mathcal{U}}$ **do** ($\pi^{\mathcal{U}}$ *chooses actions uniformly at random*)
        for a total of K times **do**
          Increment k by 1.
          Let $\overline{\pi}$ be the policy which follows π for the first $h-1$ steps and plays according to the bandit algorithm for the state visited at step h (and acts arbitrarily for steps $h' > h$).
          Draw a joint trajectory $(s_{1,k}, \boldsymbol{a}_{1,k}, \boldsymbol{r}_{1,k}, \ldots, s_{H,k}, \boldsymbol{a}_{H,k}, \boldsymbol{r}_{H,k})$ from $\overline{\pi}$.
          if $(h, s_{h,k}) \in \mathcal{V}$ **then**
            Each i updates its bandit alg. at $(h, s_{h,k})$ w/ $(a_{i,h,k}, \frac{H - r_{i,h,k} - \overline{V}^q_{i,h+1}(s_{h+1,k})}{H})$.
          else
            Each i updates its bandit alg. at $(h, s_{h,k})$ w/ $(a_{i,h,k}, \frac{H-(H+1-h)}{H})$.
          end if
        end for
      end for
      For each $s \in \mathcal{S}$, and $j \geq 1$, let $k_{j,h,s} \in [K(S+1)+1]$ denote the jth smallest value of k so that $s_{h,k} = s$, or $K(S+1)+1$ if such a jth smallest value does not exist.
      For each $s \in \mathcal{S}$, let $J_{h,s}$ denote the largest integer j so that $k_{j,h,s} \leq K(S+1)$.
      Define $\widetilde{\pi}^q_h \in \Delta(\mathcal{A})^{\mathcal{S}}$ to be the 1-step policy: $\widetilde{\pi}^q_h(\boldsymbol{a}|s) = \frac{1}{J_{h,s}} \sum_{j=1}^{J_{h,s}} \mathbb{1}[\boldsymbol{a} = \boldsymbol{a}_{h,k_{j,h,s}}]$.
      Set

$$\overline{V}^q_{i,h}(s) := \begin{cases} \frac{1}{J_{h,s}} \sum_{j=1}^{J_{h,s}} \left(r_{i,h,k_{j,h,s}} + \overline{V}^q_{i,h+1}(s_{h+1,k_{j,h,s}}) \right) & : (h,s) \in \mathcal{V} \\ (H+1-h) & : (h,s) \notin \mathcal{V}. \end{cases} \tag{11}$$

    end for
    Define the joint policy $\widetilde{\pi}^q$, which follows $\widetilde{\pi}^q_{h'}$ at each step $h' \in [H]$.
    Call EstVisitation($\widetilde{\pi}^q, N_{\text{visit}}$) (Alg. 2) to obtain estimates $\widehat{d}^q_{h'} \in \Delta(\mathcal{S})$ for each $h' \in [H]$.
    for each $s \in \mathcal{S}$ and $h' \in [H]$ **do**
      if $\widehat{d}^q_{h'}(s) \geq p$ and $(h', s) \notin \mathcal{V}$ **then**
        Set $\pi^{\text{cover}}_{h',s} \leftarrow \widetilde{\pi}^q$.
        Add (h', s) to $\mathcal{V}$.
        Set $\tau \leftarrow 0$.
      end if
    end for
  end for
  return the policy $\widehat{\pi} := \widetilde{\pi}^q$.
end procedure

Algorithm 2 `EstVisitation`

procedure $\textsc{EstVisitation}(\pi, N)$
  for $1 \le n \le N$ **do**
    Draw a trajectory from π, and let $(s_1^n, \ldots, s_H^n)$ denote the sequence of states observed.
  end for
  for $h \in [H]$ **do**
    Let $\widehat{d}_h \in \Delta(\mathcal{S})$ denote the empirical distribution over $(s_h^1, \ldots, s_h^N)$.
  end for
  return $(\widehat{d}_1, \ldots, \widehat{d}_H)$.
end procedure

criterion of `SPoCMAR`, the probability that a trajectory $(s_1, s_2, \ldots, s_H) \sim (\mathbb{G}, \widehat{\pi})$ visits any state $(h, s_h) \notin \widehat{\mathcal{V}}$ is small (Lemma 35). These arguments are worked out in detail in Lemma 36.

Reduction from infinite-horizon to finite-horizon. Finally, to derive Corollary 6, we remark that there is a simple reduction from episodic learning of an infinite-horizon discounted game with discount factor γ to episodic learning of a finite-horizon game with horizon $H := \frac{\log 1/\epsilon}{1-\gamma}$ (see Section C.3). Owing to the fact that $\gamma^H \le \epsilon$, this reduction preserves nonstationary equilibria up to an additive approximation of ϵ.

C.2. Review of adversarial bandit regret bound

`SPoCMAR` requires all players to choose their actions at certain steps of each episode according to a *no-regret* bandit algorithm at each state. For completeness, we briefly overview the setup and guarantees of adversarial no-regret bandit learning. Consider the following setting involving a bandit learner and an adversary interacting over T rounds. The learner has access to a finite set $\mathcal{B}$ of arms, with $B := |\mathcal{B}|$. For each time step $t \in [T]$:

1. The learner picks a distribution $p_t \in \Delta(\mathcal{B})$.
2. The adversary chooses a loss vector $\ell_t \in [0,1]^{\mathcal{B}}$, depending on the arms chosen by the learner at previous steps, as well as a vector $\widetilde{\ell}_t \in [0,1]^{\mathcal{B}}$, so that, if $\mathcal{F}_t$ denotes the sigma-algebra generated by all random variables in the adversary's view up to time t (including ℓ_t), for all $b \in \mathcal{B}$, we have $\mathbb{E}[\widetilde{\ell}_t(b)|\mathcal{F}_t] = \ell_t(b)$.
3. The learner takes action $b_t \sim p_t$ and sees $\widetilde{\ell}_t(b_t)$, which satisfies $\mathbb{E}[\widetilde{\ell}_t(b_t)|b_t, \mathcal{F}_t] = \ell_t(b_t)$. The learner uses the pair $(b_t, \widetilde{\ell}_t(b_t))$ to update its distribution.

Theorem 28 below gives a high-probability regret guarantee for an adversarial bandit algorithm, which may be taken to be `Exp3-IX` (Neu, 2015); the statement of the theorem differs slightly from that in prior works, so we explain how to derive it from Neu (2015) in Section E.

Theorem 28 (Neu (2015), Theorem 1) *There is an algorithm for the above adversarial bandit setting which obtains the following regret guarantee and runs in* $\mathrm{poly}(B)$ *time at each step: for any* $T_0 \in \mathbb{N}$, $\delta \in (0,1)$, *we have that, with probability at least* $1-\delta$, *for all* $T \le T_0$,

$$\max_{b \in \mathcal{B}} \sum_{t=1}^{T} (\ell_t(b_t) - \ell_t(b)) \le O\left(\sqrt{TB} \cdot \log(T_0 B/\delta)\right).$$

C.3. Preliminaries for finite-horizon stochastic games

We first introduce the requisite notation and terminology regarding finite horizon games: a finite-horizon m-player stochastic game $\mathbb{G}$ is defined as a tuple $(\mathcal{S}, (\mathcal{A}_i)_{i\in[m]}, \mathbb{P}, (r_i)_{i\in[m]}, H, \mu)$, which have the same interpretations as in the infinite-horizon discounted case, with the following exceptions:

- $H \in \mathbb{N}$ denotes the horizon (replacing the discount factor γ); in particular, a trajectory proceeds for a total of H steps, at which point it terminates.
- The reward and transitions are allows to depend on the step $h \in [H]$: in particular, r_i is to be interpreted as a tuple $r_i = (r_{i,1}, \ldots, r_{i,H})$, where each $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [-1, 1]$, and $\mathbb{P}$ is to be interpreted as a tuple $\mathbb{P} = (\mathbb{P}_1, \ldots, \mathbb{P}_H)$, where each $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$.

When discussing the finite-horizon case, we consider only *nonstationary* policies, which are sequences of maps $\pi = (\pi_1, \ldots, \pi_H)$, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$; we will therefore drop the descriptor "nonstationary". The space of such policies is denoted $\Delta(\mathcal{A})^{[H]\times\mathcal{S}}$. With a slight abuse of notation we will denote the value function of a nonstationary policy π by $V^\pi_{i,h} : \mathcal{S} \to \mathbb{R}$, $h \in [H]$, which is defined similarly to (1) except with no discount factor; in particular, in the finite-horizon setting, we have, for all $i \in [m], h \in [H], s \in \mathcal{S}$,

$$V^\pi_{i,h}(s) = \mathbb{E}_{(s_h, \boldsymbol{a}_h, \ldots, s_H, \boldsymbol{a}_H)\sim(\mathbb{G},\pi)} \left[\sum_{h'=h}^{H} r_{i,h}(s_{h'}, \boldsymbol{a}_{h'}) | s_h = s \right],$$

and $V^\pi_{i,h}(\mu) := \mathbb{E}_{s\sim\mu}\left[V^\pi_{i,h}(s)\right]$. We also write $V^\pi_i := V^\pi_{i,1}$ for simplicity, as in the infinite-horizon case. Given a policy $\pi \in \Delta(\mathcal{A})^{[H]\times\mathcal{S}}$ and $i \in [m]$, the best response policy $\pi^\dagger_i(\pi_{-i})$ is defined exactly as in the infinite-horizon case, so that, in particular, $V^{\dagger,\pi_{-i}}_{i,h}(s) = \sup_{\pi'_i\in\Delta(\mathcal{A}_i)^{[H]\times\mathcal{S}}} V^{\pi'_i\times\pi_{-i}}_{i,h}(s)$ for all $(h, s) \in [H] \times \mathcal{S}$. Finally, the notion of ϵ-(nonstationary) CCE is defined exactly as in Definitions 1, recalling that $V^\pi_i(\mu) = V^\pi_{i,1}(\mu)$ and $V^{\dagger,\pi_{-i}}_i(\mu) = V^{\dagger,\pi_{-i}}_{i,1}(\mu)$ by definition. When we wish to clarify the SG $\mathbb{G}$ that corresponds to a value function, we will write $V^{\mathbb{G},\pi}_i$ in place of V^π_i.

Given a policy $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$, the *state visitation distribution* d^π_h at step h for the policy π is defined similarly to in the infinite-horizon discounted case, except with different normalization: for all $s \in \mathcal{S}$, $d^\pi_h(s) := \mathbb{P}_{(s_1,\ldots,s_H)\sim(\mathbb{G},\pi)}(s_h = s)$. Here the trajectory $(s_1, \ldots, s_H)$ drawn from $(\mathbb{G}, \pi)$, is drawn with initial state state $s_1 \sim \mu$.

Given an infinite-horizon discounted game $\mathbb{G}'$ and a desired accuracy level ϵ, we consider the following finite-horizon game $\mathbb{G}$ with horizon $H := \frac{\log 1/\epsilon}{1-\gamma}$, so that $\gamma^H \le \epsilon$. The state space, action space, initial state distribution, and transitions at each step of $\mathbb{G}$ are the same as those of $\mathbb{G}'$. Letting $r'_i : \mathcal{S} \to [-1, 1]$ denote the reward function of $\mathbb{G}'$ and $r_{i,h} : \mathcal{S} \to [-1, 1]$ (for $h \in [H]$) denote the reward function of $\mathbb{G}$, we define $r_{i,h}(s, \boldsymbol{a}) := \gamma^{h-1} \cdot r_i(s, \boldsymbol{a})$. It is straightforward to see that for all nonstationary policies $\pi' \in \Delta(\mathcal{A})^{\mathbb{N}\times\mathcal{S}}$, the truncation of π' to the first H steps, which we denote by $\pi \in \Delta(\mathcal{A})^{[H]\times\mathcal{S}}$ satisfies, for all $i \in [m], s \in \mathcal{S}$, $\left|\frac{V^{\mathbb{G}',\pi'}_i(s)}{1-\gamma} - V^{\mathbb{G},\pi}_i(s)\right| \le \frac{\epsilon}{1-\gamma}$. Thus, given an ϵ-CCE of $\mathbb{G}$, we may readily construct a 2ϵ-CCE of $\mathbb{G}'$. Furthermore, if our algorithm is given access to $\mathbb{G}'$ in the episodic PAC-RL model of Section 2.3, we may readily simulate access to $\mathbb{G}$ by drawing the first H steps of a trajectory of $\mathbb{G}'$ and discounting the reward received at each

step $h \in [H]$ by a factor of γ^{h-1}. Thus, for the remainder of the section, we proceed to discuss the problem of learning approximate CCE in finite-horizon general-sum stochastic games (i.e., the proof of Theorem 5).

C.4. Parameters & concentration inequalities

Fix a finite-horizon stochastic game $\mathbb{G} = (\mathcal{S}, (\mathcal{A}_i)_{i\in[m]}, \mathbb{P}, (r_i)_{i\in[m]}, H, \mu)$, and an error parameter $\epsilon > 0$ as well as a failure probability $\delta > 0$. For fixed values of the above, we introduce the following notation and parameters for use throughout this section:

- Choose $p = \frac{\epsilon}{16SH^2}$.
- Choose $J = C_J \cdot \frac{H^6\iota^2 \cdot \max_{i\in[m]} A_i}{\epsilon^2}$, for some sufficiently large constant $C_J > 2$ (to be specified below).
- Choose $K = \frac{8J}{p}$.
- Choose $\varepsilon^{\mathrm{val}} = \frac{\epsilon}{4H}$.
- Choose $\varepsilon^{\mathrm{reg}} = \frac{\epsilon}{8H}$.
- Choose $\varepsilon^{\mathrm{tvd}} = p/2$.
- Choose $N_{\mathrm{visit}} = C_N \cdot \frac{S\iota}{(\varepsilon^{\mathrm{tvd}})^2}$, for some sufficiently large constant $C_N > 1$ (to be specified below).
- Set $\iota := \log\left(\frac{SH \max_{i\in[m]} A_i}{\epsilon\delta}\right)$.
- Let $\widehat{q}$ denote the value of q at termination of SPoCMAR (i.e., $\widehat{q}$ denotes the total number of stages completed by the algorithm).

Also, recall the following parameters introduced in SPoCMAR:

- For $q \geq 1$ and $k \geq 1$ we let $(s^q_{1,k}, \boldsymbol{a}^q_{1,k}, \boldsymbol{r}^q_{1,k}, \ldots, s^q_{H,k}, \boldsymbol{a}^q_{H,k}, \boldsymbol{r}^q_{H,k})$ denote the trajectory drawn in step 14 of SPoCMAR at stage q.
- For $h \in H$, we write $\widehat{s}_{h,k} := s^{\widehat{q}}_{h,k}, \widehat{\boldsymbol{a}}_{h,k} := \boldsymbol{a}^{\widehat{q}}_{h,k}, \widehat{\boldsymbol{r}}_{h,k} = \boldsymbol{r}^{\widehat{q}}_{h,k}$ to denote the trajectory at the final stage $\widehat{q}$.
- For $j \geq 1, h \in [H]$, and $q \geq 1$, let $k^q_{j,h,s}$ and $J^q_{h,s}$ denote the values of the parameters $k_{j,h,s}$ and $J_{h,s}$ defined in steps 22 and 23 at stage q.
- For all j, h, s, write $\widehat{k}_{j,h,s} := k^{\widehat{q}}_{j,h,s}$ and $\widehat{J}_{h,s} := J^{\widehat{q}}_{h,s}$ to denote the values at the final stage.
- For $q \geq 1$, let $\mathcal{V}^q$ denote the value of the set $\mathcal{V}$ at the beginning of stage q of SPoCMAR. Furthermore we will write $\widehat{\mathcal{V}}$ to denote $\mathcal{V}^{\widehat{q}}$, which is the value of $\mathcal{V}$ at the termination of SPoCMAR (by the termination criterion, the value of the set $\mathcal{V}$ does not change during the final stage $\widehat{q}$.

Throughout the proof, we let $C > 1$ denote a constant whose value may change from line to line.

Our first basic lemma states that SPoCMAR always terminates (in particular, the for loop at step 4 terminates).

Lemma 29 *The algorithm SPoCMAR terminates after at most SH stages (i.e., $\widehat{q} \le SH$).*

Proof If SPoCMAR does not terminate at some stage q, then it must add some pair (h', s) to $\mathcal{V}^q$ at that stage, which did not previously belong to $\mathcal{V}^q$. Since elements of $\mathcal{V}^q$ are never removed, the total number of stages is bounded above by SH. ■

The next lemma states that the state visitation estimates constructed in step 28 of SPoCMAR are accurate with high probability.

Lemma 30 *There is an event $\mathcal{E}^{\text{visitation}}$ that occurs with probability at least $1-\delta$ so that under the event $\mathcal{E}^{\text{visitation}}$, for all stages $q \ge 1$, and all $h' \in [H]$, it holds that*

$$\left\| d_{h'}^{\widetilde{\pi}^q} - \widehat{d}_{h'}^q \right\|_1 \le \varepsilon^{\text{tvd}}.$$

Proof Consider any call to $\texttt{EstVisitation}(\pi, N)$, which produces outputs $(\widehat{d}_1, \dots, \widehat{d}_H)$. Then by (Canonne, 2020, Theorem 1), for any $h \in [H]$, with probability $1 - \delta/(H^2 S)$, as long as $N \ge C \cdot \frac{S + \log(H^2 S/\delta)}{(\varepsilon^{\text{tvd}})^2}$ (for a sufficiently large constant C), it holds that $\left\| d_h^\pi - \widehat{d}_h \right\|_1 \le \varepsilon^{\text{tvd}}$. Taking a union bound over all $h \in [H]$ and the at most SH stages q at which EstVisitation is called at step 28 of SPoCMAR, we obtain the claim of the lemma as long as $N_{\text{visit}} \ge C \cdot \frac{S + \log(H^2 S/\delta)}{(\varepsilon^{\text{tvd}})^2}$; but this inequality is ensured by our choice of N_{visit} in Section C.4. ■

Lemma 31 *There is an event $\mathcal{E}^{\text{coverage}}$ that occurs with probability at least $1-\delta$ so that under the event $\mathcal{E}^{\text{coverage}} \cap \mathcal{E}^{\text{visitation}}$, for all stages q, $h \in [H]$, and $s \in \mathcal{S}$, then if $(h, s) \in \mathcal{V}^q$, it holds that $J_{h,s}^q \ge J$.*

Proof If $(h, s) \in \mathcal{V}^q$, then for some $q' < q$, we must have that $\widehat{d}_h^{q'}(s) \ge p$, and $\pi_{h,s}^{\text{cover}}$ was set to $\widetilde{\pi}^{q'}$. By Lemma 30 and since $\varepsilon^{\text{tvd}} \le p/2$, it must hold that $d_h^{\widetilde{\pi}^{q'}}(s) \ge p/2$ under the event $\mathcal{E}^{\text{visitation}}$. Let us now condition on the event $\mathcal{E}^{\text{visitation}}$; then in the for-loop in step 10 corresponding to the policy $\pi_{h,s}^{\text{cover}} = \widetilde{\pi}^{q'}$, for each of the K episodes (in the loop in step 11), (h, s) is visited with probability at least $p/2$. Thus, by the Chernoff bound, with probability at least $1 - e^{-Kp/8}$, the number of episodes k in the loop in step 10 at which (h, s) is visited is at least $Kp/2$, i.e., we must have $J_{h,s}^q \ge Kp/2$. Then by the union bound, since $Kp/2 \ge J$ and $Kp/8 > 2\iota \ge \log(S^2 H^2/\delta)$ (see Section C.4), under an event $\mathcal{E}^{\text{coverage}}$ occuring with probability at least $1-\delta$, for all stages q and all $(h, s) \in \mathcal{V}^q$, the state (h, s) will be visited at least $Kp/2 \ge J$ times at stage q. This completes the proof of the lemma. ■

C.5. Intermediate game

Recall that $\widehat{\mathcal{V}} = \mathcal{V}^{\widehat{q}}$ denotes the value of the set $\mathcal{V}$ at termination of SPoCMAR. Define a stochastic game $\mathbb{G}_{\widehat{\mathcal{V}}}$ as follows:

- The action space of $\mathbb{G}_{\widehat{\mathcal{V}}}$ is $\mathcal{A}$ and the state space of $\mathbb{G}_{\widehat{\mathcal{V}}}$ is $\mathcal{S} \cup s^1_{\text{sink}}$, where s^1_{sink} is a special state which always transitions to itself deterministically and at which all players receive reward 1 at each step.
- For all $(h, s) \in \widehat{\mathcal{V}}$, the transitions and reward of $\mathbb{G}_{\widehat{\mathcal{V}}}$ are identical to that of $\mathbb{G}$.
- For all $(h, s) \notin \widehat{\mathcal{V}}$, all joint actions at (h, s) yield reward 1 to all players and transition to to the state s^1_{sink}.

In the remainder of the section, we will be working with the parameters of both $\mathbb{G}$ and $\mathbb{G}_{\widehat{\mathcal{V}}}$: to avoid amibuity, we denote their value functions $V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\pi}$ and $V_{i,h}^{\mathbb{G},\pi}$; we denote their transitions as $\mathbb{P}_h^{\mathbb{G}_{\widehat{\mathcal{V}}}}$ and $\mathbb{P}_h^{\mathbb{G}}$; and we denote their reward functions as $r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}$ and $r_{i,h}^{\mathbb{G}}$.

Recall that $\widehat{\pi}$ denotes the policy output by SPoCMAR. The below lemma shows that the value functions $\overline{V}_{i,h}^{\widehat{q}}$ constructed at the final stage of SPoCMAR are close to those of $\mathbb{G}_{\widehat{\mathcal{V}}}$ under $\widehat{\pi}$.

Lemma 32 *There is an event $\mathcal{E}^{\text{val}}$ that occurs with probability at least $1 - 2\delta$, so that under the event $\mathcal{E}^{\text{val}}$, for all $s \in \mathcal{S},\ h \in [H],\ i \in [m]$, it holds that*

$$\left|\overline{V}_{i,h}^{\widehat{q}}(s) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)\right| \le \varepsilon^{\text{val}}. \tag{12}$$

Proof We use reverse induction on h, noting that the base case $h = H + 1$ is immediate since all value functions are identically 0. Now suppose that there is some constant C so that for all $h' > h$, it holds that, for some event $\mathcal{E}^{\text{val}}_{h'}$, under the event $\mathcal{E}^{\text{val}}_{h'} \cap \mathcal{E}^{\text{coverage}}$, for all $s \in \mathcal{S}$ and $i \in [m]$,

$$\left|\overline{V}_{i,h'}^{\widehat{q}}(s) - V_{i,h'}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)\right| \le (H + 1 - h') \cdot CH\sqrt{\frac{\iota}{J}}. \tag{13}$$

We will show that (13) holds with $h' = h$, for an appropriate choice of the event $\mathcal{E}^{\text{val}}_h \supset \bigcup_{h' \ge h} \mathcal{E}^{\text{val}}_{h'}$. To do so, consider any state $s \in \mathcal{S}$ and any agent $i \in [m]$; we consider the following two cases regarding s:

Case 1. $(h, s) \in \widehat{\mathcal{V}}$. By Lemma 31, under the event $\mathcal{E}^{\text{coverage}}$, we have that $\widehat{J}_{h,s} = J^{\widehat{q}}_{h,s} \ge J$. For each $j \ge 1$, recall that we have defined $\widehat{k}_{j,h,s} = k^{\widehat{q}}_{j,h,s} \in [K(S+1)+1]$ (and $k^{\widehat{q}}_{j,h,s}$ is defined in step 22 of SPoCMAR). As h, s are fixed, we will write $k_j := \widehat{k}_{j,h,s}$. It is evident that k_j is a stopping time for each j. Note that, for any $t \ge 1$, the sequence

$$\left(\mathbb{1}[k_j \le K(S+1)] \cdot \left(\widehat{r}_{i,h,k_j} + \overline{V}^{\widehat{q}}_{i,h+1}(\widehat{s}_{h+1,k_j}) - \mathbb{E}_{s' \sim \mathbb{P}_h^{\mathbb{G}}(\cdot|\widehat{s}_{h,k_j},\widehat{\boldsymbol{a}}_{h,k_j})}\left[r^{\mathbb{G}}_{i,h}(\widehat{s}_{h,k_j},\widehat{\boldsymbol{a}}_{h,k_j}) + \overline{V}^{\widehat{q}}_{i,h+1}(s')\right]\right)\right)_{1 \le j \le t} \tag{14}$$

is a martingale difference sequence with respect to the filtration $\mathcal{F}_j$, where $\mathcal{F}_j$ denotes the sigma-field generated by all random variables up to step $h+1$ of episode k_j (i.e., of stage $\widehat{q}$). By the Azuma-Hoeffding inequality and a union bound, it follows that, with probability at least $1 - \delta/(HmS)$, for

all $1 \le t \le K(S+1)$,

$$\frac{1}{t} \cdot \left| \sum_{j=1}^{t} \mathbb{1}[k_j \le K(S+1)] \cdot \left(\widehat{r}_{i,h,k_j} + \overline{V}^{\widehat{q}}_{i,h+1}(\widehat{s}_{h+1,k_j}) - \mathbb{E}_{s' \sim \mathbb{P}^{\mathbb{G}}_h(\cdot|\widehat{s}_{h,k_j}, \widehat{\boldsymbol{a}}_{h,k_j})} \left[r^{\mathbb{G}}_{i,h}(\widehat{s}_{h,k_j}, \widehat{\boldsymbol{a}}_{h,k_j}) + \overline{V}^{\widehat{q}}_{i,h+1}(s') \right] \right) \right|$$
$$\le CH\sqrt{\frac{\iota}{t}},$$

where $C > 1$ denotes some constant. Let $\mathcal{E}^{\mathrm{val}}_h$ denote the intersection of $\mathcal{E}^{\mathrm{val}}_{h+1}$ and all instances of this probability $1 - \delta/(HmS)$ event, over $i \in [m], s \in \mathcal{S}$. In particular, the above inequality holds for $t = \widehat{J}_{h,s}$ under $\mathcal{E}^{\mathrm{val}}_h$, which gives that, under the event $\mathcal{E}^{\mathrm{coverage}} \cap \mathcal{E}^{\mathrm{val}}_h$, $\widehat{J}_{h,s} \ge J$, and so

$$\left| \overline{V}^{\widehat{q}}_{i,h}(s) - \frac{1}{\widehat{J}_{h,s}} \sum_{j=1}^{\widehat{J}_{h,s}} \mathbb{E}_{s' \sim \mathbb{P}^{\mathbb{G}}_h(\cdot|s, \widehat{\boldsymbol{a}}_{h,k_j})} \left[r^{\mathbb{G}}_{i,h}(s, \widehat{\boldsymbol{a}}_{h,k_j}) + \overline{V}^{\widehat{q}}_{i,h+1}(s') \right] \right| \le CH\sqrt{\frac{\iota}{J}}. \tag{15}$$

(Here we have also used that $\widehat{s}_{h,k_j} = s$ for all $j \le \widehat{J}_{h,s}$ by the definition of k_j.)

By definition of $\widehat{\pi}$, we have that, again for the fixed value of (h, s, i),

$$V^{\mathbb{G}_{\widehat{\mathcal{V}}}, \widehat{\pi}}_{i,h}(s) = \frac{1}{\widehat{J}_{h,s}} \cdot \sum_{j=1}^{\widehat{J}_{h,s}} \mathbb{E}_{s' \sim \mathbb{P}^{\mathbb{G}}_h(\cdot|s, \widehat{\boldsymbol{a}}_{h,k_j})} \left[r^{\mathbb{G}}_{i,h}(s, \widehat{\boldsymbol{a}}_{h,k_j}) + V^{\mathbb{G}_{\widehat{\mathcal{V}}}, \widehat{\pi}}_{i,h+1}(s') \right].$$

Here we have used that since $(h, s) \in \widehat{\mathcal{V}}$, it holds that for all $\boldsymbol{a} \in \mathcal{A}$, $\mathbb{P}^{\mathbb{G}}_h(\cdot|s, \boldsymbol{a}) = \mathbb{P}^{\mathbb{G}_{\widehat{\mathcal{V}}}}_h(\cdot|s, \boldsymbol{a})$ and $r^{\mathbb{G}}_{i,h}(s, \boldsymbol{a}) = r^{\mathbb{G}_{\widehat{\mathcal{V}}}}_{i,h}(s, \boldsymbol{a})$. By the inductive hypothesis (13) with $h' = h + 1$, it holds that under the event $\mathcal{E}^{\mathrm{coverage}} \cap \mathcal{E}^{\mathrm{val}}_{h+1}$,

$$\left| V^{\mathbb{G}_{\widehat{\mathcal{V}}}, \widehat{\pi}}_{i,h+1}(s) - \overline{V}^{\widehat{q}}_{i,h+1}(s) \right| \le (H - h) \cdot CH\sqrt{\frac{\iota}{J}}. \tag{16}$$

Combining (15) and (16), we get that, under the event $\mathcal{E}^{\mathrm{coverage}} \cap \mathcal{E}^{\mathrm{val}}_h$,

$$\left| \overline{V}^{\widehat{q}}_{i,h}(s) - V^{\mathbb{G}_{\widehat{\mathcal{V}}}, \widehat{\pi}}_{i,h}(s) \right| \le (H + 1 - h) \cdot CH\sqrt{\frac{\iota}{J}},$$

thus completing the inductive step in the case that $(h, s) \in \widehat{\mathcal{V}}$.

Case 2. $(h, s) \notin \widehat{\mathcal{V}}$. Here we note that, by (11), $\overline{V}^{\widehat{q}}_{i,h}(s) = H + 1 - h$. Furthermore, it is immediate from the definition of $\mathbb{G}_{\widehat{\mathcal{V}}}$ that $V^{\mathbb{G}_{\widehat{\mathcal{V}}}}_{i,h}(s) = H + 1 - h$ since $(h, s) \notin \widehat{\mathcal{V}}$. Hence, we have $\left| \overline{V}^{\widehat{q}}_{i,h}(s) - V^{\mathbb{G}_{\widehat{\mathcal{V}}}, \widehat{\pi}}_{i,h}(s) \right| = 0$ in this case.

Thus we have verified that in all cases, (13) holds at step h, thus completing the inductive step.

Summarizing, if we set $\mathcal{E}^{\mathrm{val}} = \mathcal{E}^{\mathrm{val}}_1 \cap \mathcal{E}^{\mathrm{coverage}}$, then the guarantee (12) holds as long as we have $\varepsilon^{\mathrm{val}} \ge CH^2\sqrt{\frac{\iota}{J}}$, which is ensured by choosing the constant C_J large enough (see Section C.4). Furthermore, by a union bound (over all values of $h \in [H], i \in [m], s \in \mathcal{S}$, $\mathcal{E}^{\mathrm{val}}$ holds with probability at least $1 - 2\delta$. ■

For each $s \in \mathcal{S}, i \in [m],\ a_i \in \mathcal{A}_i$, define

$$\overline{Q}_{i,h}(s, a_i) = \mathbb{E}_{\boldsymbol{a}_{-i} \sim \widehat{\pi}_{-i,h}(s)} \mathbb{E}_{s' \sim \mathbb{P}_h^{\mathbb{G}}(\cdot|s,(a_i,\boldsymbol{a}_{-i}))} \left[r_{i,h}^{\mathbb{G}}(s, (a_i, \boldsymbol{a}_{-i})) + \overline{V}_{i,h+1}^{\widehat{q}}(s') \right]. \tag{17}$$

In Lemma 33 below, we use Theorem 28 (giving a no-regret property for the bandit learners used at each state in SPoCMAR) to bound the difference between $\overline{Q}_{i,h}$ and $\overline{V}_{i,h}^{\widehat{q}}$.

Lemma 33 *There is an event $\mathcal{E}^{\mathrm{reg}}$ that occurs with probability at least $1 - \delta$, so that under the event $\mathcal{E}^{\mathrm{reg}}$, for all $s \in \mathcal{S},\ h \in [H],\ i \in [m]$, it holds that*

$$\max_{a_i \in \mathcal{A}_i} \left(\overline{Q}_{i,h}(s, a_i) - \overline{V}_{i,h}^{\widehat{q}}(s) \right) \leq \varepsilon^{\mathrm{reg}}. \tag{18}$$

Proof Fix any $(s, h, i) \in \mathcal{S} \times [H] \times [m]$. First we treat the case that $(h, s) \notin \widehat{\mathcal{V}}$. Note that by the definition (11) we have that $\overline{V}_{i,h}^{\widehat{q}}(s) \leq H + 1 - h$ for all i, h, s. It then follows that $\overline{Q}_{i,h}(s, a_i) \leq H + 1 - h$, for all i, h, s, a_i. Furthermore, we have that $\overline{V}_{i,h}^{\widehat{q}}(s) = H + 1 - h$ when $(h, s) \notin \widehat{\mathcal{V}}$ (by the definition (11) in SPoCMAR), meaning that $\max_{a_i \in \mathcal{A}_i} \left(\overline{Q}_{i,h}(s, a_i) - \overline{V}_{i,h}^{\widehat{q}}(s) \right) \leq 0$.

For the remainder of the proof treat those pairs (h, s) so that $(h, s) \in \widehat{\mathcal{V}}$. Fix some value of $(h, s) \in \widehat{\mathcal{V}}$, and set, for each $j \geq 1$, $k_j := \widehat{k}_{j,h,s}$. It is evident that for each j, k_j is a stopping time with respect to the filtration $\mathcal{H}_k$, where $\mathcal{H}_k$ denotes the sigma-field generated by all states and actions taken up to (and including) step $h + 1$ of episode k.

Note that each agent i runs the adversarial bandit algorithm at state s and step h at each episode k_j (as $\widehat{s}_{h,k_j} = s$ by definition of k_j). Furthermore, for each j so that $k_j \leq K(S + 1)$, the expected reward that agent i would receive upon playing action $a_i \in \mathcal{A}_i$, conditioned on the actions $\widehat{\boldsymbol{a}}_{-i,h,k_j}$ taken by all other agents at step h of episode k_j, is given by

$$\ell_j(a_i) := \mathbb{E}_{s' \sim \mathbb{P}_h^{\mathbb{G}}(\cdot|s,(a_i,\widehat{\boldsymbol{a}}_{-i,h,k_j}))} \left[\frac{H - r_{i,h}^{\mathbb{G}}(s, (a_i, \widehat{\boldsymbol{a}}_{-i,h,k_j})) - \overline{V}_{i,h+1}^{\widehat{q}}(s')}{H} \right]. \tag{19}$$

Furthermore, it is evident that, for each $j \geq 1$, for the choice of action $\widehat{a}_{i,h,k_j}$ by agent i's bandit algorithm at (s, h), the feedback

$$\widetilde{\ell}_j(\widehat{a}_{i,h,k_j}) := \frac{H - r_{i,h,k_j} - \overline{V}_{i,h+1}^{\widehat{q}}(\widehat{s}_{h+1,k_j})}{H}$$

fed to the bandit algorithm satisfies to $\mathbb{E}[\widetilde{\ell}_j(\widehat{a}_{i,h,k_j}) | \mathcal{F}_{i,j-1}] = \ell_j(\widehat{a}_{i,h,k_j})$, where $\mathcal{F}_{i,j}$ denotes the the sigma-field generated by all states and actions taken up to (and including) step h of episode k_{j+1}. It is straightforward to see that $\mathcal{F}_{i,j}$ is well-defined, as k_{j+1} is a stopping time. It is also evident that $\widetilde{\ell}_j(\widehat{a}_{i,h,k_j})$ is $\mathcal{F}_{i,j}$-measurable, meaning that $\sum_{j=1}^{t} \mathbb{1}[k_j \leq K(S + 1)] \cdot (\widetilde{\ell}_j(\widehat{a}_{i,h,k_j}) - \ell_j(\widehat{a}_{i,h,k_j}))$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_{i,t}$. Thus, by the Azuma-Hoeffding inequality, with probability at least $1 - \delta/(SHm)$, for all $t \in [K(S + 1)]$, we have that for some constant $C > 0$,

$$\frac{1}{t} \cdot \left| \sum_{j=1}^{t} \mathbb{1}[k_j \leq K(S + 1)] \cdot (\widetilde{\ell}_j(\widehat{a}_{i,h,k_j}) - \ell_j(\widehat{a}_{i,h,k_j})) \right| \leq C \sqrt{\frac{\iota}{t}}. \tag{20}$$

Next, we have from Theorem 28 that for some constant C', with probability at least $1 - \delta/(SHm)$, for all $t \in [K(S+1)]$,

$$\max_{a_i \in \mathcal{A}_i} \sum_{j=1}^{t} \ell_j(\widehat{a}_{i,h,k_j}) - \ell_j(a_i) \le C' \iota \cdot \sqrt{tA_i}. \tag{21}$$

From (20) and (21), and choosing $t = \widehat{J}_{h,s}$, we get that, with probability at least $1 - 2\delta/(SHm)$,

$$\max_{a_i \in \mathcal{A}_i} \sum_{j=1}^{\widehat{J}_{h,s}} \widetilde{\ell}_j(\widehat{a}_{i,h,k_j}) - \ell_j(a_i) \le C'' \iota \cdot \sqrt{\widehat{J}_{h,s} \cdot A_i}, \tag{22}$$

for some constant $C'' > 0$. Let the event that (22) holds be denoted as $\mathcal{E}_{i,h,s}$.

By definition we have $\widehat{\pi}_h(s) \in \Delta(\mathcal{A})$ is given by the following distribution: for $a \in \mathcal{A}$,

$$\widehat{\pi}_h(a|s) = \frac{1}{\widehat{J}_{h,s}} \cdot \sum_{j=1}^{\widehat{J}_{h,s}} \mathbb{1}[\widehat{\boldsymbol{a}}_{h,k_j} = a].$$

Therefore using (17) and (19), we have that, for each $a_i \in \mathcal{A}_i$,

$$\frac{1}{\widehat{J}_{h,s}} \cdot \sum_{j=1}^{\widehat{J}_{h,s}} \ell_j(a_i) = 1 - \frac{\overline{Q}_{i,h}(s, a_i)}{H}. \tag{23}$$

From the definition of $\overline{V}^{\widehat{q}}_{i,h}(s)$ in (11), we have

$$\frac{1}{\widehat{J}_{h,s}} \cdot \sum_{j=1}^{\widehat{J}_{h,s}} \widetilde{\ell}_j(\widehat{a}_{i,h,k_j}) = \frac{1}{\widehat{J}_{h,s}} \cdot \sum_{j=1}^{\widehat{J}_{h,s}} \left(1 - \frac{\widehat{r}_{i,h,k_j} + \overline{V}^{\widehat{q}}_{i,h+1}(\widehat{s}_{h+1,k_j})}{H}\right) = 1 - \frac{\overline{V}^{\widehat{q}}_{i,h}(s)}{H}. \tag{24}$$

From (22), (23), and (24), we have that, under the event $\mathcal{E}^{\text{coverage}} \cap \mathcal{E}_{i,h,s}$,

$$\max_{a_i \in \mathcal{A}_i} \left(\overline{Q}_{i,h}(s, a_i) - \overline{V}^{\widehat{q}}_{i,h}(s)\right) \le C'' \iota \cdot H \sqrt{\frac{A_i}{J}}. \tag{25}$$

(In particular, we work under the event $\mathcal{E}^{\text{coverage}}$ to ensure that $\widehat{J}_{h,s} \ge J$). Thus, taking a union bound over all i, h, s, and letting $\mathcal{E}^{\text{reg}} := \mathcal{E}^{\text{coverage}} \cap \bigcap_{i,h,s} \mathcal{E}_{i,h,s}$, which has probability at least $1 - 3\delta$, we get that under the event $\mathcal{E}^{\text{reg}}$, for all s, h, i so that $(h, s) \in \widehat{\mathcal{V}}$, (18) holds as long as we have $\varepsilon^{\text{reg}} \ge C'' \iota H \sqrt{\frac{A_i}{J}}$, which holds as long as C_J is sufficiently large (see Section C.4). ■

Next we combine the previous lemmas in the section to show that the policy $\widehat{\pi}$ is a coarse correlated equilibrium for the game $\mathbb{G}_{\widehat{\mathcal{V}}}$.

Lemma 34 *Under the event $\mathcal{E}^{\text{reg}} \cap \mathcal{E}^{\text{val}}$, for all $i \in [m]$, for any policy π_i of player i, it holds that*

$$V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}}, (\pi_i, \widehat{\pi}_{-i})}(\mu) - V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}}, \widehat{\pi}}(\mu) \le H \cdot (\varepsilon^{\text{reg}} + 2 \cdot \varepsilon^{\text{val}}).$$

Proof For each $i \in [m], h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$, we will write, for any policy $\pi \in \Delta(\mathcal{A})^{[H]\times\mathcal{S}}$ and any $\boldsymbol{a} \in \mathcal{A}$,

$$Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\pi}(s,\boldsymbol{a}) = \mathbb{E}_{s'\sim\mathbb{P}_h^{\mathbb{G}_{\widehat{\mathcal{V}}}}(\cdot|s,\boldsymbol{a})}\left[r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s,\boldsymbol{a}) + V_{i,h+1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\pi}(s')\right].$$

Now fix any $(h,s) \in \widehat{\mathcal{V}}$, so that $\mathbb{P}_h^{\mathbb{G}}(\cdot|s,\boldsymbol{a}) = \mathbb{P}_h^{\mathbb{G}_{\widehat{\mathcal{V}}}}(\cdot|s,\boldsymbol{a})$ and $r_{i,h}^{\mathbb{G}}(s,\boldsymbol{a}) = r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s,\boldsymbol{a})$ for all $a \in \mathcal{A}$. From (17) and the fact that $(h,s) \in \widehat{\mathcal{V}}$, we have

$$\overline{Q}_{i,h}(s,a_i) = \mathbb{E}_{\boldsymbol{a}_{-i}\sim\widehat{\pi}_{-i,h}(s)}\mathbb{E}_{s'\sim\mathbb{P}_h^{\mathbb{G}_{\widehat{\mathcal{V}}}}(\cdot|s,(a_i,\boldsymbol{a}_{-i}))}\left[r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s,(a_i,\boldsymbol{a}_{-i})) + \overline{V}_{i,h+1}^{\widehat{q}}(s')\right].$$

Then for any fixed action $a_i \in \mathcal{A}_i$ of player i, we have, under the event $\mathcal{E}^{\mathrm{val}}$ (see Lemma 32),

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widehat{\pi}_{-i,h}(s)}\left[Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s,(a_i,\boldsymbol{a}_{-i}))\right] - \overline{Q}_{i,h}(s,a_i)\right| \\
=&\left|\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widehat{\pi}_{-i,h}(s)}\left[Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s,(a_i,\boldsymbol{a}_{-i})) - \mathbb{E}_{s'\sim\mathbb{P}_h^{\mathbb{G}_{\widehat{\mathcal{V}}}}(\cdot|s,(a_i,\boldsymbol{a}_{-i}))}\left[r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s,(a_i,\boldsymbol{a}_{-i})) + \overline{V}_{i,h+1}^{\widehat{q}}(s')\right]\right]\right| \\
=&\left|\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widehat{\pi}_{-i,h}(s)}\mathbb{E}_{s'\sim\mathbb{P}_h^{\mathbb{G}_{\widehat{\mathcal{V}}}}(\cdot|s,(a_i,\boldsymbol{a}_{-i}))}\left[V_{i,h+1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s') - \overline{V}_{i,h+1}^{\widehat{q}}(s')\right]\right| \\
\le&\varepsilon^{\mathrm{val}}, \qquad (26)
\end{aligned}$$

where the final inequality uses (12). Therefore, for all $(h,s) \in \widehat{\mathcal{V}}$, it holds that, under the event $\mathcal{E}^{\mathrm{val}} \cap \mathcal{E}^{\mathrm{reg}}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widehat{\pi}_{-i,h}(s)}\left[Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s,(a_i,\boldsymbol{a}_{-i})) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)\right] \\
\le&\overline{Q}_{i,h}(s,a_i) - \overline{V}_{i,h}^{\widehat{q}}(s) + 2\cdot\varepsilon^{\mathrm{val}} \qquad (27)\\
\le&\varepsilon^{\mathrm{reg}} + 2\cdot\varepsilon^{\mathrm{val}}, \qquad (28)
\end{aligned}$$

where (27) follows from (26) as well as $\left|V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s) - \overline{V}_{i,h}^{\widehat{q}}(s)\right| \le \varepsilon^{\mathrm{val}}$ under $\mathcal{E}^{\mathrm{val}}$, and (28) follows from Lemma 33.

Now consider any $(h,s) \notin \widehat{\mathcal{V}}$. Since a reward of at most 1 can be received at each step in $\mathbb{G}_{\widehat{\mathcal{V}}}$, it holds that $Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s,\boldsymbol{a}) \le H+1-h$ for all $i \in [m]$ and $\boldsymbol{a} \in \mathcal{A}$. Furthermore, since it still holds that $\left|V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s) - \overline{V}_{i,h}^{\widehat{q}}(s)\right| \le \varepsilon^{\mathrm{val}}$ under $\mathcal{E}^{\mathrm{val}}$, we see that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widehat{\pi}_{-i,h}(s)}\left[Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s,(a_i,\boldsymbol{a}_{-i})) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s)\right] \\
\le&(H+1-h) - \overline{V}_{i,h}^{\widehat{q}}(s) + \varepsilon^{\mathrm{val}} \le \varepsilon^{\mathrm{val}}, \qquad (29)
\end{aligned}$$

where the final inequality follows from $\overline{V}_{i,h}^{\widehat{q}}(s) = H+1-h$ for $(h,s) \notin \widehat{\mathcal{V}}$ (see (11)).

Now fix any player i and any policy π_i of player i. Since the policy $\widehat{\pi}_{-i}$ is a Markov policy, the value function of the game $\mathbb{G}_{\widehat{\mathcal{V}}}$ as a function of player i's policy is equivalent to that of a MDP. Thus, we may apply the finite horizon version of the performance difference lemma (Kakade and

Langford, 2002), which gives that

$$V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},(\pi_i,\widehat{\pi}_{-i})}(\mu) - V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(\mu) = \mathbb{E}_{s_{1:H},\boldsymbol{a}_{1:H}\sim(\mathbb{G}_{\widehat{\mathcal{V}}},(\pi_i,\widehat{\pi}_{-i}))}\left[\sum_{h=1}^{H} Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s_h,\boldsymbol{a}_h) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s_h)\right]. \tag{30}$$

For each $h \in [H]$, we bound the hth term in the above expression as follows:

$$\begin{aligned}
&\mathbb{E}_{s_{1:H},\boldsymbol{a}_{1:H}\sim(\mathbb{G}_{\widehat{\mathcal{V}}},(\pi_i,\widehat{\pi}_{-i}))}\left[Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s_h,\boldsymbol{a}_h) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s_h)\right] \\
=&\mathbb{E}_{s_h\sim(\mathbb{G}_{\widehat{\mathcal{V}}},(\pi_i,\widehat{\pi}_{-i}))}\mathbb{E}_{a_i\sim\pi_{i,h}(s_h)}\mathbb{E}_{\boldsymbol{a}_{-i}\sim\widehat{\pi}_{-i,h}(s_h)}\left[Q_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s_h,(a_i,\boldsymbol{a}_{-i})) - V_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(s_h)\right] \\
\leq&\varepsilon^{\mathrm{reg}} + 2\cdot\varepsilon^{\mathrm{val}},
\end{aligned} \tag{31}$$

where (31) follows from (28) and (29). Thus, from (30), we get that, under $\mathcal{E}^{\mathrm{reg}} \cap \mathcal{E}^{\mathrm{val}}$,

$$V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},(\pi_i,\widehat{\pi}_{-i})}(\mu) - V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(\mu) \leq H\cdot(\varepsilon^{\mathrm{reg}} + 2\cdot\varepsilon^{\mathrm{val}}).$$

This completes the proof. ■

C.6. Completion of proof of Theorem 5

Lemma 34 comes close to showing that $\widehat{\pi}$ is an ϵ-CCE, except that it applied to the game $\mathbb{G}_{\widehat{\mathcal{V}}}$ as opposed to the game $\mathbb{G}$. To get guarantees for the game $\mathbb{G}$, we need to bound the probability (under $\widehat{\pi}$ that a trajectory visits a state not in $\widehat{\mathcal{V}}$), which is done in Lemma 35 below.

Lemma 35 *For the output policy $\widehat{\pi}$ of* SP∘CMAR*, we have that, for all $h \in [H]$, under the event $\mathcal{E}^{\mathrm{visitation}}$,*

$$\mathbb{P}_{s_h\sim(\mathbb{G},\widehat{\pi})}\left[(h,s_h) \notin \widehat{\mathcal{V}}\right] \leq pS + \varepsilon^{\mathrm{tvd}}.$$

Proof Recall that $\widehat{q}$ denotes the index of the final stage of SP∘CMAR. Under the event $\mathcal{E}^{\mathrm{visitation}}$ of Lemma 30, since we have that $\widehat{\pi} = \widetilde{\pi}^{\widehat{q}}$, it holds that for all $h \in [H]$,

$$\left\|d_h^{\widehat{\pi}} - \widehat{d}_h^{\widehat{q}}\right\|_1 \leq \varepsilon^{\mathrm{tvd}}.$$

Since $\widehat{q}$ is the final stage, it must be the case that for all $s \in \mathcal{S}$ and $h \in [H]$ so that $(h,s) \notin \widehat{\mathcal{V}} = \mathcal{V}^{\widehat{q}}$, it holds that $\widehat{d}_h^{\widehat{q}}(s) < p$. In particular, for each $h \in [H]$, $\sum_{s\in\mathcal{S}:(h,s)\notin\widehat{\mathcal{V}}} \widehat{d}_h^{\widehat{q}}(s) < pS$. Thus, under the event $\mathcal{E}^{\mathrm{visitation}}$, it holds that $\sum_{s\in\mathcal{S}:(h,s)\notin\widehat{\mathcal{V}}} d_h^{\widehat{\pi}}(s) < pS + \varepsilon^{\mathrm{tvd}}$.

By noting that for each $h \in [H]$

$$\mathbb{P}_{s_h\sim(\mathbb{G},\widehat{\pi})}[(h,s_h) \notin \widehat{\mathcal{V}}] = \sum_{s\in\mathcal{S}:(h,s)\notin\widehat{\mathcal{V}}} d_h^{\widehat{\pi}}(s),$$

which concludes the proof. ■

Combining the previous lemmas, we now show that the policy $\widehat{\pi}$ is an approximate CCE of $\mathbb{G}$ with high probability.

Lemma 36 *Under the event $\mathcal{E}^{\text{visitation}} \cap \mathcal{E}^{\text{reg}} \cap \mathcal{E}^{\text{val}}$, the output policy $\widehat{\pi}$ of* SPoCMAR *satisfies the following: for all $i \in [m]$ and policies π_i for player i, we have*

$$V_{i,1}^{\mathbb{G},(\pi_i,\widehat{\pi}_{-i})}(\mu) - V_{i,1}^{\mathbb{G},\widehat{\pi}}(\mu) \le H \cdot (\varepsilon^{\text{reg}} + 2 \cdot \varepsilon^{\text{val}}) + 2H^2 \cdot (pS + \varepsilon^{\text{tvd}}) \le \epsilon.$$

Proof We first show the following two facts hold under the joint event $\mathcal{E}^{\text{visitation}} \cap \mathcal{E}^{\text{reg}} \cap \mathcal{E}^{\text{val}}$:

1. For all joint policies π and all players i, it holds that $V_{i,1}^{\mathbb{G},\pi}(\mu) \le V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\pi}(\mu)$.
2. It holds that $V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(\mu) \le V_{i,1}^{\mathbb{G},\widehat{\pi}}(\mu) + H^2 \cdot (pS + \varepsilon^{\text{tvd}})$.

To see the above facts, fix any joint policy π and note that, by definition,

$$V_{i,1}^{\mathbb{G},\pi}(\mu) = \mathbb{E}_{s_{1:H},\boldsymbol{a}_{1:H}\sim(\mathbb{G},\pi)}\left[\sum_{h=1}^{H} r_{i,h}^{\mathbb{G}}(s_h,\boldsymbol{a}_h)\right]$$
$$V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\pi}(\mu) = \mathbb{E}_{s'_{1:H},\boldsymbol{a}'_{1:H}\sim(\mathbb{G}_{\widehat{\mathcal{V}}},\pi)}\left[\sum_{h=1}^{H} r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h,\boldsymbol{a}'_h)\right].$$

We now construct a coupling between trajectories $(s_{1:H},\boldsymbol{a}_{1:H}) \sim (\mathbb{G},\pi)$ and $(s'_{1:H},\boldsymbol{a}'_{1:H}) \sim (\mathbb{G}_{\widehat{\mathcal{V}}},\pi)$, as follows:

1. First set $s_1 = s'_1$ to be the initial state, drawn from the initial state distribution μ.
2. Set a parameter $\tau = 1$ and $h = 1$.
3. While $\tau = 1$ and $h \le H$:
   (a) Draw a sample $\boldsymbol{a}_h = \boldsymbol{a}'_h \sim \pi_h(s_h)$.
   (b) Draw a sample $s_{h+1} = s'_{h+1} \sim \mathbb{P}_h^{\mathbb{G}}(\cdot|s_h,\boldsymbol{a}_h)$.
   (c) If $(h+1,s_{h+1}) \in \widehat{\mathcal{V}}$, then increment h by 1, and continue.
   (d) If $(h+1,s_{h+1}) \notin \widehat{\mathcal{V}}$, set $\tau = 0$ and increment h by 1.
4. If $h \le H$ (which means that the above loop was terminated early and we must have $\tau = 0$):
   (a) Draw independent samples $(\boldsymbol{a}_{h:H}, s_{h+1:H}) \sim (\mathbb{P}_h^{\mathbb{G}},\pi)$, and $(\boldsymbol{a}'_{h:H}, s'_{h+1:H}) \sim (\mathbb{P}_h^{\mathbb{G}_{\widehat{\mathcal{V}}}},\pi)$, conditioned on starting at state $s_h = s'_h$ at step h.[5]

It is immediate to see that the above joint distribution of $(s_{1:H},\boldsymbol{a}_{1:H},s'_{1:H},\boldsymbol{a}'_{1:H})$ constitutes a coupling between the trajectories induced by the pairs $(\mathbb{G},\pi)$ and $(\mathbb{G}_{\mathcal{V}},\pi)$. Let the distribution of this coupling be denoted by ν. Note that for a pair of trajectories $(s_{1:H},\boldsymbol{a}_{1:H},s'_{1:H},\boldsymbol{a}'_{1:H})$ drawn from the distribution ν, we must have, with probability 1, $s_h = s'_h, \boldsymbol{a}_h = \boldsymbol{a}'_h$ if for all $h' \le h$, $(h',s_{h'}) \in \widehat{\mathcal{V}}$. Let $\mathcal{J}_h$ denote the event that for all $h' \le h$, $(h',s_{h'}) \in \widehat{\mathcal{V}}$, and let $\chi_{\mathcal{J}_h} \in \{0,1\}$ denote the indicator of $\mathcal{J}_h$.

Next, we claim that for all $h \in [H]$ and $i \in [m]$, with probability 1, $r_{i,h}^{\mathbb{G}}(s_h,\boldsymbol{a}_h) \le r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h,\boldsymbol{a}'_h)$: this is evident under the event $\mathcal{J}_h$, since then we have $(s_h,\boldsymbol{a}_h) = (s'_h,\boldsymbol{a}'_h) \in \widehat{\mathcal{V}}$. Furthermore, if $\mathcal{J}_h$

5. In the case that $h = H$, we only draw the joint action profiles $\boldsymbol{a}_H$ and $\boldsymbol{a}'_H$.

does not hold, then for some (random) $h' \le h$, we have $(h', s_{h'}) \notin \widehat{\mathcal{V}}$, meaning that, regardless of the policy π, $r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s_h, \boldsymbol{a}_h) = 1$ (since we have either $s_h = s_{\text{sink}}^1$, and $r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s_{\text{sink}}^1, \boldsymbol{a}) = 1$ for all $\boldsymbol{a}$, or $h = h'$ in which case we have defined $r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h, \boldsymbol{a}) = 1$ for all $\boldsymbol{a}$). It follows that, for any policy π,

$$V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\pi}(\mu) - V_{i,1}^{\mathbb{G},\pi}(\mu) = \mathbb{E}_{(s_{1:H},\boldsymbol{a}_{1:H},s'_{1:H},\boldsymbol{a}'_{1:H})\sim\nu}\left[\sum_{h=1}^{H}\left(r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h,\boldsymbol{a}'_h) - r_{i,h}^{\mathbb{G}}(s_h,\boldsymbol{a}_h)\right)\right] \ge 0,$$

where ν is the joint distribution of $(s_{1:H}, \boldsymbol{a}_{1:H}, s'_{1:H}, \boldsymbol{a}'_{1:H})$ corresponding to π, establishing the first of our claims (item 1) above.

Next we establish item 2, for which we only need to consider the policy $\pi = \widehat{\pi}$ output by SPoCMAR. Under the event $\mathcal{E}^{\text{visitation}}$, we have

$$\begin{aligned}
&\left|V_{i,1}^{\mathbb{G},\widehat{\pi}}(\mu) - V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(\mu)\right| \\
=&\left|\mathbb{E}_{(s_{1:H},\boldsymbol{a}_{1:H},s'_{1:H},\boldsymbol{a}'_{1:H})\sim\nu}\left[\sum_{h=1}^{H}\left(r_{i,h}^{\mathbb{G}}(s_h,\boldsymbol{a}_h) - r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h,\boldsymbol{a}'_h)\right)\right]\right| \\
\le&\sum_{h=1}^{H}\left|\mathbb{E}_{\nu}\left[\chi_{\mathcal{J}_h}\cdot\left(r_{i,h}^{\mathbb{G}}(s_h,\boldsymbol{a}_h) - r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h,\boldsymbol{a}'_h)\right)\right]\right| + \sum_{h=1}^{H}\mathbb{E}_{\nu}\left[\left|(1-\chi_{\mathcal{J}_h})\cdot\left(r_{i,h}^{\mathbb{G}}(s_h,\boldsymbol{a}_h) - r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h,\boldsymbol{a}'_h)\right)\right|\right] \\
\le&2\sum_{h=1}^{H}\mathbb{E}_{\nu}\left[1-\chi_{\mathcal{J}_h}\right] \qquad (32)\\
\le&2H\cdot\mathbb{P}_{s_{1:H},\boldsymbol{a}_{1:H}\sim(\mathbb{G},\widehat{\pi})}\left[\exists h\in[H]\ :\ (h,s_h)\notin\widehat{\mathcal{V}}\right] \\
\le&2H\sum_{h=1}^{H}\mathbb{P}_{s_{1:H},\boldsymbol{a}_{1:H}\sim(\mathbb{G},\widehat{\pi})}\left[(h,s_h)\notin\widehat{\mathcal{V}}\right] \\
\le&2H^2\cdot(pS+\varepsilon^{\text{tvd}}), \qquad (33)
\end{aligned}$$

where (32) follows because $r_{i,h}^{\mathbb{G}}(s_h,\boldsymbol{a}_h) - r_{i,h}^{\mathbb{G}_{\widehat{\mathcal{V}}}}(s'_h,\boldsymbol{a}'_h) = 0$ whenever $\chi_{\mathcal{J}_h} = 1$ (as then $(s_h,\boldsymbol{a}_h) = (s'_h,\boldsymbol{a}'_h) \in \widehat{\mathcal{V}}$), and (33) uses the conclusion of Lemma 35 and the fact that $\mathcal{E}^{\text{visitation}}$ is assumed to hold.

Using items 1 (with the policy $(\pi_i, \widehat{\pi}_{-i})$) and 2 above, we obtain that, under the event $\mathcal{E}^{\text{visitation}} \cap \mathcal{E}^{\text{reg}} \cap \mathcal{E}^{\text{val}}$,

$$\begin{aligned}
V_{i,1}^{\mathbb{G},(\pi_i,\widehat{\pi}_{-i})}(\mu) - V_{i,1}^{\mathbb{G},\widehat{\pi}}(\mu) \le& V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},(\pi_i,\widehat{\pi}_{-i})}(\mu) - V_{i,1}^{\mathbb{G}_{\widehat{\mathcal{V}}},\widehat{\pi}}(\mu) + H^2\cdot(pS+\varepsilon^{\text{tvd}}) \\
\le& H\cdot(\varepsilon^{\text{reg}} + 2\cdot\varepsilon^{\text{val}}) + 2H^2\cdot(pS+\varepsilon^{\text{tvd}}) \le \epsilon,
\end{aligned}$$

where the second-to-last inequality follows from Lemma 34 and the final inequality follows by our choices of $\varepsilon^{\text{reg}}, \varepsilon^{\text{val}}, p, \varepsilon^{\text{tvd}}$ in Section C.4. ■

Finally, we may prove Theorem 5 as a consequence of Lemma 36 and the choices of our parameters.

Proof [Proof of Theorem 5] Consider any stochastic game $\mathbb{G}$ and any $\epsilon, \delta > 0$. Lemma 36 gives that under the event $\mathcal{E}^{\text{visitation}} \cap \mathcal{E}^{\text{reg}} \cap \mathcal{E}^{\text{val}}$ (which has probability at least $1 - 4\delta$), we have that the

output policy $\widehat{\pi}$ of SPoCMAR satisfies

$$\max_{i\in[m]}\left\{V_{i,1}^{\mathbb{G},(\dagger,\widehat{\pi}_{-i})}(\mu)-V_{i,1}^{\mathbb{G},\widehat{\pi}}(\mu)\right\}\le\epsilon,$$

which implies that $\widehat{\pi}$ is an ϵ-(nonstationary) CCE (Definition 1). It remains to bound the number of trajectories collected by SPoCMAR: it is seen by inspection to be bounded above by

$$\begin{aligned}\sum_{q=1}^{\widehat{q}}\left(\sum_{h=1}^{H}|\Pi_h^q|\cdot K+N_{\text{visit}}\right)\le&HS\cdot(SHK+N_{\text{visit}})\\ \le&HS\cdot O\left(SH\cdot\frac{J}{p}+\frac{S\iota}{p^2}\right)\\ \le&HS\cdot O\left(\frac{SH\cdot H^6\iota^2\cdot\max_i A_i\cdot SH^2}{\epsilon^3}+\frac{S\iota\cdot SH^2}{\epsilon^2}\right)\\ \le&O\left(\frac{H^{10}S^3\iota^2\max_{i\in[m]}A_i}{\epsilon^3}\right).\end{aligned}$$

Finally, the proof is completed by rescaling δ to be $\delta/4$. ■

Appendix D. On the Gates Used in GCircuit

The definition of ϵ-GCircuit in (Rubinstein, 2018) uses some gates not introduced in Definition 8, namely $G_\times(\zeta|v_1|v_2)$, $G_=(|v_1|v_2)$, $G_+(|v_1,v_2|v_3)$, $G_-(|v_1,v_2|v_3)$, $G_\vee(|v_1,v_2|v_3)$, $G_\wedge(|v_1,v_2|v_3)$, $G_\neg(|v_1|v_2)$, and $G_\leftarrow(\zeta|v)$, for $\zeta\in[0,1]$. However, it is straightforward to see that these gates may be implemented as follows:

- $G_\times, G_=, G_+, G_-$ may each be implemented using the gate $G_{\times,+}$ (for appropriate choices of ξ,ζ): in particular, we may implement $G_\times(\zeta|v_1|v_2)$ as $G_{\times,+}(\zeta/2,\zeta/2|v_1,v_1|v_2)$, $G_=(|v_1|v_2)$ as $G_{\times,+}(1/2,1/2|v_1,v_1|v_2)$, $G_-(|v_1,v_2|v_3)$ as $G_{\times,+}(1,-1|v_1,v_2|v_3)$, and G_+ as $G_{\times,+}(1,1|v_1,v_2|v_3)$.
- The gate $G_\leftarrow(\zeta|v)$, for $\zeta\in[0,1]$, may be implemented using $G_\leftarrow(1||u)$, $G_{\times,+}(\zeta/2,\zeta/2|u,u|v)$; since any ϵ-approximate assignment π must satisfy $\pi(u)=1$, we get that $\pi(v)=\zeta\cdot\pi(u)\pm\epsilon=\zeta\pm\epsilon$.
- The gate $G_\vee(|v_1,v_2|v_3)$ may be implemented using the following gates: $G_{\times,+}(1/2,1/2|v_1,v_2|u_1)$, $G_\leftarrow(1||u_2)$, $G_{\times,+}(1/8,1/8|u_2,u_2|u_3)$, and $G_<(|u_3,u_1|v_3)$, where u_1,u_2,u_3 are supplementary nodes. Any ϵ-approximate assignment π must satisfy $\pi(u_1)=\frac{\pi(v_1)+\pi(v_2)}{2}\pm\epsilon$, $\pi(u_3)=\frac{1}{4}\pm\epsilon$. Thus, when $\pi(v_1)=1\pm\epsilon$ or $\pi(v_2)=1\pm\epsilon$, as long as $1-2\epsilon>1/4+2\epsilon$ (which holds when $\epsilon<1/16$), $\pi(v_3)=1\pm\epsilon$. Furthermore, when $\pi(v_1)=0\pm\epsilon$ and $\pi(v_2)=0\pm\epsilon$, again as long as $\epsilon<1/16$, we have $\pi(v_3)=0\pm\epsilon$.
- The gate $G_\neg(|v_1|v_2)$ may be implemented using the following gates: $G_\leftarrow(1||u_1)$, $G_{\times,+}(1,-1|u_1,v_1|v_2)$, where u_1 is a supplementary node. Any ϵ-approximate assignment π must satisfy $\pi(u_1)=1$ and so $\pi(v_2)=\max\{\pi(u_1)-\pi(v_1),0\}\pm\epsilon=1-\pi(v_1)\pm\epsilon$.

- The gate $G_\wedge$ may be implemented exactly as $G_\vee(|v_1, v_2|v_3)$ above, except the gate with output node u_3 being replaced with $G_{\times,+}(3/8, 3/8|u_2, u_2|u_3)$. (Alternatively, we may use the gates $G_\neg$ and $G_\vee$.)

We also remark that our requirement that $\pi(v) = \zeta$ for $G_{\leftarrow}(\zeta||v)$ when $\zeta \in \{0, 1\}$ is stronger than that in (Rubinstein, 2018), which allows for error ϵ, and that the gate $G_{\times,+}$ is not considered in (Rubinstein, 2018). However, these modifications only make the problem harder. Summarizing, the ϵ-GCircuit problem with the set of gates listed above is still PPAD-complete for some constant ϵ.

Appendix E. Adversarial Bandit Guarantees

In the context of the adversarial no-regret bandit learning setting described in Section C.2, it was shown in (Neu, 2015, Theorem 1) (see also (Lattimore and Szepesvári, 2020, Theorem 12.1), which is not quite sufficient for us since it requires T to be known ahead of time) that for any $T_0 \in \mathbb{N}$, $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have that for all $T \leq T_0$,

$$\max_{b \in \mathcal{B}} \sum_{t=1}^{T} \left(\widetilde{\ell}_t(b_t) - \widetilde{\ell}_t(b) \right) \leq O\left(\sqrt{TB} \cdot \log(T_0 B/\delta) \right).$$

To obtain Theorem 28 as a consequence of the above, let $\mathcal{F}_t$ denote the sigma-field generated by $b_1, \ldots, b_{t+1}, \widetilde{\ell}_1, \ldots, \widetilde{\ell}_t, \ell_1, \ldots, \ell_{t+1}$. We now note that for each t, $\mathbb{E}[\widetilde{\ell}_t(b_t)|\mathcal{F}_{t-1}] = \ell_t(b_t)$ and for all t, b, $\mathbb{E}[\widetilde{\ell}_t(b)|\mathcal{F}_{t-1}] = \ell_t(b)$. We then apply the Azuma-Hoeffding inequality (followed by a union bound) to each of $\left(\widetilde{\ell}_t(b_t) - \ell_t(b_t)\right)_{t \in [T_0]}$ and, for all $b \in \mathcal{B}$, $\left(\widetilde{\ell}_t(b) - \ell_t(b)\right)_{t \in [T_0]}$, which are martingale difference sequences with respect to the filtration $\mathcal{F}_t$.

Appendix F. Omitted Proofs from Section B

In this section we give the proofs of some lemmas which were omitted in Section B.

F.1. Proofs from Section B.2

Proof [Proof of Lemma 12] Let us view $\pi \in \Delta(\mathcal{A})^{\mathcal{S}} \subset \mathbb{R}^{S \times A}$ as a vector whose components are $\pi(\boldsymbol{a}|s)$, for all $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$. Then by policy gradient theorem (Sutton et al., 1999), we have that for all π and states $s_1 \in \mathcal{S}$,

$$\begin{aligned}
\nabla_\pi V^\pi(s_1) =& \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d^\pi_{s_1}} \mathbb{E}_{\boldsymbol{a} \sim \pi(\cdot|s)} \left[\nabla_\pi \log \pi(\boldsymbol{a}|s) \cdot Q^\pi(s, \boldsymbol{a}) \right] \\
=& \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d^\pi_{s_1}} \left[\sum_{\boldsymbol{a} \in \mathcal{A}} \nabla_\pi \pi(\boldsymbol{a}|s) \cdot Q^\pi(s, \boldsymbol{a}) \right] \\
=& \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d^\pi_{s_1}} \left[\sum_{\boldsymbol{a} \in \mathcal{A}} e_{(s,\boldsymbol{a})} \cdot Q^\pi(s, \boldsymbol{a}) \right],
\end{aligned}$$

where $e_{(s,\boldsymbol{a})}$ denotes an all-zero vector except that the $(s, \boldsymbol{a})$ component is one. For a vector $v \in \mathbb{R}^{S \times A}$ and $s \in \mathcal{S}$, write $v_s := (v_{s,\boldsymbol{a}})_{\boldsymbol{a} \in \mathcal{A}} \in \mathbb{R}^A$. Since $|Q^\pi(s, \boldsymbol{a})| \leq 1$ for all $s, \boldsymbol{a}$ and $\sum_s d^\pi_{s_1}(s) =$

1, it holds that $\sum_{s\in\mathcal{S}} \|(\nabla_\pi V^\pi(s_1))_s\|_\infty \le 1/(1-\gamma)$. Furthermore, note that for any vectors $v, w \in \mathbb{R}^{S\times A}$, we have

$$|\langle v, w\rangle| \le \sum_{s\in\mathcal{S}} |\langle v_s, w_s\rangle| \le \sum_{s\in\mathcal{S}} \|v_s\|_\infty \cdot \|w_s\|_1 \le \max_{s\in\mathcal{S}}\{\|w_s\|_1\} \cdot \sum_{s\in\mathcal{S}} \|v_s\|_\infty.$$

It follows that for all $s \in \mathcal{S}$ and $i \in [m]$,

$$|V_i^\pi(s) - V_i^{\pi'}(s)| \le \max_{\widetilde{\pi}\in\Delta(\mathcal{A})^\mathcal{S}} |\langle \pi - \pi', \nabla_{\widetilde{\pi}} V^{\widetilde{\pi}}(s)\rangle| \le \frac{1}{1-\gamma} \cdot \max_{s'\in\mathcal{S}} \|\pi(\cdot|s') - \pi'(\cdot|s')\|_1,$$

verifying the first claim of the lemma. The second claim follows as a consequence of the first and the fact that for all policies π, $Q_i^\pi(s, \boldsymbol{a}) = (1-\gamma)\cdot r_i(s,\boldsymbol{a}) + \gamma \cdot \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,\boldsymbol{a})}[V_i^\pi(s')]$. ■

F.2. Proofs from Section B.3

Proof [Proof of Lemma 14] For each $i \in [m]$, $s \in \mathcal{S}$, define

$$\begin{aligned} \pi_i^\dagger(s) &\in \arg\max_{a_i\in\mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}(s)}[Q_i^\pi(s,(a_i,\boldsymbol{a}_{-i}))] \\ \rho_{i,s} &:= \max_{a_i\in\mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}(s)}[Q_i^\pi(s,(a_i,\boldsymbol{a}_{-i}))] - V_i^\pi(s). \end{aligned} \tag{34}$$

Since $V_i^\pi(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi(s)}[Q_i^\pi(s,\boldsymbol{a})]$ and $\pi_i(s)$ is a product distribution for all i, s, it holds that for all i, s,

$$\rho_{i,s} = \mathbb{E}_{\boldsymbol{a}\sim(\pi_i^\dagger\times\pi_{-i})(s)}[Q_i^\pi(s,\boldsymbol{a}) - V_i^\pi(s)] = \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}(s)}[Q_i^\pi(s,(\pi_i^\dagger(s),\boldsymbol{a}_{-i})) - V_i^\pi(s)] \ge 0. \tag{35}$$

By the performance difference lemma (Lemma 11), we have, for all $i \in [m]$, $s \in \mathcal{S}$,

$$\begin{aligned} V_i^{\pi_i^\dagger\times\pi_{-i}}(s) - V_i^\pi(s) =& \frac{1}{1-\gamma} \cdot \mathbb{E}_{s'\sim d_s^{\pi_i^\dagger\times\pi_{-i}}} \mathbb{E}_{\boldsymbol{a}\sim\pi_i^\dagger\times\pi_{-i}(s')}[Q_i^\pi(s',\boldsymbol{a}) - V_i^\pi(s')] \\ \ge& \mathbb{E}_{\boldsymbol{a}\sim(\pi_i^\dagger\times\pi_{-i})(s)}[Q_i^\pi(s,\boldsymbol{a}) - V_i^\pi(s)] = \rho_{i,s}, \end{aligned}$$

where the second inequality follows from (35) and the fact that $d_s^\pi(s) \ge 1-\gamma$ by definition of d_s^π.

If π is an ϵ-perfect NE, then we have that $V_i^{\pi_i^\dagger\times\pi_{-i}}(s) - V_i^\pi(s) \le \epsilon$ for all i, s, which therefore, implies that for all i, s, we have $\rho_{i,s} \le \epsilon$, i.e., (2) holds, and therefore π is an ϵ-PNE-SG.

If π is an ϵ-NE, then for all agents $i \in [m]$, we know from Lemma 11 that

$$\begin{aligned} \epsilon \ge \mathbb{E}_{s\sim\mu}\left[V_i^{\pi_i^\dagger\times\pi_{-i}}(s) - V_i^\pi(s)\right] =& \frac{1}{1-\gamma} \cdot \mathbb{E}_{s'\sim d_\mu^{\pi_i^\dagger\times\pi_{-i}}} \mathbb{E}_{\boldsymbol{a}\sim\pi_i^\dagger\times\pi_{-i}(s')}[Q_i^\pi(s',\boldsymbol{a}) - V_i^\pi(s')] \\ \ge& \mathbb{E}_{s\sim\mu}\mathbb{E}_{\boldsymbol{a}\sim(\pi_i^\dagger\times\pi_{-i})(s)}[Q_i^\pi(s,\boldsymbol{a}) - V_i^\pi(s)] = \mathbb{E}_{s\sim\mu}\left[\rho_{i,s}\right], \end{aligned}$$

where we use the fact that $d_\mu^\pi(s) \ge (1-\gamma)\mu(s)$ for all s. This shows that π is an ϵ-NE-SG. ■

Proof [Proof of Lemma 16] We will use the following shorthand notation in the proof: for a Markov product policy π, a state $s \in \mathcal{S}$, an agent $i \in [m]$, and an action $a_i \in \mathcal{A}_i$, we write (with a slight abuse of notation):

$$Q_i^\pi(s, a_i) := \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}(s)} \left[Q_i^\pi(s, (a_i, \boldsymbol{a}_{-i})) \right].$$

Fix a stationary product policy $\pi : \mathcal{S} \to \Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_m)$. For each $i \in [m]$ and $s \in \mathcal{S}$, define

$$\rho_{i,s} := \max_{a_i \in \mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \pi_{-i}(s)} [Q_i^\pi(s, (a_i, \boldsymbol{a}_{-i}))] - V_i^\pi(s) = \max_{a_i \in \mathcal{A}_i} Q_i^\pi(s, a_i) - \mathbb{E}_{a_i \sim \pi_i(s)} [Q_i^\pi(s, a_i)].$$

In the case that π is an ϵ-PNE-SG, we have that $\max_{i,s} \rho_{i,s} \le \epsilon$, and in the case that π is an ϵ-NE-SG, we have that $\max_i \mathbb{E}_{s \sim \mu}[\rho_{i,s}] \le \epsilon$.

Fix some $k > 1$ to be specified later. We construct a new product policy π' as follows: for each $i \in [m]$, $s \in \mathcal{S}$, and $a_i \in \mathcal{A}_i$,

$$\pi_i'(a_i|s) := \begin{cases} \frac{\pi_i(a_i|s)}{1 - \bar{\pi}_i(s)} & : \quad Q_i^\pi(s, a_i) \ge \max_{a_i' \in \mathcal{A}_i} \{Q_i^\pi(s, a_i')\} - k \cdot \rho_{i,s} \\ 0 & : \quad \text{otherwise,} \end{cases}$$

where $\bar{\pi}_i(s)$ is the sum, over all a_i so that $Q_i^\pi(s, a_i) < \max_{a_i' \in \mathcal{A}_i} \{Q_i^\pi(s, a_i')\} - k\rho_{i,s}$, of $\pi_i(a_i|s)$.

Next, using the fact that π is an ϵ-NE-SG, we have the following claim:

Lemma 37 ((Daskalakis et al., 2009), Claim 6) *For all $i \in [m],\ s \in \mathcal{S}$, it holds that*

$$\sum_{a_i \in \mathcal{A}_i} \left| \pi_i'(a_i|s) - \pi_i(a_i|s) \right| \le \frac{2}{k-1}.$$

Thus

$$\max_{s \in \mathcal{S}} \|\pi(s) - \pi'(s)\|_1 \le \max_{s \in \mathcal{S}} \sum_{i \in [m]} \|\pi_i(s) - \pi_i'(s)\|_1 \le \frac{2p}{k-1}.$$

Hence, for all $s \in \mathcal{S}, i \in [m], a \in \mathcal{A}$, by Lemma 12,

$$\left| Q_i^\pi(s, a) - Q_i^{\pi'}(s, a) \right| \le \frac{2p\gamma}{(k-1)(1-\gamma)},$$

which implies that for all $a_i \in \mathcal{A}_i$, we have

$$|Q_i^\pi(s, a_i) - Q_i^{\pi'}(s, a_i)| \le \|\pi_{-i}(\cdot|s) - \pi_{-i}'(\cdot|s)\|_1 + \frac{2p\gamma}{(k-1)(1-\gamma)} \le \frac{2p}{(k-1)(1-\gamma)}.$$

Thus, for all $s \in \mathcal{S}, i \in [m]$, and $a_i \in \mathcal{A}_i$ so that $\pi_i'(a_i|s) > 0$, we have

$$\begin{aligned} Q_i^{\pi'}(s, a_i) &\ge Q_i^\pi(s, a_i) - \frac{2p}{(k-1)(1-\gamma)} \\ &\ge \max_{a_i' \in \mathcal{A}_i} \{Q_i^\pi(s, a_i')\} - k\rho_{i,s} - \frac{2p}{(k-1)(1-\gamma)} \\ &\ge \max_{a_i' \in \mathcal{A}_i} \{Q_i^{\pi'}(s, a_i')\} - k\rho_{i,s} - \frac{8p}{k(1-\gamma)}. \end{aligned}$$

Let us now choose $k = \sqrt{\frac{8p}{(1-\gamma)\epsilon}}$. Then in the case that π is an ϵ-PNE-SG, we have $\rho_{i,s} \leq \epsilon$ for all i, s, and we immediately obtain the desired result.

If π is only an ϵ-NE-SG, then for all $i \in [m]$,

$$\begin{aligned}\mathbb{E}_{s\sim\mu}\left[\max_{a_i'\in\mathcal{A}_i} Q_i^\pi(s,a_i') - \min_{a_i\in\mathcal{A}_i:\pi_i(a_i|s)>0} Q_i^\pi(s,a_i)\right] \leq& \mathbb{E}_{s\sim\mu}\left[k\rho_{i,s} + \frac{8p}{k(1-\gamma)}\right] \\ \leq& k\epsilon + \frac{8p}{k(1-\gamma)},\end{aligned}$$

which gives that π' is an $6\cdot\sqrt{\frac{p\epsilon}{1-\gamma}}$-WSNE-SG. ■

F.3. Proofs from Section B.4

Proof [Proof of Lemma 23] Consider any policy $\pi : \mathcal{S} \to [0,1]$. We first compute $Q^\pi_{\mathtt{cr}(w)}(w,b)$ for $b \in \{0,1\}$:

- $Q^\pi_{\mathtt{cr}(w)}(w,1) = \gamma \cdot V^\pi_{\mathtt{cr}(w)}(v_2) = \gamma\cdot(\beta\cdot\pi(v_2)\pm\gamma) = \gamma\cdot\beta\cdot\pi(v_2)\pm\gamma^2$;
- $Q^\pi_{\mathtt{cr}(w)}(w,0) = \gamma \cdot V^\pi_{\mathtt{cr}(w)}(v_1) = \gamma\cdot(\beta\cdot\pi(v_1)\pm\gamma) = \gamma\cdot\beta\cdot\pi(v_1)\pm\gamma^2$.

We next compute $Q^\pi_{\mathtt{cr}(v_3)}(v_3,b)$ for $b\in\{0,1\}$ in the particular case where $\pi(w)\in\{0,1\}$:

- If $\pi(w) = 1$, then:
  - $Q^\pi_{\mathtt{cr}(v_3)}(v_3,1) = \gamma\beta\pm\gamma^2$.
  - $Q^\pi_{\mathtt{cr}(v_3)}(v_3,0) = 0$.
- If $\pi(w) = 0$, then
  - $Q^\pi_{\mathtt{cr}(v_3)}(v_3,1) = -\gamma\beta\pm\gamma^2$.
  - $Q^\pi_{\mathtt{cr}(v_3)}(v_3,0) = 0$.

Notice that $\gamma|\beta|\epsilon - 2\gamma^2 > \epsilon'$ implies $\beta \neq 0$. Suppose that $\pi(v_1) \leq \pi(v_2) - \epsilon$ and that $\beta > 0$. We have

$$Q^\pi_{\mathtt{cr}(w)}(w,1) - Q^\pi_{\mathtt{cr}(w)}(w,0) \geq \gamma\beta\cdot(\pi(v_2)-\pi(v_1)) - 2\gamma^2 \geq \gamma\beta\epsilon - 2\gamma^2 > \epsilon',$$

which implies that, since w is ϵ'-unimprovable under π, we must have that $\pi(w) = 1$. Then we have

$$Q^\pi_{\mathtt{cr}(v_3)}(v_3,1) - V^\pi_{\mathtt{cr}(v_3)}(v_3,0) \geq \gamma\beta - \gamma^2 > \epsilon',$$

meaning that $\pi(v_3) = 1$ since v_3 is ϵ'-unimprovable under π, which is what we wanted to show in this case. In a similar manner, if $\pi(v_1) \leq \pi(v_2) - \epsilon$ but $\beta < 0$, then we see that $\pi(w) = 0$ and since $-\gamma\beta - \gamma^2 > \epsilon'$, we again get that $\pi(v_3) = 1$.

Similarly, if $\pi(v_1) \geq \pi(v_2) + \epsilon$, then we have $\pi(w) = 0$ if $\beta > 0$ and $\pi(w) = 1$ if $\beta < 0$. In the case that $\beta > 0$, we get that $\pi(v_3) = 0$, and in the case that $\beta < 0$, we also get that $\pi(v_3) = 0$, as desired. ■