Hardness of Independent Learning and Sparse Equilibrium Computation in Markov Games

Dylan J. Foster [1] **Noah Golowich** [2] **Sham M. Kakade** [3]

Abstract

We consider the problem of decentralized multi-agent reinforcement learning in Markov games. A key question is whether there are algorithms that, when run independently by all agents, lead to no-regret for each player, analogous to celebrated results for no-regret learning in normal-form games. While recent work has shown that such algorithms exist for restricted settings (e.g., when regret is defined with respect to deviations to *Markov* policies), the question of whether independent no-regret learning can be achieved in the standard Markov game framework was open. We provide a decisive negative resolution to this problem, both from a computational and statistical perspective. We show that:

1. Under the assumption that PPAD-hard problems cannot be solved in polynomial time, there is no polynomial-time algorithm that attains no-regret in general-sum Markov games when executed independently by all players, even when the game is known to the algorithm designer and the number of players is a small constant.
2. When the game is unknown, no algorithm, efficient or otherwise, can achieve no-regret without observing exponentially many episodes in the number of players.

These results are proven via lower bounds for a simpler problem we refer to as SPARSECCE, in which the goal is to compute a coarse correlated equilibrium that is "sparse" in the sense that it can be represented as a mixture of a small number of product policies.

[1]Microsoft Research [2]Massachusetts Institute of Technology [3]Harvard University. Correspondence to: Dylan J. Foster<dylanfoster@microsoft.com>, Noah Golowich<nzg@mit.edu>, Sham M. Kakade<sham@seas.harvard.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

The framework of *multi-agent reinforcement learning (MARL)*, which describes settings in which multiple agents interact in a dynamic environment, has played a key role in recent breakthroughs in artificial intelligence, including the development of agents that approach or surpass human performance in games such as Go (Silver et al., 2016), Poker (Brown & Sandholm, 2018), Stratego (Perolat et al., 2022), and Diplomacy (Kramár et al., 2022; Bakhtin et al., 2022). MARL also shows promise for real-world multi-agent systems, including autonomous driving (Shalev-Shwartz et al., 2016), and cybersecurity (Malialis & Kudenko, 2015), and economic policy (Zheng et al., 2022). These applications, where reliability is critical, necessitate the development of algorithms that are practical and efficient, yet provide strong formal guarantees and robustness.

Multi-agent reinforcement learning is typically studied using the framework of *Markov games* (also known as *stochastic games*) (Shapley, 1953). In a Markov game, agents interact over a finite number of steps: at each step, each agent observes the *state* of the environment, takes an *action*, and observes a *reward* which depends on the current state as well as the other agents' actions. Then the environment transitions to a new state as a function of the current state and the actions taken. An *episode* consists of a finite number of such steps, and agents interact over the course of multiple episodes, progressively learning new information about their environment. Markov games generalize the well-known model of *Markov Decision Processes (MDPs)* (Puterman, 1994), which describe the special case in which there is a single agent acting in a dynamic environment, and we wish to find a policy that maximizes its reward. By contrast, for Markov games, we typically aim to find a distribution over agents' policies which constitutes some type of *equilibrium*.

1.1. Decentralized learning

In this paper, we focus on the problem of *decentralized* (or, independent) learning in Markov games. In decentralized MARL, each agent in the Markov game behaves independently, optimizing their policy myopically while treating the effects of the other agents as exogenous. Agents observe local information (in particular, their own actions and rewards), but do not observe the actions of the other agents directly. Decentralized learning enjoys a number of desirable properties, including

scalability, versatility, and practicality. The central question we consider is whether there exist decentralized learning algorithms which, when employed by all agents in a Markov game, lead them to play near-equilibrium strategies over time.

Decentralized equilibrium computation in MARL is not well understood theoretically, and algorithms with provable guarantees are scarce. To motivate the challenges and most salient issues, it will be helpful to contrast with the simpler problem of decentralized learning in *normal-form games*, which may be interpreted as Markov games with a single state. Much of the modern work on decentralized learning in normal-form games centers on *no-regret* learning, where agents select actions independently using *online learning* algorithms (Cesa-Bianchi & Lugosi, 2006) designed to minimize their *regret* (that is, the gap between realized payoffs and the payoff of the best fixed action in hindsight). In particular, a foundational result is that if each agent employs a no-regret learning strategy, then the average of the agents' joint action distributions approaches a *coarse correlated equilibrium (CCE)* for the normal-form game (Cesa-Bianchi & Lugosi, 2006; Hannan, 1957; Blackwell, 1956). CCE is a natural relaxation of the foundational concept of *Nash equilibrium*, which has the downside of being intractable to compute. On the other hand, there are many efficient algorithms that can achieve vanishing regret in a normal-form game, even when opponents select their actions in an arbitrary, potentially adaptive fashion, and thus converge to a CCE (Vovk, 1990; Littlestone & Warmuth, 1994; Cesa-Bianchi et al., 1997; Hart & Mas-Colell, 2000; Syrgkanis et al., 2015).

This simple connection between no-regret learning and decentralized convergence to equilibria has been influential in game theory, leading to numerous lines of research including fast rates of convergence to equilibria (Syrgkanis et al., 2015; Chen & Peng, 2020; Daskalakis et al., 2021; Anagnostides et al., 2022), price of anarchy bounds for smooth games (Roughgarden, 2015), and lower bounds on query and communication complexity for equilibrium computation (Fearnley et al., 2013; Rubinstein, 2016; Babichenko & Rubinstein, 2017). Empirically, no-regret algorithms such as regret matching (Hart & Mas-Colell, 2000) and Hedge (Vovk, 1990; Littlestone & Warmuth, 1994; Cesa-Bianchi et al., 1997) have been used to compute equilibria that can achieve state-of-the-art performance in application domains such as Poker (Brown & Sandholm, 2018) and Diplomacy (Bakhtin et al., 2022). Motivated by these successes, we ask whether an analogous theory can be developed for Markov games. In particular:

Are there efficient algorithms
for no-regret learning in Markov games?

Challenges for no-regret learning. In spite of active research effort and many promising pieces of progress (Jin et al., 2021; Song et al., 2022; Mao & Basar, 2021; Daskalakis et al., 2022; Erez et al., 2022), no-regret learning guarantees for Markov games have been elusive. A barrier faced by naive algorithms is that it is intractable to ensure no-regret against an *arbitrary* adversary, both computationally (Bai et al., 2020; Abbasi Yadkori et al., 2013) and statistically (Liu et al., 2022; Kwon et al., 2021; Foster et al., 2022). Fortunately, many of the implications of no-regret learning (in particular, convergence to equilibria) do not require the algorithm to have sublinear regret against an arbitrary adversary, but rather only against other agents who are running the same algorithm independently. This observation has been influential in normal-form games, where the line of work on fast rates of convergence to equilibrium (Syrgkanis et al., 2015; Chen & Peng, 2020; Daskalakis et al., 2021; Anagnostides et al., 2022) holds only in this more restrictive setting. This motivates the following relaxation to our central question.

Problem 1.1. *Is there an efficient algorithm that, when adopted by all agents in a Markov game and run independently, leads to sublinear regret for each individual agent?*

Attempts to address Problem 1.1. Two recent lines of research have made progress toward addressing Problem 1.1 and related questions. In one direction, several recent papers have provided algorithms, including `V-learning` (Jin et al., 2021; Song et al., 2022; Mao & Basar, 2021) and `SPoCMAR` (Daskalakis et al., 2022), that do not achieve no-regret, but can nevertheless compute and then sample from a coarse correlated equilibrium in a Markov game in a (mostly) *decentralized* fashion, with the caveat that they require a shared source of random bits as a mechanism to coordinate. Notably, `V-learning` depends only mildly on the shared randomness: agents first play policies in a fully independent fashion (i.e., without shared randomness) according to a simple learning algorithm for T episodes, and use shared random bits only once learning finishes as part of a post-processing procedure to extract a CCE policy. A question left open by these works, is whether the sequence of policies played by the `V-learning` algorithm in the initial independent phase can itself guarantee each agent sublinear regret.

Most closely related to our work, Erez et al. (2022) recently showed that Problem 1.1 can be solved positively for a restricted setting in which regret for each agent is defined as the maximum gain in value they can achieve by deviating to a fixed *Markov* policy. Markov policies are those whose choice of action depends only on the current state as opposed to the entire history of interaction. This notion of deviation is restrictive because in general, even when the opponent plays a sequence of Markov policies, the best response will be *non-Markov*. In challenging settings that abound in practice, it is standard to consider non-Markov policies (Leibo et al., 2021; Agapiou et al., 2022), since they often achieve higher value than Markov policies; we provide a simple example in Proposition B.1. Thus, while a regret guarantee with respect to the class of Markov policies (as in (Erez et al., 2022)) is certainly interesting, it may be too weak in general, and it is of great interest to understand whether Problem 1.1 can be answered positively in the general setting.[1]

[1]We remark that the `V-learning` and `SPoCMAR` algorithms

We refer the reader to Appendix B.2 for further discussion.

1.2. Our contributions

We resolve Problem 1.1 in the negative, from both a computational and statistical perspective.

Computational hardness. We provide two computational lower bounds (Theorems 1.2 and 1.3) which show that under standard complexity-theoretic assumptions, there is no efficient algorithm that runs for a polynomial number of episodes and guarantees each agent non-trivial ("sublinear") regret when used in tandem by all agents. Both results hold even if the Markov game is explicitly known to the algorithm designer; Theorem 1.3 is stronger and more general, but applies only to 3-player games, while Theorem 1.2 applies to 2-player games, but only for agents restricted to playing Markovian policies.

To state our first result, Theorem 1.2, we define a *product Markov policy* to be a joint policy in which players choose their actions independently according to Markov policies (see Sections 2 and 3 for formal definitions). Note that if all players use independent no-regret algorithms to choose Markov policies at each episode, then their joint play at each round is described by a product Markov policy, since any randomness in each player's policy must be generated independently.

Theorem 1.2 (Informal version of Corollary 3.3)**.** *If $\mathsf{PPAD} \neq \mathsf{P}$, then there is no polynomial-time algorithm that, given the description of a 2-player Markov game, outputs a sequence of joint product Markov policies which guarantees each agent sublinear regret.*

Theorem 1.2 provides a decisive negative resolution to Problem 1.1 under the assumption that $\mathsf{PPAD} \neq \mathsf{P}$,[2] which is standard in the theory of computational complexity (Papadimitriou, 1994).[3] Beyond simply ruling out the existence of fully decentralized no-regret algorithms, it rules out existence of *centralized* algorithms that compute a sequence of product policies for which each agent has sublinear regret, even if such a sequence does not arise naturally as the result of agents independently following some learning algorithm. Salient implications include:

- Theorem 1.2 provides a separation between Markov games and normal-form games, since standard no-regret algorithms for normal-form games i) run in polynomial time and ii) produce sequences of joint product policies that guarantee each agent sublinear regret. Notably, no-regret learning for normal-form games is efficient whenever the number of agents is polynomial, whereas Theorem 1.2 rules out polynomial-time algorithms for as few as two agents.
- A question left open by the work of Jin et al. (2021); Song et al. (2022); Mao & Basar (2021) was whether the sequence of policies played by the `V-learning` algorithm during its independent learning phase can guarantee each agent sublinear regret. Since `V-learning` plays product Markov policies during the independent phase and is computationally efficient, Theorem 1.2 implies that these policies *do not* enjoy sublinear regret (assuming $\mathsf{PPAD} \neq \mathsf{P}$).

Our second result, Theorem 1.3, extends the guarantee of Theorem 1.2 to the more general setting in which agents can select arbitrary, potentially *non-Markovian* policies at each episode. This comes at the cost of only providing hardness for 3-player games as opposed to 2-player games, as well as relying on the slightly stronger complexity-theoretic assumption that $\mathsf{PPAD} \nsubseteq \mathsf{RP}$.[4]

Theorem 1.3 (Informal version of Corollary 4.4)**.** *If $\mathsf{PPAD} \nsubseteq \mathsf{RP}$, then there is no polynomial-time algorithm that, given the description of a 3-player Markov game, outputs a sequence of joint product general policies (i.e., potentially non-Markov) which guarantees each agent sublinear regret.*

Statistical hardness. Theorems 1.2 and 1.3 rely on the widely-believed complexity theoretic assumption that PPAD-complete problems cannot be solved in (randomized) polynomial time. Such a restriction is inherent if we assume that the game is known to the algorithm designer. To avoid complexity-theoretic assumptions, we consider a setting in which the Markov game is *unknown* to the algorithm designer, and algorithms must learn about the game by executing policies ("querying") and observing the resulting sequences of states, actions, and rewards. Our final result, Theorem 1.4, shows *unconditionally* that, for m-player Markov games whose parameters are unknown, any algorithm computing a no-regret sequence as in Theorem 1.3 requires a number of queries that is exponential in m.

Theorem 1.4 (Informal version of Theorem 5.2)**.** *Given query access to a m-player Markov game, no algorithm that makes fewer than $2^{\Omega(m)}$ queries can output a sequence of joint product policies which guarantees each agent sublinear regret.*

Similar to our computational lower bounds, Theorem 1.4 goes far beyond decentralized algorithms, and rules out even centralized algorithms that compute a no-regret sequence by jointly controlling all players. The result provides another

mentioned above do learn equilibria that are robust to deviations to non-Markov policies, though they do not address Problem 1.1 since they do not have sublinear regret.

[2] Technically, the class we are denoting by P, namely of total search problems that have a deterministic polynomial-time algorithm, is sometimes denoted by FP, as it is a search problem. We ignore this distinction.

[3] PPAD is the most well-studied complexity class in algorithmic game theory, and is widely believed to not admit polynomial time algorithms. Notably, the problem of computing a Nash equilibrium for normal-form games with two or more players is PPAD-complete (Daskalakis et al., 2009; Chen et al., 2006; Rubinstein, 2018).

[4] We use RP to denote the class of total search problems for which there exists a polynomial-time randomized algorithm which outputs a solution with probability at least $2/3$, and otherwise outputs "fail".

separation between Markov games and normal-form games, since standard no-regret algorithms for normal-form games can achieve sublinear regret using $\mathrm{poly}(m)$ queries for any m. The $2^{\Omega(m)}$ scaling in the lower bound, which does not rule out query-efficient algorithms when m is constant, is to be expected for an unconditional result: If the game has only polynomially many parameters (which is the case for constant m), one can estimate all of the parameters using standard techniques (Jin et al., 2020), then directly find a no-regret sequence.

Proof techniques: the SPARSECCE problem. Our proofs proceed via establishing lower bounds for a computational problem we refer to as SPARSECCE. In the SPARSECCE problem, the aim is to compute a CCE that can be represented as the mixture of a small number of product policies. See Sections 3 and 4 for detailed proof overview.

Organization. Section 2 presents preliminaries, Sections 3 and 4 provide our computational lower bounds, and Section 5 presents our unconditional lower bounds for multi-player games.

Notation. For $n\in\mathbb{N}$, we write $[n]:=\{1,2,...,n\}$. For a finite set $\mathcal{T}$, $\Delta(\mathcal{T})$ denotes the space of distributions on $\mathcal{T}$. For an element $t\in\mathcal{T}$, $\mathbb{I}_t\in\Delta(\mathcal{T})$ denotes the delta distribution that places probability mass 1 on t. We adopt standard big-oh notation, and write $f=\widetilde{O}(g)$ to denote that $f=O(g\cdot\max\{1,\mathrm{polylog}(g)\})$, with $\Omega(\cdot)$ and $\widetilde{\Omega}(\cdot)$ defined analogously.

2. Preliminaries

This section contains preliminaries necessary to present our main results. We first introduce the Markov game framework (Section 2.1), then provide a brief review of normal-form games (Section 2.3), and finally introduce the concepts of coarse correlated equilibria and regret minimization (Section 2.4).

2.1. Markov games

We consider general-sum Markov games in a finite-horizon, episodic framework. For $m\in\mathbb{N}$, an m-player Markov game $\mathcal{G}$ consists of a tuple $\mathcal{G}=(\mathcal{S},H,(\mathcal{A}_i)_{i\in[m]},\mathbb{P},(R_i)_{i\in[m]},\mu)$, where:

- $\mathcal{S}$ denotes a finite state space and $H\in\mathbb{N}$ denotes a finite time horizon. We write $S:=|\mathcal{S}|$.
- For $i\in[m]$, $\mathcal{A}_i$ denotes a finite action space for agent i. We let $\mathcal{A}:=\prod_{i=1}^m\mathcal{A}_i$ denote the *joint action space* and $\mathcal{A}_{-i}:=\prod_{i'\neq i}\mathcal{A}_{i'}$. We denote joint actions in bold, e.g., $\mathbf{a}=(a_1,...,a_m)\in\mathcal{A}$. We write $A_i:=|\mathcal{A}_i|$ and $A:=|\mathcal{A}|$.
- $\mathbb{P}=(\mathbb{P}_1,\dots,\mathbb{P}_H)$ is the transition kernel, with each $\mathbb{P}_h:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ denoting the kernel for step $h\in[H]$. In particular, $\mathbb{P}_h(s'|s,\mathbf{a})$ is the probability of transitioning to s' from the state s at step h when agents play $\mathbf{a}$.
- For $i\in[m]$ and $h\in[H]$, $R_{i,h}:\mathcal{S}\times\mathcal{A}\to[-1/H,1/H]$ is the reward function for agent i:[5] the reward agent i receives in state s at step h if agents play $\mathbf{a}$ is $R_{i,h}(s,\mathbf{a})$.[6]
- $\mu\in\Delta(\mathcal{S})$ denotes the initial state distribution.

An *episode* in the Markov game proceeds as follows: the initial state s_1 is drawn from the initial state distribution μ. Then, for each $h\leqslant H$, given state s_h, each agent i plays action $a_{i,h}\in\mathcal{A}_i$, and given the joint action profile $\mathbf{a}_h=(a_{1,h},\dots,a_{m,h})$, each agent i receives reward of $r_{i,h}=R_{i,h}(s_h,\mathbf{a}_h)$ and the state of the system transitions to $s_{h+1}\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a}_h)$. We denote the tuple of agents' rewards at each step h by $\mathbf{r}_h=(r_{1,h},...,r_{m,h})$, and refer to the resulting sequence $\tau_H:=(s_1,\mathbf{a}_1,\mathbf{r}_1),...,(s_H,\mathbf{a}_H,\mathbf{r}_H)$ as a *trajectory*. For $h\in[H]$, we define the prefix of the trajectory via $\tau_h:=(s_1,\mathbf{a}_1,\mathbf{r}_1),...,(s_h,a_h,r_h)$.

We use the following notation: for some quantity x (e.g., action, reward, etc.) indexed by agents, i.e., $x=(x_1,...,x_m)$, and an agent $i\in[m]$, we write $x_{-i}=(x_1,\dots,x_{i-1},x_{i+1},\dots,x_m)$ to denote the tuple consisting of all $x_{i'}$ for $i'\neq i$.

2.2. Policies and value functions

We now introduce the notion of policies and value functions for Markov games. Policies are mappings from states (or sequences of states) to actions for the agents. We consider several different types of policies, which play a crucial role in distinguishing the types of equilibria that are tractable and those that are intractable to compute efficiently.

Markov policies. A randomized *Markov policy* for agent i is a sequence $\sigma_i=(\sigma_{i,1},...,\sigma_{i,H})$, where $\sigma_{i,h}:\mathcal{S}\to\Delta(\mathcal{A}_i)$. We denote the space of randomized Markov policies for agent i by Π_i^{markov}. We write $\Pi^{\mathrm{markov}}:=\Pi_1^{\mathrm{markov}}\times\cdots\times\Pi_m^{\mathrm{markov}}$ to denote the space of *product Markov policies*, which are joint policies in which each agent i independently follows a policy in Π_i^{markov}. In particular, a policy $\sigma\in\Pi^{\mathrm{markov}}$ is specified by a collection $\sigma=(\sigma_1,...,\sigma_H)$, where $\sigma_h:\mathcal{S}\to\Delta(\mathcal{A}_1)\times\cdots\times\Delta(\mathcal{A}_m)$. We additionally define $\Pi_{-i}^{\mathrm{markov}}:=\prod_{i'\neq i}\Pi_{i'}^{\mathrm{markov}}$, and for a policy $\sigma\in\Pi^{\mathrm{markov}}$, write σ_{-i} to denote the collection of mappings $\sigma_{-i}=(\sigma_{-i,1},...,\sigma_{-i,H})$, where $\sigma_{-i,h}:\mathcal{S}\to\prod_{i'\neq i}\Delta(\mathcal{A}_{i'})$ denotes the tuple of all but player i's policies.

When the Markov game $\mathcal{G}$ is clear from context, for a policy $\sigma\in\Pi^{\mathrm{markov}}$ we let $\mathbb{P}_\sigma[\cdot]$ denote the law of the trajectory τ when players select actions via $\mathbf{a}_h\sim\sigma(s_h)$, and let $\mathbb{E}_\sigma[\cdot]$ denote the corresponding expectation.

General (non-Markov) policies. In addition to Markov policies, we will consider general *history-dependent* (or, *non-*

[5] We assume that rewards lie in $[-1/H,1/H]$ for notational convenience, as this ensures that the cumulative reward for each episode lies in $[-1,1]$. This assumption is not important to our results.

[6] We restrict our attention to Markov games in which the rewards at each step are a deterministic function of the state and action profile. Since our goal is to prove lower bounds, this is without loss.

Markov) policies, which select actions based on the *entire sequence of states and actions* observed up the current step. To streamline notation, for $i \in [m]$, let $\tau_{i,h} = (s_1, a_{i,1}, r_{i,1}, \ldots, s_h, a_{i,h}, r_{i,h})$ denote the history of agent i's states, actions, and reward up to step h. Let $\mathscr{H}_{i,h} = (\mathcal{S} \times \mathcal{A}_i \times [0,1])^h$ denote the space of all possible histories of agent i up to step h. For $i \in [m]$, a *randomized general (i.e., non-Markov) policy of agent* i is a collection of mappings $\sigma_i = (\sigma_{i,1}, \ldots, \sigma_{i,H})$ where $\sigma_{i,h} : \mathscr{H}_{i,h-1} \times \mathcal{S} \to \Delta(\mathcal{A}_i)$ is a mapping that takes the history observed by agent i up to step $h-1$ and the current state and outputs a distribution over actions for agent i.

We denote by $\Pi_i^{\mathrm{gen,rnd}}$ the space of randomized general policies of agent i, and further write $\Pi^{\mathrm{gen,rnd}} := \Pi_1^{\mathrm{gen,rnd}} \times \cdots \times \Pi_m^{\mathrm{gen,rnd}}$ to denote the space of product general policies; note that $\Pi_i^{\mathrm{markov}} \subset \Pi_i^{\mathrm{gen,rnd}}$ and $\Pi^{\mathrm{markov}} \subset \Pi^{\mathrm{gen,rnd}}$. In particular, a policy $\sigma \in \Pi^{\mathrm{gen,rnd}}$ is specicfied by a collection $(\sigma_{i,h})_{i \in [m], h \in [H]}$, where $\sigma_{i,h} : \mathscr{H}_{i,h-1} \times \mathcal{S} \to \Delta(\mathcal{A}_i)$. When agents play according to a general policy $\sigma \in \Pi^{\mathrm{gen,rnd}}$, at each step h, each agent, given the current state s_h and their history $\tau_{i,h-1} \in \mathscr{H}_{i,h-1}$, chooses to play an action $a_{i,h} \sim \sigma_{i,h}(\tau_{i,h-1}, s_h)$, independently from all other agents. For a policy $\sigma \in \Pi^{\mathrm{gen,rnd}}$, we let $\mathbb{P}_\sigma[\cdot]$ and $\mathbb{E}_\sigma[\cdot]$ denote the law and expectation operator for the trajectory τ when players select actions via $\mathbf{a}_h \sim \sigma(\tau_{h-1}, s_h)$, and write σ_{-i} to denote the collection of policies of all agents but i, i.e., $\sigma_{-i} = (\sigma_{j,h})_{h \in [H], j \in [m] \setminus \{i\}}$.

We will also consider distributions over product randomized general policies, namely elements of $\Delta(\Pi^{\mathrm{gen,rnd}})$.[7] We will refer to elements of $\Delta(\Pi^{\mathrm{gen,rnd}})$ as *distributional policies*. To play a distributional policy $P \in \Delta(\Pi^{\mathrm{gen,rnd}})$, agents draw a randomized policy $\sigma \sim P$ (so that $\sigma \in \Pi^{\mathrm{gen,rnd}}$) and then play σ.

Value functions. For a general policy $\sigma \in \Pi^{\mathrm{gen,rnd}}$, we define the value function for agent $i \in [m]$ as $V_i^\sigma := \mathbb{E}_\sigma \left[\sum_{h=1}^H R_{i,h}(s_h, \mathbf{a}_h) \mid s_1 \sim \mu \right]$; this represents the expected reward that agent i receives when each agent chooses their actions via $a_{i,h} \sim \sigma_h(\tau_{i,h-1}, s_h)$. For a distributional policy $P \in \Delta(\Pi^{\mathrm{gen,rnd}})$, we extend this notation by defining $V_i^P := \mathbb{E}_{\sigma \sim P}[V_i^\sigma]$.

2.3. Normal-form games

To motivate the solution concepts we consider for Markov games, let us first revisit the notion of normal-form games, which may be interpreted as Markov games with a single state. For $m, n \in \mathbb{N}$, an *m-player n-action normal-form game* G is specified by a tuple of m *reward tensors* $M_1, \ldots, M_m \in [0,1]^{n \times \cdots \times n}$, where each tensor is of order m (i.e., has n^m entries). We will write $G = (M_1, \ldots, M_m)$. We assume for simplicity that each player has the same number n of actions, and identify each player's action space with $[n]$. Then an an action profile is specified by $\mathbf{a} \in [n]^m$; if each player acts according to $\mathbf{a}$, then the reward for player $i \in [m]$ is given by $(M_i)_{\mathbf{a}} \in [0,1]$. Our hardness results will use the standard notion of *Nash equilibrium* in normal-form games. We define the *m-player (n,ϵ)-*NASH *problem* to be the problem of computing an ϵ-approximate Nash equilibrium of a given m-player n-action normal-form game. (See Definition C.2 for a formal definition of ϵ-Nash equilibrium.) A celebrated result is that Nash equilibria are PPAD-hard to approximate, i.e., the 2-player (n, n^{-c})-NASH problem is PPAD-hard for any constant $c > 0$ (Daskalakis et al., 2009; Chen et al., 2006). We refer the reader to Section C.2 for further background on these concepts.

2.4. Markov games: Equilibria and no-regret

We now turn our focus back to Markov games, and introduce the main solution concepts we consider, as well as the notion of no-regret. Since computing Nash equilibria is intractable even for normal-form games, much of the work on efficient equilibrium computation has focused on alternative notions of equilibrium, notably *coarse correlated equilibria*.

For a distributional policy $P \in \Delta(\Pi^{\mathrm{gen,rnd}})$ and a randomized policy $\sigma_i' \in \Pi_i^{\mathrm{gen,rnd}}$ of player i, we let $\sigma_i' \times P_{-i} \in \Delta(\Pi^{\mathrm{gen,rnd}})$ denote the distributional policy which is given by the distribution of $(\sigma_i', \sigma_{-i}) \in \Pi^{\mathrm{gen,rnd}}$ for $\sigma \sim P$ (and σ_{-i} denotes the marginal of σ on all players but i). For $\sigma \in \Pi^{\mathrm{gen,rnd}}$, we write $\sigma_i' \times \sigma_{-i}$ to denote the policy given by $(\sigma_i', \sigma_{-i}) \in \Pi^{\mathrm{gen,rnd}}$. Let us fix a Markov game $\mathcal{G}$, which in particular determines the players' value functions V_i^σ.

Definition 2.1 (Coarse correlated equilibrium)**.** For $\epsilon > 0$, a distributional policy $P \in \Delta(\Pi^{\mathrm{gen,rnd}})$ is defined to be an *ϵ-coarse correlated equilibrium (CCE)* if for each $i \in [m]$, it holds that $\max_{\sigma_i' \in \Pi_i^{\mathrm{gen,rnd}}} V_i^{\sigma_i' \times P_{-i}} - V_i^P \leqslant \epsilon$.

Coarse correlated equilibria can be computed efficiently for both normal-form games and Markov games, and are fundamentally connected to the notion of no-regret and independent learning, which we now introduce.

Regret. For a policy $\sigma \in \Pi^{\mathrm{gen,rnd}}$, we denote the distributional policy which puts all its mass on σ by $\mathbb{I}_\sigma \in \Delta(\Pi^{\mathrm{gen,rnd}})$. Thus $\frac{1}{T} \sum_{t=1}^T \mathbb{I}_{\sigma^{(t)}} \in \Delta(\Pi^{\mathrm{gen,rnd}})$ denotes the distributional policy which randomizes uniformly over the $\sigma^{(t)}$. We define *regret* as follows.

Definition 2.2 (Regret)**.** Consider a sequence of policies $\sigma^{(1)}, \ldots, \sigma^{(T)} \in \Pi^{\mathrm{gen,rnd}}$. For $i \in [m]$, the *regret of agent* i with respect to this sequence is defined as:

$$\mathrm{Reg}_{i,T}(\sigma^{(1)}, \ldots, \sigma^{(T)}) = \max_{\sigma_i \in \Pi_i^{\mathrm{gen,rnd}}} \sum_{t=1}^T V_i^{\sigma_i \times \sigma_{-i}^{(t)}} - V_i^{\sigma^{(t)}}. \quad (1)$$

It is immediate from the above definitions that a sequence of policies $\sigma^{(1)}, \ldots, \sigma^{(T)} \in \Pi^{\mathrm{gen,rnd}}$ satisfies $\mathrm{Reg}_{i,t}(\sigma^{(1)}, \ldots, \sigma^{(T)}) \leqslant$

[7] When $\mathcal{T}$ is not a finite set, we take $\Delta(\mathcal{T})$ to be the set of Radon probability measures over $\mathcal{T}$ equipped with the Borel σ-algebra.

$\epsilon \cdot T$ if and only if the distributional policy $\overline{\sigma} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\sigma^{(t)}}$ is an ϵ-CCE (stated formally in Fact C.1 in the appendix).

No-regret learning. A standard approach to decentralized equilibrium computation, which exploits Fact C.1, is to select $\sigma^{(1)}, \dots, \sigma^{(T)} \in \Pi^{\mathrm{gen,rnd}}$ using independent *no-regret learning* algorithms. A no-regret learning algorithm for player i selects $\sigma_i^{(t)} \in \Pi_i^{\mathrm{gen,rnd}}$ based on the realized trajectories $\tau_{i,H}^{(1)},\dots,\tau_{i,H}^{(t-1)} \in \mathscr{H}_{i,H}$ that player i observes over the course of play,[8] but with no knowledge of $\sigma_{-i}^{(t)}$, so as to ensure that no-regret is achieved: $\mathrm{Reg}_{i,T}(\sigma^{(1)},\dots,\sigma^{(T)}) \leq \epsilon \cdot T$. If each player i uses their own, independent no-regret learning algorithm, this approach yields *product policies* $\sigma^{(t)} = \sigma_1^{(t)} \times \cdots \times \sigma_m^{(t)}$, and the uniform average of the $\sigma^{(t)}$ yields a CCE as long as all of the players can keep their regret small.[9]

For the special case of normal-form games, there are several efficient algorithms, which—when run independently—ensure that each player's regret after T episodes is bounded above by $O(\sqrt{T})$ (that is $\epsilon = O(1/\sqrt{T})$), even when the other players' actions are chosen *adversarially*.

3. Lower bound for Markovian algorithms

In this section we prove Theorem 1.2 (restated formally below as Theorem 3.2), establishing that in two-player Markov games, there is no computationally efficient algorithm that computes a sequence $\sigma^{(1)},\dots,\sigma^{(T)}$ of product Markov policies so that each player has small regret under this sequence. This section serves as a warm-up for our results in Section 4, which remove the assumption that $\sigma^{(1)},\dots,\sigma^{(T)}$ are Markovian.

3.1. SparseMarkovCCE and computational model

As discussed in the introduction, our lower bounds for no-regret learning are a consequence of lower bounds for the SparseCCE problem. In what follows, we formalize this problem (specifically, the Markovian variant, which we refer to as SparseMarkovCCE), as well as our computational model.

Description length for Markov games (constant m). Given a Markov game $\mathcal{G}$, we let $\beta(\mathcal{G})$ denote the maximum number of bits needed to describe any of the rewards $R_{i,h}(s,\mathbf{a})$ or transition probabilities $\mathbb{P}_h(s'|s,\mathbf{a})$ in binary.[10] We define $|\mathcal{G}| := \max\{S, \max_{i\in[m]} A_i, H, \beta(\mathcal{G})\}$. The interpretation of $|\mathcal{G}|$ depends on the number of players m: If m is a constant (as will be the case in the current section and Section 4), then $|\mathcal{G}|$ should be interpreted as the description length of the game $\mathcal{G}$, up to polynomial factors. In particular, for constant m, the game $\mathcal{G}$ can be described using $|\mathcal{G}|^{O(1)}$ bits. In Section 5, we discuss the interpretation of $|\mathcal{G}|$ when m is large.

The SparseMarkovCCE problem. From Fact C.1, we know that the problem of computing a sequence $\sigma^{(1)},\dots,\sigma^{(T)}$ of joint product Markov policies for which each player has at most $\epsilon \cdot T$ regret is equivalent to computing a sequence $\sigma^{(1)},\dots,\sigma^{(T)}$ for which the uniform mixture forms an ϵ-approximate CCE. We define (T,ϵ)-SparseMarkovCCE as the computational problem of computing such a CCE directly.

Definition 3.1 (SparseMarkovCCE problem)**.** For an m-player Markov game $\mathcal{G}$ and parameters $T \in \mathbb{N}$ and $\epsilon > 0$ (which may depend on the size of the game $\mathcal{G}$), (T,ϵ)-SparseMarkovCCE is the problem of finding a sequence $\sigma^{(1)},\dots,\sigma^{(T)}$, with each $\sigma^{(t)} \in \Pi^{\mathrm{markov}}$, such that the distributional policy $\overline{\sigma} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\sigma^{(t)}} \in \Delta(\Pi^{\mathrm{gen,rnd}})$ is an ϵ-CCE of $\mathcal{G}$ (or equivalently, such that for all $i \in [m]$, $\mathrm{Reg}_{i,T}(\sigma^{(1)},\dots,\sigma^{(T)}) \leq \epsilon \cdot T$).

Decentralized learning algorithms naturally lead to solutions to the SparseMarkovCCE problem. In particular, consider any decentralized protocol which runs for T episodes, where at each timestep $t \in [T]$, each player $i \in [m]$ chooses a Markov policy $\sigma_i^{(t)} \in \Pi_i^{\mathrm{markov}}$ to play, without knowledge of the other players' policies $\sigma_{-i}^{(t)}$ (but possibly using the history); any strategy in which players independently run online learning algorithms falls under this protocol. If each player experiences overall regret at most $\epsilon \cdot T$, then the sequence $\sigma^{(1)},\dots,\sigma^{(T)}$ is a solution to the (T,ϵ)-SparseMarkovCCE problem. However, one might expect the (T,ϵ)-SparseMarkovCCE problem to be much easier than decentralized learning, since it allows for algorithms that produce $(\sigma^{(1)},\dots,\sigma^{(T)})$ satisfying the constraints of Definition 3.1 in a centralized manner. The main result of this section, Theorem 3.2, rules out the existence of *any* efficient algorithms, including centralized ones, that solve the SparseMarkovCCE problem.

Before moving on, let us give a sense for what sort of scaling one should expect for the parameters T and ϵ in the (T,ϵ)-SparseMarkovCCE problem. First, we note that there always exists a solution to the $(1,0)$-SparseMarkovCCE problem in a Markov game, which is given by a (Markov) Nash equilibrium of the game; of course, Nash equilibria are intractable to compute in general.[11] For the special case of normal-form games (where there is only a single state, and $H = 1$), no-regret learning (e.g., Hedge) yields a computationally efficient solution to the $(T,\widetilde{O}(1/\sqrt{T}))$-SparseMarkovCCE problem, where the $\widetilde{O}(\cdot)$ hides a $\max_i \log|A_i|$ factor. Refined convergence guarantees of Daskalakis et al. (2021); Anagnostides et al. (2022) improve upon this result, and yield an efficient solution

[8] An alternative model allows for player i to have knowledge of the previous joint policies $\sigma^{(1)},\dots,\sigma^{(t-1)}$, when selecting $\sigma_i^{(t)}$.

[9] In Appendix B, we discuss the implications of relaxing the stipulation that $\sigma^{(t)}$ be product policies (for example, by allowing the use of shared randomness, as in `V-learning`). In short, allowing $\sigma^{(t)}$ to be non-product essentially trivializes the problem.

[10] We emphasize that $\beta(\mathcal{G})$ is defined as the maximum number of bits required by any particular $(s,\mathbf{a})$ pair, not the total number of bits required for *all* $(s,\mathbf{a})$ pairs.

[11] Such a Nash equilibrium can be seen to exist by using backwards induction to specify the player's joint distribution of play at each state at steps $H, H-1, \dots, 1$.

to the $(T,\widetilde{O}(1/T))$-SPARSEMARKOVCCE problem.

3.2. Main result

Theorem 3.2. *There is a constant $C_0>1$ so that the following holds. Let $n\in\mathbb{N}$ be given, and let $T\in\mathbb{N}$ and $\epsilon>0$ satisfy $T<\exp(\epsilon^2\cdot n^{1/2}/2^5)$. Suppose there is an algorithm that, given the description of any 2-player Markov game $\mathcal{G}$ with $|\mathcal{G}|\leqslant n$, solves the (T,ϵ)-SPARSEMARKOVCCE problem in time U, for some $U\in\mathbb{N}$. Then, for each $n\in\mathbb{N}$, the 2-player $(\lfloor n^{1/2}\rfloor,4\cdot\epsilon)$-NASH problem (Definition C.2) can be solved in time $(nTU)^{C_0}$.*

We emphasize that the range $T<\exp(n^{O(1)})$ ruled out by Theorem 3.2 is the most natural parameter regime, since the runtime of any decentralized algorithm which runs for T episodes and produces a solution to the SPARSEMARKOVCCE problem is at least linear in T. Using that 2-player (n,ϵ)-NASH is PPAD-complete for $\epsilon=n^{-c}$ (for any $c>0$) (Daskalakis et al., 2009; Chen et al., 2006; Rubinstein, 2018), we obtain the following corollary.

Corollary 3.3 (SPARSEMARKOVCCE is PPAD-complete)**.** *For any constant $C>4$, if there is an algorithm which, given the description of a 2-player Markov game $\mathcal{G}$, solves the $(|\mathcal{G}|^C,|\mathcal{G}|^{-\frac{1}{C}})$-SPARSEMARKOVCCE problem in time $\mathrm{poly}(|\mathcal{G}|)$, then PPAD$=$P.*

The condition $C>4$ in Corollary 3.3 is set to ensure that $|\mathcal{G}|^C<\exp(|\mathcal{G}|^{-2/C}\cdot\sqrt{|\mathcal{G}|}/2^6)$ for sufficiently large $|\mathcal{G}|$, so as to satisfy the condition of Theorem 3.2. Corollary 3.3 rules out the existence of a polynomial-time algorithm that solves the SPARSEMARKOVCCE problem with accuracy ϵ polynomially small and T polynomially large in $|\mathcal{G}|$.

Proof overview. The proof of Theorem 3.2 is based on a reduction, which shows that any algorithm that efficiently solves the (T,ϵ)-SPARSEMARKOVCCE problem, for T not too large, can be used to efficiently compute an approximate Nash equilibrium of any given normal-form game. In particular, fix $n_0\in\mathbb{N}$, and let a 2-player normal form game G with n_0 actions be given. We construct a Markov game $\mathcal{G}=\mathcal{G}(G)$ with horizon $H=n_0$ and action sets identical to those of the game G, i.e., $\mathcal{A}_1=\mathcal{A}_2=[n_0]$. The state space of $\mathcal{G}$ consists n_0^2 states, which are indexed by joint action profiles; the transitions are defined so that the value of the state at step h encodes the action profile taken by the agents at step $h-1$.[12] At each state of $\mathcal{G}$, the reward functions are given by the payoff matrices of G, scaled down by a factor of $1/H$ (which ensures that the rewards received at each step belong to $[0,1/H]$). In particular, the rewards and transitions out of a given state do not depend on the identity of the state, and so $\mathcal{G}$ can be thought of as a repeated game where G is played H times. The formal definition of $\mathcal{G}$ is given in Definition D.3.

Fix any algorithm for the SPARSEMARKOVCCE problem, and recall that for each step h and state s for $\mathcal{G}$, $\sigma_h^{(t)}(s)\in\Delta(\mathcal{A}_1)\times\Delta(\mathcal{A}_2)$ denotes the joint action distribution taken in s at step h for the sequence of $\sigma^{(1)},\ldots,\sigma^{(T)}$ produced by the algorithm. The bulk of the proof of Theorem 3.2 consists of proving a key technical result, Lemma D.4, which states that if $\sigma^{(1)},\ldots,\sigma^{(T)}$ indeed solves (T,ϵ)-SPARSEMARKOVCCE, then there exists some tuple (h,s,t) such that $\sigma_h^{(t)}(s)$ is an approximate Nash equilibrium for G. With this established, it follows that we can find a Nash equilibrium efficiently by simply trying all HST choices for (h,s,t).

To prove Lemma D.4, we reason as follows. Assume that $\overline{\sigma}:=\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{\sigma^{(t)}}\in\Delta(\Pi^{\mathrm{gen,rnd}})$ is an ϵ-CCE. If, by contradiction, none of the distributions $\left\{\sigma_h^{(t)}(s)\right\}_{h\in[H],s\in\mathcal{S},t\in[T]}$ are approximate Nash equilibria for G, then it must be the case that for each t, one of the players has a profitable deviation in G with respect to the product strategy $\sigma_h^{(t)}(s)$, at least for a constant fraction of the tuples (s,h). We will argue that if this were to be the case, it would imply that there exists a non-Markov deviation policy for at least one player i in Definition 2.1, meaning that $\overline{\sigma}$ is not in fact an ϵ-CCE.

To sketch the idea, recall that to draw a trajectory from $\overline{\sigma}$, we first draw a random index $t^\star\sim[T]$ uniformly at random, and then execute $\sigma^{(t^\star)}$ for an episode. We will show (roughly) that for each player i, it is possible to compute a non-Markov deviation policy $\pi_i^\dagger$ which, under the draw of a trajectory from $\overline{\sigma}$, can "infer" the value of the index $t^\star$ within the first few steps of the episode. Then policy $\pi_i^\dagger$ then, at each state s and step h after the first few steps, play a best response to their opponent's portion of the strategy $\sigma_h^{(t^\star)}(s)$. If, for each possible value of $t^\star$, none of the distributions $\sigma_h^{(t^\star)}(s)$ are approximate Nash equilibria of G, this means that at least one of the players i can significantly increase their value in $\mathcal{G}$ over that of $\overline{\sigma}$ by playing $\pi_i^\dagger$, which contradicts the assumption that $\overline{\sigma}$ is an ϵ-CCE.

It remains to explain how we can construct a non-Markov policy $\pi_i^\dagger$ which "infers" the value of $t^\star$. Unfortunately, exactly inferring the value of $t^\star$ in the fashion described above is impossible: for instance, if there are $t_1\neq t_2$ so that $\sigma^{(t_1)}=\sigma^{(t_2)}$, then clearly it is impossible to distinguish between the cases $t^\star=t_1$ and $t^\star=t_2$. Nevertheless, by using the fact that each player observes the full joint action profile played at each step h, we can construct a non-Markov policy which employs *Vovk's aggregating algorithm* for online density estimation (Vovk, 1990; Cesa-Bianchi & Lugosi, 2006) in order to compute a distribution which is *close* to $\sigma_h^{(t^\star)}(s)$ for most $h\in[H]$.[13] This guarantee is stated formally in an abstract setting in Proposition D.2, and is instantiated in the proof of Theorem 3.2 in ((5)). As we show in Section D.2, approximating $\sigma_h^{(t^\star)}(s)$ as we have described is sufficient to carry out the reasoning from the previous paragraph.

[12] For technical reasons, this only is the case for even values of h; we discuss further details in the full proof in Section D.2.

[13] Vovk's aggregating algorithm is essentially the exponential weights algorithm with the logarithmic loss. A detailed background for the algorithm is provided in Section D.1.

4. Lower bound for non-Markov algorithms

In this section, we prove Theorem 1.3 (restated formally below as Theorem 4.3), which strengthens Theorem 3.2 by allowing the sequence $\sigma^{(1)},\dots,\sigma^{(T)}$ of product policies to be non-Markovian. This additional strength comes at the cost of our lower bound only applying to *3-player* Markov games (as opposed to Theorem 3.2, which applied to 2-player games).

4.1. SPARSECCE problem and computational model

To formalize the computational model for the SPARSECCE problem, we must first describe how the non-Markov product policies $\sigma^{(t)} = (\sigma_1^{(t)},\dots,\sigma_m^{(t)})$ are represented. Recall that a non-Markov policy $\sigma_i^{(t)} \in \Pi_i^{\text{gen,rnd}}$ is, by definition, a mapping from agent i's history and current state to a distribution over their next action. Since there are exponentially many possible histories, it is information-theoretically impossible to express an arbitrary policy in $\Pi_i^{\text{gen,rnd}}$ with polynomially many bits. As our focus is on computing a sequence of such policies $\sigma^{(t)}$ in polynomial time, certainly a prerequisite is that $\sigma^{(t)}$ can be expressed in polynomial space. Thus, we adopt the representational assumption, stated formally in Definition 4.1, that each of the policies $\sigma_i^{(t)} \in \Pi_i^{\text{gen,rnd}}$ is described by a bounded-size circuit that can compute the conditional distribution of each next action given the history. This assumption is satisfied by essentially all empirical and theoretical work concerning non-Markov policies (e.g., (Leibo et al., 2021; Agapiou et al., 2022; Jin et al., 2021; Song et al., 2022)).

Definition 4.1 (Computable policy)**.** Given a m-player Markov game $\mathcal{G}$ and $N \in \mathbb{N}$, we say that a policy $\sigma_i \in \Pi_i^{\text{gen,rnd}}$ is N-*computable* if for each $h \in [H]$, there is a circuit of size N that,[14] on input $(\tau_{i,h-1}, s) \in \mathscr{H}_{i,h-1} \times \mathcal{S}$, outputs the distribution $\sigma_i(\tau_{i,h-1}, s) \in \Delta(\mathcal{A}_i)$. A policy $\sigma = (\sigma_1,\dots,\sigma_m) \in \Pi^{\text{gen,rnd}}$ is N-computable if each constituent policy σ_i is.

Our lower bound applies to algorithms that produce sequences $\sigma^{(1)},\dots,\sigma^{(T)}$ for which each $\sigma^{(t)}$ is N-computable, where the value N is taken to be polynomial in the description length of the game $\mathcal{G}$. For example, Markov policies whose probabilities can be expressed with β bits are $O(HSA_i\beta)$-computable for each player i, since one can simply store each of the probabilities $\sigma_{i,h}(s_h, a_{i,h})$, each of which takes β bits to represent.

The SPARSECCE problem. SPARSECCE is the problem of computing a sequence of non-Markov product policies $\sigma^{(1)},\dots,\sigma^{(T)}$ such that the uniform mixture forms an ϵ-approximate CCE. The problem generalizes SPARSEMARKOVCCE (Definition 3.1) by relaxing the condition that the policies $\sigma^{(t)}$ be Markov.

[14] For concreteness, we suppose that "circuit" means "boolean circuit" as in Definition 6.1 of (Arora & Barak, 2006), where probabilities are represented in binary. The precise model of computation we use does not matter, though, and we could equally assume that the policies σ_i may be computed by Turing machines that terminate after N steps.

Definition 4.2 (SPARSECCE Problem)**.** For an m-player Markov game $\mathcal{G}$ and parameters $T, N \in \mathbb{N}$ and $\epsilon > 0$ (which may depend on the size of the game $\mathcal{G}$), (T,ϵ,N)-SPARSECCE is the problem of finding a sequence $\sigma^{(1)},\dots,\sigma^{(T)} \in \Pi^{\text{gen,rnd}}$, with each $\sigma^{(t)}$ being N-computable, such that the distributional policy $\overline{\sigma} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\sigma^{(t)}} \in \Delta(\Pi^{\text{gen,rnd}})$ is an ϵ-CCE for $\mathcal{G}$ (equivalently, such that for all $i \in [m]$, $\text{Reg}_{i,T}(\sigma^{(1)},\dots,\sigma^{(T)}) \leqslant \epsilon \cdot T$).

4.2. Main result

Our main theorem for this section, Theorem 4.3, shows that for appropriate values of T, ϵ, and N, solving the (T,ϵ,N)-SPARSECCE problem is at least as hard as computing Nash equilibria in normal-form games.

Theorem 4.3. *Fix $n \in \mathbb{N}$, and let $T, N \in \mathbb{N}$, and $\epsilon > 0$ satisfy $1 < T < \exp\left(\frac{\epsilon^2 \cdot n}{16}\right)$. Suppose there exists an algorithm that, given the description of any 3-player Markov game $\mathcal{G}$ with $|\mathcal{G}| \leqslant n$, solves the (T,ϵ,N)-SPARSECCE problem in time U, for some $U \in \mathbb{N}$. Then, for any $\delta > 0$, the 2-player $(\lfloor n/2 \rfloor, 50\epsilon)$-NASH problem can be solved in randomized time $(nTNU\log(1/\delta)/\epsilon)^{C_0}$ with failure probability δ, where $C_0 > 0$ is an absolute constant.*

By analogy to Corollary 3.3, we obtain the following immediate consequence.

Corollary 4.4 (SPARSECCE is hard under $\mathsf{PPAD} \nsubseteq \mathsf{RP}$)**.** *For any $C > 4$, if there is an algorithm which, given the description of a 3-player Markov game $\mathcal{G}$, solves the $(|\mathcal{G}|^C, |\mathcal{G}|^{-\frac{1}{C}}, |\mathcal{G}|^C)$-*SPARSECCE *problem in time* $\text{poly}(|\mathcal{G}|)$*, then $\mathsf{PPAD} \subseteq \mathsf{RP}$.*

Proof overview for Theorem 4.3. The proof of Theorem 4.3 has a similar high-level structure to that of Theorem 3.2: given an m-player normal-form G, we define an $(m+1)$-player Markov game $\mathcal{G} = \mathcal{G}(G)$ which has $n_0 := \lfloor n/m \rfloor$ actions per player and horizon $H \approx n_0$. The key difference in the proof of Theorem 4.3 is the structure of the players' reward functions. To motivate this difference and the addition of an $(m+1)$-th player, we explain why the proof of Theorem 3.2 fails to extend: a sequence $\sigma^{(1)},\dots,\sigma^{(T)}$ can hypothetically solve the SPARSECCE problem by attempting to punish any one player's deviation policy, and thus avoid having to compute a Nash equilibrium of G. In particular, if player i plays according to the policy $\pi_i^{\dagger}$ that we described in Section 3.2, then other players $j \neq i$ can use the non-Markov property of $\sigma_j^{(t)}$ to adjust their choice of actions in later rounds to decrease player i's value.

This behavior is reminiscent of "tit-for-tat" strategies which are used to establish the *folk theorem* in the theory of repeated games (Maskin & Fudenberg, 1986). The folk theorem describes how Nash equilibria are more numerous in repeated games than in single-shot normal form games. As it turns out, the folk theorem does not yield to worst-case speedups in repeated games, when the number of players is at least 3. Indeed, Borgs et al. (2008) gave an "anti-folk theorem", showing that computing Nash equilibria in $(m+1)$-player

repeated games is PPAD-hard for $m \geqslant 2$, via a reduction to m-player normal-form games. We adapt their reduction to our setting: roughly speaking, this approach adds an $(m+1)$-th player whose actions represent potential deviations for each of the m players. The structure of the rewards ensures that if $\overline{\sigma} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\sigma^{(t)}}$ is an ϵ-CCE, then for some policy $\pi^{\dagger}_{m+1}$ of the $(m+1)$-th player, the first m players will play an approximate Nash of G with constant probability, under a trajectory drawn from the joint policy $\overline{\sigma}_{-(m+1)} \times \pi^{\dagger}_{m+1}$. Thus, in order to efficiently find a Nash (see Algorithm 2), we need to simulate the policy $\overline{\sigma}_{-(m+1)} \times \pi^{\dagger}_{m+1}$, which involves running Vovk's algorithm. This approach is in contrast to the proof of Theorem 3.2, which used Vovk's algorithm as an ingredient in the proof but not in the Nash computation algorithm.

Two-player games. One intriguing question we leave open is whether the SPARSECCE problem remains hard for two-player Markov games. Interestingly, as shown by Littman & Stone (2005), there is a polynomial time algorithm to find an exact Nash equilibrium for the special case of repeated two-player normal-form games. Though their result only applies in the infinite-horizon setting, it is possible to extend their results to the finite-horizon setting, which rules out naive approaches to extend the proof of Theorem 4.3 and Corollary 4.4 to two players.

5. Multi-player games: lower bounds

In this section we present Theorem 1.4 (restated formally below as Theorem 5.2), which gives a statistical lower bound for the SPARSECCE problem. The lower bound applies to any algorithm, regardless of computational cost, that accesses the underlying Markov game through a *generative model*.

Definition 5.1 (Generative model)**.** For an m-player Markov game $\mathcal{G} = (\mathcal{S}, H, (\mathcal{A}_i)_{i\in[m]}, \mathbb{P}, (R_i)_{i\in[m]}, \mu)$, a *generative model oracle* is defined as follows: given a query described by a tuple $(h, s, \mathbf{a}) \in [H] \times \mathcal{S} \times \mathcal{A}$, the oracle returns the distribution $\mathbb{P}_h(\cdot|s,\mathbf{a}) \in \Delta(\mathcal{S})$ and the tuple of rewards $(R_{i,h}(s,\mathbf{a}))_{i\in[m]}$.

From the perspective of lower bounds, the assumption that the algorithm has access to a generative model is quite reasonable, as it encompasses most standard access models in RL, including the online access model, in which the algorithm repeatedly queries a policy and observes a trajectory drawn from it, as well as the *local access generative model* used in from (Yin et al., 2022; Weisz et al., 2021). We remark that it is slightly more standard to assume that queries to the generative model only return a *sample* from the distribution $\mathbb{P}_h(\cdot|s,\mathbf{a})$ as opposed to the distribution itself (Kakade, 2003; Kearns et al., 1999), but since our goal is to prove lower bounds, the notion in Definition 5.1 only makes our results stronger.

To state our main result, we recall the definition $|\mathcal{G}| = \max\{S, \max_{i\in[m]} A_i, H, \beta(\mathcal{G})\}$. In the present section, we consider the setting where the number of players m is large. Here, $|\mathcal{G}|$ does not necessarily correspond to the description length for $\mathcal{G}$, and should be interpreted, roughly speaking, as a measure of the description complexity of $\mathcal{G}$ $|\mathcal{G}|$ with respect to *decentralized* learning algorithms. In particular, from the perspective of an individual agent implementing a decentralized learning algorithm, their sample complexity should depend only on the size of their *individual action set* (as well as the global parameters $S, H, \beta(\mathcal{G})$), as opposed to the size of the *joint action set*, which grows exponentially in m; the former is captured by $|\mathcal{G}|$, while the latter is not. Indeed, a key advantage shared by much prior work on decentralized RL (Jin et al., 2021; Song et al., 2022; Mao & Basar, 2021; Daskalakis et al., 2022) is their avoidance of the *curse of multi-agents*, which describes the situation where an algorithm has sample and computational costs that scale exponentially in m.

Our main result for this section, Theorem 5.2, states that for m-player Markov games, exponentially many generative model queries (in m) are necessary to produce a solution to the (T,ϵ,N)-SPARSECCE problem, unless T is exponential in m.

Theorem 5.2. *Let $m \geqslant 2$ be given. There are constants $c, \epsilon > 0$ so that the following holds. Suppose there is an algorithm $\mathscr{B}$ which, given access to a generative model for a $(m+1)$-player Markov game $\mathcal{G}$ with $|\mathcal{G}| \leqslant 2m^6$, solves the $(T,\epsilon/(10m),N)$-SPARSECCE problem for $\mathcal{G}$ for some T satisfying $1 < T < \exp(cm)$, and* any *$N \in \mathbb{N}$. Then $\mathscr{B}$ must make at least $2^{\Omega(m)}$ queries to the generative model.*

Theorem 5.2 establishes that there are m-player Markov games, where the number of states, actions per player, and horizon are bounded by $\mathrm{poly}(m)$, but any algorithm with regret $o(T/m)$ must make $2^{\Omega(m)}$ queries (via Fact C.1). In particular, if there are $\mathrm{poly}(m)$ queries per episode, as is standard in the online simulator model where a trajectory is drawn from the policy $\sigma^{(t)}$ at each episode $t \in [T]$, then $T > 2^{\Omega(m)}$ episodes are required to have regret $o(T/m)$. This is in stark contrast to the setting of normal-form games, where even for the case of bandit feedback (which is a special case of the generative model setting), standard no-regret algorithms have the property that each player's regret scales as $\widetilde{O}(\sqrt{Tn})$ (i.e., independently of m), where n denotes the number of actions per player (Lattimore & Szepesvári, 2020). As with our computational lower bounds, Theorem 5.2 is not limited to decentralized algorithms, and also rules out *centralized* algorithms which have access to a generative model.

Acknowledgements

This work was performed in part while Noah Golowich was an intern at Microsoft Research. Noah Golowich is supported at MIT by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. Sham Kakade acknowledges funding from the Office of Naval Research under award N00014-22-1-2377 and the National Science Foundation Grant under award #CCF-2212841.

References

Abbasi Yadkori, Y., Bartlett, P. L., Kanade, V., Seldin, Y., and Szepesvari, C. Online learning in markov decision processes with adversarially chosen transition probability distributions. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/4f284803bd0966cc24fa8683a34afc6e-Paper.pdf.

Agapiou, J. P., Vezhnevets, A. S., Duéñez-Guzmán, E. A., Matyas, J., Mao, Y., Sunehag, P., Köster, R., Madhushani, U., Kopparapu, K., Comanescu, R., Strouse, D., Johanson, M. B., Singh, S., Haas, J., Mordatch, I., Mobbs, D., and Leibo, J. Z. Melting pot 2.0, 2022. URL https://arxiv.org/abs/2211.13746.

Anagnostides, I., Farina, G., Kroer, C., Lee, C.-W., Luo, H., and Sandholm, T. Uncoupled learning dynamics with $o(\log t)$ swap regret in multiplayer games. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=CZwh1XdAhNv.

Arora, S. and Barak, B. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2006. ISBN 978-0-521-42426-4.

Babichenko, Y. Query complexity of approximate nash equilibria. *J. ACM*, 63(4), oct 2016. ISSN 0004-5411.

Babichenko, Y. and Rubinstein, A. Communication complexity of approximate nash equilibria. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pp. 878–889, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345286. doi: 10.1145/3055399.3055407. URL https://doi.org/10.1145/3055399.3055407.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., Rowe, D., Shi, W., Spisak, J., Wei, A., Wu, D., Zhang, H., and Zijlstra, M. Human-level play in the game of ¡i¿diplomacy¡/i¿ by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/abs/10.1126/science.ade9097.

Blackwell, D. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

Borgs, C., Chayes, J., Immorlica, N., Kalai, A. T., Mirrokni, V., and Papadimitriou, C. The myth of the folk theorem. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 365–372, 2008.

Brown, G. W. Some notes on computation of games solutions. 1949.

Brown, N. and Sandholm, T. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. doi: 10.1126/science.aao1733. URL https://www.science.org/doi/abs/10.1126/science.aao1733.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., and Warmuth, M. K. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

Chen, X. and Peng, B. Hedging in games: Faster convergence of external and swap regrets. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18990–18999. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/db346ccb62d491029b590bbbf0f5c412-Paper.pdf.

Chen, X., Deng, X., and Teng, S.-h. Computing nash equilibria: Approximation and smoothed complexity. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 603–612, 2006.

Chen, X., Cheng, Y., and Tang, B. Well-Supported vs. Approximate Nash Equilibria: Query Complexity of Large Games. In Papadimitriou, C. H. (ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 57:1–57:9, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.57. URL http://drops.dagstuhl.de/opus/volltexte/2017/8163.

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

Daskalakis, C., Golowich, N., and Zhang, K. The complexity of markov equilibrium in stochastic games, 2022. URL https://arxiv.org/abs/2204.03991.

Daskalakis, C. C., Fishelson, M., and Golowich, N. Near-optimal no-regret learning in general

games. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=cVwc7IHWEWi.

Erez, L., Lancewicki, T., Sherman, U., Koren, T., and Mansour, Y. Regret minimization and convergence to equilibria in general-sum markov games, 2022. URL https://arxiv.org/abs/2207.14211.

Fearnley, J., Gairing, M., Goldberg, P., and Savani, R. Learning equilibria of games via payoff queries. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pp. 397–414, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450319621. doi: 10.1145/2492002.2482558. URL https://doi.org/10.1145/2492002.2482558.

Foster, D. J., Rakhlin, A., Sekhari, A., and Sridharan, K. On the complexity of adversarial decision making. *arXiv preprint arXiv:2206.13063*, 2022.

Fudenberg, D., Levine, D., and Maskin, E. The folk theorem with imperfect public information. *Econometrica*, 62(5): 997–1039, 1994. ISSN 00129682, 14680262.

Hannan, J. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Hart, S. and Mas-Colell, A. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000. doi: https://doi.org/10.1111/1468-0262.00153. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00153.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 4870–4879. PMLR, 2020.

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning–a simple, efficient, decentralized algorithm for multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021.

Jin, Y., Muthukumar, V., and Sidford, A. The complexity of infinite-horizon general-sum stochastic games, 2022.

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

Kearns, M., Mansour, Y., and Ng, A. Y. A sparse sampling algorithm for near-optimal planning in large markov decision processes. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pp. 1324–1331, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

Kramár, J., Eccles, T., Gemp, I., Tacchetti, A., McKee, K. R., Malinowski, M., Graepel, T., and Bachrach, Y. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, 13(1):7214, December 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34473-5. URL https://www.nature.com/articles/s41467-022-34473-5. Number: 1 Publisher: Nature Publishing Group.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. RL for latent MDPs: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Leibo, J. Z., Dueñez-Guzman, E. A., Vezhnevets, A., Agapiou, J. P., Sunehag, P., Koster, R., Matyas, J., Beattie, C., Mordatch, I., and Graepel, T. Scalable evaluation of multi-agent reinforcement learning with melting pot. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6187–6199. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/leibo21a.html.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Littman, M. L. and Stone, P. A polynomial-time Nash equilibrium algorithm for repeated games. *Decision Support Systems*, 39:55–66, 2005.

Liu, Q., Wang, Y., and Jin, C. Learning Markov games with adversarial opponents: Efficient algorithms and fundamental limits. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14036–14053. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22r.html.

Malialis, K. and Kudenko, D. Distributed response to network intrusions using multiagent reinforcement learning. *Engineering Applications of Artificial Intelligence*, 41:270–284, 2015. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2015.01.013. URL https://www.sciencedirect.com/science/article/pii/S095219761500024X.

Mao, W. and Basar, T. Provably efficient reinforcement learning in decentralized general-sum markov games. *CoRR*, abs/2110.05682, 2021. URL https://arxiv.org/abs/2110.05682.

Maskin, E. and Fudenberg, D. The folk theorem in repeated games with discounting or with incomplete information.

Econometrica, 53(3):533–554, 1986. Reprinted in A. Rubinstein (ed.), Game Theory in Economics, London: Edward Elgar, 1995. Also reprinted in D. Fudenberg and D. Levine (eds.), A Long-Run Collaboration on Games with Long-Run Patient Players, World Scientific Publishers, 2009, pp. 209-230.

Nash, J. Non-cooperative games. *Annals of Mathematics*, 54 (2):286–295, 1951. ISSN 0003486X.

Papadimitriou, C. H. On the complexity of the parity argument and other inefficient proofs of existence. *J. Comput. Syst. Sci.*, 48(3):498–532, 1994.

Perolat, J., Vylder, B. D., Hennes, D., Tarassov, E., Strub, F., de Boer, V., Muller, P., Connor, J. T., Burch, N., Anthony, T., McAleer, S., Elie, R., Cen, S. H., Wang, Z., Gruslys, A., Malysheva, A., Khan, M., Ozair, S., Timbers, F., Pohlen, T., Eccles, T., Rowland, M., Lanctot, M., Lespiau, J.-B., Piot, B., Omidshafiei, S., Lockhart, E., Sifre, L., Beauguerlange, N., Munos, R., Silver, D., Singh, S., Hassabis, D., and Tuyls, K. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022. doi: 10.1126/science.add4679. URL https://www.science.org/doi/abs/10.1126/science.add4679.

Puterman, M. *Markov Decision Processes*. John Wiley & Sons, Ltd, 1 edition, 1994. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470316887.

Roughgarden, T. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 2015.

Rubinstein, A. Settling the complexity of computing approximate two-player Nash equilibria. In *Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 258–265. IEEE, 2016.

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

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *CoRR*, abs/1610.03295, 2016. URL http://arxiv.org/abs/1610.03295.

Shapley, L. Stochastic Games. *PNAS*, 1953.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Song, Z., Mei, S., and Bai, Y. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=6MmiS0HUJHR.

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2989–2997, 2015.

Vovk, V. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

Weisz, G., Amortila, P., Janzer, B., Abbasi-Yadkori, Y., Jiang, N., and Szepesvari, C. On query-efficient planning in mdps under linear realizability of the optimal state-value function. In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4355–4385. PMLR, 15–19 Aug 2021.

Yin, D., Hao, B., Abbasi-Yadkori, Y., Lazić, N., and Szepesvári, C. Efficient local planning with linear function approximation. In Dasgupta, S. and Haghtalab, N. (eds.), *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 1165–1192. PMLR, 29 Mar–01 Apr 2022.

Zhan, W., Lee, J. D., and Yang, Z. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. *arXiv preprint arXiv:2206.01588*, 2022.

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., and Socher, R. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607, 2022. doi: 10.1126/sciadv.abk2607. URL https://www.science.org/doi/abs/10.1126/sciadv.abk2607.

Contents of Appendix

Part I

Additional results and discussion

A. Tighter computational lower bounds under ETH for PPAD

Recall that Corollary 3.3 states that if PPAD $\neq$ P, then there is no constant $C > 4$ and $\mathrm{poly}(|\mathcal{G}|)$-time algorithm which solves the $(|\mathcal{G}|^C, |\mathcal{G}|^{-1/C})$-SPARSEMARKOVCCE problem for any 2-player Markov game $\mathcal{G}$. Using a stronger complexity-theoretic assumption, the Exponential Time Hypothesis for PPAD (Rubinstein, 2016), we can obtain a hardness result which rules out

efficient algorithms even when 1) the accuracy ϵ is constant, as opposed to being $|\mathcal{G}|^{-1/C}$, and 2) T is quasipolynomially large, as opposed to only being of polynomial size, i.e., $|\mathcal{G}|^C$.

Corollary A.1 (ETH-hardness of SPARSEMARKOVCCE)**.** *There is a constant $\epsilon_0 > 0$ such that if there exists an algorithm that solves the $(|\mathcal{G}|^{o(\log|\mathcal{G}|)},\epsilon_0)$-SPARSEMARKOVCCE problem in $|\mathcal{G}|^{o(\log|\mathcal{G}|)}$ time, then the Exponential Time Hypothesis for PPAD fails to hold.*

Corollary A.1 is an immediate consequence of Theorem 3.2 and the fact that for some absolute constant $\epsilon_0 > 0$, there are no polynomial-time algorithms for computing ϵ_0-Nash equilibria in 2-player normal-form games under the Exponential Time Hypothesis for PPAD (as shown in (Rubinstein, 2016)).

B. Discussion and interpretation

Theorems 3.2, 4.3, and 5.2 present barriers—both computational and statistical—toward developing efficient decentralized no-regret guarantees for multi-agent reinforcement learning. We emphasize that no-regret algorithms are the only known approach for obtaining fully decentralized learning algorithms (i.e., those which do not rely even on shared randomness) in normal-form games, and it seems unlikely that a substantially different approach would work in Markov games. Thus, these lower bounds for finding subexponential-length sequences of policies with the no-regret property represent a significant obstacle for fully decentralized multi-agent reinforcement learning. Moreover, these results rule out even the prospect of developing efficient *centralized* algorithms that produce no-regret sequences of policies, i.e., those which "resemble" independent learning. In this section, we compare our lower bounds with recent upper bounds for decentralized learning in Markov games, and explain how to reconcile these results.

B.1. Comparison to `V-learning`

The `V-learning` algorithm (Jin et al., 2021; Song et al., 2022; Mao & Basar, 2021) is a polynomial-time decentralized learning algorithm that proceeds in two phases. In the first phase, the m agents interact over the course of K episodes in a decentralized fashion, playing product Markov policies $\sigma^{(1)},\ldots,\sigma^{(K)} \in \Pi^{\text{markov}}$. In the second phase, the agents use data gathered during the first phase to produce a distributional policy $\widehat{\sigma} \in \Delta(\Pi^{\text{gen,rnd}})$, which we refer to as the *output policy* of `V-learning`. As discussed in Section 1, one implication of Theorem 3.2 is that the first phase of `V-learning` cannot guarantee each agent sublinear regret. Indeed if K is of polynomial size (and PPAD $\neq$ P), this follows because a bound of the form $\text{Reg}_{i,K}(\sigma^{(1)},\ldots,\sigma^{(K)}) \leqslant \epsilon K$ for all i implies that $(\sigma^{(1)},\ldots,\sigma^{(K)})$ solves the (K,ϵ)-SPARSEMARKOVCCE problem.

The output policy $\widehat{\sigma} \in \Delta(\Pi^{\text{gen,rnd}})$ produced by `V-learning` is an approximate CCE (per Definition 2.1), and it is natural to ask how many product policies it takes to represent $\widehat{\sigma}$ as a uniform mixture (that is, whether $\widehat{\sigma}$ solves the (T,ϵ)-SPARSEMARKOVCCE problem for a reasonable value of T). First, recall that `V-learning` requires $K = \text{poly}(H,S,\max_i A_i)/\epsilon^2$ episodes to ensure that $\widehat{\sigma}$ is an ϵ-CCE. It is straightforward to show that $\widehat{\sigma}$ can be expressed as a *non-uniform* mixture of at most K^{KHS+1} policies in $\Pi^{\text{gen,rnd}}$ (we prove this fact in detail below). By discretizing the non-uniform mixture, one can equivalently represent it as *uniform* mixture of $O(1/\epsilon)\cdot K^{KHS+1}$ product policies, up to ϵ error. Recalling the value of K, we conclude that we can express $\widehat{\sigma}$ as a uniform mixture of $T = \exp(\widetilde{O}(1/\epsilon^2)\cdot\text{poly}(H,S,\max_i A_i))$ product policies in $\Pi^{\text{gen,rnd}}$. Note that the lower bound of Theorem 4.3 rules out the efficient computation of an ϵ-CCE represented as a uniform mixture of $T \ll \exp(\epsilon^2 \cdot \max\{H,S,\max_i A_i\})$ efficiently computable policies in $\Pi^{\text{gen,rnd}}$. Thus, in the regime where $1/\epsilon$ is polynomial in $H,S,\max_i A_i$, this upper bound on the sparsity of the policy $\widehat{\sigma}$ produced `V-learning` matches that from Theorem 4.3, up to a polynomial in the exponent.

The sparsity of the output policy from `V-learning`. We now sketch a proof of the fact that the output policy $\widehat{\sigma}$ produced by `V-learning` can be expressed as a (non-uniform) average of K^{KHS+1} policies in $\Pi^{\text{gen,rnd}}$, where K is the number of episodes in the algorithm's initial phase. We adopt the notation and terminology from Jin et al. (2021).

Consider Algorithm 3 of Jin et al. (2021), which describes the second phase of `V-learning`, which produces the output policy $\widehat{\sigma}$. We describe how to write $\widehat{\sigma}$ as a weighted average of a collection of product policies, each of which is indexed by a function $\phi : [H]\times\mathcal{S}\times[K] \to [K]$ and a parameter $k_0 \in [K]$: in particular, we will write $\widehat{\sigma} = \sum_{k_0,\phi} w_{k_0,\phi}\cdot\sigma_{k_0,\phi} \in \Delta(\Pi^{\text{gen,rnd}})$, where $w_{k_0,\phi} \in [0,1]$ are mixing weights summing to 1 and $\sigma_{k_0,\phi} \in \Pi^{\text{gen,rnd}}$. The number of tuples (k_0,ϕ) is K^{1+KHS}.

We define the mixing weight allocated $w_{k_0,\phi}$ to any tuple (k_0,ϕ) to be:

$$\frac{1}{K}\cdot\prod_{(h,s,k)\in[H]\times\mathcal{S}\times[K]} \mathbb{1}\{\phi(h,s,k) \in [N_h^k(s)]\}\cdot\alpha_{N_h^k(s)}^{\phi(h,s,k)},$$

where $N_h^k(s) \in [K]$ and $\alpha_{N_h^k(s)}^i \in [0,1]$ (for $i \in [N_h^k(s)]$) are defined as in (Jin et al., 2021).

Next, for each k_0,ϕ, we define $\sigma_{k_0,\phi}\in\Pi^{\mathrm{gen,rnd}}$ to be the following policy: it maintains a parameter $k\in[K]$ over the first $h\leqslant H$ steps of the episode (as in Algorithm 3 of (Jin et al., 2021)), but upon reaching state s at step h, given the present value of $k\in[K]$, sets $i:=\phi(h,s,k)$, and updates $k\leftarrow k_h^i(s)$, and then samples an action $\mathbf{a}\sim\pi_h^k(\cdot|s)$ (where $k_h^i(s),\pi_h^k(\cdot|s)$ are defined in (Jin et al., 2021)). Since the mixing weights $w_{k_0,\phi}$ defined above exactly simulate the random draws of the parameter k in Line 1 and the parameters i in Algorithm 3, Line 4 of (Jin et al., 2021), it follows that the distributional policy $\widehat{\sigma}$ defined by Algorithm 3 of (Jin et al., 2021) is equal to $\sum_{k_0,\phi}w_{k_0,\phi}\cdot\sigma_{k_0,\phi}\in\Delta(\Pi^{\mathrm{gen,rnd}})$.

B.2. No-regret learning against Markov deviations

As discussed in Section 1, Erez et al. (2022) showed the existence of a learning algorithm with the property that if each agent plays it independently for T episodes, then no player can achieve regret more than $O(\mathrm{poly}(m,H,S,\max_i A_i)\cdot T^{3/4})$ by deviating to any fixed *Markov policy*. This notion of regret corresponds to, in the context of Definition 2.2, replacing $\max_{\sigma_i\in\Pi_i^{\mathrm{gen,rnd}}}$ with the smaller quantity $\max_{\sigma_i\in\Pi_i^{\mathrm{markov}}}$. Thus, the result of Erez et al. (2022) applies to a weaker notion of regret than that of the SPARSECCE problem, and so does not contradict any of our lower bounds. One may wonder which of these two notions of regret (namely, best possible gain via deviation to a Markov versus non-Markov policy) is the "right" one. We do not believe that there is a definitive answer to this question, but we remark that in many empirical applications of multi-agent reinforcement learning it is standard to consider non-Markov policies (Leibo et al., 2021; Agapiou et al., 2022). Furthermore, as shown in the proposition below, there are extremely simple games, e.g., of constant size, in which Markov deviations lead to "vacuous" behavior: in particular, all Markov policies have the same (suboptimal) value but the best non-Markov policy has much greater value:

Proposition B.1. *There is a 2-player, 2-action, 1-state Markov game with horizon* 2 *and a non-Markov policy* $\sigma_2\in\Pi_2^{\mathrm{gen,rnd}}$ *for player 2 so that for all* $\sigma_1\in\Pi_1^{\mathrm{markov}}$, $V_1^{\sigma_1\times\sigma_2}=1/2$ *yet* $\max_{\sigma_1\in\Pi_1^{\mathrm{gen,rnd}}}\left\{V_1^{\sigma_1\times\sigma_2}\right\}=3/4$.

The proof of Proposition B.1 is provided in Section B.5 below.

Other recent work has also proved no-regret guarantees with respect to deviations to restricted policy classes. In particular, Zhan et al. (2022) studies a setting in which each agent i is allowed to play policies in an arbitrary restricted policy class $\Pi_i'\subseteq\Pi_i^{\mathrm{gen,rnd}}$ in each episode, and regret is measured with respect to deviations to any policy in Π_i'. Zhan et al. (2022) introduces an algorithm, DORIS, with the property that when all agents play it independently, each agent i experiences regret $O\left(\mathrm{poly}(m,A,S,H)\cdot\sqrt{T\sum_{i=1}^m\log|\Pi_i'|}\right)$ to their respective class Π_i'.[15]

DORIS is not computationally efficient, since it involves performing exponential weights over the class Π_i', which requires space complexity $|\Pi_I'|$. Nonetheless, one can compare the statistical guarantees the algorithm provides to our own results. Let $\Pi_i^{\mathrm{markov,det}}\subset\Pi_i^{\mathrm{markov}}$ denote the set of deterministic Markov policies of agent i, namely sequences $\pi_i=(\pi_{i,1},\ldots,\pi_{i,H})$ so that $\pi_{i,h}:\mathcal{S}\to\mathcal{A}_i$. In the case that $\Pi_i'=\Pi_i^{\mathrm{markov,det}}$, Π_i', we have $\log|\Pi_i'|=O(SH\log A_i)$, which means that DORIS obtains no-regret against Markov deviations when m is constant, comparable to Erez et al. (2022).[16] However, we are interested in the setting in which each player's regret is measured with respect to all deviations in $\Pi_i^{\mathrm{gen,rnd}}$ (equivalently, $\Pi_i^{\mathrm{gen,det}}$). Accordingly, if we take $\Pi_i'=\Pi_i^{\mathrm{gen,det}}\subset\Pi_i^{\mathrm{gen,rnd}}$,[17] then $\log|\Pi_i'|>(SA_i)^{H-1}$, meaning that DORIS does not imply any sort of sample-efficient guarantee, even for $m=2$.

Finally, we remark that the algorithm DORIS (Zhan et al., 2022), as well as the similar algorithm OPMD from earlier work of Liu et al. (2022), obtains the same regret bound stated above even when the opponents are controlled by (possibly adaptive) adversaries. However, this guarantee crucially relies on the fact that any agent implementing DORIS must observe the policies played by opponents following each episode; this feature is the reason that the regret bound of DORIS does not contradict the exponential lower bound of Liu et al. (2022) for no-regret learning against an adversarial opponent. As a result of being restricted to this "revealed-policy" setting, DORIS is not a fully decentralized algorithm in the sense we consider in this paper.

[15] Note that in the tabular setting, the sample complexity of DORIS (Corollary 1) scales with the size A of the *joint* action set, since each player's value function class consists of the class of all functions $f:\mathcal{S}\times\mathcal{A}\to[0,1]$, which has Eluder dimension scaling with $S\cdot A$, i.e., exponential in m.

[16] Erez et al. (2022) has the added bonus of computational efficiency, even for polynomially large m, though has the significant drawback of assuming that the Markov game is known.

[17] DORIS plays distributions over policies in $\Pi_i'=\Pi_i^{\mathrm{gen,det}}$ at each episode, whereas in our lower bounds we consider the setting where a policy in $\Pi_i^{\mathrm{gen,rnd}}$ is played each episode; Facts F.2 and F.3 shows that these two settings are essentially equivalent, in that any policy in $\Pi_1^{\mathrm{gen,rnd}}\times\cdots\times\Pi_m^{\mathrm{gen,rnd}}$ can be simulated by one in $\Delta(\Pi_1^{\mathrm{gen,det}})\times\cdots\times\Delta(\Pi_m^{\mathrm{gen,det}})$, and vise versa.

B.3. On the role of shared randomness

A key assumption in our lower bounds for no-regret learning is that each of the joint policies $\sigma^{(1)},\ldots,\sigma^{(T)}$ produced by the algorithm is a *product policy*; such an assumption is natural, since it subsumes independent learning protocols in which each agent i selects $\sigma_i^{(t)}$ without knowledge of $\sigma_{-i}^{(t)}$. Compared to general (stochastic) joint policies, product policies have the desirable property that, to sample a trajectory from $\sigma^{(t)}=(\sigma_1^{(t)},\ldots,\sigma_m^{(t)})\in\Pi_1^{\mathrm{gen,rnd}}\times\cdots\times\Pi_m^{\mathrm{gen,rnd}}=\Pi^{\mathrm{gen,rnd}}$, the agents do no require access to shared randomness. In particular, each agent i can independently sample its action from $\sigma_i^{(t)}$ at each of the h steps of the episode. It is natural to ask how the situation changes if we allow the agents to use shared random bits when sampling from their policies, which corresponds to allowing $\sigma^{(1)},\ldots,\sigma^{(T)}$ to be non-product policies. In this case, `V-learning` yields a positive result via a standard "batch-to-online" conversion: by applying the first phase of `V-learning` during the first $T^{2/3}$ episodes and playing trajectories sampled i.i.d. from the output policy produced by `V-learning` during the remaining $T-T^{2/3}$ episodes (which requires shared randomness), it is straightforward to see that a regret bound of order $\mathrm{poly}(H,S,\max_i A_i)\cdot T^{2/3}$ can be obtained. Similar remarks apply to `SPoCMAR` (Daskalakis et al., 2022), which can obtain a slightly worse regret bound of order $\mathrm{poly}(H,S,\max_i A_i)\cdot T^{3/4}$ in the same fashion. In fact, the batch-to-online conversion approach gives a generic solution for the setting in which shared randomness is available. That is, *the assumption of shared randomness eliminates any distinction between no-regret algorithms and (non-sparse) equilibrium computation algorithms*, modulo slight loss in rates. For this reason, the shared randomness assumption is too strong to develop any sort of distinct theory of no-regret learning.

B.4. Comparison to lower bounds for finding stationary CCE

A separate line of work Daskalakis et al. (2022); Jin et al. (2022) has recently shown PPAD-hardness for the problem of finding stationary Markov CCE in infinite-horizon discounted stochastic games. These results are incomparable with our own: stationary Markov CCE are not sparse (in the sense of Definition 3.1), whereas we do not require stationarity of policies (as is standard in the finite-horizon setting).

B.5. Proof of Proposition B.1

Below we prove Proposition B.1.

Proof of Proposition B.1. We construct the claimed Markov game $\mathcal{G}$ as follows. The single state is denoted by $\mathfrak{s}$; as there is only a single state, the transitions are trivial. We denote each player's action space as $\mathcal{A}_1=\mathcal{A}_2=\{1,2\}$. The rewards to player 1 are given as follows: for all $(a_1,a_2)\in\mathcal{A}$,

$$R_{1,1}(\mathfrak{s},(a_1,a_2))=\frac{1}{2}\cdot\mathbb{I}_{a_2=1},\qquad R_{1,2}(\mathfrak{s},(a_1,a_2))=\frac{1}{2}\cdot\mathbb{I}_{a_1=a_2}.$$

We allow the rewards of player 2 to be arbitrary; they do not affect the proof in any way.

We let $\sigma_2=(\sigma_{2,1},\sigma_{2,2})\in\Pi_2^{\mathrm{gen,rnd}}$ be the policy which plays a uniformly random action at step 1 and then plays the same action at step 2: formally, $\sigma_{2,1}(s_1)=\mathrm{Unif}(\mathcal{A}_2)$, and $\sigma_{2,2}((s_1,a_{2,1},r_{2,1}),s_2)=\mathbb{I}_{a_{2,1}}$. Then for any Markov policy $\sigma_1\in\Pi_1^{\mathrm{markov}}$ of player 1, we must have $\mathbb{P}_{\sigma_1\times\sigma_2}(a_{1,2}=a_{2,2})=1/2$, which means that $V_1^{\sigma_1\times\sigma_2}=\frac{1}{2}\cdot\mathbb{E}_{\sigma_1\times\sigma_2}[\mathbb{I}_{a_{2,1}=1}+\mathbb{I}_{a_{1,2}=a_{2,2}}]=1/2\cdot(1/2+1/2)=1/2$.

On the other hand, any general (non-Markov) policy $\sigma_1\in\Pi_1^{\mathrm{gen,rnd}}$ which satisfies

$$\sigma_{1,2}((s_1,a_{1,1},r_{1,1}),s_2)=\begin{cases}\mathbb{I}_1: & r_{1,1}=1/2\\ \mathbb{I}_2: & r_{1,1}=0\end{cases}$$

has $V_1^{\sigma_1\times\sigma_2}=1/2\cdot(1/2+1)=3/4$. □

Part II

Proofs

C. Additional preliminaries

C.1. Additional preliminaries for Markov games

Deterministic policies. It will be helpful to introduce notation for *deterministic* general (non-Markov) policies, which correspond to the special case of randomized policies where each policy $\sigma_{i,h}$ exclusively maps to singleton distributions. In particular, a deterministic general policy of agent i is a collection of mappings $\pi_i = (\pi_{i,1},\ldots,\pi_{i,H})$, where $\pi_{i,h} : \mathscr{H}_{i,h-1} \times \mathcal{S} \to \mathcal{A}_i$. We denote by $\Pi_i^{\mathrm{gen,det}}$ the space of deterministic general policies of agent i, and further write $\Pi^{\mathrm{gen,det}} := \Pi_1^{\mathrm{gen,det}} \times \cdots \times \Pi_m^{\mathrm{gen,det}}$ to denote the space of *joint deterministic policies*. We use the convention throughout that deterministic policies are denoted by the letter π, whereas randomized policies are denoted by σ.

Additional facts on regret and CCE. The following facts regarding deterministic policies and the definition of coarse correlated equilibria and regret are well-known:

- In the context of Definition 2.1 (defining an ϵ-CCE), the maximizing policy σ_i' can always be chosen to be determinimistic, so $P \in \Delta(\Pi^{\mathrm{gen,rnd}})$ is an ϵ-CCE if and only if $\max_{\pi_i \in \Pi_i^{\mathrm{gen,det}}} V_i^{\pi_i \times P_{-i}} - V_i^P \leqslant \epsilon$.
- In the context of (1) in the definition of regret, the maximum over $\sigma_i \in \Pi_i^{\mathrm{gen,rnd}}$ is always achieved by a deterministic general policy, so we have $\mathrm{Reg}_{i,T} = \max_{\pi_i \in \Pi_i^{\mathrm{gen,det}}} \sum_{t=1}^T \left(V_i^{\pi_i \times \sigma_{-i}^{(t)}} - V_i^{\sigma^{(t)}}\right)$.

Next, the following standard result shows that the uniform average of any no-regret sequence forms an approximate coarse correlated equilibrium.

Fact C.1 (No-regret is equivalent to CCE)**.** *Suppose that a sequence of policies $\sigma^{(1)}, \ldots, \sigma^{(T)} \in \Pi^{\mathrm{gen,rnd}}$ satisfies $\mathrm{Reg}_{i,T}(\sigma^{(1)}, \ldots, \sigma^{(T)}) \leqslant \epsilon \cdot T$ for each $i \in [m]$. Then the uniform average of these T policies, namely the distributional policy $\overline{\sigma} := \frac{1}{T}\sum_{t=1}^T \mathbb{I}_{\sigma^{(t)}} \in \Delta(\Pi^{\mathrm{gen,rnd}})$, is an ϵ-CCE.*

Likewise if a sequence of policies $\sigma^{(1)}, \ldots, \sigma^{(T)} \in \Pi^{\mathrm{gen,rnd}}$ has the property that the distributional policy $\overline{\sigma} := \frac{1}{T}\sum_{t=1}^T \mathbb{I}_{\sigma^{(t)}} \in \Delta(\Pi^{\mathrm{gen,rnd}})$, is an ϵ-CCE, then we have $\mathrm{Reg}_{i,T}(\sigma^{(1)},\ldots,\sigma^{(T)}) \leqslant \epsilon \cdot T$ for all $i \in [m]$.

Fact C.1 is an immediate consequence of Definitions 2.1 and 2.2.

C.2. Nash equilibria and computational hardness.

The most foundational and well known solution concept for normal-form games is the *Nash equilibrium* (Nash, 1951).

Definition C.2 ((n,ϵ)-NASH problem)**.** For a normal-form game $G = (M_1, \ldots, M_m)$ and $\epsilon > 0$, a product distribution $p \in \prod_{j=1}^m \Delta([n])$ is said to be an ϵ-Nash equilibrium for G if for all $i \in [n]$,

$$\max_{a_i' \in [n]} \mathbb{E}_{\mathbf{a}\sim p}[(M_i)_{a_i', \mathbf{a}_{-i}}] - \mathbb{E}_{\mathbf{a}\sim p}[(M_i)_{\mathbf{a}}] \leqslant \epsilon.$$

We define the *m-player (n,ϵ)-NASH problem* to be the problem of computing an ϵ-Nash equilibrium of a given m-player n-action normal-form game.[18]

Informally, p is an ϵ-Nash equilibrium if no player i can gain more than ϵ in reward by deviating to a single fixed action a_i', while all other players randomly choose their actions according to p. Despite the intuitive appeal of Nash equilibria, they are intractable to compute: for any $c > 0$, it is PPAD-hard to solve the (n, n^{-c})-NASH problem, namely, to compute n^{-c}-approximate Nash

[18] One must also take care to specify the bit complexity of representing a normal-form game. We assume that the payoffs of any normal-form game given as an instance to the (n,ϵ)-NASH problem can each be expressed with $\max\{n,m\}$ bits; this assumption is without loss of generality as long as $\epsilon \geqslant 2^{-\max\{n,m\}}$ (which it will be for us).

equilibria in 2-player n-action normal-form games (Daskalakis et al., 2009; Chen et al., 2006; Rubinstein, 2018). We recall that the complexity class PPAD consists of all total search problems which have a polynomial-time reduction to the End-of-The-Line (EOTL) problem. PPAD is the most well-studied complexity class in algorithmic game theory, and it is widely believed that PPAD $\neq$ P. We refer the reader to (Daskalakis et al., 2009; Chen et al., 2006; Rubinstein, 2018; Papadimitriou, 1994) for further background on the class PPAD and the EOTL problem.

C.3. Query complexity of Nash equilibria

Our statistical lower bound for the SPARSECCE problem in Theorem 5.2 relies on existing query complexity lower bounds for computing approximate Nash equilibria in m-player normal-form games. We first review the query complexity model for normal-form games.

Oracle model for normal-form games. For $m,n\in\mathbb{N}$, consider an m-player n-action normal form game G, specified by payoff tensors $M_1,...,M_m$. Since the tensors $M_1,...,M_m$ contain a total of mn^m real-valued payoffs, in the setting when m is large, it is unrealistic to assume that an algorithm is given the full payoff tensors as input. Therefore, prior work on computing equilibria in such games has studied the setting in which the algorithm makes adaptive *oracle queries* to the payoff tensors.

In particular, the algorithm, which is allowed to be randomized, has access to a *payoff oracle* $\mathcal{O}_G$ for the game G, which works as follows. At each time step, the algorithm can choose to specify an action profile $\mathbf{a}\in[n]^m$ and then query $\mathcal{O}_G$ at the action profile $\mathbf{a}$. The oracle $\mathcal{O}_G$ then returns the payoffs $(M_1)_{\mathbf{a}},...,(M_m)_{\mathbf{a}}$ for each player if the action profile $\mathbf{a}$ is played.

Query complexity lower bound for approximate Nash equilibrium. The following theorem gives a lower bound on the number of queries any randomized algorithm needs to make to compute an approximate Nash equilibrium in an m-player game.

Theorem C.3 (Corollary 4.5 of (Rubinstein, 2016))**.** *There is a constant $\epsilon_0>0$ so that any randomized algorithm which solves the $(2,\epsilon_0)$-*NASH *problem for m-player normal-form games with probability at least $2/3$ must use at least $2^{\Omega(m)}$ payoff queries.*

We remark that (Babichenko, 2016; Chen et al., 2017) provide similar, though quantitatively weaker, lower bounds to that in Theorem C.3. We also emphasize that the lower bound of Theorem C.3 applies to *any* algorithm, i.e., including those which require extremely large computation time.

D. Proofs of lower bounds for SPARSEMARKOVCCE (Section 3)

D.1. Preliminaries: Online density estimation

Our proof makes use of tools for online learning with the logarithmic loss, also known as conditional density estimation. In particular, we use a variant of the exponential weights algorithm known as *Vovk's aggregating algorithm* in the context of density estimation (Vovk, 1990; Cesa-Bianchi & Lugosi, 2006). We consider the following setting with two players, a *Learner* and *Nature*. Furthermore, there is a set $\mathcal{Y}$, called the *outcome space*, and a set $\mathcal{X}$, called the context space; for our applications it suffices to assume $\mathcal{Y}$ and $\mathcal{X}$ are finite. For some $T\in\mathbb{N}$, there are T time steps $t=1,2,...,T$. At each time step $t\in[T]$:

- Nature reveals a context $x^{(t)}\in\mathcal{X}$;
- Having seen the context $x^{(t)}$, the learner predicts a distribution $\widehat{q}^{(t)}\in\Delta(\mathcal{Y})$;
- Nature chooses an outcome $y^{(t)}\in\mathcal{Y}$, and the learner suffers loss $\ell^{(t)}_{\log}(\widehat{q}^{(t)}):=\log\left(\frac{1}{\widehat{q}^{(t)}(y^{(t)})}\right)$.

For each $t\in[T]$, we let $\mathcal{H}^{(t)}=\{(x^{(1)},y^{(1)},\widehat{q}^{(1)}),...,(x^{(t)},y^{(t)},\widehat{q}^{(t)})\}$ denote the history of interaction up to step t; we emphasize that each context $x^{(t)}$ may be chosen adaptively as a function of $\mathcal{H}^{(t-1)}$. Let $\mathscr{F}^{(t)}$ denote the sigma-algebra generated by $(\mathcal{H}^{(t)},x^{(t+1)})$. We measure performance in terms of regret against a set $\mathcal{I}$ of *experts*, also known as the *expert setting*. Each expert $i\in\mathcal{I}$ consists of a function $p_i:\mathcal{X}\to\Delta(\mathcal{Y})$. The *regret* of an algorithm against the expert class $\mathcal{I}$ when it receives contexts $x^{(1)},...,x^{(T)}$ and observes outcomes $y^{(1)},...,y^{(T)}$ is defined as

$$\mathrm{Reg}_{\mathcal{I},T}=\sum_{t=1}^{T}\ell^{(t)}_{\log}(\widehat{q}^{(t)})-\min_{i\in\mathcal{I}}\sum_{t=1}^{T}\ell^{(t)}_{\log}(p_i(x^{(t)})).$$

Note that the learner can observe the expert predictions $\{p_i(x^{(t)})\}_{i\in\mathcal{I}}$ and use them to make its own prediction at each round t.

Proposition D.1 (Vovk's aggregating algorithm)**.** *Consider Vovk's aggregating algorithm, which predicts via*

$$\widehat{q}^{(t)}(y) := \mathbb{E}_{i\sim\widetilde{q}^{(t)}}[p_i(x^{(t)})], \quad \textit{where} \quad \widetilde{q}^{(t)}(i) := \frac{\exp\left(-\sum_{s=1}^{t-1}\ell_{\log}^{(s)}(p_i(x^{(s)}))\right)}{\sum_{j\in\mathcal{I}}\exp\left(-\sum_{s=1}^{t-1}\ell_{\log}^{(s)}(p_j(x^{(s)}))\right)}. \tag{2}$$

This algorithm guarantees a regret bound of $\mathrm{Reg}_{\mathcal{I},T} \leqslant \log|\mathcal{I}|$.

Recall that for probability distributions p,q on a finite set $\mathcal{B}$, their total variation distance is defined as

$$D_{\mathsf{TV}}(p,q) = \max_{\mathcal{E}\subset\mathcal{B}} |p(\mathcal{E}) - q(\mathcal{E})|. \tag{3}$$

As a (standard) consequence of Proposition D.1, in the *realizable* setting in which the distribution of $y^{(t)}|x^{(t)}$ follows $p_{i^\star}(x^{(t)})$ for some fixed (unknown) expert $i^\star \in \mathcal{I}$, we can obtain a bound on the total variation distance between the algorithm's predictions and those of $p_{i^\star}(x^{(t)})$.

Proposition D.2. *If the distribution of outcomes is* realizable*, i.e., there exists an expert $i^\star \in \mathcal{I}$ so that $y^{(t)} \sim p_{i^\star}(x^{(t)}) \mid x^{(t)}, \mathcal{H}^{(t-1)}$ for all $t \in [T]$, then the predictions $\widehat{q}^{(t)}$ of the aggregation algorithm (2) satisfy*

$$\sum_{t=1}^{T} \mathbb{E}[D_{\mathsf{TV}}(\widehat{q}^{(t)}, p_{i^\star}(x^{(t)}))] \leqslant \sqrt{T\log|\mathcal{I}|}.$$

For completeness, we provide the proof of Proposition D.2 here.

Proof of Proposition D.2. To simplify notation, for an expert $i \in \mathcal{I}$, a context $x \in \mathcal{X}$, and an outcome $y \in \mathcal{Y}$, we write $p_i(y|x)$ to denote $p_i(x)(y)$.

Proposition D.1 gives that the following inequality holds (almost surely):

$$\mathrm{Reg}_{\mathcal{I},T} = \sum_{t=1}^{T} \log\left(\frac{1}{\widehat{q}^{(t)}(y^{(t)})}\right) - \sum_{t=1}^{T} \log\left(\frac{1}{p_{i^\star}(y^{(t)}|x^{(t)})}\right) \leqslant \log|\mathcal{I}|.$$

For each $t \in [T]$, note that $\widehat{q}^{(t)}$ and $x^{(t)}$ are $\mathscr{F}^{(t-1)}$-measurable (by definition). Then

$$\begin{aligned}
\sum_{t=1}^{T} D_{\mathsf{TV}}(\widehat{q}^{(t)}, p_{i^\star}(x^{(t)}))^2 &\leqslant \sum_{t=1}^{T} D_{\mathsf{KL}}(p_{i^\star}(x^{(t)}) \| \widehat{q}^{(t)}) \\
&= \sum_{t=1}^{T} \sum_{y\in\mathcal{Y}} p_{i^\star}(y|x^{(t)}) \cdot \log\left(\frac{p_{i^\star}(y|x^{(t)})}{\widehat{q}^{(t)}(y)}\right) \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\log\left(\frac{1}{\widehat{q}^{(t)}(y^{(t)})}\right) - \log\left(\frac{1}{p_{i^\star}(y^{(t)}|x^{(t)})}\right) \mid \mathscr{F}^{(t-1)}\right],
\end{aligned}$$

where the first inequality uses Pinsker's inequality and the final equality uses the fact that $y^{(t)} \sim p_{i^\star}(x^{(t)})|x^{(t)}, \mathcal{H}^{(t-1)}$. It follows that

$$\mathbb{E}\left[\sum_{t=1}^{T} D_{\mathsf{TV}}(\widehat{q}^{(t)}, p_{i^\star}(x^{(t)}))^2\right] \leqslant \mathbb{E}[\mathrm{Reg}_{\mathcal{I},T}] \leqslant \log|\mathcal{I}|.$$

Jensen's inequality now gives that

$$\mathbb{E}\left[\sum_{t=1}^{T} D_{\mathsf{TV}}(\widehat{q}^{(t)}, p_{i^\star}(x^{(t)}))\right] \leqslant \sqrt{T} \cdot \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} D_{\mathsf{TV}}(\widehat{q}^{(t)}, p_{i^\star}(x^{(t)}))^2\right]} \leqslant \sqrt{T\log|\mathcal{I}|}.$$

□

D.2. Proof of Theorem 3.2

Proof of Theorem 3.2. Fix $n \in \mathbb{N}$, which we recall represents an upper bound on the description length of the Markov game. Assume that we are given an algorithm $\mathscr{B}$ that solves the (T,ϵ)-SPARSEMARKOVCCE problem for Markov games $\mathcal{G}$ satisfying $|\mathcal{G}| \leqslant n$ in time U. We proceed to describe an algorithm which solves the 2-player $(\lfloor n^{1/2}/2 \rfloor, 4\cdot\epsilon)$-NASH problem in time $(nTU)^{C_0}$, as long as $T < \exp(\epsilon^2 \cdot n^{1/2}/2^5)$. First, define $n_0 := \lfloor n^{1/2}/2 \rfloor$, and consider an arbitrary 2-player n_0-action normal form G, which is specified by payoff matrices $M_1, M_2 \in [0,1]^{n_0 \times n_0}$, so that all entries of the game can be written in binary using at most n_0 bits (recall, per footnote 18, that we may assume that the entries of an instance of $(n_0, 4\cdot\epsilon)$-NASH can be specified with n_0 bits). Based on G, we construct a 2-player Markov game $\mathcal{G} := \mathcal{G}(G)$ as follows:

Definition D.3. We define the game $\mathcal{G}(G)$ to consist of the tuple $\mathcal{G}(G) = (\mathcal{S}, H, (\mathcal{A}_i)_{i\in[2]}, \mathbb{P}, (R_i)_{i\in[2]}, \mu)$, where:

- The horizon of $\mathcal{G}$ is $H = 2\lfloor n_0/2 \rfloor$ (i.e., the largest even number at most n_0).
- Let $A = n_0$; the action spaces of the 2 agents are given by $\mathcal{A}_1 = \mathcal{A}_2 = [A]$.
- There are a total of $A^2 + 1$ states: in particular, there is a state $\mathfrak{s}_{(a_1,a_2)}$ for each $(a_1,a_2) \in [A]^2$, as well as a distinguished state $\mathfrak{s}$, so we have:

$$\mathcal{S} = \{\mathfrak{s}\} \cup \{\mathfrak{s}_{(a_1,a_2)} : (a_1,a_2) \in [A]^2\}.$$

- For all odd $h \in [H]$, the reward to agents $j \in [2]$ given that the action profile (a_1,a_2) is played at step h is given by $R_{j,h}(s,(a_1,a_2)) := \frac{1}{H} \cdot (M_j)_{a_1,a_2}$, for all $s \in \mathcal{S}$. All agents receive 0 reward at even steps $h \in [H]$.
- At odd steps $h \in [H]$, if actions $a_1, a_2 \in [A]$ are taken, the game transitions to the state $\mathfrak{s}_{(a_1,a_2)}$. At even steps $h \in [H]$, the game always transitions to the state $\mathfrak{s}$.
- The initial state (i.e., at step $h = 1$) is $\mathfrak{s}$ (i.e., μ is a singleton distribution supported on $\mathfrak{s}$).

It is evident that this construction takes polynomial time, and satisfies $|\mathcal{G}| \leqslant A^2 + 1 \leqslant n_0^2 + 1 \leqslant n$. We will now show by applying the algorithm $\mathscr{B}$ to $\mathcal{G}$, we can efficiently compute $4\cdot\epsilon$-approximate Nash equilibrium for the original game G. To do so, we appeal to Algorithm 1.

Algorithm 1 Algorithm to compute Nash equilibrium used in proof of Theorem 3.2.

Input: 2-player, n_0-action normal form game G.
Construct the 2-player Markov game $\mathcal{G} = \mathcal{G}(G)$ per Definition D.3, which satisfies $|\mathcal{G}| \leqslant n$.
Call the algorithm $\mathscr{B}$ on the game $\mathcal{G}$, which produces a sequence $\sigma^{(1)}, \ldots, \sigma^{(T)}$, where each $\sigma^{(t)} \in \Pi^{\mathrm{markov}}$.
for $t \in [T]$ and odd $h \in [H]$: **do**
  if $\sigma_h^{(t)}(\mathfrak{s}) \in \Delta(\mathcal{A}_1) \times \Delta(\mathcal{A}_2)$ is a $(4\cdot\epsilon, n)$-Nash equilibrium of G: **then**
    return $\sigma_h^{(t)}(\mathfrak{s})$.
  end if
end for
if the for loop terminates without returning: return **fail**.

Algorithm 1 proceeds as follows. First, it constructs the 2-player Markov game $\mathcal{G}(G)$ as defined above, and calls the algorithm $\mathscr{B}$, which returns a sequence $\sigma^{(1)}, \ldots, \sigma^{(T)} \in \Pi^{\mathrm{markov}}$ of product Markov policies with the property that the average $\overline{\sigma} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\sigma^{(t)}}$ is an ϵ-CCE of $\mathcal{G}$. It then enumerates over the distributions $\sigma_h^{(t)}(\mathfrak{s}) \in \Delta(\mathcal{A}_1) \times \Delta(\mathcal{A}_2)$ for each $t \in [T]$ and $h \in [H]$ odd, and checks whether each one is a $4\cdot\epsilon$-approximate Nash equilibrium of G. If so, the algorithm outputs such a Nash equilibrium, and otherwise, it fails. The proof of Theorem 3.2 is thus completed by the following lemma, which states that as long as $\overline{\sigma}$ is an ϵ-CCE of $\mathcal{G}$, Algorithm 1 never fails.

Lemma D.4 (Correctness of Algorithm 1)**.** *Consider the normal form game G and the Markov game $\mathcal{G} = \mathcal{G}(G)$ as constructed above, which has horizon H. For any $\epsilon_0 > 0$, $T \in \mathbb{N}$, if $T < \exp(H \cdot \epsilon_0^2/2^8)$ and $\sigma^{(1)}, \ldots, \sigma^{(T)} \in \Pi^{\mathrm{markov}}$ are product Markov policies so that $\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\sigma^{(t)}}$ is an $(\epsilon_0/4)$-CCE of $\mathcal{G}$, then there is some odd $h \in [H]$ and $t \in [T]$ so that $\sigma_h^{(t)}(\mathfrak{s})$ is an ϵ_0-Nash equilibrium of G.*

The proof of Lemma D.4 is given below. Applying Lemma D.4 with $\epsilon_0 = 4\epsilon$ (which is a valid application since $T < \exp(n_0 \cdot (4\epsilon)^2/2^8)$ by our assumption on T,ϵ), yields that Algorithm 1 always finds a 4ϵ-Nash equilibrium of the n_0-action normal form game G, thus solving the given instance of the $(n_0, 4\cdot\epsilon)$-NASH problem. Furthermore, it is straightforward to see that Algorithm 1 runs in time $U + (nT)^{C_0} \leqslant (UnT)^{C_0}$, for some constant $C_0 \geqslant 1$.

□

Proof of Lemma D.4. Consider a sequence of product Markov policies $\sigma^{(1)}, \ldots, \sigma^{(T)}$ with the property that the average $\overline{\sigma} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\sigma^{(T)}}$ is an $(\epsilon_0/4)$-CCE of $\mathcal{G}$. For all odd $h \in [H]$ and $j \in [2]$, let $p_{j,h}^{(t)} := \sigma_{j,h}^{(t)}(\mathfrak{s}) \in \Delta(\mathcal{A}_j)$, which is the distribution played under $\sigma^{(t)}$ by player j at step h (at the unique state $\mathfrak{s}$ with positive probability of being reached at step h). For odd h, we have $\sigma_h^{(t)}(\mathfrak{s}) = p_{1,h}^{(t)} \times p_{2,h}^{(t)}$, and our goal is to show that for some odd $h \in [H]$ and $t \in [T]$, $p_{1,h}^{(t)} \times p_{2,h}^{(t)}$ is an ϵ_0-Nash equilibrium of G. To proceed, suppose for the sake of contradiction that this is not the case.

Let us write $\mathcal{O}_H := \{h \in [H] : h \text{ odd}\}$ to denote the set of odd-numbered steps, and $\mathcal{E}_H = [H] \backslash \mathcal{O}_H$ to denote the set of even-numbered steps. Let $H_0 = |\mathcal{O}_H| = |\mathcal{E}_H| = H/2$. We first note that for $j \in [2]$, agent j's value under the mixture policy $\overline{\sigma}$ is given as follows:

$$V_j^{\overline{\sigma}} = \frac{1}{TH} \sum_{t=1}^{T} \sum_{h \in \mathcal{O}_H} \mathbb{E}_{a_1 \sim p_{1,h}^{(t)}, a_2 \sim p_{2,h}^{(t)}} [(M_j)_{a_1,a_2}].$$

For each $j \in [2]$, we will derive a contradiction by constructing a (non-Markov) deviation policy for player j in $\mathcal{G}$, denoted $\pi_j^\dagger \in \Pi_j^{\text{gen,det}}$, which will give player j a significant gain in value against the policy $\overline{\sigma}$. To do so, we need to specify $\pi_{j,h}^\dagger(\tau_{j,h-1}, s_h) \in \mathcal{A}_j$, for all $\tau_{j,h-1} \in \mathscr{H}_{j,h-1}$ and $s_h \in \mathcal{S}$; note that we may restrict our attention only to histories τ_{j,h_0-1} that occur with positive probability under the transitions of $\mathcal{G}$.

Fix any $h_0 \in [H]$, $\tau_{j,h_0-1} \in \mathscr{H}_{j,h_0-1}$, and $s_{h_0} \in \mathcal{S}$. If τ_{j,h_0-1} occurs with positive probability under the transitions of $\mathcal{G}$, then for each $h \in \mathcal{O}_H$, $h < h_0 - 1$ and both $j' \in [2]$, the action played by agent j' at step h is determined by $\tau_{j,h}$. Namely, if the state at step $h+1$ of τ_{j,h_0-1} is $\mathfrak{s}_{(a_1', a_2')}$, then player j' played action a_j' at step h. So, for each $h \in \mathcal{O}_H$ with $h < h_0 - 1$, we may define $(a_{1,h}, a_{2,h})$ as the action profile played at step h, which is a measurable function of τ_{j,h_0-1}. With this in mind, we define $\pi_{j,h_0}^\dagger(\tau_{j,h_0-1}, s_{h_0})$ by applying Vovk's aggregating algorithm (Proposition D.2) as follows.

1. If h_0 is even, play an arbitrary action (note that the actions at even-numbered steps have no influence on the transitions or rewards).

2. If h_0 is odd, define $\widehat{q}_{j,h_0} \in \Delta(\mathcal{A}_j)$, by $\widehat{q}_{j,h_0} := \mathbb{E}_{t \sim \widetilde{q}_{j,h_0}}[p_{-j,h}^{(t)}]$, where $\widetilde{q}_{j,h_0} \in \Delta([T])$ is defined as follows: for $t \in [T]$,

$$\widetilde{q}_{j,h_0}(t) := \frac{\exp\left(-\sum_{h < h_0 :\, h \in \mathcal{O}_H} \log\left(\frac{1}{p_{-j,h}^{(t)}(a_{-j,h})}\right)\right)}{\sum_{t'=1}^{T} \exp\left(-\sum_{h < h_0 :\, h \in \mathcal{O}_H} \log\left(\frac{1}{p_{-j,h}^{(t')}(a_{-j,h})}\right)\right)}.$$

   Note that $\widehat{q}_{j,h_0}$ is a function of τ_{j,h_0-1} via the action profiles $\{(a_{1,h}, a_{2,h})\}_{h < h_0 : h \in \mathcal{O}_H}$; to simplify notation, we suppress this dependence.

3. Then for any state $s_{h_0} \in \mathcal{S}$, define $\pi_{j,h_0}^\dagger(\tau_{j,h_0-1}, s_{h_0})$ to be a best response to $\widehat{q}_{j,h_0}$, namely

$$\pi_{j,h_0}^\dagger(\tau_{j,h_0-1}, s_{h_0}) := \operatorname*{argmax}_{a_j \in \mathcal{A}_j} \mathbb{E}_{a_{-j} \sim \widehat{q}_{j,h_0}} [R_{j,h}(\mathfrak{s}_{h_0}, (a_1, a_2))] = \operatorname*{argmax}_{a_j \in \mathcal{A}_j} \mathbb{E}_{a_{-j} \sim \widehat{q}_{j,h_0}} [(M_j)_{a_1,a_2}]. \tag{4}$$

Note that, for odd h_0, the distribution $\widehat{q}_{j,h_0} \in \Delta(\mathcal{A}_j)$ defined above can be viewed as an application of Vovk's online aggregation algorithm at step $(h_0+1)/2$ in the following setting: the number of steps (T, in the notation of Proposition D.2; note that T plays a different role in the present proof) is $H_0 = H/2$, the context space is $\mathcal{O}_H$, and the outcome space is $\mathcal{A}_{-j}$.[19] There are T experts $\widetilde{p}^{(1)}, \ldots, \widetilde{p}^{(T)}$ (i.e., we have $\mathcal{I} = \{\widetilde{p}^{(t)}\}_{t \in [T]}$), whose predictions on a context $h \in \mathcal{O}_H$ are defined as follows: the expert $\widetilde{p}^{(t)}$ predicts

[19] Here $-j$ denotes the index of the player who is not j.

$\widetilde{p}^{(t)}(h) := p^{(t)}_{-j,h}$. Then, the distribution $\widehat{q}_{j,h_0}$ is obtained by updating the aggregation algorithm with the context-observation pairs $(h, a_{-j,h})$, for *odd* values of $h < h_0$.

We next analyze the value of $V_j^{\pi^\dagger_j, \overline{\sigma}_{-j}}$ for $j \in [2]$ to show that the deviation strategy we have defined indeed obtains significant gain. To do so, recall that this value represents the payoff for player j under the process in which we draw an index $t^\star \in [T]$ uniformly at random, then for each step $h \in [H]$, player j plays according to $\pi^\dagger_j$ and player $-j$ plays according to $\sigma^{(t^\star)}_{-j}$. (In particular, at odd-numbered steps, player $-j$ plays according to $p^{(t^\star)}_{-j,h}$.) We recall that $\mathbb{E}_{\pi^\dagger_j \times \overline{\sigma}_{-j}}[\cdot]$ denotes the expectation under this process. We let $\tau_{j,h-1} \in \mathscr{H}_{j,h-1}$ denote the random variable which is the history observed by player j in this setup, i.e., when the policy played is $\pi^\dagger_j \times \overline{\sigma}_{-j}$, and let $\{(a_{1,h}, a_{2,h})\}_{h \in \mathcal{O}_H}$ denote the action profiles for odd rounds, which are a measurable function of each player's trajectory.

We apply Proposition D.2 with the time horizon as H_0, and with the set of experts set to $\mathcal{I} := \{\widetilde{p}^{(1)}, \ldots, \widetilde{p}^{(T)}\}$ as defined above. The context sequence the sequence of increasing values of $h \in \mathcal{O}_H$, and for each $h \in \mathcal{O}_H$, the outcome at step $(h+1)/2$ (for which the context is h) is distributed as $a_{-j,h} \sim \widetilde{p}^{(t^\star)}(h) = p^{(t^\star)}_{-j,h}$ conditioned on $t^\star$, which in particular satisfies the realizability assumption stated in Proposition D.2. Then, since (as remarked above), the distributions $\widehat{q}_{j,h}$, for $h \in \mathcal{O}_H$, are exactly the predictions made by Vovk's aggregating algorithm, Proposition D.2 gives that[20]

$$\mathbb{E}_{\pi^\dagger_j \times \overline{\sigma}_{-j}}\left[\sum_{h \in \mathcal{O}_H} D_{\mathsf{TV}}(\widehat{q}_{j,h}, p^{(t^\star)}_{-j,h})\right] = \mathbb{E}_{\pi^\dagger_j \times \overline{\sigma}_{-j}}\left[\sum_{h \in \mathcal{O}_H} D_{\mathsf{TV}}(\widehat{q}_{j,h}, \widetilde{p}^{(t^\star)}(h))\right] \leqslant \sqrt{H_0 \log T}. \tag{5}$$

Recall that we have assumed for the sake of contradiction that $p^{(t)}_{1,h} \times p^{(t)}_{2,h}$ is not an ϵ_0-Nash equilibrium of G for each $h \in [H]$ and $t \in [T]$. Consider a fixed draw of the random variable $t^\star \in [T]$ defined above. Then it holds that for $j \in [2]$ and $h \in [H]$, defining

$$\epsilon_{0,j,h} := \max_{a_j \in [A]} \mathbb{E}_{a_{-j} \sim p^{(t^\star)}_{-j,h}}[(M_j)_{a_1,a_2}] - \mathbb{E}_{a_1 \sim p^{(t^\star)}_{1,h}, a_2 \sim p^{(t^\star)}_{2,h}}[(M_j)_{a_1,a_2}], \tag{6}$$

we have $\epsilon_{0,1,h} + \epsilon_{0,2,h} \geqslant \epsilon_0$. Consider any $j \in [2]$, $h \in \mathcal{O}_H$, and a history $\tau_{j,h-1} \in \mathscr{H}_{j,h-1}$ of agent j up to step $h-1$ (conditioned on $t^\star$). Let us write $\delta^{(t^\star)}_{-j,h} := D_{\mathsf{TV}}(p^{(t^\star)}_{-j,h}, \widehat{q}_{j,h})$; note that $\delta^{(t^\star)}_{-j,h}$ is a function of $\tau_{j,h-1}$, through its dependence on $\widehat{q}_{j,h}$. We have, by the definition of $\pi^\dagger_{j,h}(\tau_{j,h-1}, s_h)$ in (4) and the definition of $\delta^{(t^\star)}_{-j,h}$,

$$\begin{aligned}
\mathbb{E}_{a_{-j} \sim p^{(t^\star)}_{-j,h}}\left[(M_j)_{\pi^\dagger_{h,j}(\tau_{j,h-1}, \mathfrak{s}), a_{-j}} \mid t^\star, \tau_{j,h-1}\right] &\geqslant \mathbb{E}_{a_{-j} \sim \widehat{q}_{j,h}}\left[(M_j)_{\pi^\dagger_{h,j}(\tau_{j,h-1}, \mathfrak{s}), a_{-j}} \mid t^\star, \tau_{j,h-1}\right] - \delta^{(t^\star)}_{-j,h} \\
&= \max_{a_j \in [A]} \mathbb{E}_{a_{-j} \sim \widehat{q}_{j,h}}\left[(M_j)_{a_j, a_{-j}} \mid t^\star, \tau_{j,h-1}\right] - \delta^{(t^\star)}_{h,-j} \\
&\geqslant \max_{a_j \in [A]} \mathbb{E}_{a_{-j} \sim p^{(t^\star)}_{h,-j}}\left[(M_j)_{a_j, a_{-j}}\right] - 2\delta^{(t^\star)}_{-j,h}.
\end{aligned} \tag{7}$$

Combining (6) and (7), we get that for any fixed $h \in \mathcal{O}_H$, $j \in [2]$, and $\tau_{j,h-1} \in \mathscr{H}_{j,h-1}$,

$$\mathbb{E}_{a_{-j} \sim p^{(t^\star)}_{-j,h}}\left[(M_j)_{\pi^\dagger_{j,h}(\tau_{j,h-1}, \mathfrak{s}), a_{-j}} \mid t^\star, \tau_{j,h-1}\right] - \mathbb{E}_{a_1 \sim p^{(t^\star)}_{1,h}, a_2 \sim p^{(t^\star)}_{2,h}}[(M_j)_{a_1,a_2}] > \epsilon_{0,j,h} - 2\delta^{(t^\star)}_{-j,h}. \tag{8}$$

[20] In fact, Proposition D.2 implies that a similar bound holds uniformly for each possible realization of $t^\star$, but (5) suffices for our purposes.

Averaging over the draw of $t^\star \in [T]$, which we recall is chosen uniformly, we see that

$$\begin{aligned}
&\sum_{j\in[2]} V_j^{\pi_j^\dagger \times \overline{\sigma}_{-j}} - V_j^{\overline{\sigma}} \\
=&\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in[2]} V_j^{\pi_j^\dagger \times \sigma_{-j}^{(t)}} - V_j^{\sigma^{(t)}} && (9)\\
=&\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in[2]} \mathbb{E}_{\pi_j^\dagger \times \sigma_{-j}^{(t)}}\left[\sum_{h\in\mathcal{O}_H} \mathbb{E}_{a_{-j}\sim p_{-j,h}^{(t)}}\left[R_{j,h}(\mathfrak{s},(\pi_{j,h}^\dagger(\tau_{j,h-1},\mathfrak{s}),a_{-j})) \mid t, \tau_{j,h-1}\right] - \mathbb{E}_{a_1\sim p_{1,h}^{(t)}, a_2\sim p_{2,h}^{(t)}}\left[R_{j,h}(\mathfrak{s},(a_1,a_2))\right]\right] \\
=&\frac{1}{TH}\sum_{t=1}^{T}\sum_{j\in[2]} \mathbb{E}_{\pi_j^\dagger \times \sigma_{-j}^{(t)}}\left[\sum_{h\in\mathcal{O}_H} \mathbb{E}_{a_{-j}\sim p_{-j,h}^{(t)}}\left[(M_j)_{\pi_{j,h}^\dagger(\tau_{j,h-1},\mathfrak{s}),a_{-j}} \mid t, \tau_{j,h-1}\right] - \mathbb{E}_{a_1\sim p_{1,h}^{(t)}, a_2\sim p_{2,h}^{(t)}}\left[(M_j)_{a_1,a_2}\right]\right] \\
\geqslant&\frac{1}{TH}\sum_{t=1}^{T}\sum_{j\in[2]} \mathbb{E}_{\pi_j^\dagger \times \sigma_{-j}^{(t)}}\left[\sum_{h\in\mathcal{O}_H}\left(\epsilon_{0,j,h} - 2\delta_{-j,h}^{(t)}\right)\right] && (10)\\
\geqslant&\frac{\epsilon_0}{2} - \frac{2}{TH}\sum_{t=1}^{T} 2\sqrt{H_0 \log T} \geqslant \frac{\epsilon_0}{2} - 4\sqrt{\log(T)/H}, && (11)
\end{aligned}$$

where (9) follows from the definition $\overline{\sigma} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{\sigma^{(t)}}$, (10) follows from (8), and (11) uses (5). As long as $T < \exp(H\cdot(\epsilon_0/16)^2)$, the this expression is bounded below by $\epsilon_0/4$, meaning that $\overline{\sigma}$ is not an $\epsilon_0/4$-approximate CCE. This completes the contradiction. □

E. Proofs of lower bounds for SPARSECCE (Sections 4 and 5)

In this section we prove our computational lower bounds for solving the SPARSECCE problem with $m = 3$ players (Theorem 4.3 and Corollary 4.4), as well as our statistical lower bound for solving the SPARSECCE problem with a general number m of players (Theorem 5.2).

Both theorems are proven as consequences of a more general result given in Theorem E.1 below, which reduces the NASH problem in m-player normal-form games to the SPARSECCE problem in $(m+1)$-player Markov games. In more detail, the theorem shows that (a) if an algorithm for SPARSECCE makes few calls to a generative model oracle, then we get an algorithm for the NASH problem with few calls to a payoff oracle (see Section C.3 for background on the payoff oracle for the NASH problem), and (b) if the algorithm for SPARSECCE is *computationally* efficient, then so is the algorithm for the NASH problem.

Theorem E.1. *There is a constant $C_0 > 0$ so that the following holds. Consider $n, m \in \mathbb{N}$, and suppose $T, N, Q \in \mathbb{N}$ and $\epsilon > 0$ satisfy $1 < T < \exp\left(\frac{\epsilon^2\cdot\lfloor n/m\rfloor}{m^2}\right)$. Suppose there is an algorithm $\mathscr{B}$ which, given a generative model oracle for a $(m+1)$-player Markov game $\mathcal{G}$ with $|\mathcal{G}| \leqslant n$, solves the (T,ϵ,N)-SPARSECCE problem for $\mathcal{G}$ using Q generative model oracle queries. Then the following conclusions hold:*

- *For any $\delta > 0$, the m-player $(\lfloor n/m\rfloor, 16(m+1)\cdot\epsilon)$-NASH problem for any normal-form game G can be solved, with failure probability δ, using at most $C_0\cdot(Q\cdot\log(1/\delta)) + (\log(1/\delta)\cdot nm/\epsilon)^{C_0}$ queries to a payoff oracle $\mathcal{O}_G$ for G.*
- *If the algorithm $\mathscr{B}$ additionally runs in time U for some $U \in \mathbb{N}$, then the algorithm solving NASH from the previous bullet point runs in time $(nmTNU\log(1/\delta)/\epsilon)^{C_0}$.*

Theorem 4.3 follows directly from Theorem E.1 by taking $m = 2$.

Proof of Theorem 4.3. Suppose there is an algorithm which, given the description of any 3-player Markov game $\mathcal{G}$ with $|\mathcal{G}| \leqslant n$, solves the (T,ϵ,N)-SPARSECCE problem in time U. Such an algorithm immediately yields an algorithm which can solve the (T,ϵ,N)-SPARSECCE problem in time $U + |\mathcal{G}|^{O(1)}$ using only a generative model oracle, since the exact description of the Markov game can be obtained with $HS|\mathcal{A}| \leqslant HS(\max_i A_i)^3 \leqslant |\mathcal{G}|^5$ queries to the generative model (across all $(h,s,\mathbf{a})$ tuples). We can now solve the problem of computing a $50\cdot\epsilon$-Nash equilibrium of a given 2-player $\lfloor n/2\rfloor$-action normal form game G as follows. We simply apply the algorithm of Theorem E.1 with $m = 2$, noting that the oracle $\mathcal{O}_G$ in the theorem statement can be implemented by reading the corresponding bits of input of the input game G. The second bullet point yields that this algorithm

takes time $(nTNU\log(1/\delta)/\epsilon)^{C_0}$, for some constant C_0. Furthermore, the assumption $T<\exp(\epsilon^2\cdot\lfloor n/m\rfloor/m^2)$ of Theorem E.1 is implied by the assumption that $T<\exp(\epsilon^2 n/16)$ of Theorem 4.3. □

In a similar manner, Theorem 5.2 follows from Theorem E.1 by applying Theorem C.3, which states that there is no randomized algorithm that finds approximate Nash equilibria of m-player, 2-action normal form games in time $2^{o(m)}$.

Proof of Theorem 5.2. Let ϵ_0 be the constant from Theorem C.3, and consider any $m\geqslant 3$. Suppose there is an algorithm which, for any m-player Markov game $\mathcal{G}$ with $|\mathcal{G}|\leqslant 2m^6$, makes Q oracle queries to a generative model oracle for $\mathcal{G}$, and solves the $(T,\epsilon_0/(10m),N)$-SPARSECCE problem for $\mathcal{G}$ for some $T,N\in\mathbb{N}$ so that $T<\exp(cm)$, for a sufficiently small absolute constant c. Then, by Theorem E.1 with $\epsilon=\epsilon_0/(10m)$ and $n=m^6$ (which ensures that $T<\exp((\epsilon_0/(10m))^2\cdot\lfloor n/m\rfloor/m^2)$ as long as c is sufficiently small), there is an algorithm which solves the (m^5,ϵ_0)-NASH problem—and thus the $(2,\epsilon_0)$-NASH problem—for $(m-1)$-player games with failure probability $1/3$, using $O(Q)+m^{O(1)}$ queries to a payoff oracle. But by Theorem C.3, any such algorithm requires $2^{\Omega(m)}$ queries to a payoff oracle. It follows that $Q\geqslant 2^{\Omega(m)}$, as desired. □

E.1. Proof of Theorem E.1

Proof of Theorem E.1. Fix any $m\geqslant 2$, $n\in\mathbb{N}$. Suppose we are given an algorithm $\mathscr{B}$ that solves the $(m+1)$-player (T,ϵ,N)-SPARSECCE problem for Markov games $\mathcal{G}$ satisfying $|\mathcal{G}|\leqslant n$, running in time U and using at most Q generative model queries. We proceed to describe an algorithm which solves the m-player $(\lfloor n/m\rfloor,16(m+1)\cdot\epsilon)$-NASH problem using $C_0\cdot(Q\cdot\log(1/\delta))+(\log(1/\delta)\cdot nm/\epsilon)^{C_0}$ queries to a payoff oracle, and running in time $(nmTNU\log(1/\delta)/\epsilon)^{C_0}$, where δ represents the failure probability. Define $n_0:=\lfloor n/m\rfloor$, and assume we are given an arbitrary m-player n_0-action normal form G, which is specified by payoff matrices $M_1,\ldots,M_m\in[0,1]^{n_0\times\cdots\times n_0}$. We assume that all entries of each of the matrices M_j have only the most significant $\max\{n_0,\lceil\log 1/\epsilon\rceil\}$ bits nonzero; this assumption is without loss of generality, since by truncating the utilities to satisfy this assumption, we change all payoffs by at most ϵ, which degrades the quality of any approximate equilibrium by at most 2ϵ (in addition, we have $\lceil\log 1/\epsilon\rceil\leqslant n_0$ since we have assumed $1<T<\exp(\epsilon^2 n_0/m^2)$). We assume $\epsilon\leqslant 1/2$ without loss of generality. Based on G, we construct an $(m+1)$-player Markov game $\mathcal{G}:=\mathcal{G}(G)$ as follows.

Definition E.2. We define the Markov game $\mathcal{G}(G)$ as the tuple $\mathcal{G}(G)=(\mathcal{S},H,(\mathcal{A}_i)_{i\in[2]},\mathbb{P},(R_i)_{i\in[2]},\mu)$, where:

- The horizon of $\mathcal{G}$ is chosen to be the power of 2 satisfying $n_0\leqslant H<2n_0$.
- Let $A:=n_0$. The action spaces of agents $1,2,\ldots,m$ are given by $\mathcal{A}_1=\cdots=\mathcal{A}_m=[A]$. The action space of agent $m+1$ is

  $$\mathcal{A}_{m+1}=\{(j,a_j)\,:\,j\in[m],a_j\in\mathcal{A}_j\},$$

  so that $|\mathcal{A}_{m+1}|=Am\leqslant n$.

  We write $\mathcal{A}=\prod_{j=1}^{m}\mathcal{A}_j$ to denote the joint action space of the first m agents, and $\overline{\mathcal{A}}:=\prod_{j=1}^{m+1}\mathcal{A}_j$ to denote the joint action space of all agents.
- There is a single state, denoted by $\mathfrak{s}$, i.e., $\mathcal{S}=\{\mathfrak{s}\}$ (in particular, μ is a singleton distribution supported on $\mathfrak{s}$).
- For all $h\in[H]$, the reward for agent $j\in[m+1]$, given an action profile $\mathbf{a}=(a_1,\ldots,a_{m+1})$ at the unique state $\mathfrak{s}$, is as follows: writing $a_{m+1}=(j',a'_{j'})$, we have

  $$R_{j,h}(\mathfrak{s},\mathbf{a})=\overline{R}_{j,h}(\mathfrak{s},\mathbf{a})+\frac{1}{H}\cdot 2^{-3\lceil\log 1/\epsilon\rceil}\cdot\mathsf{enc}(\mathbf{a}),\tag{12}$$

  where $\overline{R}_{j,h}(\mathfrak{s},\mathbf{a})$ is defined per the kibitzer construction of (Borgs et al., 2008):

  $$\overline{R}_{j,h}(\mathfrak{s},\mathbf{a}):=\begin{cases}0 & :j\notin\{j',m+1\}\\ \frac{1}{H}\cdot\left((M_j)_{a_1,\ldots,a_m}-(M_j)_{a_1,\ldots,a'_{j'},\ldots,a_m}\right) & :j=j'\\ \frac{1}{H}\cdot\left((M_j)_{a_1,\ldots,a'_{j'},\ldots,a_m}-(M_j)_{a_1,\ldots,a_m}\right) & :j=m+1.\end{cases}\tag{13}$$

  In (12) above, $\mathsf{enc}(\mathbf{a})\in[0,1]$ is the binary representation of a binary encoding of the action profile $\mathbf{a}$. In particular, if the binary encoding of $\mathbf{a}$ is $(b_1,\ldots,b_N)$, with $b_i\in\{0,1\}$, then $\mathsf{enc}(\mathbf{a})=\sum_{i=1}^{N}2^{-i}\cdot b_i$. Note that $\mathsf{enc}(\mathbf{a})$ takes $N=O(m\log n_0)\leqslant O(m\log n)$ bits to specify.

Algorithm 2 Algorithm to compute Nash equilibrium used in proof of Theorem E.1.

Input:

Parameters $n,n_0,m,T\in\mathbb{N}$, $\delta=\epsilon/(6H)$, $K=\lceil 4\log(mn_0/\delta)/\epsilon^2\rceil$.

An m-player, n_0-action normal form game G, with utilies accessible by oracle $\mathcal{O}_G$.

An algorithm $\mathscr{B}$ for computing approximate CCE of Markov games.

Call the algorithm $\mathscr{B}$ on the $(m+1)$-player Markov game $\mathcal{G}=\mathcal{G}(G)$ constructed as in Definition E.2, which produces a sequence $\sigma^{(1)},\ldots,\sigma^{(T)}$, where each $\sigma^{(t)}=(\sigma_1^{(t)},\ldots,\sigma_{m+1}^{(t)})$ with $\sigma_j^{(t)}\in\Pi_j^{\mathrm{gen,rnd}}$. Here, we use the oracle $\mathcal{O}_G$ to simulate generative model oracle queries made by $\mathscr{B}$.

Draw $t^\star\in[T]$ uniformly at random.

For each $j\in[m]$, initialize $\tau_{j,0}$ to be an empty trajectory.

for $h\in[H]$: **do**

Set $s_h=\mathfrak{s}$ (per the transitions of $\mathcal{G}$).

For each $j\in[m]$, define $\widehat{q}_{j,h}:=\mathbb{E}_{t\sim\widetilde{q}_{j,h}}\left[\sigma_{j,h}^{(t)}(\tau_{j,h-1},s_h)\right]\in\Delta(\mathcal{A}_j)$, where $\widetilde{q}_{j,h}\in\Delta([T])$ is defined as follows: for $t\in[T]$,

$$\widetilde{q}_{j,h}(t):=\frac{\exp\left(-\sum_{g<h}\log\left(\frac{1}{\sigma_{j,g}^{(t)}(a_{j,g}|\tau_{j,g-1},s_g)}\right)\right)}{\sum_{t'=1}^{T}\exp\left(-\sum_{g<h}\log\left(\frac{1}{\sigma_{j,g}^{(t')}(a_{j,g}|\tau_{j,g-1},s_g)}\right)\right)}.$$

Draw K i.i.d. samples $\mathbf{a}_h^1,\ldots,\mathbf{a}_h^K\sim\times_{j\in[m]}\widehat{q}_{j,h}$.

For each $a'\in\mathcal{A}_{m+1}$, define $\widehat{R}_{m+1,h}(a'):=\frac{1}{K}\sum_{k=1}^{K}R_{m+1,h}(s_h,(\mathbf{a}_h^k,a'))$. Here, we use the oracle $\mathcal{O}_G$ to compute $R_{m+1,h}(s_h,(\mathbf{a}_h^k,a'))$ for each tuple $(\mathbf{a}_h^k,a')$.

For each $j\in[m]$, draw $a_{j,h}\sim\sigma_{j,h}^{(t^\star)}(\cdot|\tau_{j,h-1},s_h)$.

Choose the action $a_{m+1,h}$ of player $m+1$ as follows: *(Action $a_{m+1,h}$ is corresponds to the action selected by the policy $\pi_{m+1}^{\dagger}$ of player $m+1$ defined within the proof of Lemma E.3; this policy is well-defined because the action profiles of all players $i\in[m]$ can be extracted from the lower-order bits of player $m+1$'s reward)*

$$a_{m+1,h}:=\operatorname*{argmax}_{a'\in\mathcal{A}_{m+1}}\left\{\widehat{R}_{m+1,h}(a')\right\}. \tag{14}$$

For each $j\in[m+1]$, let $r_{j,h}=R_{j,h}(s_h,(a_{1,h},\ldots,a_{m+1,h}))$.

Each player j constructs $\tau_{j,h}$ by updating $\tau_{j,h-1}$ with $(s_h,a_{j,h},r_{j,h})$.

if $\widehat{R}_{m+1,h}(a_{m+1,h})\leqslant 14(m+1)\cdot\epsilon/H$ **then**

return $\widehat{q}_h:=\times_{j\in[m]}\widehat{q}_{j,h}$ as a candidate approximate Nash equilibrium for G.

end if

end for

if the for loop terminates without returning: return **fail**.

It is evident that this construction takes polynomial time and satisfies $|\mathcal{G}|\leqslant mn_0\leqslant n$. Furthermore, it is clear that a single generative model oracle call for the Markov game $\mathcal{G}$ (per Definition 5.1) can be implemented using at most 2 calls to the oracle $\mathcal{O}_G$ for the normal-form game G. We will now show by applying the algorithm $\mathscr{B}$ to $\mathcal{G}$, we can efficiently (in terms of runtime and oracle calls) compute a $16(m+1)\cdot\epsilon$-approximate Nash equilibrium for the original game G. To do so, we appeal to Algorithm 2.

Algorithm 2 proceeds as follows. First, it calls the algorithm $\mathscr{B}$ on the $(m+1)$-player Markov game $\mathcal{G}(G)$, using the oracle $\mathcal{O}_G$ to simulate $\mathscr{B}$'s calls to the generative model oracle for $\mathcal{G}$. By assumption, the algorithm $\mathscr{B}$ returns a sequence $\sigma^{(1)},\ldots,\sigma^{(T)}$ of product policies of the form $\sigma^{(t)}=(\sigma_1^{(t)},\ldots,\sigma_{m+1}^{(t)})$, so that each $\sigma_j^{(t)}\in\Pi_j^{\mathrm{gen,rnd}}$ is N-computable, and so that the average $\overline{\sigma}:=\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{\sigma^{(t)}}$

is an ϵ-CCE of $\mathcal{G}$. Next, Algorithm 2 samples a trajectory from $\mathcal{G}$ in which:

- Players 1,...,m each play according to a policy $\sigma^{(t^\star)}$ for an index $t^\star \in [T]$ chosen uniformly at the start of the episode.
- Player $m+1$ plays according to a strategy that, at each step $h \in [H]$, computes distributions $\widehat{q}_{j,h}$ representing its "belief" of what action each player $j \in [m]$ will play at step h (Line 10), and plays an approximate best response to the product of the strategies $\widehat{q}_{j,h}$, $j \in [m]$ (Line 14).

In order avoid exponential dependence on the number of players m when computing an approximate best response to $\times_{j\in[m]}\widehat{q}_{j,h}$, we draw $K := \lceil 4\log(mn_0/\delta)/\epsilon^2 \rceil$ (for $\delta = \epsilon/(6H)$) samples from $\times_{j\in[m]}\widehat{q}_{j,h}$ and use these samples to compute the best response. In particular, letting $\mathbf{a}_h^K \in \mathcal{A}$ denote the kth sampled action profile, we construct a function $\widehat{R}_{m+1,h} : \mathcal{A}_{m+1} \to \mathbb{R}$ in Lines 11 and 14 which, for each $a' \in \mathcal{A}_{m+1}$, is defined as the average over samples $\{\mathbf{a}_h^k\}_{k\in[K]}$ of the realized payoffs $R_{m+1,h}(s_h, (\mathbf{a}_h^k, a'))$; note that to compute the payoffs for each sample, Algorithm 2 needs only two oracle calls to $\mathcal{O}_G$.

The following lemma, proven in the sequel, gives a correctness guarantee for Algorithm 2.

Lemma E.3 (Correctness of Algorithm 2)**.** *Given any m-player n_0-action normal form game G, if the algorithm $\mathscr{B}$ solves the (T,ϵ,N)-*SparseCCE *problem for the game $\mathcal{G}(G)$ with T,ϵ,N satisfying $T \leqslant \exp(n_0\epsilon^2/m^2)$, then Algorithm 2 outputs a $16(m+1)\cdot\epsilon$-approximate Nash equilibrium of G with probability at least $1/3$, and otherwise fails.*

The assumption that $T < \exp\left(\frac{\epsilon^2 \cdot \lfloor n/m \rfloor}{m^2}\right)$ from the statement of Theorem E.1 yields that $T \leqslant \exp(n_0\epsilon^2/m^2)$, so Lemma E.3 yields that Algorithm 2 outputs a $16(m+1)\cdot\epsilon$-Nash equilibrium of G with probability at least $1/3$ (and otherwise fails). By iterating Algorithm 2 for $\log(1/\delta)$ times, we may thus compute a $16(m+1)\cdot\epsilon$-Nash equilibrium of G with failure probability $1-\delta$.

We now analyze the oracle cost and computational cost of Algorithm 2. It takes $2Q$ oracle calls to $\mathcal{O}_G$ to simulate the Q generative model oracle calls of $\mathscr{B}$, and therefore, if $\mathscr{B}$ runs in time U, then the call to $\mathscr{B}$ on Line 5, using oracle calls to $\mathcal{O}_G$ to simulate simulate the generative model oracle calls, runs in time $O(U)$. Next, the computations of $\widetilde{q}_{j,h}$ (and thus $\widehat{q}_{j,h}$) in Line 10 can be performed in $(nmTN)^{O(1)}$ time, the computation of $\widehat{R}_{m+1,h} : \mathcal{A}_{m+1} \to \mathbb{R}$ in Line 14 requires time (and oracle calls to $\mathcal{O}_G$) bounded above by $O(|\mathcal{A}_{m+1}|\cdot K) \leqslant (nm\log(1/\delta)/\epsilon)^{O(1)}$, constructing the actions $a_{j,h}$ (for $j \in [m+1]$) in Lines 13 and 14 takes time $(Nmn)^{O(1)}$ (using the fact that the policies $\sigma_{j,h}^{(t^\star)}$ are N-computable), and constructing the rewards $r_{j,h}$ on Line 15 requires another $2(m+1)$ oracle calls to $\mathcal{O}_G$. Altogether, Algorithm 2 requires $2Q + (nm\log(1/\delta)/\epsilon)^{C_0}$ oracle calls to $\mathcal{O}_G$ and, if $\mathscr{B}$ runs in time U, then Algorithm 2 takes time $(nmTNU\log(1/\delta)/\epsilon)^{C_0}$, for some absolute constant C_0.

□

Remark E.4 (Bit complexity of exponential weights updates). In the above proof we have noted that $\widetilde{q}_{j,h}$ (as defined in Line 10 of Algorithm 2) can be computed in time $(nmTN)^{O(1)}$. A detail we do not handle formally is that, since the values of $\widetilde{q}_{j,h}(t)$ are in general irrational, only the $(nmTN)^{O(1)}$ most significant bits of each real number $\widetilde{q}_{j,h}(t)$ can be computed in time $(nmTN)^{O(1)}$. To give a truly polynomial-time implementation of Algorithm 2, one can compute only the $(nmTN)^{O(1)}$ most significant bits of each distribution $\widetilde{q}_{j,h}$, which is sufficient to approximate the true value of $\widehat{q}_{j,h}$ to within $\exp(-(nmTN)^{O(1)})$ in total variation distance. Since $\widehat{q}_{j,h}$ only influences the subsequent execution of Algorithm 2 via the samples $\mathbf{a}_h^1,\ldots,\mathbf{a}_h^K \sim \times_{j\in[m]}\widehat{q}_{j,h}$ drawn in Line 11, by a union bound, the approximation of $\widehat{q}_{j,h}$ we have described perturbs the execution of the algorithm by at most $O(KH)\cdot\exp(-(nmTN)^{O(1)})$ in total variation distance. In particular, the correctness guarantee of Lemma E.3 still holds, with sucess probability at least $1/3 - \exp(-(nmTN)^{O(1)}) > 1/4$.

It remains to prove Lemma E.3, which is the bulk of the proof of Theorem E.1.

Proof of Lemma E.3. We will establish the following two facts:

1. First, the choices of $a_{m+1,h}$ in Line 14 (i.e., Eq. 14) of Algorithm 2 correspond to a valid policy $\pi_{m+1}^\dagger \in \Pi^{\mathrm{gen,rnd}}$ for player $m+1$ (representing a strategy for deviating from the equilibrium $\overline{\sigma}$), in that they can be expressed as a function of player $(m+1)$'s history, $(\tau_{m+1,h-1}, s_h)$ at each step h.
2. Second, we will show that, since $\overline{\sigma}$ is an ϵ-CCE of $\mathcal{G}$, the strategy $\pi_{m+1}^\dagger$ cannot not lead to a large increase of value for player $m+1$, which will imply that Algorithm 2 must return a Nash equilibrium with high enough probability.

Defining $\pi_i^\dagger$ for $i \in [m+1]$. We begin by constructing the policy $\pi_{m+1}^\dagger$ described; for later use in the proof, it will be convenient to construct a collection of closely related policies $\pi_i^\dagger \in \Pi^{\mathrm{gen,rnd}}$ for $i \in [m]$, also representing strategies for deviating from the equilibrium $\overline{\sigma}$.

Let $i \in [m+1]$ be fixed. For $h \in [H]$, the mapping $\pi_{i,h}^\dagger : \mathscr{H}_{i,h-1} \times \mathcal{S} \to \mathcal{A}_i$ is defined as follows. Given a history $\tau_{i,h-1} = (s_1, a_{i,1}, r_{i,1}, \ldots, s_{h-1}, a_{i,h-1}, r_{i,h-1}) \in \mathscr{H}_{i,h-1}$ (we assume without loss of generality that $\tau_{i,h-1}$ occurs with positive probability under some sequence of general policies) and a current state s_h, we define $\pi_{i,h}^\dagger(\tau_{i,h-1}, s_h) \in \mathcal{A}_i$ through the following process.

1. First, we claim that for all players $j \in [m+1] \setminus \{i\}$, it is possible to extract the trajectory $\tau_{j,h-1}$ from the trajectory $\tau_{i,h-1}$ of player i.
   (a) Recall that for each $g < h$, from the definition in (12) and the function $\mathsf{enc}(\mathbf{a})$, the bits following position $3\lceil \log 1/\epsilon \rceil$ of the reward $r_{i,g}$ given to player i at step g of the trajectory $\tau_{i,g-1}$ encode an action profile $\mathbf{a}_g \in \overline{\mathcal{A}}$. Since $\tau_{i,h-1}$ occurs with positive probability, this is precisely the action profile which was played by agents at step g. Note we also use here that by definition of the rewards $R_{j,h}(s, \mathbf{a})$ in (12), the component $\overline{R}_{j,h}(s, \mathbf{a})$ of the reward only affects the first $2\lceil \log 1/\epsilon \rceil$ bits.
   (b) For $g < h$ and $j \in [m+1] \setminus \{i\}$, define $r_{j,g} := R_{j,g}(s_g, \mathbf{a}_g)$.
   (c) For $j \in [m+1] \setminus \{i\}$, write $\tau_{j,h-1} := (s_1, a_{j,1}, r_{j,1}, \ldots, s_{h-1}, a_{j,h-1}, r_{j,h-1})$; in particular, $\tau_{j,h-1}$ is a deterministic function of $(\tau_{i,h-1}, s_h)$. (Note that, since $\tau_{i,h-1}$ occurs with positive probability, the history $\tau_{j,h-1}$ observed by player j up to step $h-1$ can be computed from it via Steps (a) and (b)). Going forward, for $g < h-1$, we let $\tau_{j,g}$ denote the prefix of $\tau_{j,h-1}$ up to step g.
2. Now, using that player i can compute all players' trajectories, for each $j \in [m+1]$ we define

$$\widehat{q}_{j,h} := \mathbb{E}_{t \sim \widetilde{q}_{j,h}} \left[\sigma_{j,h}^{(t)}(\tau_{j,h-1}, s_h) \right] \in \Delta(\mathcal{A}_j), \tag{15}$$

   where $\widetilde{q}_{j,h} \in \Delta([T])$ is defined as follows: for $t \in [T]$,

$$\widetilde{q}_{j,h}(t) := \frac{\exp\left(-\sum_{g<h} \log\left(\frac{1}{\sigma_{j,g}^{(t)}(a_{j,g} | \tau_{j,g-1}, s_g)} \right) \right)}{\sum_{t'=1}^{T} \exp\left(-\sum_{g<h} \log\left(\frac{1}{\sigma_{j,g}^{(t')}(a_{j,g} | \tau_{j,g-1}, s_g)} \right) \right)}. \tag{16}$$

   Note that $\widehat{q}_{j,h}$ is a random variable which depends on the trajectory $(\tau_{j,h-1}, s_h)$ (which can be computed from $(\tau_{i,h-1}, s_h)$). In addition, the definition of $\widehat{q}_{j,h}$ (for each $j \in [m]$) is exactly as is defined in Line 10 of Algorithm 2.
3. For $i \in [m]$, define $\pi_{i,h}^\dagger(\tau_{i,h-1}, s_h)$ as follows:

$$\pi_{i,h}^\dagger(\tau_{i,h-1}, s_h) := \operatorname*{argmax}_{a' \in \mathcal{A}_i} \mathbb{E}_{\mathbf{a}_{-i} \sim \times_{j \neq i} \widehat{q}_{j,h}} \left[R_{m+1,h}(s_h, (a', \mathbf{a}_{-i})) \right]. \tag{17}$$

   For the case $i = m+1$, define $\pi_{m+1,h}^\dagger(\tau_{m+1,h-1}, s_h) \in \Delta(\mathcal{A}_{m+1})$ (implicitly) to be the following distribution over $a_{m+1,h}^\dagger \in \mathcal{A}_{m+1}$: draw $\mathbf{a}_h^1, \ldots, \mathbf{a}_h^K \sim \times_{j \in [m]} \widehat{q}_{j,h}$, define $\widehat{R}_{m+1,h}(a') := \frac{1}{K} \sum_{k=1}^K R_{m+1,h}(s_h, (\mathbf{a}_h^k, a'))$ for $a' \in \mathcal{A}_{m+1}$, and finally set

$$a_{m+1,h}^\dagger := \operatorname*{argmax}_{a' \in \mathcal{A}_{m+1}} \left\{ \widehat{R}_{m+1,h}(a') \right\}. \tag{18}$$

   Note that, for each choice of $(\tau_{m+1,h-1}, s_h)$, the distribution $\pi_{m+1,h}^\dagger(\tau_{m+1,h-1}, s_h)$ as defined above coincides with the distribution of the action $a_{m+1,h}^\dagger$ defined in Eq. 14 in Algorithm 2, when player $m+1$'s history is $\tau_{m+1,h-1}$ and the state at step h is s_h. The following lemma, for use later in the proof, bounds the approximation error incurred in sampling $\mathbf{a}_h^1, \ldots, \mathbf{a}_h^K \sim \times_{j \in [m]} \widehat{q}_{j,h}$.

Lemma E.5. *Fix any $(\tau_{m+1,h-1}, s_h) \in \mathscr{H}_{j,h-1}$. With probability at least $1-\delta$ over the draw of $\mathbf{a}_h^1, \ldots, \mathbf{a}_h^K \sim \times_{j \in [m]} \widehat{q}_{j,h}$, it holds that for all $a' \in \mathcal{A}_{m+1}$,*

$$\left| \widehat{R}_{m+1,h}(a') - \mathbb{E}_{a_j \sim \widehat{q}_{j,h} \ \forall j \in [m]} [R_{m+1,h}(s_h, (a_1, \ldots, a_m, a'))] \right| \leq \frac{\epsilon}{H},$$

which implies in particular that with probability at least $1-\delta$ *over the draw of* $a^\dagger_{m+1,h} \sim \pi^\dagger_{m+1,h}(\tau_{m+1,h-1}, s_h)$,

$$\max_{a' \in \mathcal{A}_{m+1}} \left\{ \mathbb{E}_{a_j \sim \widehat{q}_{j,h} \ \forall j \in [m]} [R_{m+1,h}(s_h, (a_1, \ldots, a_m, a'))] \right\} - \frac{2\epsilon}{H} \leqslant \mathbb{E}_{a_j \sim \widehat{q}_{j,h} \ \forall j \in [m]} [R_{m+1,h}(s_h, (a_1, \ldots, a_m, a^\dagger_{m+1,h}))]. \quad (19)$$

It is immediate from our construction above that the following fact holds.

Lemma E.6. *The joint distribution of* $\tau_{j,h}$*, for* $j \in [m+1]$ *and* $h \in [H]$*, as computed by Algorithm 2, coincides with the distribution of* $\tau_{j,h}$ *in an episode of* $\mathcal{G}$ *when players follow the policy* $\pi^\dagger_{m+1} \times \overline{\sigma}_{-(m+1)}$.

Analyzing the distributions $\widehat{q}_{j,h}$**.** Fix any $i \in [m+1]$. We next prove some facts about the distributions $\widehat{q}_{j,h}$ defined above (as a function of $(\tau_{i,h-1}, s_h)$) in the process of computing $\pi^\dagger_{i,h}(\tau_{i,h-1}, s_h)$.

For each $h \in [H]$, consider any choice of $(\tau_{i,h-1}, s_h) \in \mathscr{H}_{i,h-1} \times \mathcal{S}$; note that for each $j \in [m+1]$, the distributions $\widehat{q}_{j,h} \in \Delta(\mathcal{A}_j)$ for $h \in [H]$ may be viewed as an application Vovk's aggregating algorithm (Proposition D.2) in the following setting: the number of steps (T, in the context of Proposition D.2; note that T has a different meaning in the present proof) horizon is H, the context space is $\bigcup_{h=1}^{H} \mathscr{H}_{j,h-1} \times \mathcal{S}$, and the output space is $\mathcal{A}_j$. The expert set is $\mathcal{I} = \{\rho_j^{(1)}, \ldots, \rho_j^{(T)}\}$ (which has $|\mathcal{I}| = T$), and the experts' predictions on a context $(\tau_{j,h-1}, s) \in \mathscr{H}_{j,h-1} \times \mathcal{S}$ are defined via $\rho_j^{(t)}(\cdot | \tau_{j,h-1}, s) := \sigma_{j,h}^{(t)}(\cdot | \tau_{j,h-1}, s) \in \Delta(\mathcal{A}_j)$. Then for each $h \in [H]$, the distribution $\widehat{q}_{j,h}$ is obtained by updating the aggregating algorithm with the context-observation pairs $(\tau_{j,h'-1}, a_{j,h'})$ for $h' = 1, 2, \ldots, h-1$.

In more detail, fix any $t^\star \in [T]$ and $j \in [m+1]$ with $i \neq j$. We may apply Proposition D.2 with the number of steps set to H, the set of experts as $\mathcal{I} = \{\rho_j^{(1)}, \ldots, \rho_j^{(T)}\}$, and contexts and outcomes generated according to the distribution induced by running the policy $\pi^\dagger_i \times \sigma^{(t^\star)}_{-i}$ in the Markov game $\mathcal{G}$ as follows:

- For each $h \in [H]$, we are given, at steps $h' < h$, the actions $a_{k,h'}$ rewards $r_{k,h'}$ for all agents $k \in [m+1]$, as well as the states $s_1, \ldots, s_h$.
  - For each $k \in [m+1]$, set $\tau_{k,h-1} = (s_1, a_{k,1}, r_{k,1}, \ldots, s_{h-1}, a_{k,h-1}, r_{k,h-1})$ to be agent k's history.
  - The *context* fed to the aggregation algorithm at step h is $(\tau_{j,h-1}, s_h)$.
  - The *outcome* at step h is given by $a_{j,h} \sim \sigma^{(t^\star)}_{j,h}(\cdot | \tau_{j,h-1}, s_h)$; note that this choice satisfies the realizability assumption in Proposition D.2.
  - To aid in generating the next context at step $h+1$, choose $a_{k,h} \sim \sigma^{t^\star}_{k,h}(\tau_{k,h-1}, s_h)$ for all $k \in [m+1] \setminus \{i, j\}$ and $a_{i,h} = \pi^\dagger_{i,h}(\tau_{i,h-1}, s_h)$. Then set s_{h+1} to be the next state given the transitions of $\mathcal{G}$ and the action profile $\mathbf{a}_h = (a_{1,h}, \ldots, a_{m+1,h})$.

By Proposition D.2, it follows that for any fixed $t^\star \in [T]$ and $j \in [m+1]$ with $j \neq i$, under the process described above we have

$$\mathbb{E}_{\pi^\dagger_i \times \sigma^{(t^\star)}_{-i}} \left[\sum_{h=1}^{H} D_{\mathsf{TV}}(\sigma^{(t^\star)}_{j,h}(\tau_{j,h-1}, s_h), \widehat{q}_{j,h}) \right] \leqslant \sqrt{H \cdot \log T}. \quad (20)$$

Analyzing the value of $\pi^\dagger_{m+1}$**.** Next, using the development above, we show that if Algorithm 2 successfully computes a Nash equilibrium with constant probability (via $\pi^\dagger_{m+1}$) whenever $\bar{\sigma}$ is an ϵ-CCE. We first state the following claim, which is proven in the sequel by analyzing the values $V_i^{\pi^\dagger_i \times \overline{\sigma}_{-i}}$ for $i \in [m]$.

Lemma E.7. *If* $\overline{\sigma}$ *is an* ϵ*-CCE of* $\mathcal{G}$*, then it holds that for all* $i \in [m]$,

$$V_i^{\overline{\sigma}} \geqslant -\epsilon - m\sqrt{\log(T)/H}.$$

Note that in the game $\mathcal{G}$, since for all $h \in [H]$, $s \in \mathcal{S}$ and $\mathbf{a} \in \overline{\mathcal{A}}$, it holds that $\left| \sum_{j=1}^{m+1} R_{j,h}(s, \mathbf{a}) \right| \leqslant \frac{(m+1)\epsilon^2}{H}$ (which holds since in (12), $\mathsf{enc}(\mathbf{a})$ is multiplied by $\frac{1}{H} \cdot 2^{-3\lceil \log 1/\epsilon \rceil}$), it follows that $\left| \sum_{j=1}^{m+1} V_j^{\overline{\sigma}} \right| \leqslant (m+1)\epsilon^2$. Thus, by Lemma E.7, we have $V_{m+1}^{\overline{\sigma}} \leqslant (m+1)\epsilon^2 + m \cdot (\epsilon + m\sqrt{\log(T)/H})$, and since $\overline{\sigma}$ is an ϵ-CCE of $\mathcal{G}$ it follows that

$$V_{m+1}^{\pi^\dagger_{m+1} \times \overline{\sigma}_{-(m+1)}} \leqslant 2(m+1) \cdot \epsilon + m^2 \cdot \sqrt{\log(T)/H}. \quad (21)$$

To simplify notation, we will write $\widehat{q}_h := \widehat{q}_{1,h} \times \cdots \times \widehat{q}_{m,h}$ in the below calculations, where we recall that each $\widehat{q}_{j,h}$ is determined given the history up to step h, $(\tau_{j,h-1}, s_h)$, as defined in (15) and (16). An action profile drawn from $\widehat{q}_h$ is denoted as $\mathbf{a} \sim \widehat{q}_h$, with $\mathbf{a} \in \mathcal{A}$. We may now write $V_{m+1}^{\pi_{m+1}^\dagger \times \overline{\sigma}_{-(m+1)}}$ as follows:

$$
\begin{aligned}
&V_{m+1}^{\pi_{m+1}^\dagger \times \overline{\sigma}_{-(m+1)}} \\
=&\mathbb{E}_{t^\star \sim [T]} \sum_{h=1}^{H} \mathbb{E}_{\pi_{m+1}^\dagger \times \sigma_{-(m+1)}^{(t^\star)}} \mathbb{E}_{\substack{a_{j,h} \sim \sigma_{j,h}^{(t^\star)}(\tau_{j,h-1}, s_h) \ \forall j \in [m] \\ a_{m+1,h} \sim \pi_{m+1,h}^\dagger(\tau_{m+1,h-1}, s_h) \\ \mathbf{a} := (a_{1,h}, \ldots, a_{m+1,h})}} [R_{m+1,h}(s_h, \mathbf{a})] \\
\geqslant&\mathbb{E}_{t^\star \sim [T]} \sum_{h=1}^{H} \mathbb{E}_{\pi_{m+1}^\dagger \times \sigma_{-(m+1)}^{(t^\star)}} \left(\mathbb{E}_{\substack{a_{j,h} \sim \widehat{q}_{j,h} \ \forall j \in [m] \\ a_{m+1,h} \sim \pi_{m+1,h}^\dagger(\tau_{m+1,h-1}, s_h) \\ \mathbf{a} := (a_{1,h}, \ldots, a_{m+1,h})}} [R_{m+1,h}(s_h, \mathbf{a})] \right. \\
&\qquad \left. - \frac{1}{H} \sum_{j \in [m]} D_{\mathsf{TV}}(\sigma_{j,h}^{(t^\star)}(\tau_{j,h-1}, s_h), \widehat{q}_{j,h}) \right) \\
\geqslant&\mathbb{E}_{t^\star \sim [T]} \sum_{h=1}^{H} \mathbb{E}_{\pi_{m+1}^\dagger \times \sigma_{-(m+1)}^{(t^\star)}} \left(\max_{a'_{m+1,h} \in \mathcal{A}_{m+1}} \mathbb{E}_{\mathbf{a} \sim \widehat{q}_h} \left[R_{m+1,h}(s_h, (\mathbf{a}, a'_{m+1,h})) \right] - \frac{2\epsilon}{H} - \frac{\delta}{H} \right. \\
&\qquad \left. - \frac{1}{H} \sum_{j \in [m]} D_{\mathsf{TV}}(\sigma_{j,h}^{(t^\star)}(\tau_{j,h-1}, s_h), \widehat{q}_{j,h}) \right) \\
\geqslant&\frac{1}{H} \cdot \mathbb{E}_{t^\star \sim [T]} \sum_{h=1}^{H} \mathbb{E}_{\pi_{m+1}^\dagger \times \sigma_{-(m+1)}^{(t^\star)}} \left(\max_{j \in [m], a'_{j,h} \in \mathcal{A}_j} \mathbb{E}_{\mathbf{a} \sim \widehat{q}_h} [(M_j)_{a'_j, \mathbf{a}_{-j}} - (M_j)_{\mathbf{a}}] \right) - \frac{m}{H} \cdot \sqrt{H \log T} - 2\epsilon - \delta - \epsilon^2,
\end{aligned}
$$

where:

- The first inequality follows from the fact that $R_{m+1,h}(\cdot)$ takes values in $[-1/H, 1/H]$ and the fact that the total variation between product distributions is bounded above by the sum of total variation distances between each of the pairs of component distributions.
- The second inequality follows from the inequality (19) of Lemma E.5.
- The final equality follows from the definition of the rewards in (12) and (13), and by summing (20) over $j \in [m]$. We remark that the $-\epsilon^2$ term in the final line comes from the term $\frac{1}{H} \cdot 2^{-3\lceil \log 1/\epsilon \rceil} \cdot \mathsf{enc}(\mathbf{a})$ in (12).

Rearranging and using (21) as well as the fact that $\delta + \epsilon^2 = \epsilon/(6H) + \epsilon^2 \leqslant \epsilon$ (as $\epsilon \leqslant 1/2$), we get that

$$
\begin{aligned}
&\mathbb{E}_{t^\star \sim [T]} \mathbb{E}_{\pi_{m+1}^\dagger \times \sigma_{-(m+1)}^{(t^\star)}} \sum_{h=1}^{H} \left(\max_{j \in [m], a'_{j,h} \in \mathcal{A}_j} \mathbb{E}_{\mathbf{a} \sim \widehat{q}_h} [(M_j)_{a'_j, \mathbf{a}_{-j}} - (M_j)_{\mathbf{a}}] \right) \\
\leqslant&2H \cdot \epsilon \cdot (m+1) + (m+1)m \cdot \sqrt{H \log T} + 3H\epsilon.
\end{aligned}
$$

Since $\widehat{q}_h$ is a product distribution a.s., we have that

$$
\max_{j \in [m], a'_{j,h} \in \mathcal{A}_j} \mathbb{E}_{\mathbf{a} \sim \widehat{q}_h} [(M_j)_{a'_j, \mathbf{a}_{-j}} - (M_j)_{\mathbf{a}}] \geqslant 0.
$$

Therefore, by Markov's inequality, with probability at least $1/2$ over the choice of $t^\star \sim [T]$ and the trajectories $(\tau_{j,h-1}, s_h) \sim \pi_{m+1}^\dagger \times \sigma_{-(m+1)}^{(t^\star)}$ for $j \in [m]$ (which collectively determine $\widehat{q}_h$), there is some $h \in [H]$ so that

$$
\max_{j \in [m], a'_{j,h} \in \mathcal{A}_j} \mathbb{E}_{\mathbf{a} \sim \widehat{q}_h} [(M_j)_{a'_j, \mathbf{a}_{-j}} - (M_j)_{\mathbf{a}}] \leqslant 10(m+1) \cdot \epsilon + 2(m+1)m \cdot \sqrt{\log(T)/H} \leqslant 12(m+1) \cdot \epsilon, \tag{22}
$$

where the final inequality follows as long as $H \cdot \epsilon^2 \geqslant m^2 \log T$, i.e., $T \leqslant \exp\left(\frac{H \cdot \epsilon^2}{m^2}\right)$, which holds since $H \geqslant n_0$ and we have assumed that $T \leqslant \exp(\epsilon^2 \cdot n_0 / m^2)$.

Note that (22) implies that with probability at least $1/2$ under an episode drawn from $\pi^{\dagger}_{m+1} \times \overline{\sigma}_{-(m+1)}$, there is some $h \in [H]$ so that $\widehat{q}_h$ is a $12(m+1)\cdot\epsilon$-Nash equilibrium of the stage game G. Thus, by Lemma E.6, with probability at least $1/2$ under an episode drawn from the distribution of Algorithm 2, there is some $h \in [H]$ so that $\widehat{q}_h$ is a $12(m+1)\cdot\epsilon$-Nash equilibrium of G.

Finally, the following two observations conclude the proof of Lemma E.3.

- If $\widehat{q}_h$ is a $12(m+1)\cdot\epsilon$-Nash equilibrium of G, then by definition of the reward function $R_{m+1,h}(\cdot)$ in (12), upper bounding $\frac{1}{H}\cdot 2^{-3\lceil \log 1/\epsilon \rceil}\cdot \mathsf{enc}(\mathbf{a})$ by ϵ^2/H,

$$\max_{a' \in \mathcal{A}_{m+1}} \mathbb{E}_{\mathbf{a}\sim\widehat{q}_h}\left[R_{m+1,h}(\mathfrak{s},(\mathbf{a},a'))\right] \leqslant \frac{1}{H}\cdot 12(m+1)\cdot\epsilon + \frac{\epsilon^2}{H},$$

  which implies, by Lemma E.5, that with probability at least $1-\delta$ over the draw of $\mathbf{a}^1_h,\ldots,\mathbf{a}^K_h$,

$$\max_{a' \in \mathcal{A}_{m+1}} \left\{\widehat{R}_{m+1,h}(a')\right\} \leqslant \frac{1}{H}\cdot 12(m+1)\cdot\epsilon + \frac{\epsilon^2}{H} + \frac{\epsilon}{H} \leqslant \frac{1}{H}\cdot 14(m+1)\cdot\epsilon,$$

  i.e., the check in Line 17 of Algorithm 2 will pass and the algorithm will return $\widehat{q}_h$ (if step h is reached).

- Conversely, if $\max_{a' \in \mathcal{A}_{m+1}} \left\{\widehat{R}_{m+1,h}(a')\right\} \leqslant 14(m+1)\cdot\epsilon$, i.e., the check in Line 17 passes, then by Lemma E.5, with probability at least $1-\delta$ over $\mathbf{a}^1_h,\ldots,\mathbf{a}^K_h$,

$$\max_{a' \in \mathcal{A}_{m+1}} \mathbb{E}_{\mathbf{a}\sim\widehat{q}_h}\left[R_{m+1,h}(\mathfrak{s},(\mathbf{a},a'))\right] \leqslant \frac{1}{H}\cdot 14(m+1)\cdot\epsilon + \frac{\epsilon}{H} \leqslant \frac{1}{H}\cdot 15(m+1)\cdot\epsilon,$$

  which implies, by the definition of $R_{m+1,h}(\cdot)$ in (12) and (13), that $\widehat{q}_h$ is a $16(m+1)\cdot\epsilon$-Nash equilibrium of G.

Taking a union bound over all H of the probability-δ failure events from Lemma E.5 for the sampling $\mathbf{a}^1_h,\ldots,\mathbf{a}^K_h \sim \widehat{q}_h$ (for $h \in [H]$), as well as over the probability-$1/2$ event that there is no $\widehat{q}_h$ which is a $12(m+1)\cdot\epsilon$-Nash equilibrium of G, we obtain that with probability at least $1-1/2-H\cdot\epsilon/(6H) \geqslant 1/3$, Algorithm 2 outputs a $16(m+1)\cdot\epsilon$-Nash equilibrium of G. □

Finally, we prove the remaining claims stated without proof above.

Proof of Lemma E.5. Since $R_{m+1,h}(\mathfrak{s},\mathbf{a}) \in [-1/H, 1/H]$ for each $\mathbf{a} \in \overline{\mathcal{A}}$, by Hoeffding's inequality, for any fixed $a' \in \mathcal{A}_{m+1}$, with probability at least $1-\delta/|\mathcal{A}_{m+1}| = 1-\delta/(mn_0)$ over the draw of $\mathbf{a}^1_h,\ldots,\mathbf{a}^K_h \sim \times_{j\in[m]} \widehat{q}_{j,h}$, it holds that

$$\left|\widehat{R}_{m+1,h}(a') - \mathbb{E}_{a_j\sim\widehat{q}_{j,h}\ \forall j\in[m]}[R_{m+1,h}(s_h,(a_1,\ldots,a_m,a'))]\right| \leqslant \frac{2}{H}\cdot\sqrt{\frac{\log mn_0/\delta}{K}} \leqslant \frac{\epsilon}{H},$$

where the final inequality follows from the choice of $K = \lceil 4\log(mn_0/\delta)/\epsilon^2 \rceil$. The statement of the lemma follows by a union bound over all $|\mathcal{A}_{m+1}|$ actions $a' \in \mathcal{A}_{m+1}$. □

Proof of Lemma E.7. Fix any agent $i \in [m]$. We will argue that the policy $\pi^{\dagger}_i \in \Pi^{\mathrm{gen,det}}_i$ defined within the proof of Lemma E.3 satisfies $V_i^{\pi^{\dagger}_i,\overline{\sigma}_{-i}} \geqslant -m\sqrt{\log(T)/H}$. Since $\overline{\sigma}$ is an ϵ-CCE of $\mathcal{G}$, it follows that

$$\epsilon \geqslant V_i^{\pi^{\dagger}_i,\overline{\sigma}_{-i}} - V_i^{\overline{\sigma}} \geqslant -m\sqrt{\log(T)/H} - V_i^{\overline{\sigma}},$$

from which the result of Lemma E.7 follows after rearranging terms. To simplify notation, let us write $\widehat{q}_{-i,h} := \times_{j\neq i}\widehat{q}_{j,h}$, where we recall that each $\widehat{q}_{j,h}$ is determined given the history up to step h, $(\tau_{j,h-1}, s_h)$, as defined in (15) and (16). An action profile

drawn from $\widehat{q}_{-i,h}$ is denoted by $\mathbf{a}_{-i} \sim \widehat{q}_{-i,h}$, with $\mathbf{a}_{-i} \in \overline{\mathcal{A}}_{-i}$. We compute

$$\begin{aligned}
&V_i^{\pi_i^\dagger \times \overline{\sigma}_{-i}} \\
=&\mathbb{E}_{t^\star \sim [T]} \sum_{h=1}^{H} \mathbb{E}_{\pi_i^\dagger \times \sigma_{-i}^{(t^\star)}} \mathbb{E}_{\mathbf{a}_{-i} \sim \times_{j \neq i} \sigma_{j,h}^{(t^\star)}(\tau_{j,h-1}, s_h)} \Big[R_{i,h}(s_h, (\pi_{i,h}^\dagger(\tau_{i,h-1}, s_h), \mathbf{a}_{-i})) \Big] \\
\geqslant&\mathbb{E}_{t^\star \sim [T]} \sum_{h=1}^{H} \mathbb{E}_{\pi_i^\dagger \times \sigma_{-i}^{(t^\star)}} \left(\mathbb{E}_{\mathbf{a}_{-i} \sim \widehat{q}_{-i,h}} \Big[R_{i,h}(s_h, (\pi_{i,h}^\dagger(\tau_{i,h-1}, s_h), \mathbf{a}_{-i})) \Big] - \frac{1}{H} \sum_{j \neq i} D_{\mathsf{TV}}(\sigma_{j,h}^{(t^\star)}(\tau_{j,h-1}, s_h), \widehat{q}_{j,h}) \right) \\
\geqslant&\mathbb{E}_{t^\star \sim [T]} \sum_{h=1}^{H} \mathbb{E}_{\pi_i^\dagger \times \sigma_{-i}^{(t^\star)}} \left(\max_{a_i' \in \mathcal{A}_i} \mathbb{E}_{\mathbf{a}_{-i} \sim \widehat{q}_{-i,h}} \big[R_{i,h}(s_h, (a_i', \mathbf{a}_{-i})) \big] \right) - \frac{m}{H} \cdot \sqrt{H \log T} \\
\geqslant&-m \sqrt{\log(T)/H},
\end{aligned}$$

where:

- The first inequality follows from the fact that the rewards $R_{i,h}(\cdot)$ take values in $[-1/H, 1/H]$ and that the total variation between product distributions is bounded above by the sum of total variation distances between each of the pairs of component distributions.
- The second inequality follows from the definition of $\pi_{i,h}^\dagger(\tau_{i,h-1}, s_h)$ in terms of $\widehat{q}_{-i,h}$ in (17) as well as (20) applied to each $j \neq i$ and each $t^\star \in [T]$.
- The final inequality follows by Lemma E.8 below, applied to agent i and to the distribution $\widehat{q}_{-i,h}$, which we recall is a product distribution almost surely.

□

Lemma E.8. *For any $i \in [m]$, $s \in \mathcal{S}, h \in [H]$, and any product distribution $q \in \Delta(\overline{\mathcal{A}}_{-i})$, it holds that*

$$\max_{a_i' \in \mathcal{A}_i} \mathbb{E}_{\mathbf{a} \sim q} \big[R_{i,h}(s, (a_i', \mathbf{a})) \big] \geqslant 0.$$

Proof. Choose $a_i^\star := \operatorname{argmax}_{a_i' \in \mathcal{A}} \mathbb{E}_{\mathbf{a} \sim q} \big[(M_i)_{a_i', \mathbf{a}} \big]$. Now we compute

$$\begin{aligned}
H \cdot \mathbb{E}_{\mathbf{a} \sim q} \big[R_{i,h}(s, (a_i^\star, \mathbf{a})) \big] &\geqslant H \cdot \min_{a_{m+1}' \in \mathcal{A}_{m+1}} \mathbb{E}_{\mathbf{a} \sim q} \big[R_{i,h}(s, (a_i^\star, a_{m+1}', \mathbf{a}_{-(m+1)})) \big] \\
&\geqslant \min_{(j, a_j') \in \mathcal{A}_{m+1}} \mathbb{1}\{j = i\} \cdot \mathbb{E}_{\mathbf{a} \sim q} \big[(M_i)_{a_i^\star, \mathbf{a}} - (M_i)_{a_i', \mathbf{a}} \big] \\
&\geqslant 0,
\end{aligned}$$

where the first inequality follows since q is a product distribution, the second inequality uses that $\mathsf{enc}(\cdot)$ is non-negative, and the final inequality follows since by choice of $a_i^\star$ we have $\mathbb{E}_{\mathbf{a} \sim q} \big[(M_i)_{a_i^\star, \mathbf{a}} \big] \geqslant \mathbb{E}_{\mathbf{a} \sim q} \big[(M_i)_{a_i', \mathbf{a}} \big]$ for all $a_i' \in \mathcal{A}_i$. □

E.2. Remarks on bit complexity of the rewards

The Markov game $\mathcal{G}(G)$ constructed to prove Theorem E.1 uses lower-order bits of the rewards to record the action profile taken each step. These lower order bits may be used by each agent to infer what actions were taken by other agents at the previous step, and we use this idea to construct the best-response policies $\pi_i^\dagger$ defined in the proof. As a result of this aspect of the construction, the rewards of the game $\mathcal{G}(G)$ each take $O(m \cdot \log(n) + \log(1/\epsilon))$ bits to specify. As discussed in the proof of Theorem E.1, it is without loss of generality to assume that the payoffs of the given normal-form game G take $O(\log 1/\epsilon)$ bits each to specify, so when either $m \gg 1$ or $n \gg 1/\epsilon$, the construction of $\mathcal{G}(G)$ uses more bits to express its rewards than what is used for the normal-form game G.

It is possible to avoid this phenomenon by instead using the state transitions of the Markov game to encode the action profile taken at each step, as was done in the proof of Theorem 3.2. The idea, which we sketch here, is to replace the game $\mathcal{G}(G)$ of Definition E.2 with the following game $\mathcal{G}'(G)$:

Definition E.9 (Alternative construction to Definition E.2)**.** Given an m-player, n_0-action normal-form game G, we define the game $\mathcal{G}'(G)=(\mathcal{S},H,(\mathcal{A}_i)_{i\in[2]},\mathbb{P},(R_i)_{i\in[2]},\mu)$ as follows.

- The horizon of $\mathcal{G}$ is $H=n_0$.
- Let $A=n_0$. The action spaces of agents $1,2,...,m$ are given by $\mathcal{A}_1=\cdots=\mathcal{A}_m=[A]$. The action space of agent $m+1$ is

  $$\mathcal{A}_{m+1}=\{(j,a_j)\,:\,j\in[m],a_j\in\mathcal{A}_j\},$$

  so that $|\mathcal{A}_{m+1}|=Am\leqslant n$.

  We write $\mathcal{A}=\prod_{j=1}^{m}\mathcal{A}_j$ to denote the joint action space of the first m agents, and $\overline{\mathcal{A}}:=\prod_{j=1}^{m+1}\mathcal{A}_j$ to denote the joint action space of all agents. Then $|\overline{\mathcal{A}}|=A^m\cdot(mA)=mA^{m+1}\leqslant n$.
- The state space $\mathcal{S}$ is defined as follows. There are $|\overline{\mathcal{A}}|$ states, one for each action tuple $\mathbf{a}\in\overline{\mathcal{A}}$. For each $\mathbf{a}\in\overline{\mathcal{A}}$, we denote the corresponding state by $\mathfrak{s}_{\mathbf{a}}$.
- For all $h\in[H]$, the reward to agent $j\in[m+1]$ given action profile $\mathbf{a}=(a_1,...,a_{m+1})$ at any state $s\in\mathcal{S}$ is as follows: writing $a_{m+1}=(j',a'_{j'})$,

$$R_{j,h}(s,\mathbf{a}):=\begin{cases}0 & :j\notin\{j',m+1\}\\ \frac{1}{H}\cdot\left((M_j)_{a_1,...,a_m}-(M_j)_{a_1,...,a'_{j'},...,a_m}\right) & :j=j'\\ \frac{1}{H}\cdot\left((M_j)_{a_1,...,a'_{j'},...,a_m}-(M_j)_{a_1,...,a_m}\right) & :j=m+1.\end{cases}\tag{23}$$

- At each step $h\in[H]$, if action profile $\mathbf{a}\in\overline{\mathcal{A}}$ is taken, the game transitions to the state $\mathfrak{s}_{\mathbf{a}}$.

Note that the number of states of $\mathcal{G}'(G)$ is equal to $|\overline{\mathcal{A}}|=mn_0^{m+1}$, and so $|\mathcal{G}'(G)|=mn_0^{m+1}$. As a result, if we were to use the game $\mathcal{G}'(G)$ in place of $\mathcal{G}(G)$ in the proof of Theorem E.1, we would need to define $n_0:=\lfloor n^{1/(m+1)}/m\rfloor$ to ensure that $|\mathcal{G}'(G)|\leqslant n$, and so the condition $T<\exp(\epsilon^2\cdot\lfloor n/m\rfloor/m^2)$ would be replaced by $T<\exp(\epsilon^2\cdot\lfloor n^{1/(m+1)}/m\rfloor/m^2)$. This would only lead to a small quantitative degradement in the statement of Theorem 4.3, with the condition in the statement replaced by $T<\exp(c\cdot\epsilon^2\cdot n^{1/3})$ for some constant $c>0$. However, it would render the statement of Theorem 5.2 essentially vacuous. For this reason, we opt to go with the approach of Definition E.2 as opposed to Definition E.9.

We expect that the construction of Definition E.2 can nevertheless still be modified to use $O(\log 1/\epsilon)$ bits to express each reward in the Markov game $\mathcal{G}$. In particular, one could introduce stochastic transitions to encode in the state of the Markov game a small number of random bits of the full action profile played at each step. We leave such an approach for future work.

F. Equivalence between $\Pi_j^{\mathrm{gen,rnd}}$ and $\Delta(\Pi_j^{\mathrm{gen,det}})$

In this section we consider an alternate definition of the space $\Pi_i^{\mathrm{gen,rnd}}$ of randomized general policies of player i, and show that it is equivalent to the one we gave in Section 2.

In particular, suppose we were to define a randomized general policy of agent i as a distribution over deterministic general policies of agent i: we write $\widetilde{\Pi}_i^{\mathrm{gen,rnd}}:=\Delta(\Pi_i^{\mathrm{gen,det}})$ to denote the space of such distributions. Moreover, write $\widetilde{\Pi}^{\mathrm{gen,rnd}}:=\widetilde{\Pi}_1^{\mathrm{gen,rnd}}\times\cdots\times\widetilde{\Pi}_m^{\mathrm{gen,rnd}}=\Delta(\Pi_1^{\mathrm{gen,det}})\times\cdots\times\Delta(\Pi_m^{\mathrm{gen,det}})$ to denote the space of product distributions over agents' deterministic policies. Our goal in this section is to show that policies in $\widetilde{\Pi}^{\mathrm{gen,rnd}}$ are equivalent to those in $\Pi^{\mathrm{gen,rnd}}$ in the following sense: there is an embedding map $\mathsf{Emb}:\Pi^{\mathrm{gen,rnd}}\to\widetilde{\Pi}^{\mathrm{gen,rnd}}$, not depending on the Markov game, so that the distribution of a trajectory drawn from any $\sigma\in\Pi^{\mathrm{gen,rnd}}$, for any Markov game, is the same as the distribution of a trajectory drawn from $\mathsf{Emb}(\sigma)$ (Fact F.2). Furthermore, Emb is surjective in the following sense: any policy $\widetilde{\sigma}\in\widetilde{\Pi}^{\mathrm{gen,rnd}}$ produces trajectories that are distributed identically to those of $\mathsf{Emb}(\sigma)$ (and thus of σ), for some $\sigma\in\Pi^{\mathrm{gen,rnd}}$ (Fact F.3). In Definition F.1 below, we define Emb.

Definition F.1. For $j\in[m]$ and $\sigma_j\in\Pi_j^{\mathrm{gen,rnd}}$, define $\mathsf{Emb}_j(\sigma_j)\in\widetilde{\Pi}_j^{\mathrm{gen,rnd}}=\Delta(\Pi_j^{\mathrm{gen,det}})$ to put the following amount of mass on each $\pi_j\in\Pi_j^{\mathrm{gen,det}}$:

$$(\mathsf{Emb}_j(\sigma_j))(\pi_j):=\prod_{h=1}^{H}\prod_{(\tau_{j,h-1},s_h)\in\mathscr{H}_{j,h-1}\times\mathcal{S}}\sigma_j(\pi_{j,h}(\tau_{j,h-1},s_h)\mid\tau_{j,h-1},s_h)\tag{24}$$

Furthermore, for $\sigma=(\sigma_1,\ldots,\sigma_m)\in\Pi^{\mathrm{gen,rnd}}$, define $\mathsf{Emb}(\sigma)=(\mathsf{Emb}(\sigma_1),\ldots,\mathsf{Emb}(\sigma_m))$.

Note that, in the special case that $\sigma_j\in\Pi_j^{\mathrm{gen,det}}$, $\mathsf{Emb}_j(\sigma_j)$ is the point mass on σ_j.

Fact F.2 (Embedding equivalence)**.** *Fix a m-player Markov game $\mathcal{G}$ and, arbitrary policies $\sigma_j\in\Pi_j^{\mathrm{gen,rnd}}$. Then a trajectory drawn from the product policy $\sigma=(\sigma_1,\ldots,\sigma_m)\in\Pi_1^{\mathrm{gen,rnd}}\times\cdots\times\Pi_m^{\mathrm{gen,rnd}}$ is distributed identically to a trajectory drawn from $\mathsf{Emb}(\sigma)\in\widetilde{\Pi}^{\mathrm{gen,rnd}}$.*

The proof of Fact F.2 is provided in Section F.1. Next, we show that the mapping Emb is surjective in the following sense:

Fact F.3 (Right inverse of Emb_j)**.** *There is a mapping $\mathsf{Fac}:\widetilde{\Pi}^{\mathrm{gen,rnd}}\to\Pi^{\mathrm{gen,rnd}}$ so that for any Markov game $\mathcal{G}$ and any $\widetilde{\sigma}\in\widetilde{\Pi}^{\mathrm{gen,rnd}}$, the distribution of a trajectory drawn from $\widetilde{\sigma}$ is identical to the distribution of a trajectory drawn from $\mathsf{Emb}\circ\mathsf{Fac}(\widetilde{\sigma})$.*

We will write $\mathsf{Fac}((\widetilde{\sigma}_1,\ldots,\widetilde{\sigma}_m)):=(\mathsf{Fac}_1(\widetilde{\sigma}_1),\ldots,\mathsf{Fac}_m(\widetilde{\sigma}_m))$. Fact F.3 states that the policy $\mathsf{Fac}(\widetilde{\sigma})$ maps, under Emb, to a policy in $\widetilde{\Pi}^{\mathrm{gen,rnd}}$ which is equivalent to $\widetilde{\sigma}$ (in the sense that their trajectories are identically distributed for any Markov game).

An important consequence of Fact F.2 is that the expected reward (i.e., value) under any $\sigma\in\Pi^{\mathrm{gen,rnd}}$ is the same as that of $\mathsf{Emb}(\sigma)$. Thus given a Markov game, the induced normal-form game in which the players' pure action sets are $\Pi_1^{\mathrm{gen,rnd}},\ldots,\Pi_m^{\mathrm{gen,rnd}}$ is equivalent to the normal-form game in which the players' pure action sets are $\Pi_1^{\mathrm{gen,det}},\ldots,\Pi_m^{\mathrm{gen,det}}$, in the following sense: for any mixed strategy in the former, namely a product distributional policy $P\in\Delta(\Pi_1^{\mathrm{gen,rnd}})\times\cdots\times\Delta(\Pi_m^{\mathrm{gen,rnd}})$, the policy $\mathbb{E}_{\sigma\sim P}[\mathsf{Emb}(\sigma)]\in\Delta(\Pi_1^{\mathrm{gen,det}})\times\cdots\times\Delta(\Pi_m^{\mathrm{gen,det}})=\widetilde{\Pi}^{\mathrm{gen,rnd}}$ is a mixed strategy in the latter which gives each player the same value as under P. (Note that $\mathbb{E}_{\sigma\sim P}[\mathsf{Emb}(\sigma)]$ is indeed a product distribution since P is a product distribution and Emb factors into individual coordiantes.) Furthermore, by Fact F.3, any distributional policy in $\widetilde{\Pi}^{\mathrm{gen,rnd}}$ arises in this manner, for some $P\in\Delta(\Pi_1^{\mathrm{gen,rnd}})\times\cdots\times\Delta(\Pi_m^{\mathrm{gen,rnd}})$; in fact, P may be chosen to place all its mass on a single $\sigma\in\Pi_1^{\mathrm{gen,rnd}}\times\Pi_m^{\mathrm{gen,rnd}}$. Since Emb factors into individual coordinates, it follows that Emb yields a one-to-one mapping between the coarse correlated equilibria (or any other notion of equilibria, e.g., Nash equilibria or correlated equilibria) of these two normal-form games.

F.1. Proofs of the equivalence

Proof of Fact F.2. Consider any trajectory $\tau=(s_1,\mathbf{a}_1,\mathbf{r}_1,\ldots,s_H,\mathbf{a}_H,\mathbf{r}_H)$ consisting of a sequence of H states and actions and rewards for each of the m agents. Assume that $r_{i,h}=R_{i,h}(s,\mathbf{a}_h)$ for all i,h (as otherwise τ has probability 0 under any policy). Write:

$$p_\tau:=\prod_{h=1}^{H-1}\mathbb{P}_h(s_{h+1}|s_h,\mathbf{a}_h).$$

Then the probability of observing τ under σ is

$$p_\tau\cdot\prod_{h=1}^{H-1}\prod_{j=1}^{m}\sigma_{j,h}(a_{j,h}|\tau_{j,h-1},s_h) \tag{25}$$

where, per usual, $\tau_{j,h-1}=(s_1,a_{j,1},r_{j,1},\ldots,s_{h-1},a_{j,h-1},r_{j,h-1})$. Write $\sigma=(\sigma_1,\ldots,\sigma_m)=\mathsf{Emb}(\sigma)$. The probability of observing τ under σ is

$$p_\tau\cdot\prod_{j\in[m]}\sum_{\pi_j\in\Pi_j^{\mathrm{gen,det}}:\ \forall h,\ \pi(\tau_{j,h-1},s_h)=a_{j,h}}\sigma_j(\pi_j) \tag{26}$$

It is now straightforward to see from the definition of $\sigma_j(\pi_j)$ in (24) that the quantities in (25) and (26) are equal. □

Proof of Fact F.3. Fix a policy $\widetilde{\sigma}_j\in\widetilde{\Pi}_j^{\mathrm{gen,rnd}}=\Delta(\Pi_j^{\mathrm{gen,det}})$. We define $\mathsf{Fac}_j(\widetilde{\sigma}_j)$ to be the policy $\sigma_j\in\Pi_j^{\mathrm{gen,rnd}}$, which is defined as follows: for $\tau_{j,h-1}=(s_{j,1},a_{j,1},r_{j,1},\ldots,s_{j,h-1},a_{j,h-1},r_{j,h-1})\in\mathscr{H}_{j,h-1}$, $s_h\in\mathcal{S}$, we have, for $a_{j,h}\in\mathcal{A}_j$,

$$\sigma_j(\tau_{j,h-1},s_h)(a_{j,h})=\frac{\widetilde{\sigma}_j\Big(\{\pi_j\in\Pi_j^{\mathrm{gen,det}}\ :\ \pi_j(\tau_{j,g},s_g)=a_{j,g}\ \forall g\leqslant h\}\Big)}{\widetilde{\sigma}_j\Big(\{\pi_j\in\Pi_j^{\mathrm{gen,det}}\ :\ \pi_j(\tau_{j,g},s_g)=a_{j,g}\ \forall g\leqslant h-1\}\Big)}.$$

If the denominator of the above expression is 0, then $\sigma_j(\tau_{j,h-1}, s_h)$ is defined to be an arbitrary distribution on $\Delta(\mathcal{A}_j)$. (For concreteness, let us say that it puts all its mass on a fixed action in $\mathcal{A}_j$.) Furthermore, for $\widetilde{\sigma} \in \widetilde{\Pi}^{\mathrm{gen,rnd}}$, define $\mathsf{Fac}(\widetilde{\sigma}) := (\mathsf{Fac}_1(\widetilde{\sigma}_1),...,\mathsf{Fac}_m(\widetilde{\sigma}_m)) \in \Pi^{\mathrm{gen,rnd}}$.

Next, fix any $\widetilde{\sigma} = (\widetilde{\sigma}_1,...,\widetilde{\sigma}_m) \in \widetilde{\Pi}_1^{\mathrm{gen,rnd}} \times \cdots \times \widetilde{\Pi}_m^{\mathrm{gen,rnd}}$. Let $\sigma = \mathsf{Fac}(\widetilde{\sigma})$. By Fact F.2, it suffices to show that the distribution of trajectories under σ is the same as the distribution of trajectories drawn from σ.

So consider any trajectory $\tau = (s_1,\mathbf{a}_1,\mathbf{r}_1,...,s_H,\mathbf{a}_H,\mathbf{r}_H)$ consisting of a sequence of H states and actions and rewards for each of the m agents. Assume that $r_{i,h} = R_{i,h}(s,\mathbf{a}_h)$ for all i,h (as otherwise τ has probability 0 under any policy). Write:

$$p_\tau := \prod_{h=1}^{H-1} \mathbb{P}_h(s_{h+1}|s_h,\mathbf{a}_h).$$

Then the probability of observing τ under σ is

$$\begin{aligned}
&p_\tau \cdot \prod_{h=1}^{H}\prod_{j=1}^{m} \sigma_{j,h}(a_{j,h}|\tau_{j,h-1},s_h) \\
=&p_\tau \cdot \prod_{j=1}^{m}\prod_{h=1}^{H} \frac{\widetilde{\sigma}_j\Big(\{\pi_j \in \Pi_j^{\mathrm{gen,det}} \,:\, \pi_j(\tau_{j,g},s_g) = a_{j,g}\ \forall g \leqslant h\}\Big)}{\widetilde{\sigma}_j\Big(\{\pi_j \in \Pi_j^{\mathrm{gen,det}} \,:\, \pi_j(\tau_{j,g},s_g) = a_{j,g}\ \forall g \leqslant h-1\}\Big)} \\
=&p_\tau \cdot \prod_{j=1}^{m} \widetilde{\sigma}_j\Big(\{\pi_j \in \Pi_j^{\mathrm{gen,det}} \,:\, \pi_j(\tau_{j,g},s_g) = a_{j,g}\ \forall g \leqslant H\}\Big),
\end{aligned}$$

which is equal to the probability of observing τ under $\widetilde{\sigma}$. □