On the Complexity of Multi-Agent Decision Making:
From Learning in Games to Partial Monitoring

Dylan J. Foster   DYLANFOSTER@MICROSOFT.COM
Microsoft Research

Dean P. Foster   DEAN@FOSTER.NET
Amazon

Noah Golowich[*]   NZG@MIT.EDU
MIT

Alexander Rakhlin   RAKHLIN@MIT.EDU
MIT

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

A central problem in the theory of multi-agent reinforcement learning (MARL) is to understand what structural conditions and algorithmic principles lead to sample-efficient learning guarantees, and how these considerations change as we move from few to many agents. We study this question in a general framework for interactive decision making with multiple agents, encompassing Markov games with function approximation and normal-form games with bandit feedback. We focus on equilibrium computation, in which a centralized learning algorithm aims to compute an equilibrium by controlling multiple agents that interact with an (unknown) environment. Our main contributions are:

- We provide upper and lower bounds on the optimal sample complexity for multi-agent decision making based on a multi-agent generalization of the *Decision-Estimation Coefficient*, a complexity measure introduced by Foster et al. (2021) in the single-agent counterpart to our setting. Compared to the best results for the single-agent setting, our upper and lower bounds have additional gaps. We show that no "reasonable" complexity measure can close these gaps, highlighting a striking separation between single and multiple agents.

- We show that characterizing the statistical complexity for multi-agent decision making is equivalent to characterizing the statistical complexity of *single-agent* decision making, but with *hidden (unobserved) rewards*, a framework that subsumes variants of the partial monitoring problem. As a consequence of this connection, we characterize the statistical complexity for hidden-reward interactive decision making to the best extent possible.

Building on this development, we provide several new structural results, including 1) conditions under which the statistical complexity of multi-agent decision making can be reduced to that of single-agent, and 2) conditions under which the so-called *curse of multiple agents* can be avoided.

---

[*] Work done in part while interning at Microsoft Research.

1. Introduction

Many of the most exciting frontiers for artificial intelligence are game-theoretic in nature, and involve multiple agents with differing incentives interacting and making decisions in dynamic environments, either in cooperation or in competition. Numerous recent approaches, adopting the framework of *multi-agent reinforcement learning* (MARL), have achieved human-level performance in multi-agent game-playing domains (Silver et al., 2016; Brown and Sandholm, 2018; Perolat et al., 2022; Kramár et al., 2022; Bakhtin et al., 2022), and while there is great potential to apply MARL further in domains such as cybersecurity (Malialis and Kudenko, 2015), autonomous driving (Shalev-Shwartz et al., 2016), and economic policy (Zheng et al., 2022), sample-efficiency and reliability are obstacles for real-world deployment. Consequently, a central question is to understand what modeling assumptions and algorithm design principles lead to robust, sample-efficient learning guarantees. This issue is particularly salient in domains with high-dimensional feedback and decision spaces, where the use of flexible models such as neural networks is critical.

For reinforcement learning in single-agent settings, an extensive line of research identifies modeling assumptions (or, structural conditions) under which sample-efficient learning is possible (Russo and Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020; Du et al., 2021; Jin et al., 2021a; Foster et al., 2021). Notably, Foster et al. (2021, 2022b, 2023) provide a notion of statistical complexity, the *Decision-Estimation Coefficient* (DEC), which is both necessary and sufficient for low sample complexity, and leads to unified principles for algorithm design. For multi-agent reinforcement learning, structural conditions for sample-efficient learning have also received active investigation (Chen et al., 2022b; Li et al., 2022; Xie et al., 2020; Jin et al., 2022; Huang et al., 2021; Zhan et al., 2022; Liu et al., 2022), drawing inspiration from the single agent setting. However, insights from single agents do not always transfer to multiple agents in intuitive ways (Daskalakis et al., 2022), and development has largely proceeded on a case-by-case basis. As such, the problem of developing a unified understanding or *necessary* conditions for sample-efficient multi-agent reinforcement learning remained open.

Contributions. We consider a general framework, *Multi-Agent Decision Making with Structured Observations* (MA-DMSO), which generalizes the single-agent DMSO framework of Foster et al. (2021) and subsumes multi-agent reinforcement learning with general function approximation, as well as normal-form games with bandit feedback and structured action spaces. We focus on *centralized* equilibrium computation, where a centralized learning algorithm with control of all agents aims to compute an equilibrium by interacting with the (unknown) environment. Our main results are:

- **Complexity of multi-agent decision making.** We introduce a new complexity measure, the *Multi-Agent Decision-Estimation Coefficient*, generalizing the Decision-Estimation Coefficient of Foster et al. (2021, 2023), and show that it leads to upper and lower bounds on the optimal sample complexity for multi-agent decision making. Compared to the best results for the single-agent setting (Foster et al., 2023), our upper and lower bounds have additional gaps, which we show that *no (reasonable) complexity measure can close*.

- **Complexity of hidden-reward decision making.** We show that characterizing the statistical complexity for multi-agent decision making is equivalent to characterizing the statistical complexity of *single-agent decision making*, but with *hidden (unobserved) rewards*, a framework that we refer to as *Hidden-Reward Decision Making with Structured Observations* (HR-DMSO).

Leveraging this connection, we characterize the statistical complexity of the HR-DMSO framework, which encompasses PAC variants of the stochastic partial monitoring problem (Bartók et al., 2014), to the best extent possible (for any reasonable complexity measure).

- **Additional insights for multiple agents.** Building on the results above, we provide a number of new structural results and algorithmic insights for multi-agent decision making and RL, including 1) general conditions under which the complexity of multi-agent decision making can be reduced to that of single agent decision making, and 2) general conditions under which the so-called *curse of multiple agents* (Jin et al., 2021b) can be removed.

Organization. *Due to space constraints, the main body presents an informal overview of our results, with detailed statements deferred to Part I of the appendix. Examples are given in Part II. See Appendix A for an overview of appendix organization.*

1.1. Multi-agent interactive decision making (MA-DMSO)

We introduce a multi-agent generalization of the *Decision Making with Structured Observations* framework of Foster et al. (2021), which we refer to as *Multi-Agent Decision Making with Structured Observations* (MA-DMSO). The framework consists of T rounds of interaction between K agents and the environment. For each round $t = 1, 2, \ldots, T$:

1. The agents collectively select a *joint decision* $\pi^t \in \Pi$, where Π is the *joint decision space*.

2. Each agent $k \in [K]$ receives a *reward* $r_k^t \in \mathcal{R} \subseteq \mathbb{R}$ and a *pure observation* $o_\circ^t \in \mathcal{O}_\circ$ sampled via $(r_1^t, \ldots, r_K^t, o_\circ^t) \sim M^\star(\pi^t)$, where $M^\star : \Pi \to \Delta(\mathcal{R}^K \times \mathcal{O}_\circ)$ is the underlying *model*. We refer to $\mathcal{R}$ as the *reward space* and to $\mathcal{O}_\circ$ as the *pure observation space*. We call the tuple $(r_1^t, \ldots, r_K^t, o_\circ^t)$ consisting of all information revealed to agents on round t the *full observation*.

After the T rounds of interaction, the agents collectively output a joint decision $\widehat{\pi} \in \Pi$, which may be chosen in an arbitrary fashion based on the data observed over the T rounds, and may be randomized according to a distribution $p \in \Delta(\Pi)$. Their goal, which we formalize in the sequel, is to choose $\widehat{\pi}$ to be an equilibrium (e.g., Nash or CCE) for the average reward function induced by $M^\star$. The model $M^\star$, which is formalized as a probability kernel from decisions to full observations (Appendix B), is unknown to the agents, and is to be interpreted as the underlying environment.

The DMSO framework captures most online decision making problems in which a *single agent* interacts with an unknown environment, and the MA-DMSO framework further generalizes it to capture a wide variety of problems in *multi-agent* reinforcement learning. Examples include learning in normal-form games with bandit feedback (Rakhlin and Sridharan, 2013; Foster et al., 2016; Heliou et al., 2017; Wei and Luo, 2018; Giannou et al., 2021), where $M^\star$ represents the distribution over rewards for each entry in the game, and learning in Markov games with function approximation (Chen et al., 2022b; Li et al., 2022; Xie et al., 2020; Jin et al., 2022; Huang et al., 2021; Zhan et al., 2022; Liu et al., 2022), where $M^\star$ represents the underlying Markov game. Additional examples include normal-form games with structured (e.g., convex-concave) rewards and high-dimensional action spaces (Bravo et al., 2018; Maheshwari et al., 2022; Lin et al., 2021).

Realizability. While the model $M^\star$ is unknown, we make a standard *realizability* assumption.

Assumption 1.1 (Realizability for MA-DMSO). *The agents have access to a model class $\mathcal{M}$ consisting of probability kernels $M : \Pi \to \Delta(\mathcal{R}^K \times \mathcal{O}_\circ)$ that contains the true model $M^\star$.*

For normal-form games, the class $\mathcal{M}$ encodes structure in the rewards (e.g., linearity or convexity) or decision space, and for Markov games it encodes structure in transition probabilities or value functions. See Part II of the appendix for examples, as well as Foster et al. (2021) for with $K = 1$.

1.1.1. EQUILIBRIA

The goal of the agents in the MA-DMSO framework is to produce an equilibrium for the underlying game/model $M^\star$. We formalize the notion of equilibrium in a general fashion which encompasses several standard game-theoretic equilibria. To keep notation compact, we define $\mathcal{O} := \mathcal{R}^K \times \mathcal{O}_\circ$ to be the *full observation space*, and will write $o^t := (r_1^t, \ldots, r_K^t, o_\circ^t)$ to denote the (full) observation. For $M \in \mathcal{M}$ and $\pi \in \Pi$, let $\mathbb{E}^{M,\pi}[\cdot]$ denote expectation under the process $(r_1, \ldots, r_K, o_\circ) \sim M(\pi)$; in light of our notation $\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ$, we will sometimes denote this process as $o \sim M(\pi)$. For each $k \in [K]$ and $M \in \mathcal{M}$, define the mapping $f_k^M : \Pi \to \mathbb{R}$ by $f_k^M(\pi) = \mathbb{E}^{M,\pi}[r_k]$, which denotes agent k's expected reward under M when the joint decision π is played.

For each agent k, we assume they are given a *deviation space* Π_k', together with a *switching function*, $U_k : \Pi_k' \times \Pi \to \Pi$. Given a joint decision $\pi \in \Pi$, each agent k can choose a deviation $\pi_k' \in \Pi_k'$, which will have the effect that the joint policy played by agents is $U_k(\pi_k', \pi)$ instead of π. We aim for the output policy $\widehat{\pi} \sim p$ produced in the MA-DMSO setup to have the property that no agent can significantly increase their value by deviating. We quantify this via

$$\mathbf{Risk}(T) := \mathbb{E}_{\widehat{\pi} \sim p} \left[\sum_{k=1}^K \sup_{\pi_k' \in \Pi_k'} f_k^{M^\star}(U_k(\pi_k', \pi)) - f_k^{M^\star}(\pi) \right]. \tag{1}$$

For $M \in \mathcal{M}$ and $\pi \in \Pi$, we abbreviate $h^M(\pi) = \sum_{k=1}^K \sup_{\pi_k' \in \Pi_k'} f_k^M(U_k(\pi_k', \pi)) - f_k^M(\pi)$, so that $\mathbf{Risk}(T) := \mathbb{E}_{\widehat{\pi} \sim p}[h^{M^\star}(\widehat{\pi})]$. The quantity $h^M(\pi)$ measures the sum of players' incentives to deviate from the joint decision π under M; we say that π is an *equilibrium* for M if $h^M(\pi) = 0$.

The notion (1) captures standard notions of equilibria, including Nash equilibria, correlated equilibria (CE), and coarse correlated equilibria (CCE). As we have strived to make the setup in this section as general as possible, we make two regularity assumptions to rule out other, potentially pathological notions of equilibria.

Assumption 1.2 (Existence of equilibria). *For all $M \in \mathcal{M}$, there exists $\pi \in \Pi$ with $h^M(\pi) = 0$.*

Assumption 1.3 (Monotonicity of the optimal deviation). *For any model $M \in \mathcal{M}$, agent $k \in [K]$, and joint decision $\pi \in \Pi$, there is some deviation $\pi_k' \in \Pi_k'$ such that $f_k^M(U_k(\pi_k', \pi)) \geq f_k^M(\pi)$.*

Assumption 1.3 implies that, up to a factor of K, the notion of risk in (1) is equivalent to the maximal gain any agent can achieve by deviating. Both assumptions are satisfied by Nash equilibria, CE, and CCE (see Definitions 1.1, 1.2 and G.1).

Summarizing, the MA-DMSO framework captures the problem of equilibrium computation: the agents aim to find an (ε-approximate) equilibrium $\widehat{\pi}$ so that $\mathbf{Risk}(T) \leq \varepsilon$, but the underlying game is unknown, so they must gather information by interacting with it and exploring. We refer to the tuple $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ as an *instance* for the MA-DMSO framework. The instance $\mathscr{M}$ specifies all information known a-priori to the agents before the learning process begins.

Remark 1.1. As described, the MA-DMSO framework allows *centralized* learning protocols, in which a single learning algorithm may control all agents in a centralized fashion (equivalently,

unlimited communication and coordination is permitted amongst agents throughout the learning process). Lower bounds against centralized learning algorithms certainly apply to decentralized algorithms, being a special case of the former. However, in general there may be gaps between the minimax sample complexity for centralized and decentralized algorithms, and we leave a detailed investigation of decentralized multi-agent interactive decision-making for future work.

Remark 1.2. Our presentation of the MA-DMSO framework captures settings in which (multi-agent) learning algorithms are evaluated only on the proximity of output decision $\widehat{\pi}$ to equilibrium, as opposed to, say, the average proximity to equilibrium for the decisions played throughout the T rounds of learning. In the single-agent setting, such guarantees are often referred as PAC (Probability Approximately Correct) guarantees, as opposed to regret guarantees (Foster et al., 2023). It is fairly straightforward to extend many of our results to the regret setting.

1.1.2. EXAMPLES OF INSTANCES FOR MA-DMSO

We now highlight basic multi-agent bandit and MARL problems captured by the MA-DMSO framework. We describe how to capture concrete notions of equilibria, then give examples of instances $\mathscr{M}$.

Examples of equilibria. In Definitions 1.1 and 1.2 below, we specify the decision spaces, deviation spaces, and switching functions that can be used to capture *Nash equilibria* and *coarse correlated equilibria (CCE)*; see Appendix G.1 for further examples, including *correlated equilibria (CE)* and variants of CCE and CCE which have been studied in the context of Markov games.

Definition 1.1 (Nash equilibrium instance). *An MA-DMSO instance* $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ *is a* Nash equilibrium (NE) instance *if the following holds:*

1. *For sets* $\Pi_1, \ldots, \Pi_K$, *we have* $\Pi = \Pi_1 \times \cdots \times \Pi_K$. *For each* $k \in [K]$, *we have* $\Pi'_k = \Pi_k$.

2. *For each* $k \in [K]$, $\pi \in \Pi$, *and* $\pi'_k \in \Pi'_k$, *it holds that* $U_k(\pi'_k, \pi) = (\pi'_k, \pi_{-k})$.[1]

We say that the NE instance $\mathscr{M}$ *is a* two-player zero-sum NE instance *if* $K = 2$, *and for all* $M \in \mathcal{M}, \pi \in \Pi$, *it holds that* $f_1^M(\pi) + f_2^M(\pi) = 0$.

The notion of Nash equilibrium in Definition 1.1 encompasses, but goes well beyond the standard notion of mixed Nash equilibria in normal-form games (e.g., (Nisan et al., 2007)). In particular, Definition 1.1 does not assume that the decision spaces Π_k are distributions over a pure action space of player k. Therefore, it captures refined solution concepts including *pure* Nash equilibria in normal-form games (Daskalakis and Papadimitriou, 2006) and Markov Nash equilibria in Markov games (Example 1.2). As a result of this generality, an NE instance per Definition 1.1 is not guaranteed to satisfy Assumption 1.2, i.e., to have equilibria; nevertheless, we will ensure that all examples of NE instances we consider are constructed in such a way so that Assumption 1.2 is satisfied.

Definition 1.2 (Coarse correlated equilibrium instance). *An instance* $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ *for MA-DMSO is a* coarse correlated equilibrium (CCE) instance *if the following holds:*

1. *For some sets* $\Sigma_1, \ldots, \Sigma_K$ *(called* pure decisions*), we have* $\Pi = \Delta(\Sigma_1 \times \cdots \times \Sigma_K)$. *We will write* $\Sigma := \Sigma_1 \times \cdots \times \Sigma_K$. *For each* $k \in [K]$, *we have* $\Pi'_k = \Sigma_k \cup \{\bot\}$.

---

1. We adopt the convention that $\pi_{-k} = (\pi_1, \ldots, \pi_{k-1}, \pi_{k+1}, \ldots)$ and $(\pi_k, \pi_{-k}) = (\pi_1, \ldots, \pi_k, \ldots, \pi_K)$.

2. *For each $\pi \in \Pi$ and $M \in \mathcal{M}$, it holds that $M(\pi) = \mathbb{E}_{\sigma \sim \pi}[M(\sigma)]$. Further, there is a measurable function $\varphi : \mathcal{O} \to \Sigma$ so that $\mathbb{P}_{o \sim M(\sigma)}(\varphi(o) = \sigma) = 1$ for each $M \in \mathcal{M}$ and $\sigma \in \Sigma$ (i.e., $M(\sigma)$ reveals σ).*[2]

3. *For each $k \in [K]$, $\pi \in \Pi$, and $\pi'_k \in \Pi'_k$, it holds that*

$$U_k(\pi'_k, \pi) = \begin{cases} \mathbb{I}_{\pi'_k} \times \pi_{-k} & : \pi'_k \neq \perp \\ \pi & : \pi'_k = \perp \end{cases},$$

where $\mathbb{I}_{\pi'_k} \times \pi_{-k} \in \Pi$ denotes the product distribution whereby agent k plays π'_k and the other agents play according to their joint marginal under $\pi \in \Pi$.

In Definition 1.2, the inclusion of $\perp \in \Pi'_k$ corresponds to player k choosing not to deviate. This is necessary to satisfy Assumption 1.3 since there can be distributions $\pi \in \Pi$ so that if player k deviates to any fixed option in Σ_k, their value decreases.[3] We also remark that Definition 1.2 captures the notion of CCE in *normal-form* games (with pure action sets Σ_k); in Appendix G.1 we give an example of an instance capturing a slightly different notion of CCE in Markov games.

Examples of equilibria. We now provide concrete examples for the NE and CCE instances in Definitions 1.1 and 1.2; see Appendix G for additional examples (including CE) and discussion.

Example 1.1 (Learning Nash, and CCE in normal-form games). We begin by describing the problem of learning in normal-form games with bandit feedback. Suppose that each player $k \in [K]$ has a finite *action set* $\mathcal{A}_k$, with joint action set denoted by $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$. Upon playing a joint action profile $a \in \mathcal{A}$, the (unknown) ground truth model $M^\star$ samples $(r_1, \ldots, r_K) \sim M^\star(a)$, where r_k denotes the reward received by player k. The goal is to compute a distribution over joint action profiles which is some type of *equilibrium* of the game whose payoffs are given by expected rewards under $M^\star$.

- To express the problem of Nash equilibrium computation, set $\Pi_k := \Delta(\mathcal{A}_k)$ for each k, let $\Pi = \Pi_1 \times \cdots \Pi_K$ be the space of product distributions on $\mathcal{A}$, and define Π'_k, U_k as in Definition 1.1. Moreover, let $\mathcal{R} = [0,1]$ and $\mathcal{O}_\circ = \mathcal{A}$, $\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ$. Let $\mathcal{M}$ be the class of models so that: (a) for all singleton distributions $\mathbb{I}_a = \mathbb{I}_{a_1} \times \cdots \times \mathbb{I}_{a_K} \in \Pi$, $M(\mathbb{I}_a) \in \Delta(\mathcal{R}^K) \times \{\mathbb{I}_a\}$, and (b) for all $\pi \in \Pi$, $M(\pi) = \mathbb{E}_{a \sim \pi}[M(\mathbb{I}_a)]$. In words, $M(\pi)$ samples an action profile $a \sim \pi$ (in particular, $a_k \sim \pi_k$ for each k), reveals the action profile a sampled,[4] as well as K $[0,1]$-valued rewards drawn from an arbitrary distribution. Then the instance $\mathcal{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is an NE instance per Definition 1.1. For $\widehat{\pi} \in \Pi$, $h^{M^\star}(\widehat{\pi})$ measures the sum of the players' incentives to deviate from $\widehat{\pi}$ under the true model $M^\star$; in particular, $h^{M^\star}(\widehat{\pi}) = 0$ if and only if $\widehat{\pi}$ is a Nash equilibrium of the game whose payoff functions are given by $a \mapsto f_k^{M^\star}(a) := \mathbb{E}^{M^\star, a}[r_k]$.

- To express the problem of CCE computation, set $\Pi = \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$, and define Π'_k, U_k as in Definition 1.2 with $\Sigma_k = \mathcal{A}_k$ for each k. Moreover, let $\mathcal{R} = [0,1]$, and $\mathcal{O}_\circ = \mathcal{A}$,

---

2. When clear from context, we associate singleton distributions $\mathbb{I}_\sigma$ with the element σ the distribution places its mass on.

3. In some contexts, CCE are defined without such an option $\perp \in \Pi'_k$; in settings where the only goal is to establish *upper bounds*, the addition of $\perp$ does not make a material difference (since its only effect is to guarantee that the suboptimality of a decision is non-negative), but since we aim to prove *lower bounds* as well, it is crucial to have the option $\perp \in \Pi'_k$.

4. We assume that the model reveals the action profile played for technical reasons (see Assumption E.1); this is a very mild assumption, satisfied in essentially all (centralized) settings, since agents know which action they play.

$\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ$. Let $\mathcal{M}$ be the class of models so that: (a) for all singleton distributions $\mathbb{I}_a \in \Pi$, $M(\mathbb{I}_a) \in \Delta(\mathcal{R}^K) \times \{\mathbb{I}_a\}$, and (b), for $\pi \in \Pi$, $M(\pi) = \mathbb{E}_{a \sim \pi}[M(\mathbb{I}_a)]$. Then the instance $\mathcal{M} :=$ $(\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is a CCE instance per Definition 1.2. For $\widehat{\pi} \in \Pi$, $h^{M^\star}(\widehat{\pi})$ measures the sum of players' incentives to deviate from $\widehat{\pi}$ under the true model $M^\star$; in particular, $h^{M^\star}(\widehat{\pi}) = 0$ if and only if $\widehat{\pi}$ is a CCE of the game whose payoff functions are given by $a \mapsto f_k^{M^\star}(a) := \mathbb{E}^{M^\star, a}[r_k]$.

In the most basic ("finite-action") version of the normal-form game setup, we allow $M^\star(a)$ to be arbitrary, subject to the constraint that $r_k \in [0, 1]$, but assume that $A_k := |\mathcal{A}_k| < \infty$ for all k. Beyond finite-action normal-form games, the MA-DMSO framework captures structured normal-form games with bandit feedback (equivalently, multi-agent variants of the structured bandit problem), in which the players' action spaces are large or infinite, but rewards have additional structure. Examples include linear, convex, or concave payoffs (generalizing bandit convex optimization) (Bravo et al., 2018; Maheshwari et al., 2022; Lin et al., 2021), and many others (Cui et al., 2022). ◁

Example 1.2 (Learning Nash equilibria in Markov games). For reinforcement learning, each model $M \in \mathcal{M}$ is a Markov game $M = (H, \{\mathcal{S}_h\}_{h \in [H]}, \{\mathcal{A}_k\}_{k \in [K]}, \{P_h^M\}_{h \in [H]}, \{R_{k,h}^M\}_{k \in [K], h \in [H]}, d_1)$, where $H \in \mathbb{N}$ denotes the *horizon*, $\mathcal{S}_h$ denotes the state space for layer h, $\mathcal{A}_k$ denotes the action space for player k, $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ denotes the joint action space, $P_h^M : \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$ denotes the probability transition kernel for layer h, $R_{k,h}^M : \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathbb{R})$ denotes player k's reward distribution for layer h, and $d_1 \in \Delta(\mathcal{S}_1)$ denotes the initial state distribution. Each agent's decision space Π_k is the space of their *randomized Markov policies* $\pi_k = (\pi_{k,1}, \ldots, \pi_{k,H})$, where $\pi_{k,h} : \mathcal{S}_h \to \Delta(\mathcal{A}_k)$, and the joint decision space is $\Pi = \Pi_1 \times \cdots \times \Pi_K$. Given a joint decision $\pi \in \Pi$, an observation is drawn from $M(\pi)$ according to the following process, called an *episode*. First, an initial state is drawn according to $s_1 \sim d_1$. Then, for $h \in [H]$, the state evolves via:

1. $a_{k,h} \sim \pi_{k,h}(s_h)$, and $r_{k,h} \sim R_{k,h}^M(s_h, (a_{1,h}, \ldots, a_{K,h})) \ \forall k \in [K]$.   2. $s_{h+1} \sim P_h^M(\cdot | s_h, (a_{1,h}, \ldots, a_{K,h}))$

The distribution of $(r_1, \ldots, r_K, o_\circ) \sim M(\pi)$ is given by $o_\circ = \{(s_h, (a_{1,h}, \ldots, a_{K,h}), (r_{1,h}, \ldots, r_{K,h})\}_{h \in [H]}$ and $r_k = \sum_{h=1}^H r_{k,h}$. With Π'_k, U_k defined as in Definition 1.1, $\mathcal{M} := (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is an NE instance of MA-DMSO, and $h^{M^\star}(\widehat{\pi}) = 0$ if and only if $\widehat{\pi}$ is a *Markov Nash equilibrium* of $M^\star$ (e.g., Daskalakis et al. (2022)). By restricting $\mathcal{M}$ appropriately, this formulation captures complex settings with function approximation (Chen et al., 2022b; Li et al., 2022; Xie et al., 2020; Jin et al., 2022; Huang et al., 2021; Zhan et al., 2022; Liu et al., 2022); see Appendix G. ◁

1.2. MA-DMSO: Overview of results

We provide upper and lower bounds on the minimax sample complexity for the MA-DMSO framework using a new complexity measure, the *Multi-Agent Decision-Estimation Coefficient*, which generalizes the *Constrained Decision-Estimation Coefficient* introduced by Foster et al. (2023) in the single agent setting.

The Multi-Agent Decision-Estimation Coefficient. For measures $\mathbb{P}$ and $\mathbb{Q}$ with a common dominating measure ν, define squared Hellinger distance by $D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) = \int \left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu$. Consider an instance $\mathcal{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ for the MA-DMSO framework, as well as a *reference model* $\overline{M} : \Pi \to \Delta(\mathcal{O})$.[5] For a scale parameter $\varepsilon > 0$, the Multi-Agent Decision-

---

5. The reference model $\overline{M}$ may be arbitrary, and is not required to lie in $\mathcal{M}$.

Estimation Coefficient for the instance $\mathscr{M}$ with reference model $\overline{M}$ at scale ε is defined by

$$\mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M}) := \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[h^M(\pi)] \mid \mathbb{E}_{\pi \sim q}[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)] \le \varepsilon^2 \right\}; \qquad (2)$$

whenever the set $\mathcal{H}_{q,\varepsilon}(\overline{M}) := \{M \in \mathcal{M} \mid \mathbb{E}_{\pi \sim q}[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)] \le \varepsilon^2\}$ is empty, we adopt the convention that $\mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M}) = 0$. In addition, we define $\mathsf{dec}_\varepsilon(\mathscr{M}) := \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M})$, where $\mathrm{co}(\mathcal{M})$ denotes the convex hull of the class $\mathcal{M}$.

The interpretation of the definition (2), which is a min-max game, is as follows. The model $M \in \mathcal{M}$ selected by max-player represents a worst-case choice for the underlying model. The joint distributions $p, q \in \Delta(\Pi)$ selected by the min-player represent strategies for a centralized learning algorithm controlling all agents. The distribution $q \in \Delta(\Pi)$ is an *exploration distribution* which acts as a strategy for acquiring information, with the quantity $\mathbb{E}_{\pi \sim q}[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)]$ acting as their average "information gain" (that is, the amount information that allows to distinguish between $M \in \mathcal{M}$ and the reference model $\overline{M}$). The distribution $p \in \Delta(\Pi)$ is an *exploitation distribution* which aims to be near equilibrium for the model $M \in \mathcal{M}$ selected by the max-player, with $\mathbb{E}_{\pi \sim p}[h^M(\pi)]$ representing the distance from equilibrium. Thus, to summarize, the value (2) captures, for a best-case choice of $p, q \in \Delta(\Pi)$, the worst-case distance to equilibrium for p for models $M \in \mathcal{M}$ that are "close" to $\overline{M}$ in the sense that their information gain under q is small.

For familiar readers, we recall that the (single-agent) constrained DEC generalizes the earlier *offset* DEC of Foster et al. (2021) (which acts as a Lagrangian relaxation), and always leads to tighter guarantees (Foster et al., 2023). Our definition (2) generalizes the so-called PAC variant of the constrained DEC in Foster et al. (2023), as opposed the *regret* variant, which restricts to $p = q$.

Main results. The first of our results gives upper and lower bounds on the minimax sample complexity for the MA-DMSO framework based on the Multi-Agent Decision-Estimation Coefficient. To state the result in the simplest form, we assume that $|\mathcal{M}| < \infty$; see Appendix D for more general results.

Theorem 1.1 (Informal version of Corollaries D.1 and D.2). *For any instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ for the MA-DMSO framework and $T \in \mathbb{N}$:*

- *Upper bound: Under Assumption 1.1, there exists an algorithm that achieves*

$$\mathbb{E}[\mathbf{Risk}(T)] \le \widetilde{O}(1) \cdot \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M}), \quad \text{where} \quad \bar{\varepsilon}(T) \le \widetilde{\Theta}\big(\sqrt{\log|\mathcal{M}|/T}\big). \qquad (3)$$

- *Lower bound: For a worst-case model $M \in \mathcal{M}$, any algorithm must have*

$$\mathbb{E}[\mathbf{Risk}(T)] \ge \widetilde{\Omega}(1) \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{M}), \quad \text{where} \quad \underline{\varepsilon}(T) \text{ solves } \mathsf{dec}_\varepsilon(\mathscr{M}) \ge \widetilde{\Omega}\big(\varepsilon^2 KT\big). \qquad (4)$$

This result shows that the MA-DEC is a fundamental limit for equilibrium computation in the MA-DMSO framework, and is sufficient for low sample complexity whenever $\log|\mathcal{M}| < \infty$. The upper bound is an immediate corollary of an upper bound given by Foster et al. (2023) in the single-agent setting, while the lower bound requires a new approach; this is due to fundamental differences between the single and multiple agents, which we highlight in the sequel.

To build intuition, let us start with a basic example. Suppose that $\mathscr{M}$ is a CCE instance consisting of two-player $A_1 \times A_2$ normal-form games (that is, $|\mathcal{A}_1| = A_1$ and $|\mathcal{A}_2| = A_2$) with bandit feedback (Example 1.1) and Bernoulli noise. In this case, one can show that $\mathsf{dec}_\varepsilon(\mathscr{M}) \propto \varepsilon \cdot \sqrt{A_1 + A_2}$, so that

the upper bound (3) gives $\mathbb{E}[\mathbf{Risk}(T)] \lesssim \sqrt{\frac{(A_1+A_2)\log|\mathcal{M}|}{T}}$, or equivalently, $\frac{(A_1+A_2)\log|\mathcal{M}|}{\varepsilon^2}$ rounds of interaction are sufficient to find an ε-CCE. For this class, one can take $\log|\mathcal{M}| \lesssim \widetilde{O}(A_1 \cdot A_2)$. We give more refined results (Appendix F) which allow one to replace $\log|\mathcal{M}|$ by $\max_k \log|\Pi'_k| \lesssim \log(A_1 + A_2)$, so that we achieve sample complexity $\widetilde{O}\big(\frac{A_1+A_2}{\varepsilon^2}\big)$, which is optimal.

Turning to lower bounds, for the same normal-form game instance $\mathscr{M}$, one can choose $\underline{\varepsilon}(T) \gtrsim \frac{\sqrt{A_1+A_2}}{T}$, so that (4) gives $\mathbb{E}[\mathbf{Risk}(T)] \gtrsim \frac{A_1+A_2}{T}$, or equivalently, $\widetilde{\Omega}\big(\frac{A_1+A_2}{\varepsilon}\big)$ rounds of interaction are necessary to find an ε-CCE. Comparing the upper and lower bounds, there are two gaps. The first is the term $\log|\mathcal{M}|$ appearing in the upper bound, which represents the sample complexity required to perform statistical estimation with the class $\mathcal{M}$, and in general scales poorly with the number of agents. This can be refined (cf. Appendix F), but is not possible to completely remove in general, even in the single-agent setting; see Foster et al. (2021, 2023) and Appendix D for further discussion.

The second gap is the difference between the values $\bar{\varepsilon}(T)$ and $\underline{\varepsilon}(T)$ appearing in the upper and lower bound; we set $\bar{\varepsilon}(T) \propto 1/\sqrt{T}$, while $\underline{\varepsilon}(T)$ is chosen to solve the fixed-point equation $\mathsf{dec}_\varepsilon(\mathscr{M}) \geq \widetilde{\Omega}\big(\varepsilon^2 T\big)$ (we focus on the case of constant K in this discussion). For normal-form games, this causes the lower bound to scale with $\frac{1}{\varepsilon}$ instead of $\frac{1}{\varepsilon^2}$. This gap is not present in the single-agent setting (Foster et al., 2023), where the best upper and lower bounds based on the constrained DEC have $\bar{\varepsilon}(T) \approx \underline{\varepsilon}(T)$ (up to dependence on $\log|\mathcal{M}|$). We show (Proposition D.1) that for most parameter regimes,

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M}) \lesssim \big(K^2 \log|\mathcal{M}| \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{M})\big)^{1/2},$$

i.e., the gap between the upper and lower bounds is no worse than quadratic generically. This gap turns out to be fundamental: We show (Propositions D.2 and D.3) that there exist instances for which each bound (upper and lower) is tight, and—somewhat surprisingly—the following result shows that *no complexity measure* satisfying fairly general conditions can fully characterize the sample complexitymaking beyond a quadratic gap, even when $\log|\mathcal{M}| = \widetilde{O}(1)$.

Theorem 1.2 (Informal version of Theorem D.4). *For any $\varepsilon \in \mathbb{N}$, there exist two-player zero-sum Nash equilibrium MA-DMSO instances $\mathscr{M}_1 = (\mathcal{M}_1, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ and $\mathscr{M}_2 = (\mathcal{M}_2, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ with $\log|\mathcal{M}_1| = \log|\mathcal{M}_2| = \widetilde{O}(1)$ and a one-to-one mapping $\mathscr{E} : \mathcal{M}_1 \rightarrow \mathcal{M}_2$ satisfying:*

1. *For all $M \in \mathcal{M}_1$, $f_k^M \equiv f_k^{\mathscr{E}(M)}$ for all $k \in [2]$.*

2. *For all $M, M' \in \mathcal{M}_1$ and all $\pi \in \Pi$, $D_{\mathsf{H}}^2(M(\pi), M'(\pi)) = D_{\mathsf{H}}^2(\mathscr{E}(M)(\pi), \mathscr{E}(M')(\pi))$.*

3. *There exists an algorithm that finds an ε-NE for any model in $\mathscr{M}_1$ using $\widetilde{O}\big(\frac{1}{\varepsilon}\big)$ rounds, yet any algorithm requires $\widetilde{\Omega}\big(\frac{1}{\varepsilon^2}\big)$ rounds to find an ε-NE for a worst-case model in $\mathscr{M}_2$.*

Informally, this result states that if a complexity measure depends on the instance $\mathscr{M}$ only through 1) reward functions and 2) pairwise Hellinger distances for models in $\mathcal{M}$, then it cannot characterize the optimal sample complexity for every instance beyond the gap in the prequel. In addition, the full result is not limited to Hellinger distance, and applies to general f-divergences including KL- and χ^2-divergence. This rules out tighter guarantees based on various variants of the DEC, as well as most other general-purpose complexity measures; see Appendix D.2.2 for details.[6]

---

6. Directly applying Theorem 1.2 to the constrained DEC presents complications due to $\overline{M} \in \mathrm{co}(\mathcal{M})$; see App. D.2.2.

We emphasize a fundamental separation between the single and multi-agent frameworks. In the single-agent setting, the constrained DEC characterizes, up to logarithmic factors, the optimal number of samples required to learn an ε-optimal decision, as long as $\log|\mathcal{M}| = \widetilde{O}(1)$ (Foster et al., 2023). For two or more agents, Theorem 1.2 and Propositions D.2 and D.3 rule out such a characterization.

1.3. Hidden-reward interactive decision making (HR-DMSO)

To prove the results in the prequel, we establish a certain equivalence between the MA-DMSO framework and another *single-agent* setting we refer to as *Hidden-Reward Decision Making with Structured Observations* (HR-DMSO), which generalizes the single-agent DMSO framework (MA-DMSO with $K = 1$) by allowing rewards to be hidden from the agent. This setting is of interest in its own right, and can be thought of as a stochastic, PAC variant of the partial monitoring problem (Bartók et al., 2014). In what follows, we introduce the framework, then show that 1) MA-DMSO can be viewed as a special case of the HR-DMSO framework via a simple reduction, and 2) a converse holds, thus showing a sort of equivalence. We then discuss implications for minimax rates in both frameworks.

Formally, the HR-DMSO framework proceeds in T rounds, where for each round $t = 1, 2, \ldots, T$:

1. The learner selects a *decision* $\pi^t \in \Pi$, where Π is the *decision space*, and gains (but does not observe) reward $f^{M^\star}(\pi^t)$.

2. The learner receives an observation $o^t \in \mathcal{O}$ sampled via $o^t \sim M^\star(\pi^t)$, where $M^\star : \Pi \to \Delta(\mathcal{O})$ is the underlying *model*. We refer to $\mathcal{O}$ as the *observation space*.

After this process finishes, the learner uses the data collected throughout the T rounds of interaction to produce an output decision $\widehat{\pi} \in \Pi$, which may be randomized according to a distribution $p \in \Delta(\Pi)$. The learner's goal is to choose the decision $\widehat{\pi}$ so as to maximize its (unobserved) reward $f^{M^\star}(\widehat{\pi})$. Formally, writing $\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$, we define the *risk* of an algorithm as: $\mathbf{Risk}(T) := \mathbb{E}_{\widehat{\pi} \sim p}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi})]$.

We assume that every model M is associated a (known) function $f^M : \Pi \to \mathbb{R}$, where $f^M(\pi)$ specifies the learner's value under decision $\pi \in \Pi$ when the underlying model is M. We make the following realizability assumption, analogous to Assumption 1.1.

Assumption 1.4 (Realizability for HR-DMSO). *The learner has access to a model class $\mathcal{M}$ consisting of probability kernels $M : \Pi \to \Delta(\mathcal{O})$ that contains the true model $M^\star$.*

We refer to the tuple $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_{M \in \mathcal{M}})$ as an *instance* for the HR-DMSO framework. We extend the constrained Decision-Estimation Coefficient of Foster et al. (2023) to HR-DMSO as follows. For an instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$, reference model $\overline{M} : \Pi \to \Delta(\mathcal{O})$, and scale parameter $\varepsilon > 0$, the constrained Decision-Estimation Coefficient is given by[7]

$$\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) = \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim q}[D^2_{\mathsf{H}}(M(\pi), \overline{M}(\pi))] \leq \varepsilon^2 \right\}. \quad (5)$$

We define the Decision-Estimation Coefficient (DEC) of the instance $\mathscr{H}$ at scale ε to be $\mathsf{dec}_\varepsilon(\mathscr{H}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M})$. This definition is identical to the constrained PAC DEC (Foster et al., 2023); this is natural, as the only difference between the HR-DMSO framework and the DMSO framework (Foster et al., 2023) is that we relax the constraint that the agent observes its reward.

---

7. Note that we use the same notation for the DEC in the HR-DMSO and MA-DMSO settings; we will typically use the letter $\mathscr{H}$ to denote HR-DMSO instances and $\mathscr{M}$ to denote MA-DMSO instances to avoid ambiguity.

Remark 1.3. The HR-DMSO framework is related to the partial monitoring problem (Bartók et al., 2014), which typically considers regret guarantees; in contrast, we consider PAC guarantees.

1.4. HR-DMSO: Overview of results

It is fairly immediate to see that the HR-DMSO framework generalizes the MA-DMSO framework. For any MA-DMSO instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ satisfying Assumption 1.3 and Assumption 1.2, by choosing the value function $f^M(\cdot) = -h^M(\cdot)$, the instance of the HR-DMSO framework specified by the tuple $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M\}_M)$ (recalling that $\mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}^K$) is statistically equivalent to $\mathscr{M}$.[8] In particular, letting $\mathfrak{M}(\mathscr{M}, T)$ denote the minimax risk for an instance $\mathscr{M}$ in the MA-DMSO framework, and let $\mathfrak{M}(\mathscr{H}, T)$ denote the minimax risk for the corresponding HR-DMSO instance $\mathscr{H}$ (see Appendix B for formal definitions), we have:

1. For all $\overline{M}$ and $\varepsilon > 0$, $\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) = \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M})$.   2. For all T, $\mathfrak{M}(\mathscr{H}, T) = \mathfrak{M}(\mathscr{M}, T)$.

It is natural to ask whether the HR-DMSO framework is *strictly* more general than the MA-DMSO framework. Indeed, by allowing rewards to be hidden, one might imagine that HR-DMSO can capture problems outside of MA-DMSO, which forces rewards to be observed. The next result shows that this is not the case: any HR-DMSO instance can be embedded in a two-player zero-sum NE instance for MA-DMSO, with minimal increase in statistical complexity.

Theorem 1.3 (Informal version of Theorem C.2). *Consider any HR-DMSO instance specified by the tuple $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$. For any $\delta > 0$, there exists a two-player zero-sum NE MA-DMSO instance $\mathscr{M} = (\widetilde{\mathcal{M}}, \widetilde{\Pi}, \widetilde{\mathcal{O}}, \Pi'_k, U_k)$ (Definition 1.1) such that:*

1. For all $\varepsilon > 0$, $\mathsf{dec}_\varepsilon(\mathscr{H}) \leq \mathsf{dec}_\varepsilon(\mathscr{M}) \leq \delta + \mathsf{dec}_{\varepsilon+\delta}(\mathscr{H})$.

2. For all $T \in \mathbb{N}$, it holds that $\mathfrak{M}(\mathscr{H}, T) \leq \mathfrak{M}(\mathscr{M}, T) \leq \mathfrak{M}(\mathscr{H}, T) + \delta$.

3. If $\mathcal{M}$ is finite, then $\log|\widetilde{\mathcal{M}}| \leq \log|\mathcal{M}| + \mathrm{polylog}(T, \delta^{-1})$.

This result establishes that the MA-DMSO and HR-DMSO frameworks satisfy a sort of equivalence, and shows that characterizing the minimax sample complexity for MA-DMSO is no easier than characterizing the minimax sample complexity for the HR-DMSO framework. The proof proceeds by embedding a given instance for the HR-DMSO framework into a two-player game: the first of the two agents in the game plays the role of the HR-DMSO agent, and the second agent selects actions to ensure that optimal actions for the original HR-DMSO instance are Nash equilibria for the new instance, and vice-versa. The key idea is that even though rewards in the game are observed, by making the game polynomially large, we can ensure that discovering them requires a prohibitively large amount of exploration, rendering them effectively hidden.

HR-DMSO: Minimax rates.   To prove the multi-agent minimax rates in Theorems 1.1 and 1.2, we first prove analogous bounds for the HR-DMSO framework, then use the equivalence above to extend them to MA-DMSO. These results can be found in Theorems D.1 to D.3 in Appendix D.

---

8. It is essential for this reduction that the rewards in $\mathscr{H}$ be hidden, since it is in general impossible to simulate a reward whose mean is $-h^M(\pi)$ using samples from $M(\pi)$.

1.5. MA-DMSO: Additional results

Beyond minimax rates, we provide a number of structural results for the MA-DMSO framework, including: (1) conditions under which the multi-agent DEC can be controlled by the single-agent DEC, and (2) conditions under which the so-called *curse of multiple agents* can be avoided.

From multi-agent to single-agent. We show that it is generically possible to upper bound the MA-DEC in terms of the single-agent DEC for each player k. This result is most easily stated in terms of a multi-agent analogue of the *offset* version of the DEC introduced in Foster et al. (2021). Specifically, we consider a *regret* variant of the offset DEC that restricts $p = q$, coupling exploration and exploitation: For an instance $\mathscr{M}$, reference model $\overline{M}$, and scale parameter $\gamma > 0$, we define

$$\mathsf{r\text{-}dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}, \overline{M}) := \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[h^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi \sim p}[D_{\mathsf{H}}^2(M(\pi), \overline{M}(\pi))] \right\}. \tag{6}$$

It follows immediately from the results of Foster et al. (2023) (see Proposition E.1) that $\mathsf{dec}_{\varepsilon}(\mathscr{M}, \overline{M}) \leq \inf_{\gamma > 0} \{\mathsf{r\text{-}dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}, \overline{M}) \vee 0 + \gamma \varepsilon^2\}$, so upper bounds on $\mathsf{r\text{-}dec}_{\gamma}^{\mathrm{o}}(\mathscr{M})$ yield upper bounds on $\mathsf{dec}_{\varepsilon}(\mathscr{M})$, which can in turn be inserted into Theorem 1.1 to yield upper bounds on minimax risk. While it is also possible to directly upper bound $\mathsf{dec}_{\varepsilon}(\mathscr{M}, \overline{M})$ without going through $\mathsf{r\text{-}dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}, \overline{M})$, using $\mathsf{r\text{-}dec}_{\gamma}^{\mathrm{o}}(\cdot)$ is more convenient and does not lead to any significant quantitative loss in the resulting upper bounds.

We prove an upper bound on the multi-agent DEC of the instance $\mathscr{M}$, in terms of the (single-agent) DEC of K different model classes $\widetilde{\mathcal{M}}_k$, defined in terms of $\mathscr{M}$. To define these model classes, for $M \in \mathcal{M}$ and $k \in [K]$, we first define an induced *single-agent* model $M|_k$ as follows: a pure observation drawn from $M|_k(\pi)$ has the distribution of the pure observation $o_{\circ}$ when $o_{\circ} \sim M(\pi)$, and the reward drawn from $M|_k(\pi)$ has the distribution of r_k when $(r_1, \ldots, r_K) \sim M(\pi)$. In short, the model $M|_k$ is identical to M but simply ignores the rewards of all agents except k. Next, the model class $\widetilde{\mathcal{M}}_k$ is defined to have policy space Π_k, so that models in $\widetilde{\mathcal{M}}_k$ are mappings $\widetilde{M} : \Pi_k \to \Delta(\mathcal{R} \times \mathcal{O}_{\circ})$. Finally, we define the class $\widetilde{\mathcal{M}}_k$, which is indexed by $\Pi_{-k} \times \mathcal{M}$, via $\widetilde{\mathcal{M}}_k = \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) : \pi_{-k} \in \Pi_{-k}, M \in \mathcal{M}\}$.

Theorem 1.4. *Let $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ be a NE MA-DMSO instance satisfying Assumption E.1. Then for any $\gamma > 0$, it holds that[9]*

$$\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{r\text{-}dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}, \overline{M}) \leq \sum_{k=1}^{K} \sup_{\overline{M}_k \in \mathrm{co}(\widetilde{\mathcal{M}}_k)} \mathsf{r\text{-}dec}_{\gamma/K}^{\mathrm{o}}(\widetilde{\mathcal{M}}_k, \overline{M}_k).$$

This result allows us to bound the MA-DEC using standard bounds on the single-agent DEC (Foster et al., 2021). For example, for normal-form games, where each player has A_k actions, it yields $\mathsf{r\text{-}dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}, \overline{M}) \lesssim K \cdot \sum_{k=1}^{K} \frac{A_k}{\gamma}$. See Appendix E for refinements concerning Markov games.

The proof of Theorem E.1 employs a novel fixed-point argument: For each agent k, if all other agents commit to some joint distribution, this induces a single-agent DMSO instance, and it is natural for k to play the strategy that minimizes the single-agent DEC for this instance. Using Kakutani's fixed point theoerem, we show that it is possible for all K agents to apply this strategy simultaneously.

---

9. Here, the notation $\mathsf{r\text{-}dec}_{\gamma/K}^{\mathrm{o}}(\widetilde{\mathcal{M}}_k, \overline{M}_k)$ refers to the single-agent DEC for the model class $\widetilde{\mathcal{M}}_k$; see Appendix B.

On the curse of multiple agents. In multi-agent reinforcement learning, the *curse of multiple agents* refers to the situation in which the sample complexity required to learn an equilibrium scales exponentially in the number of players (Jin et al., 2021b). In general, our upper bounds on sample complexity for the MA-DMSO framework (Theorem 1.1) suffer from the curse of multiple agents due to the presence of the estimation complexity term $\log|\mathcal{M}|$. For example, in a K-player normal-form game with A actions per player, one has $\log|\mathcal{M}| \approx A^K$ (using an appropriate discretization of $\mathcal{M}$). Our final result shows that it is possible to avoid the curse of multiple agents by replacing the estimation complexity $\log|\mathcal{M}|$ with the maximum size $\max_k \log|\Pi'_k|$ for each player's deviation set, which is usually polynomial in the number of agents; the tradeoff is that the result scales with the MA-DEC for the MA-DMSO instance in which the model class $\mathcal{M}$ is *convexified* via $\mathcal{M} \leftarrow \mathrm{co}(\mathcal{M})$.

Theorem 1.5 (Informal version of Theorem F.1). *Let $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ be a CCE instance (Definition 1.2) or a CE instance (Definition G.1) of the MA-DMSO framework. Then, for any $T \in \mathbb{N}$, Algorithm 1 outputs $\widehat{\pi} \in \Pi$ such that with probability at least $1 - \delta$,*

$$\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \leq \widetilde{O}(K) \cdot \inf_{\gamma > 0} \left\{ \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})) + \frac{\gamma}{T} \cdot \log\left(\frac{\max_k |\Pi'_k|}{\delta} \right) \right\},$$

where we adopt the convention that $\mathrm{co}(\mathscr{M}) \equiv (\mathrm{co}(\mathcal{M}), \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$.

In normal-form games with K players and A actions per player, we have $\mathsf{dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})) \lesssim \frac{A}{\gamma}$ and $\max_k \log|\Pi'_k| = \log(A)$, so this result gives $\mathbf{Risk}(T) \lesssim \sqrt{\mathrm{poly}(K)A/T}$. More broadly, Theorem 1.5 shows that it is generically possible to avoid the curse of multiple agents for convex classes, including structured classes of normal-form games with bandit feedback such as games with linear or convex payoffs. In general though, it does not lead to tight guarantees for non-convex classes such as Markov games, and is best thought of as complementary to results for this setting (Jin et al., 2021b; Song et al., 2021; Mao and Basar, 2022). The result is proven by adapting the powerful *exploration-by-optimization* algorithm from the single-agent setting (Lattimore, 2022; Foster et al., 2022b) so as to exploit the unique feedback structure of the multi-agent setting. We refer to Appendix F for details, as well as additional results which highlight settings in which the curse of multiple agents cannot be avoided in the sense of Theorem 1.5.

Acknowledgments

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proc. of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

Ioannis Anagnostides, Gabriele Farina, and Tuomas Sandholm. Near-optimal phi-regret learning in extensive-form games, 2022.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

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

Yu Bai, Chi Jin, and Tiancheng Yu. 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.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 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.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4): 967–997, 2014.

Gerald Beer. *Topologies on Closed and Closed Convex Sets*. Kluwer Academic Publishers, 1993.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

Noam Brown and Tuomas Sandholm. 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.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 41–1. JMLR Workshop and Conference Proceedings, 2012.

Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022a.

Xi Chen, Yu Cheng, and Bo Tang. Well-Supported vs. Approximate Nash Equilibria: Query Complexity of Large Games. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 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.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022b.

Qiwen Cui, Zhihan Xiong, Maryam Fazel, and Simon S Du. Learning in congestion games with bandit feedback. *arXiv preprint arXiv:2206.01880*, 2022.

Varsha Dani, Thomas P Hayes, and Sham Kakade. The price of bandit information for online optimization. 2007.

Constantinos Daskalakis and Christos H Papadimitriou. Computing pure nash equilibria in graphical games via markov random fields. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 91–99, 2006.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. *International Conference on Machine Learning*, 2021.

Eyal Even-Dar, Yishay Mansour, and Uri Nadav. On the convergence of regret minimization dynamics in concave games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 523–532, 2009.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Dylan J Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250*, 2022a.

Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. *arXiv preprint arXiv:2206.13063*, 2022b.

Dylan J. Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023.

Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. *Advances in Neural Information Processing Systems*, 34:22655–22666, 2021.

Geoffrey J. Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 360–367, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054.

Venkatesan Guruswami, Atri Ruda, and Madhu Sudan. *Essential Coding Theory*. 2022.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pages 1704–1713, 2017.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Neural Information Processing Systems*, 2021a.

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

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR, 2022.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17:697–704, 2004.

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R. McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. 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.

Tor Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3):311–334, 2020.

Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. *arXiv preprint arXiv:2202.10997*, 2022.

Tor Lattimore and Andras György. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

Chris Junchi Li, Dongruo Zhou, Quanquan Gu, and Michael I Jordan. Learning two-player mixture markov games: Kernel function approximation and correlated equilibrium. *arXiv preprint arXiv:2208.05363*, 2022.

Tianyi Lin, Zhengyuan Zhou, Wenjia Ba, and Jiawei Zhang. Optimal no-regret learning in strongly monotone games with bandit feedback. *arXiv preprint arXiv:2112.02856*, 2021.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7001–7010. PMLR, 18–24 Jul 2021.

Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable markov games. *arXiv preprint arXiv:2206.01315*, 2022.

Chinmay Maheshwari, Chih-Yuan Chiu, Eric Mazumdar, Shankar Sastry, and Lillian Ratliff. Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 6702–6734. PMLR, 2022.

Kleanthis Malialis and Daniel Kudenko. 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.

Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against learners in bayesian games. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5221–5252. PMLR, 02–05 Jul 2022.

Weichao Mao and Tamer Basar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge, 2007.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair,

Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. 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.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. 2014.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3066–3074, 2013.

J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3):520–534, 1965.

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

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

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

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

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

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Salil Vadhan. Pseudorandomness. *Foundations and Trends in Theoretical Computer Science*, 7: 1–336, 2012.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

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

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. 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
Main results

Appendix A. Organization of appendix

Part I of the appendix presents our main results, with preliminaries in Appendix B. In particular, Appendix C and Appendix D present our main sample complexity guarantees for the MA-DMSO and HR-DMSO frameworks.

- Appendix C establishes a certain equivalence between the MA-DMSO and HR-DMSO frameworks.

- Appendix D establishes upper and lower bounds on the minimax rates for both frameworks based on the Decision-Estimation Coefficient, and highlights barriers to obtaining sharper guarantees analogous to those found in the single-agent, reward-observed DMSO framework (Foster et al., 2023).

Appendix E and Appendix F present additional results concerning the MA-DMSO framework.

- Appendix E gives general conditions under which it is possible to bound the MA-DEC in terms of the single-agent DEC.

- Appendix F gives conditions under which one can obtain sample complexity guarantees in the MA-DMSO framework that avoid the so-called *curse of multiple agents*.

Part II provides examples (instances, as well as upper and lower bounds on the DEC and minimax rates) and discussion for the MA-DMSO framework, and Part III contains proofs for all results.

Appendix B. Preliminaries

Below we provide additional technical preliminaries which will be used throughout our proofs.

Probability kernels. For probability spaces $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$, a *probability kernel* $P(\cdot|\cdot)$ from $(\mathcal{X}, \mathscr{X})$ to $(\mathcal{Y}, \mathscr{Y})$ is a mapping $P : \mathscr{Y} \times \mathcal{X} \to [0, 1]$ which satisfies (1) for all $x \in \mathcal{X}$, $P(\cdot|x)$ is a probability measure on $(\mathcal{Y}, \mathscr{Y})$, and (2) for all $Y \in \mathscr{Y}$, the mapping $x \mapsto P(Y|x)$ is measurable with respect to $\mathscr{X}$. To simplify notation we often denote probability kernels as $P : \mathcal{X} \to \Delta(\mathcal{Y})$.

MA-DMSO framework. We adopt the same formalism for probability spaces as in Foster et al. (2021, 2023). Decisions are associated with a measure space $(\Pi, \mathscr{P})$, and observations are associated with the measure space $(\mathcal{O}, \mathscr{O})$. In the MA-DMSO framework, pure observations are associated with the measure space $(\mathcal{O}_\circ, \mathscr{O}_\circ)$ and rewards are associated with a measure space $(\mathcal{R}, \mathscr{R})$, and furthermore, we have $\mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}^K$ and $\mathscr{O} = \mathscr{O}_\circ \otimes \mathscr{R}^{\otimes K}$. Formally, a *model* $M(\cdot \mid \cdot)$ is a probability kernel from $(\Pi, \mathscr{P})$ to $(\mathcal{O}, \mathscr{O})$. We denote the set of all models as $\mathcal{M}^+$. Note that $\mathcal{M}^+$ depends on the measure spaces $(\Pi, \mathscr{P}), (\mathcal{O}, \mathscr{O})$; when we wish to make this dependence explicit, we will write $\mathcal{M}^+_{\Pi, \mathcal{O}}$. The *history* up to time t is given by $\mathfrak{H}^t = (\pi^1, o^1), \ldots, (\pi^t, o^t)$. We define

$$\Omega^t = \prod_{i=1}^t (\Pi \times \mathcal{O}), \qquad \mathscr{F}^t = \bigotimes_{i=1}^T (\mathscr{P} \otimes \mathscr{O}),$$

22

so that $\mathfrak{H}^t$ is associated with the space $(\Omega^t, \mathscr{F}^t)$.

We assume throughout the paper that $\mathcal{R} = [0,1]$ (which implies in particular that $h^M(\pi) \in [0,K]$ for all M, π) unless otherwise stated. To simplify notation, for each $\pi \in \Pi$ and $M \in \mathcal{M}$, we write $h_k^M(\pi) := \sup_{\pi_k' \in \Pi_k} f_k^M(U_k(\pi_k', \pi)) - f_k^M(\pi)$, so that $h^M(\pi) = \sum_{k=1}^K h_k^M(\pi)$.

Remark B.1 (Notation for MA-DMSO framework). We use the following convention throughout the paper: When convenient, we associate any singleton distribution with the element that the distribution places its mass on. For instance, for a pure decision $\sigma = (\sigma_1, \ldots, \sigma_K) \in \Sigma_1 \times \cdots \times \Sigma_K$ in the context of Definition 1.2, we will denote its corresponding singleton distribution $\mathbb{I}_\sigma \in \Delta(\Sigma) = \Pi$ as just $\sigma \in \Pi$. In addition, when possible, we use the convention that Σ denotes a pure decision set, whereas Π denotes a decision set that may be pure or mixed (this will be clear from context).

The canonical single-agent instance. Given a decision space Π, an observation space $\mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}$, and a model class $\mathcal{M} \subset (\Pi \to \Delta(\mathcal{O}))$, there is a canonical single-agent instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ corresponding to the model class $\mathcal{M}$: we take $\Pi_1' = \Pi$ and $U_1(\pi_1', \pi) = \pi_1'$, which ensures that $h^M(\pi) = \max_{\pi' \in \Pi} f^M(\pi') - f^M(\pi)$ for all $\pi \in \Pi, M \in \mathcal{M}$. The single-agent instance $\mathscr{M}$ of the 1-player MA-DMSO framework exactly captures the DMSO framework in Foster et al. (2021, 2023) for the model class $\mathcal{M}$. Furthermore, for any model $\overline{M}$, we will write $\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) = \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M})$ (and similarly we will write $\mathsf{r\text{-}dec}_\gamma^\circ(\mathcal{M}, \overline{M}) = \mathsf{r\text{-}dec}_\gamma^\circ(\mathscr{M}, \overline{M})$ for regret variant of the offset DEC introduced in Appendix E); the quantity $\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) = \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M})$ is identical to the constrained (PAC) DEC of the model class $\mathcal{M}$ as defined in Foster et al. (2023), and the quantity $\mathsf{r\text{-}dec}_\gamma^\circ(\mathcal{M}, \overline{M}) = \mathsf{r\text{-}dec}_\gamma^\circ(\mathscr{M}, \overline{M})$ is identical to the offset (regret) DEC of the model class $\mathcal{M}$ as defined in Foster et al. (2021).

HR-DMSO framework. As in the MA-DMSO framework, decisions are associated with a measure space $(\Pi, \mathscr{P})$, observations are associated with the measure space $(\mathcal{O}, \mathscr{O})$, and models $M(\cdot \mid \cdot)$ are probability kernels from $(\Pi, \mathscr{P})$ to $(\mathcal{O}, \mathscr{O})$. The history up to time t is given by $\mathfrak{H}^t = (\pi^1, o^1), \ldots, (\pi^t, o^t)$, and is associated with the space $(\Omega^t, \mathscr{F}^t)$ given by

$$\Omega^t = \prod_{i=1}^t (\Pi \times \mathcal{O}), \qquad \mathscr{F}^t = \bigotimes_{i=1}^T (\mathscr{P} \otimes \mathscr{O}).$$

We denote the set of all models as $\mathcal{M}^+$. Unless stated otherwise, we will assume throughout that $f^M(\pi) \in [0,1]$ for all $M \in \mathcal{M}^+$ and $\pi \in \Pi$.

For a model M and decision $\pi \in \Pi$, $\mathbb{E}^{M,\pi}[\cdot]$ denotes expectation under the process $o \sim M(\pi)$. To simplify notation, we often abbreviate $g^M(\pi) := f^M(\pi_M) - f^M(\pi)$, so that $\mathbf{Risk}(T) = \mathbb{E}_{\widehat{\pi} \sim p}[g^{M^\star}(\widehat{\pi})]$.

Remark B.2 (Alternative formulation for HR-DMSO). An equivalent formulation of the HR-DMSO framework would be to consider models $M : \Pi \to \Delta(\mathcal{O} \times \mathcal{R})$ that specify joint distributions over observations and rewards and define $f^M(\pi) = \mathbb{E}^{M,\pi}[r]$, but only allow o to be observed by the learner under $(o, r) \sim M(\pi)$.

Remark B.3 (Contrast with *reward-free* DMSO). Despite the similar name, the HR-DMSO framework is distinct from the "reward-free" DMSO framework considered in the recent work of Chen et al. (2022a); in the latter framework, which is specialized to Markov decision processes, a reward-function is given to the learner explicitly, but only after the learning process ends.

Density ratios. For both the MA-DMSO and HR-DMSO, we define

$$V(\mathcal{M}) := \sup_{M,M' \in \mathcal{M}} \sup_{\pi \in \Pi} \sup_{A \in \mathscr{O}} \left\{ \frac{M(A \mid \pi)}{M'(A \mid \pi)} \right\} \vee e. \tag{7}$$

Finiteness of $V(\mathcal{M})$ is not necessary for our results to hold, but improves several of our bounds by a $\log(T)$ factor.

Divergences. Total variation distance is given by

$$D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int |d\mathbb{P} - d\mathbb{Q}|,$$

and the Kullback Leibler divergence is given by

$$D_{\mathsf{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \begin{cases} \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}, & \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Minimax sample complexity. Formally, for $T \in \mathbb{N}$, an *algorithm* (for either the HR-DMSO or MA-DMSO frameworks) is a collection of probability kernels $(p, q) = \left(p(\cdot \mid \cdot), \{q^t(\cdot \mid \cdot)\}_{t=1}^T\right)$, where each $q^t : \Omega^{t-1} \to \Delta(\Pi)$ is a probability kernel from $(\Omega^{t-1}, \mathscr{F}^{t-1})$ to $(\Pi, \mathscr{P})$, and $p : \Omega^T \to \Delta(\Pi)$ is a probability kernel from $(\Omega^T, \mathscr{F}^T)$ to $(\Pi, \mathscr{P})$. We let $\mathbb{P}^{M,(p,q)}$ denote the law of $(\mathfrak{H}^T, \widehat{\pi})$ under the process:

$$\pi^t \sim q^t(\cdot \mid \mathfrak{H}^{t-1}), \ o^t \sim M(\cdot \mid \pi^t), \ \forall t \in [T], \qquad \widehat{\pi} \sim p(\cdot \mid \mathfrak{H}^T),$$

and we use $\mathbb{E}^{M,(p,q)}$ to denote the corresponding expectation. Our main goal is to characterize the *minimax PAC sample complexity* of an instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ of the MA-DMSO framework or $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ of the HR-DMSO framework. The minimax sample complexities for both cases are defined in an identical manner, spelled out below:

$$\mathfrak{M}(\mathscr{M}, T) := \inf_{(p,q)} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,(p,q)} \mathbb{E}_{\widehat{\pi} \sim p(\cdot \mid \mathfrak{H}^T)} \left[\sum_{k=1}^K \sup_{\pi'_k \in \Pi'_k} f_k^M(U_k(\pi'_k, \widehat{\pi})) - f_k^M(\widehat{\pi}) \right],$$

$$\mathfrak{M}(\mathscr{H}, T) := \inf_{(p,q)} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,(p,q)} \mathbb{E}_{\widehat{\pi} \sim p(\cdot \mid \mathfrak{H}^T)} [f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi})].$$

Additional notation. For an integer $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, \ldots, n\}$. For a set $\mathcal{X}$, we let $\Delta(\mathcal{X})$ denote the set of all probability distributions over $\mathcal{X}$. For $x \in \mathcal{X}$, we use $\mathbb{I}_x \in \Delta(\mathcal{X})$ to denote the distribution which places probability mass 1 on x. We adopt standard big-oh notation, and write $f = \widetilde{O}(g)$ to denote that $f = O(g \cdot \max\{1, \mathrm{polylog}(g)\})$. We use $\lesssim$ only in informal statements to emphasize the most relevant aspects of an inequality. For a set $\mathcal{X}$, let $\mathcal{P}(\mathcal{X})$ denote the power set of i.e., the set of all subsets of $\mathcal{X}$.

Appendix C. Equivalence of MA-DMSO and HR-DMSO frameworks

In this section, which forms the starting point for our main results, we show that the MA-DMSO and HR-DMSO frameworks satisfy a certain statistical equivalence. First, in Theorem C.1, we formalize the trivial direction of this equivalence: namely, any instance of the MA-DMSO framework can be viewed as an instance of the HR-DMSO framework. To state the result, recall that per our convention, the full observation space in a MA-DMSO instance is denoted by $\mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}^K$.

24

Theorem C.1 (Reducing MA-DMSO to HR-DMSO). *Consider any instance of the MA-DMSO framework satisfying Assumption 1.3 and Assumption 1.2 and specified by the tuple $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$. Then for some choice of value functions $\widetilde{f}^M$, the instance of the HR-DMSO framework specified by the tuple $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{\widetilde{f}^M\}_M)$, satisfies:*

1. *For all models $\overline{M}$ and $\varepsilon > 0$, $\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) = \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M})$.*

2. *For all $T \in \mathbb{N}$, $\mathfrak{M}(\mathscr{H}, T) = \mathfrak{M}(\mathscr{M}, T)$.*

This result proceeds by choosing the value function $\widetilde{f}^M(\pi) = K - h^M(\pi)$. Note that for this reduction to be admissible, it is critical that rewards are hidden: the function $h^M(\pi)$ is not observed directly in the MA-DMSO framework, and as we will see, this is a source of fundamental hardness.

Theorem C.1 is a fairly immediate result, and it is natural to imagine that the HR-DMSO framework might truly be more general than the MA-DMSO framework, especially since rewards *are* observed in the latter. The following result, which is the formal version of Theorem 1.3, shows that if one allows for small approximation, any instance of the HR-DMSO framework can be embedded in a two-player, zero-sum NE instance for MA-DMSO with minimal increase in complexity.

Theorem C.2 (Reducing HR-DMSO to MA-DMSO). *Consider any instance of the HR-DMSO framework specified by the tuple $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$. Then for any $V \in \mathbb{N}$, there is a two-player zero-sum NE instance $\mathscr{M} = (\widetilde{\mathcal{M}}, \widetilde{\Pi}, \widetilde{\mathcal{O}}, \Pi'_k, U_k)$ for the MA-DMSO framework (Definition 1.1) such that:*

1. *For all $\varepsilon > 0$, $\mathsf{dec}_\varepsilon(\mathscr{M}) \leq \mathsf{dec}_\varepsilon(\mathscr{H}) \leq 6/\sqrt{V} + \mathsf{dec}_{\varepsilon + (6/V)^{-1/2}}(\mathscr{M})$.*

2. *For all $T \in \mathbb{N}$, it holds that $\mathfrak{M}(\mathscr{M}, T) \leq \mathfrak{M}(\mathscr{H}, T) \leq \mathfrak{M}(\mathscr{M}, T) + O((T \log(T)/V)^{1/4})$.*

3. *$\widetilde{\mathcal{M}}$ is indexed by tuples $(M, i) \in \mathcal{M} \times [V]$. In particular, if $\mathcal{M}$ is finite, then $\log |\widetilde{\mathcal{M}}| = \log |\mathcal{M}| + \log V$.*

The main consequence of this result is that characterizing the minimax sample complexity for the MA-DMSO is no easier than characterizing the minimax sample complexity for the HR-DMSO framework; this will allow us to restrict our attention to the latter task for the results that follow. Let us make some additional remarks.

- As we increase the parameter V, the approximation to the minimax rate in Theorem C.2 improves. Choosing $V = \mathrm{poly}(T)$ suffices for all settings of interest, the only tradeoff is that the size of the model class $\mathcal{M}$ increases from $\log|\mathcal{M}|$ to $\log|\mathcal{M}| + \log V$. For the results we consider in subsequent sections, this increase will be inconsequential (beyond $\log(T)$ factors).

- Beyond preserving the minimax risk, both reductions preserve the value of the Decision-Estimation Coefficient, which is a consequence of preserving rewards and Hellinger distances for models in the class. This will become relevant for our results in the sequel (Appendix D), where we show that the DEC is closely connected to minimax risk, yet not completely equivalent.

- Both reductions are algorithmic in nature. For example, suppose that we start with a HR-DMSO instance $\mathscr{H}$ and produce a MA-DMSO instance $\mathscr{M}$ via the reduction in Theorem C.2. Then any algorithm that achieves low risk for every model in $\mathscr{M}$ can be efficiently lifted to an algorithm for the original class $\mathscr{H}$.

25

Theorem C.2 is proven by embedding a given instance $\mathscr{H}$ for the HR-DMSO framework into a two-player zero game instance $\mathscr{M}$, where the first of the two agents plays the role of the HR-DMSO agent. The key properties of the embedding are that:

1. The second agent selects actions to ensure that near-optimal decisions for the original HR-DMSO instance form Nash equilibria for the new instance, and vice-versa.

2. Even though rewards in the game instance $\mathscr{M}$ are observed, by increasing the size of the game (as a function of the parameter V), we can ensure that discovering an action with non-zero reward requires a prohibitively large amount of exploration, rendering them hidden (up to small approximation error).

Appendix D. Upper and lower bounds on minimax rates

This section presents our results regarding minimax rates for the MA-DMSO and HR-DMSO frameworks. We work in the HR-DMSO framework for the majority of the section, and give implications for the MA-DMSO at the end, using the equivalence from Appendix C. In more detail:

- In Appendix D.1, we give upper and lower bounds on the minimax rates for interactive decision making in the HR-DMSO framework, which scale with the constrained DEC.

- Next, we establish in Appendix D.2 that, under mild regularity assumptions on the constrained DEC, the upper and lower bounds on the minimax rate are separated by at most a polynomial factor (ignoring the estimation error term); for most parameter regimes, the gap between the bounds is at most quadratic. We then show—perhaps surprisingly—that neither the upper or lower bounds can be improved, in that there are instances where each is nearly tight. In other words, in contrast to the DMSO framework (Foster et al., 2023), in the HR-DMSO framework, the constrained DEC cannot not give a characterization of the minimax sample complexity which is tight beyond a quadratic factor. We show further that this gap is not limited to the constrained DEC, and in fact holds for an entire family of complexity measures based on pairwise f-divergences between models. As a result, any characterization of the minimax rate for HR-DMSO which is tight up to polylogarithmic factors must use a complexity measure substantially different from those considered in recent works (Foster et al., 2021, 2022b, 2023).

- Finally, using the equivalence shown in the previous section, we establish (Appendix D.3) that all of the results above hold verbatim in the MA-DMSO framework.

All of the results in this section are presented in a general form. We refer to Part II of the appendix for applications to specific instances of interest.

D.1. HR-DMSO: Upper and lower bounds on minimax rates

We now give upper and lower bounds on the minimax risk for the HR-DMSO framework. We obtain upper bounds as an immediate corollary of regret bounds for the *Estimation-to-Decisions$^+$* (E2D$^+$ for PAC) algorithm from recent work of Foster et al. (2023). The E2D$^+$ for PAC algorithm was introduced in the (single-agent/non-hidden-reward) DMSO framework, where it leads to tight upper bounds on minimax risk based on the constrained DEC (Foster et al., 2023). We observe that it provides identical guarantees for the more general HR-DMSO framework without modification; this can be

seen by inspecting the proof of correctness of the E2D$^+$ for PAC algorithm in Foster et al. (2023) and noting that it does not make use of the fact that the learning agent observes the rewards $r^1, \dots, r^T$. Further background on the algorithm may be found in Appendix J.1.

Our main upper bound is stated for the case in which $\mathcal{M}$ is finite ($|\mathcal{M}| < \infty$); more general guarantees for infinite classes are given in Appendix J.1.

Theorem D.1 (Minimax upper bound for HR-DMSO (Foster et al., 2023))**.** *Fix $\delta \in \left(0, \frac{1}{10}\right)$ and $T \in \mathbb{N}$, and consider any instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$. Suppose that Assumption 1.4 holds. Letting $\bar{\varepsilon}(T) := 16\sqrt{\frac{\lceil \log 2/\delta \rceil}{T} \cdot \log \frac{|\mathcal{M}|}{\delta}}$, the E2D$^+$ for PAC algorithm, when configured appropriately, guarantees that with probability at least $1 - \delta$,*

$$\mathbf{Risk}(T) \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}).$$

In addition, if $f^M(\cdot) \in [0, R]$ for all $M \in \mathcal{M}$ and some $R > 0$, then the expected risk is bounded as $\mathbb{E}[\mathbf{Risk}(T)] \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) + \delta R$.

Before interpreting this result, we complement it with our main lower bound, Theorem D.2, which shows that the minimax risk for any algorithm is lower bounded by the constrained DEC for an appropriate choice of the scale parameter $\varepsilon > 0$. The statement of this result uses the definition $C(T) := \log(T \wedge V(\mathcal{M}))$. In addition, we recall that $g^M(\pi) := f^M(\pi_M) - f^M(\pi)$.

Theorem D.2 (Minimax lower bound for HR-DMSO)**.** *Consider any instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ and write $R := \sup_{\pi \in \Pi, M \in \mathcal{M}} g^M(\pi)$. Given $T \in \mathbb{N}$, let $\underline{\varepsilon}(T) > 0$ be chosen as large as possible such that*

$$\underline{\varepsilon}(T)^2 \cdot C(T) \cdot R \cdot T \leq \frac{1}{8} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H}). \tag{8}$$

Then for any algorithm, there exists a a model in $\mathcal{M}$ for which

$$\mathbb{E}[\mathbf{Risk}(T)] \geq \frac{1}{6} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H}).$$

Understanding the bounds. We now give a sense for the behavior of the lower bound of Theorem D.2 and the upper bound of Theorem D.1 through several examples. For simplicity we consider the case that $R = \sup_{\pi \in \Pi, M \in \mathcal{M}} g^M(\pi) = 1$ (in the context of Theorem D.2).

- $\sqrt{T}$-*rates.* Most of the classes studied in the literature on bandits and reinforcement learning have the property that the optimal rate is $O(\sqrt{T})$. Many of these problems have the property that rewards are observed (i.e., they lie in the DMSO framework), but such rates also arise for problems in HR-DMSO for which rewards are not observed; a notable example is *locally observable finite partial monitoring problems* (Bartók et al., 2014). For such classes, it holds that $\mathsf{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon \cdot \sqrt{C_{\text{prob}}}$, for some problem-dependent constant $C_{\text{prob}} > 0$ reflecting the complexity of the model class $\mathcal{M}$ (see Foster et al. (2021, 2023) for examples). In this case, by choosing a failure probability of $\delta = 1/T$, we have $\bar{\varepsilon}(T) \lesssim \sqrt{\log(T) \log(T|\mathcal{M}|)/T}$, so that Theorem D.1 gives an upper bound of

$$\mathbb{E}[\mathbf{Risk}(T)] \leq \widetilde{O}\left(\sqrt{\frac{C_{\text{prob}} \log |\mathcal{M}|}{T}}\right)$$

on the minimax risk. For lower bounds, if $\mathsf{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}}$, then the solution to the fixed point equation (8) is $\underline{\varepsilon}(T) \gtrsim \sqrt{C_{\mathrm{prob}}}/(T \cdot C(T))$. This translates, via Theorem D.2, into a lower bound of

$$\mathbb{E}[\mathbf{Risk}(T)] \geq \widetilde{\Omega}\left(\frac{C_{\mathrm{prob}}}{T}\right)$$

on the minimax risk, which differs from the upper bound by a quadratic factor (ignoring the $\log|\mathcal{M}|$ factor). By the results of Foster et al. (2023), for the special case where rewards are observed (i.e., the DMSO framework), the upper bound of $\widetilde{O}\left(\sqrt{\frac{C_{\mathrm{prob}}\log|\mathcal{M}|}{T}}\right)$ is the correct rate (up to the $\log|\mathcal{M}|$ factor and $\log T$ factors). We will show in the sequel that for general settings where rewards are not observed, this is not necessarily the case, and the lower bound can be tight.

- *Nonparametric rates.* For nonparametric model classes, for which the optimal regret is $\omega(\sqrt{T})$, it is typically the case that $\mathsf{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon^{1-\rho}$ for some $\rho \in (0,1)$. For such problems, Theorem D.1 yields an upper bound of $\mathbb{E}[\mathbf{Risk}(T)] \leq \widetilde{O}\left((\log|\mathcal{M}|/T)^{(1-\rho)/2}\right)$ on the minimax risk. In contrast, the best possible solution to the fixed point equation in (8) is $\underline{\varepsilon}(T) \gtrsim 1/(T \cdot C(T))^{\frac{1}{1+\rho}}$, which translates, via Theorem D.2, into a lower bound of $\mathbb{E}[\mathbf{Risk}(T)] \geq \widetilde{\Omega}\left(1/T^{\frac{1-\rho}{1+\rho}}\right)$ on the minimax risk. Here the lower bound is off from the upper bound (ignoring the $\log|\mathcal{M}|$ factor) by a power of $\frac{2}{1+\rho} \leq 2$. By the results of Foster et al. (2023), for the special case where rewards are observed, the upper bound of $\widetilde{O}\left((\log|\mathcal{M}|/T)^{(1-\rho)/2}\right)$ is the correct rate (up to the $\log|\mathcal{M}|$ factor and $\log T$ factors).

We refer to Foster et al. (2023) for concrete examples exhibiting the growth rates sketched above for the special case where rewards are observed (DMSO), and to Part II of the appendix for examples arising from MA-DMSO.

D.2. HR-DMSO: Gaps between bounds and impossibility of tight characterizations

We now investigate the nature of the gap between the upper and lower bounds in Theorems D.1 and D.2. We first give a generic bound on the gap, then show that it is not possible—in a fairly strong sense—to close the gap further.

D.2.1. ON THE GAP BETWEEN THE UPPER AND LOWER BOUNDS

Ignoring constant factors, the only difference between the upper and lower bounds of Theorems D.1 and D.2 is the scale ε at which the DEC is computed. The upper bound of Theorem D.1 uses scale $\bar{\varepsilon}(T) = 8\sqrt{\frac{\lceil\log 2/\delta\rceil}{T} \cdot \log|\mathcal{M}|}$, whereas the lower bound of Theorem D.2 (with $R = 1$) uses the scale $\underline{\varepsilon}(T)$, which is defined implicitly to be as large as possible subject to the constraint $\underline{\varepsilon}(T)^2 \cdot C(T) \cdot T \leq \frac{1}{8} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})$. Thus, the size of the gap between $\bar{\varepsilon}(T)$ and $\underline{\varepsilon}(T)$ controls the degree of tightness of these upper and lower bounds. In what follows, we give a bound on the size of this gap that holds whenever the constrained DEC satisfies the following regularity assumption.

Assumption D.1 (Regularity). *An instance $\mathscr{H}$ (of either HR-DMSO or MA-DMSO) is said to satisfy the regularity condition with constants $C_{\mathrm{reg}}, c_{\mathrm{reg}} > 1$ at scale $\varepsilon \in (0,2)$ if*

$$\mathsf{dec}_\varepsilon(\mathscr{H}) \leq c_{\mathrm{reg}}^2 \cdot \mathsf{dec}_{\varepsilon/C_{\mathrm{reg}}}(\mathscr{H}).$$

Most natural classes satisfy Assumption D.1 for some constants $c_{\mathrm{reg}}, C_{\mathrm{reg}}$ (in particular, the condition is satisfied whenever $\mathsf{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon^p$ for $p < 2$). We note that a similar assumption used in Foster et al. (2023) to give upper bounds on the optimal rates attainable in the DMSO framework.

Under Assumption D.1, the following result shows that our upper bound on minimax risk, which scales with $\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H})$, is bounded above by a quantity that is a polynomial of our lower bound, namely $\mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})$.

Proposition D.1. *Suppose that an instance $\mathscr{H}$ (for either HR-DMSO or MA-DMSO) satisfies Assumption D.1 for some values $C_{\mathrm{reg}} > c_{\mathrm{reg}} > 1$ and for all $\varepsilon \in (\underline{\varepsilon}(T) \cdot \frac{c_{\mathrm{reg}}}{C_{\mathrm{reg}}}, \bar{\varepsilon}(T))$. Choose any $\beta \geq \frac{\log c_{\mathrm{reg}}}{\log(C_{\mathrm{reg}}/c_{\mathrm{reg}})}$. Then for any $T \in \mathbb{N}$,*

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \leq (C \log 1/\delta \cdot \log|\mathcal{M}| \cdot C(T) \cdot C_{\mathrm{reg}}/c_{\mathrm{reg}})^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{\frac{1}{1+\beta}}.$$

We remark that Proposition D.1 is a purely algebraic fact that makes no use of the structure of the DEC, and in particular holds for instances of both the HR-DMSO and MA-DMSO frameworks. To make the result concrete, we consider, we revisit each of the situations we discussed in Appendix D.1, and describe how applying Proposition D.1 allows us to conclude that our upper and lower bounds are related by a polynomial factor.

- $\sqrt{T}$*-rates.* Suppose that $\mathsf{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}}$, for some problem-dependent constant $C_{\mathrm{prob}} > 0$. Then, for any constant $\beta > 1$, there is a sufficiently large absolute constant $C_{\mathrm{reg}} > 1$ so that, for all $\varepsilon > 0$, $\mathsf{dec}_\varepsilon(\mathscr{H}) \leq C_{\mathrm{reg}}^\beta \cdot \mathsf{dec}_{\varepsilon/C_{\mathrm{reg}}}(\mathscr{H})$. It follows that Assumption D.1 is satisfied with the constants C_{reg} and $c_{\mathrm{reg}} := C_{\mathrm{reg}}^{\beta/2}$ (which satisfy $\beta \geq \frac{\log c_{\mathrm{reg}}}{\log(C_{\mathrm{reg}}/c_{\mathrm{reg}})}$), and Proposition D.1 gives that

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \leq \widetilde{O}(\log|\mathcal{M}|)^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{\frac{1}{1+\beta}}.$$

  Disregarding the estimation error and taking $\beta \to 1$, we conclude that $\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \lesssim \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{1/2-o(1)}$, i.e., there is a (roughly) quadratic gap between our upper and lower bounds.

- *Nonparametric rates.* Suppose that $\mathsf{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon^{1-\rho}$ for some $\rho \in (0,1)$. Then for any constant $\beta > \frac{1-\rho}{1+\rho}$, there is a sufficiently large constant $C_{\mathrm{reg}} > 1$ so that, for all $\varepsilon > 0$, $\mathsf{dec}_\varepsilon(\mathscr{H}) \leq C_{\mathrm{reg}}^{\beta(1+\rho)} \cdot \mathsf{dec}_{\varepsilon/C_{\mathrm{reg}}}(\mathscr{H})$. Thus, Assumption D.1 is satisfied with the constants C_{reg} and $c_{\mathrm{reg}} := C_{\mathrm{reg}}^{\beta(1+\rho)/2}$, which satisfy $\beta \geq \frac{\log c_{\mathrm{reg}}}{\log(C_{\mathrm{reg}}/c_{\mathrm{reg}})}$, and Proposition D.1 gives that

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \leq \widetilde{O}(\log|\mathcal{M}|)^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{\frac{1}{1+\beta}}.$$

  Disregarding the estimation error and taking $\beta \to \frac{1-\rho}{1+\rho}$ (so that $\frac{1}{1+\beta} \to \frac{1+\rho}{2}$), we conclude that $\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \lesssim \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{\frac{1+\rho}{2}-o(1)}$, i.e., the gap between the upper and lower bounds is smaller than quadratic.

Of course, the arguments in Appendix D.1 already allowed us to draw these conclusions directly; the purpose here is to exhibit how this conclusion can obtained as a special case of the more general Proposition D.1.

D.2.2. ON TIGHT CHARACTERIZATIONS FOR THE MINIMAX RISK

It is natural to wonder whether the polynomial gap between our upper and lower bounds can be tightened to give a characterization of the minimax risk up that is only loose by polylogarithmic factors. In this section, we show that this is not possible in several senses.

Tightness of the upper and lower bounds. In Propositions D.2 and D.3, we give two instances $\mathscr{H}_1$ $\mathscr{H}_2$, so that, up to $\log \frac{1}{\varepsilon}$ factors, we have both $\mathsf{dec}_\varepsilon(\mathscr{H}_1) \asymp \varepsilon$ and $\mathsf{dec}_\varepsilon(\mathscr{H}_2) \asymp \varepsilon$. Despite having the same behavior for the DEC, the minimax rates for the instances are different: For the instance $\mathscr{H}_1$, the upper bound from Theorem D.1 is tight ($\mathfrak{M}(\mathscr{H}_1, T) \gtrsim 1/\sqrt{T}$), yet for $\mathscr{H}_2$, the lower bound from Theorem D.2 is tight ($\mathfrak{M}(\mathscr{H}_2, T) \lesssim \log(T)/T$).

Proposition D.2 (An instance where the upper bound is tight). *For any sufficiently $L, A \in \mathbb{N}$, there is an instance $\mathscr{H}_1 = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ with $\log|\mathcal{M}| \leq \log(LA)$ and which satisfies the following properties:*

1. *For all $T \leq 2^{L/2}$, the minimax rate for $\mathscr{H}_1$ is given by $\mathfrak{M}(\mathscr{H}_1, T) = \Theta(\sqrt{A/T})$.*

2. *For all $\varepsilon \in (2^{-L}, 1/\sqrt{A})$, it holds that $c \cdot \varepsilon\sqrt{A} \leq \mathsf{dec}_\varepsilon(\mathscr{H}_1) \leq C \cdot \varepsilon\sqrt{A}$, for some constants $c, C > 0$.*

The instance $\mathscr{H}_1$ in Proposition D.2 has model class given by a subclass of multi-armed bandit problems with A arms and Bernoulli rewards, and the bounds in the proposition are an immediate consequence of prior work. We provide a proof in Appendix D.2 for completeness.

Proposition D.3 (An instance where the lower bound is tight). *For any sufficiently large $L \in \mathbb{N}$ and any $C_{\mathrm{prob}} \geq 1$, there exists an instance $\mathscr{H}_2 = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ with $\log|\mathcal{M}| \leq L^2$, satisfying the following properties:*

1. *For all $T \leq 2^L$, the minimax rate for the instance $\mathscr{H}_2$ is bounded as $\mathfrak{M}(\mathscr{H}_2, T) \leq \frac{8C_{\mathrm{prob}}^2 \log T}{T}$.*

2. *For all $\varepsilon \geq \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot 2^L}$, we have $\frac{C_{\mathrm{prob}}}{\sqrt{8} \cdot L} \cdot \varepsilon \leq \mathsf{dec}_\varepsilon(\mathscr{H}_2) \leq 2C_{\mathrm{prob}} \cdot \varepsilon$. In particular, $\underline{\varepsilon}(T) \geq \Omega\left(\frac{C_{\mathrm{prob}}}{T \log(T) \cdot L}\right)$ as long as $T \leq 2^L/L^3$.*

In particular, for any $T \in \mathbb{N}$, by choosing $L = 100 \log T$, we have that for all $\varepsilon \geq \Omega\left(\frac{1}{C_{\mathrm{prob}} \cdot T^{100}}\right)$, the instance $\mathscr{H}_2$ satisfies, $\Omega(\varepsilon \cdot C_{\mathrm{prob}}/\log \frac{1}{\varepsilon}) \leq \mathsf{dec}_\varepsilon(\mathscr{H}_2) \leq O(\varepsilon \cdot C_{\mathrm{prob}})$, yet the minimax risk is bounded as $\mathfrak{M}(\mathscr{H}_2, T) \leq O(C_{\mathrm{prob}}^2 \log(T)/T)$.

Let us compare the instances for Proposition D.2 and Proposition D.3. First, note that for both instances, the estimation complexity $\log|\mathcal{M}|$ scales as $\widetilde{O}(1)$. Thus:

- Theorem D.1, using the radius $\bar{\varepsilon}(T)$, yields an upper bound on the minimax risk of $\widetilde{O}(1/\sqrt{T})$, which is tight for $\mathscr{H}_1$.

- Theorem D.2, using the radius $\underline{\varepsilon}(T)$, yields a lower bound on the minimax risk of $\widetilde{\Omega}(1/T)$, which is tight for $\mathscr{H}_2$.

That is, the instance $\mathscr{H}_1$ establishes that our upper bound cannot be improved to use the radius $\underline{\varepsilon}(T)$, and the instance $\mathscr{H}_2$ establishes that our lower bound cannot be improved to use the radius $\bar{\varepsilon}(T)$. More generally, since $\mathsf{dec}_\varepsilon(\mathscr{H}_1)$ and $\mathsf{dec}_\varepsilon(\mathscr{H}_2)$ have the same behavior, yet $\mathscr{H}_1$ and $\mathscr{H}_2$ have different minimax rates, the constrained DEC cannot give a tight characterization of the minimax risk for the HR-DMSO framework. This contrasts the situation for the (reward-observed) DMSO framework in Foster et al. (2023), where the constrained DEC characterizes the minimax rates up to logarithmic factors whenever $\log|\mathcal{M}| = \widetilde{O}(1)$.

We remark in passing that the instances constructed in Propositions D.2 and D.3 satisfy the regularity condition of Assumption D.1 for $C_{\mathrm{reg}}, c_{\mathrm{reg}} \leq O(\log T)$ and all $\varepsilon \geq \underline{\varepsilon}(T)$. Thus, the regularity condition is not sufficient to close the gap between the upper and lower bounds.

Ruling out more general characterizations. We now show that the gaps highlighted above are not limited to the DEC, and are in fact intrinsic to a broad class of complexity measures. Our main result, Theorem D.3 shows that for any f-divergence $D(\cdot \,\|\, \cdot)$ satisfying a mild assumption, it is possible to construct two HR-DMSO instances $\mathscr{H}_1$ and $\mathscr{H}_2$ for which the minimax risk differs by a polynomial factor, yet 1) the value functions associated with $\mathscr{H}_1$ and $\mathscr{H}_2$ are identical, and 2) the pairwise $D(\cdot \,\|\, \cdot)$-divergences between all models in $\mathscr{H}_1$ and $\mathscr{H}_2$ are identical. In other words:

It is impossible to obtain a tight characterization for minimax risk that depends only on value functions and pairwise f-divergences.

Definition D.1 gives our main technical assumption regarding f-divergences: roughly speaking, it states that the function defining the f-divergence exhibits at most polynomial growth near 0 and ∞.

Definition D.1 (Bounded f-divergence). *Consider a convex function $\phi : [0, \infty) \to [0, \infty]$ so that $\phi(1) = 0$ and $\phi(x)$ is finite for all $x > 0$, and let*

$$D_\phi(\mathbb{P} \,\|\, \mathbb{Q}) := \mathbb{E}_\mathbb{Q}\left[\phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right] \tag{9}$$

denote the associated f-divergence for probability measures $\mathbb{P}$ and $\mathbb{Q}$ with $\mathbb{P} \ll \mathbb{Q}$. For constants $\alpha, \beta \geq 0$, we say that ϕ is (α, β)-bounded if, for all $x \geq 1$,

$$\phi(1/x) + \frac{\phi(x)}{x} \leq \beta \cdot x^\alpha.$$

In such a case, we say that the f-divergence D_ϕ is (α, β)-bounded.

Essentially all commonly used f-divergences satisfy Definition D.1 for small values of α and β. For the Hellinger divergence, we have $\phi(x) = (\sqrt{x} - 1)^2$, so that $D_\mathsf{H}^2(\cdot, \cdot)$ is $(0, 2)$-bounded; for the KL-divergence, we have $\phi(x) = x \ln(x) + 1 - x$, so that $D_\mathsf{KL}(\cdot \,\|\, \cdot)$ is $(0, 2)$-bounded; and for the χ^2-divergence, we have $\phi(x) = (x - 1)^2$, so that $D_{\chi^2}(\cdot, \cdot)$ is $(1, 1)$-bounded.

Remark D.1 (Non-negativity of ϕ). We remark that often, when f-divergences are presented, it is assumed that the function ϕ maps to $[-\infty, \infty]$ (as opposed to $[0, \infty]$). Assuming that ϕ maps to $[0, \infty]$ is without loss of generality, for the following reason. It is well-known that for any $c \in \mathbb{R}$, and for any convex function ϕ satisfying $\phi(1) = 0$, letting $\widetilde{\phi}(x) = \phi(x) + c \cdot (x - 1)$, we have $D_\phi = D_{\widetilde{\phi}}$. Thus, given any $\phi : [0, \infty) \to [-\infty, \infty]$, we may choose any $c \in -\partial\phi(1)$, so that $0 \in \partial(\phi(x) + c \cdot (x - 1))$, which in particular implies that $\phi(x) + c \cdot (x - 1) \geq 0$ for all x, and the f-divergence induced by $\phi(x) + c \cdot (x - 1)$ is equivalent to D_ϕ.

Theorem D.3. *For some constants $\alpha, \beta \geq 0$, suppose D_ϕ is an (α, β)-bounded f-divergence. Then for any $T \in \mathbb{N}$, $\epsilon \in (0,1)$, and $C_{\mathrm{prob}} \geq 1$, there are instances $\mathscr{H}_1 = (\mathcal{M}_1, \Pi_1, \mathcal{O}_1, \{f_1^M(\cdot)\}_{M \in \mathcal{M}_1})$, $\mathscr{H}_2 = (\mathcal{M}_2, \Pi_2, \mathcal{O}_2, \{f_2^M(\cdot)\}_{M \in \mathcal{M}_2})$ of the HR-DMSO framework, so that $\Pi_1 = \Pi_2$, $\mathcal{O}_1 = \mathcal{O}_2$, and there is a one-to-one mapping $\mathscr{E} : \mathcal{M}_1 \to \mathcal{M}_2$ satisfying:*

1. *For all $M \in \mathcal{M}_1$, $f_1^M \equiv f_2^{\mathscr{E}(M)}$.*

2. *For all $M, M' \in \mathcal{M}_1$, and $\pi \in \Pi_1$, $D_\phi(M(\pi) \parallel M'(\pi)) = D_\phi(\mathscr{E}(M)(\pi) \parallel \mathscr{E}(M')(\pi))$.*

3. *There is some constant C_ϕ depending only on ϕ so that for all T' with $T \leq T' \leq T^{3/2 - 2\epsilon} \cdot (C_\phi C_{\mathrm{prob}}^{1/2+\epsilon} \ln T)^{-1}$, it holds that*

$$\mathfrak{M}(\mathscr{H}_1, T') \leq \frac{1}{T} + 2 \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2 + \epsilon/(2\alpha)}, \qquad \mathfrak{M}(\mathscr{H}_2, T') \geq 2^{-2 - 2/\epsilon} \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2}.$$

In the event that $\alpha = 0$, the quantity $(C_{\mathrm{prob}}/T)^{1/2+\epsilon/(2\alpha)}$ in the statement of Theorem D.3 is to be interpreted as 0. In particular, if $D(\cdot \parallel \cdot)$ is the Hellinger divergence or the KL divergence, then we have $\mathfrak{M}(\mathscr{H}_1, T') \leq 1/T$ in Item 3, giving a quadratic separation. If $D(\cdot \parallel \cdot)$ is the χ^2-divergence, then we have $\mathfrak{M}(\mathscr{H}_1, T') \leq O(1/T^{1/2+\epsilon/2})$, which leads to a smaller, yet still polynomial separation for any choice of the constant $\epsilon > 0$.

Several variants of the DEC and related complexity measures depend only on the value functions $f^M(\cdot)$ (for $M \in \mathcal{M}$) and pairwise f-divergences between models in the class $\mathcal{M}$, and thus cannot provide a characterization for minimax risk in the HR-DMSO framework that is tight up to polylogarithmic factors. Below, we highlight a few notable examples.

- The distributional offset DEC (Foster et al., 2021; Chen et al., 2022a; Foster et al., 2023), is defined for $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ as:[10]

$$\mathsf{dec}_\gamma^{\mathrm{o},\mathrm{rnd}}(\mathscr{H}) = \sup_{\nu \in \Delta(\mathcal{M})} \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi \sim q}\left[\mathbb{E}_{\overline{M} \sim \nu}\left[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right]\right].$$

Clearly, this definition depends only on value functions $\{f^M\}_{M \in \mathcal{M}}$ and pairwise Hellinger distances for models in $\mathcal{M}$, and hence can only characterize minimax risk up to a quadratic factor.

- The offset DEC (Foster et al., 2021, 2023) is defined for $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ as:

$$\mathsf{dec}_\gamma^{\mathrm{o}}(\mathscr{H}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi \sim q}[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)].$$

Note that $\mathsf{dec}_\gamma^{\mathrm{o}}(\mathscr{H})$ depends on the divergence between models in $\mathcal{M}$ and those in $\mathrm{co}(\mathcal{M})$, which is not covered by Theorem D.3. However, Foster et al. (2023, Proposition D.2) show that $\mathsf{dec}_\gamma^{\mathrm{o}}(\mathscr{H}) \leq \mathsf{dec}_{\gamma/4}^{\mathrm{o},\mathrm{rnd}}(\mathscr{H}) \leq \mathsf{dec}_{\gamma/4}^{\mathrm{o}}(\mathscr{H})$ (that is, $\mathsf{dec}_\gamma^{\mathrm{o}}(\mathscr{H})$ and $\mathsf{dec}_\gamma^{\mathrm{o},\mathrm{rnd}}(\mathscr{H})$ are equivalent up to constant factors), so it follows from the previous bullet point that this complexity measure can only characterize minimax risk up to a quadratic factor.

---

10. We consider the PAC variant of the offset DEC here (Foster et al., 2023), but it is clear that our argument applies identically to the regret version of the DEC (Foster et al., 2021).

- Foster et al. (2021, 2023) consider variants of the DEC that are applied to *localized* subsets of the model class $\mathcal{M}$. In particular, the following two notions of localization have been considered in Foster et al. (2021): for some localization radius $\alpha > 0$, a model class $\mathcal{M}$, and a reference model $\overline{M}$,

$$\mathcal{M}_\alpha(\overline{M}) := \{M \in \mathcal{M} \ : \ f^M(\pi_M) \leq f^{\overline{M}}(\pi_{\overline{M}}) + \alpha\}, \quad \text{and}$$
$$\mathcal{M}_\alpha^\infty(\overline{M}) := \left\{M \in \mathcal{M} \ : \ \left|(f^M(\pi_M) - f^M(\pi)) - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))\right| \leq \alpha \ \forall \pi \in \Pi\right\}.$$

  Since these definitions only depend on the value functions $\{f^M\}_{M \in \mathcal{M}}$, Theorem D.3 implies that incorporating localization into the variants of the DEC considered above cannot help to provide a characterization of the minimax risk.

- The *information ratio* (Russo and Van Roy, 2014, 2018; Lattimore and György, 2021) was introduced to bound the Bayesian regret for posterior sampling and a more general algorithm known as *information-directed sampling*. The information ratio of a model class $\mathcal{M}$ is closely related to the DEC of the convex hull of $\mathcal{M}$; in particular, Foster et al. (2022b) showed that a parametrized version of the information ratio of $\mathcal{M}$ is equivalent to the DEC of the convex hull of $\mathcal{M}$, up to constant factors. As the DEC of $\mathrm{co}(\mathcal{M})$ involves pairwise Hellinger distances between models in the *convex hull of $\mathcal{M}$*, Theorem D.3 does not definitively rule it out as providing a characterization of minimax risk. However, the DEC of $\mathrm{co}(\mathcal{M})$ is known to be exponentially larger than the minimax risk for many natural examples (e.g., tabular reinforcement learning (Foster et al., 2022b)), so it seems unlikely to provide a tight characterization.

  There are also variants of the information ratio which Theorem D.3 *does* rule out: given a reference model $\overline{M} \in \mathcal{M}$ and a distribution $\mu \in \Delta(\mathcal{M})$, one can define (Foster et al., 2021)

$$\mathcal{I}(\mathscr{H}, \overline{M}, \mu) := \underset{p,q \in \Delta(\Pi)}{\arg\min} \frac{(\mathbb{E}_{\pi \sim p}\mathbb{E}_{M \sim \mu}[f^M(\pi_M) - f^M(\pi)])^2}{\mathbb{E}_{\pi \sim q}\mathbb{E}_{M \sim \mu}[D_{\mathsf{KL}}(M(\pi) \,\|\, \overline{M}(\pi))]}.$$

  As this definition depends only on value functions and pairwise KL-divergences for models in $\mathcal{M}$, Theorem D.3, no function of $\mathcal{I}(\mathscr{H}, \overline{M}, \mu)$ (such as a worst-case version of the information ratio defined by $\max_{\overline{M} \in \mathcal{M}} \max_{\mu \in \Delta(\mathcal{M})} \mathcal{I}(\mathscr{H}, \overline{M}, \mu)$) can provide a characterization of minimax risk.

- Note that in general, the constrained DEC $\mathsf{dec}_\varepsilon(\mathscr{H}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M})$ depends on Hellinger divergences between models in $\mathcal{M}$ and those in $\mathrm{co}(\mathcal{M})$, so Theorem D.3 does not directly rule out a characterization in terms of $\mathsf{dec}_\varepsilon(\mathscr{H})$. However, we have already ruled out such a characterization separately in Propositions D.2 and D.3. Of course, the variant $\sup_{\overline{M} \in \mathcal{M}} \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M})$, which restricts to $\overline{M} \in \mathcal{M}$, only depends on the value functions and pairwise Hellinger divergences of models in $\mathcal{M}$, and hence is covered by Theorem D.3.

Let us remark that one complexity measure not currently ruled out by our results is the generalized information ratio considered in the work of Lattimore (2022) on adversarial partial monitoring, which uses an unnormalized KL-like divergence based on the logarithmic barrier, and cannot be written in terms of f-divergences. The upper and lower bounds on regret given by Lattimore (2022) are loose by $\mathrm{poly}(|\Pi|)$ factors, and as such we find it to be unlikely that this complexity measure can give tight

guarantees in the "large decision-space/model class" regime where $T \ll \min\{|\mathcal{M}|, |\Pi|\}$, which is the focus of our work.

Remark D.2. While this is out of scope for the present paper, we remark that it is possible to establish similar impossibility results for the regret (as opposed to PAC) framework.

D.3. Implications for MA-DMSO framework

Up to this point, all of the results in this section concerned the HR-DMSO framework. Using Theorems C.1 and C.2, we can immediately derive analogous results for the MA-DMSO framework. In what follows, we state these analogues (in particular, upper and lower bounds on minimax risk, and impossibility of tighter results), all of which are corollaries the results in the prequel. We refer to Part II of the appendix for applications of these results.

Upper and lower bounds on minimax risk. We begin by stating upper and lower bounds for the minimax risk for instance of MA-DMSO in terms of the Multi-Agent DEC; these results are corollaries of Theorems D.1 and D.2.

Corollary D.1 (Minimax upper bound for MA-DMSO). Fix $\delta \in \left(0, \frac{1}{10}\right)$ and $T \in \mathbb{N}$, and consider any K-player MA-DMSO instance $\mathcal{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$. Suppose that $f_k^M(\cdot) \in [0, 1]$ for all $k \in [K]$ and $M \in \mathcal{M}$, and let $\bar{\varepsilon}(T) := 16\sqrt{\frac{\lceil \log 2/\delta \rceil}{T} \cdot \frac{\log |\mathcal{M}|}{\delta}}$. Then we have

$$\mathfrak{M}(\mathcal{M}, T) \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M}) + K\delta. \tag{10}$$

Proof of Corollary D.1. Given an instance $\mathcal{M}$ of MA-DMSO, consider the instance $\mathcal{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{\widetilde{f}^M\}_M)$ as per Theorem C.1. We have that $h^M(\cdot) \in [0, K]$ for all $M \in \mathcal{M}$, meaning that $\widetilde{f}^M(\cdot) \in [-K+1, 1]$ for all $M \in \mathcal{M}$ under the construction in the proof of Theorem C.1. By rescaling $\widetilde{f}^M(\cdot)$, the guarantee from Theorem D.1 ensures that $\mathfrak{M}(\mathcal{H}, T) \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{H}) + K\delta$, from which (10) follows using Theorem C.1. We have also used here that both $\mathfrak{M}(\mathcal{H}, T)$ and $\mathsf{dec}_\varepsilon(\mathcal{H})$ scale linearly under rescaling of the value functions $\widetilde{f}^M(\cdot)$. ∎

As we discuss further in Remark J.1, the high-probability guarantee from Theorem J.1 applies also in the MA-DMSO setting, i.e., in the contex of Corollary D.1.

Corollary D.2 (Minimax lower bound for MA-DMSO). Consider any instance $\mathcal{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ for the MA-DMSO framework with $\mathcal{R} = [0, 1]$. Given $T \in \mathbb{N}$, let $\underline{\varepsilon}(T) > 0$ be chosen as large as possible such that $\underline{\varepsilon}(T)^2 \cdot C(T) \cdot K \cdot T \leq \frac{1}{8} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})$. Then

$$\mathfrak{M}(\mathcal{M}, T) \geq \frac{1}{6} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}).$$

Corollary D.2. Given an instance $\mathcal{M}$ of MA-DMSO, consider the instance $\mathcal{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{\widetilde{f}^M\}_M)$ as per Theorem C.1. By definition of $\widetilde{f}^M$, we have that $\sup_{\pi \in \Pi, M \in \mathcal{M}} \sup_{\pi' \in \Pi} \widetilde{f}^M(\pi') - \widetilde{f}^M(\pi) \leq K$. Then we have $\mathfrak{M}(\mathcal{M}, T) = \mathfrak{M}(\mathcal{H}, T) \geq \frac{1}{6} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{H}) = \frac{1}{6} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})$, where the two equalities use Theorem C.1 and the inequality uses Theorem D.2. ∎

As we have already remarked, Proposition D.1, which bounds the gap between our upper and lower bounds based on the DEC, already applies to instances of MA-DMSO whenever Assumption D.1 is satisfied. In particular, this means that whenever $\mathsf{dec}_\varepsilon(\mathscr{M}) \propto \varepsilon^{1-\rho}$ for $\rho \in [0,1)$, we have

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M}) \leq \widetilde{O}(K^{\frac{1-\rho}{1+\rho}} \log^{\frac{1-\rho}{2}} |\mathcal{M}|) \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{M})^{\frac{1+\rho}{2}}.$$

Tightness of the gaps. Next, we provide analogues of Propositions D.2 and D.3 for the MA-DMSO. The results construct MA-DMSO instances $\mathscr{M}_1$ (Proposition D.4) and $\mathscr{M}_2$ (Proposition D.5) that exhibit the same DEC behavior, in that $\mathsf{dec}_\varepsilon(\mathscr{M}_1) \asymp \varepsilon$ and $\mathsf{dec}_\varepsilon(\mathscr{M}_2) \asymp \varepsilon$, yet have minimax rates: $\mathfrak{M}(\mathscr{M}_1, T) \gtrsim 1/\sqrt{T}$ and $\mathfrak{M}(\mathscr{M}_2, T) \lesssim \log(T)/T$. In particular, Proposition D.4 below shows that in the upper bound Corollary D.1, the scale $\bar{\varepsilon}(T)$ cannot be decreased, and Proposition D.5 below shows that in the lower bound Corollary D.2, the scale $\bar{\varepsilon}(T)$ cannot be increased.

Proposition D.4. *For any sufficiently large $L, A \in \mathbb{N}$, there is an instance $\mathscr{M}_1 = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ with $\log |\mathcal{M}| \leq \log(LA)$ and which satisfies the following properties:*

1. *For all $T \leq 2^{L/2}$, the minimax rate for the instance $\mathscr{M}_1$ is given by $\mathfrak{M}(\mathscr{M}_1, T) = \Theta(\sqrt{A/T})$.*

2. *For all $\varepsilon \in (2^{-L}, 1/\sqrt{A})$, it holds that $c \cdot \varepsilon\sqrt{A} \leq \mathsf{dec}_\varepsilon(\mathscr{M}_1) \leq C \cdot \varepsilon\sqrt{A}$, for some constants $c, C > 0$.*

Proof of Proposition D.4. We observe that the instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ used to prove Proposition D.2 immediately yields the 1-player instance of MA-DMSO given by $\mathscr{M}_1 = (\mathcal{M}, \Pi, \mathcal{O}, \Pi'_1, U_1)$, with $\Pi'_1 = \Pi$ and $U_1(\pi'_1, \pi) = \pi'_1$, since rewards are observed under all models in $\mathcal{M}$. The result then follows immediately from Proposition D.2. ∎

Proposition D.5. *For any sufficiently large $L \in \mathbb{N}$ and any $C_{\mathrm{prob}} \geq 1$, there exists an instance $\mathscr{M}_2 = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ with $\log |\mathcal{M}| \leq O(L^2 + \log C_{\mathrm{prob}})$, satisfying the following properties:*

1. *For all $T \leq 2^L$, the minimax rate for the instance $\mathscr{M}_2$ is bounded as $\mathfrak{M}(\mathscr{M}_2, T) \leq \frac{8 C_{\mathrm{prob}}^2 \log T}{T}$.*

2. *For all $\varepsilon \geq \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot 2^L}$, we have $\frac{C_{\mathrm{prob}}}{\sqrt{8} \cdot L} \cdot \varepsilon \leq \mathsf{dec}_\varepsilon(\mathscr{M}_2) \leq 2 C_{\mathrm{prob}} \cdot \varepsilon$. In particular, $\underline{\varepsilon}(T) \geq \Omega\left(\frac{C_{\mathrm{prob}}}{T \log(T) \cdot L}\right)$ as long as $T \leq 2^L / L^3$.*

Proof of Proposition D.5. Given L and C_{prob}, let $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ be the instance given per Proposition D.3. Next, let $\mathscr{M}_2 = (\widetilde{\mathcal{M}}, \widetilde{\Pi}, \widetilde{\mathcal{O}}, \{\Pi'_k\}_k, \{U_k\}_k)$ be the instance constructed per Theorem C.2 for the instance $\mathscr{H}$ with $V = 100 \cdot C_{\mathrm{prob}}^2 \cdot 2^{2L}$. We have $\log |\widetilde{\mathcal{M}}| = \log |\mathcal{M}| + \log V \leq O(L^2 + \log C_{\mathrm{prob}})$. Using the guarantees of Proposition D.3 and Theorem C.2, we have $\mathfrak{M}(\mathscr{M}_2, T) \leq \mathfrak{M}(\mathscr{H}, T) \leq \frac{8 C_{\mathrm{prob}}^2 \log T}{T}$ for $T \leq 2^L$, and for all $\varepsilon \geq \frac{2}{C_{\mathrm{prob}} \cdot 2^L}$ (which ensures that $\varepsilon - \sqrt{6/V} \geq \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot 2^L}$),

$$\frac{C_{\mathrm{prob}}}{\sqrt{8} \cdot L} \cdot \left(\varepsilon - \sqrt{6/V}\right) \leq \mathsf{dec}_{\varepsilon - (6/V)^{-1/2}}(\mathscr{H}) - 6/\sqrt{V} \leq \mathsf{dec}_\varepsilon(\mathscr{M}_2) \leq \mathsf{dec}_\varepsilon(\mathscr{H}) \leq 2 C_{\mathrm{prob}} \cdot \varepsilon.$$

Since $\varepsilon \geq \frac{1}{C_{\mathrm{prob}} \cdot 2^L}$ implies that $\sqrt{6/V} \leq \varepsilon/2$, it follows that $\frac{C_{\mathrm{prob}}}{2\sqrt{8}L} \cdot \varepsilon \leq \mathsf{dec}_\varepsilon(\mathscr{M}_2) \leq 2 C_{\mathrm{prob}} \cdot \varepsilon$. ∎

Ruling out more general characterizations. Finally, we state an analogue of Theorem D.3 for the MA-DMSO framework, which shows that any complexity measure that dependence on the instance $\mathscr{M}$ only through value functions and pairwise f-divergences can only characterize the minimax risk up to polynomial factors.

Theorem D.4. *For some constants $\alpha, \beta \geq 0$, suppose that D_ϕ is an (α, β)-bounded f-divergence (Definition D.1). Then for any $T \in \mathbb{N}$, $\epsilon > 0$, and $C_{\mathrm{prob}} \geq 1$, there are instances $\mathscr{M}_1 = (\mathcal{M}_1, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$, $\mathscr{M}_2 = (\mathcal{M}_2, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ of the MA-DMSO framework, so that there is a one-to-one mapping $\mathscr{E} : \mathcal{M}_1 \to \mathcal{M}_2$ satisfying:*

1. *For all $M \in \mathcal{M}_1$, $f_1^M \equiv f_2^{\mathscr{E}(M)}$.*

2. *For all $M, M' \in \mathcal{M}_1$, and $\pi \in \Pi$, $D_\phi(M(\pi) \parallel M'(\pi)) = D_\phi(\mathscr{E}(M)(\pi) \parallel \mathscr{E}(M')(\pi))$.*

3. *There is some constant C_ϕ depending only on ϕ so that for all T' with $T \leq T' \leq T^{3/2-2\epsilon} \cdot (C_\phi C_{\mathrm{prob}}^{1/2+\epsilon} \ln T)^{-1}$, it holds that*

$$\mathfrak{M}(\mathscr{M}_1, T') \leq \frac{1}{T} + 2 \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2+\epsilon/(2\alpha)}, \quad yet \quad \mathfrak{M}(\mathscr{M}_2, T') \geq 2^{-3-2/\epsilon} \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2}.$$

The proof uses the equivalence of Theorem C.2 to translate the construction of HR-DMSO instances in Theorem D.3 to the MA-DMSO framework. Since Theorem D.3 makes a claim about pairwise f-divergences as opposed to the constrained DEC of the instance, $\mathrm{dec}_\varepsilon(\mathscr{H})$, we cannot apply Theorem C.2 in an entirely black-box manner, yet most of the reasoning from the proof of Theorem C.2 carries over.

Appendix E. MA-DMSO: From Multi-Agent to Single-Agent

Having established upper and lower bounds on the minimax risk for the MA-DMSO framework based on the Multi-Agent Decision-Estimation Coefficient, we spend the remainder of the paper providing structural results which can be used to apply our main risk bounds to concrete settings of interest. To this end, in section we provide generic results which allow the conditions under which the multi-agent DEC can be controlled by the single-agent DEC, thereby allowing one to lift the plethora of existing results for the single-agent setting (Foster et al., 2021, 2023) to multiple agents.

Induced single-agent model classes. Consider a Nash equilibrium instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ for the MA-DMSO framework (Definition 1.1), recalling that $\Pi = \Pi_1 \times \cdots \times \Pi_K$. We will prove upper bounds on the multi-agent DEC of the instance $\mathscr{M}$ in terms of the single-agent DEC for a collection of *induced single-agent* model classes $\widetilde{\mathcal{M}}_k$ defined based on $\mathscr{M}$. To define the model classes $\widetilde{\mathcal{M}}_k$, for $M \in \mathcal{M}$ and $k \in [K]$, we first define a *single-agent model* $M|_k$ as follows: a pure observation drawn from $M|_k(\pi)$ has the distribution of the pure observation $o_\circ$ when $o_\circ \sim M(\pi)$, and the reward drawn from $M|_k(\pi)$ has the distribution of r_k when $(r_1, \ldots, r_K) \sim M(\pi)$. In other words, the model $M|_k$ is identical to M but ignores the rewards of all agents except k.

The single-agent model class $\widetilde{\mathcal{M}}_k$ is defined to have policy space Π_k, so that models in $\widetilde{\mathcal{M}}_k$ are mappings $\widetilde{M} : \Pi_k \to \Delta(\mathcal{O}_\circ \times \mathcal{R})$. In addition, $\widetilde{\mathcal{M}}_k$ is indexed by $\Pi_{-k} \times \mathcal{M}$ and its models are given as follows:

$$\widetilde{\mathcal{M}}_k = \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \, : \, \pi_{-k} \in \Pi_{-k}, \, M \in \mathcal{M}\}. \tag{11}$$

The intuition behind this definition is that for each agent k, if other agents commit to playing π_{-k}, this induces a "single-agent" environment for k. If $M \in \mathcal{M}$ is the original environment, then the model $M|_k(\cdot, \pi_{-k}) \in \widetilde{\mathcal{M}}_k$ is precisely the induced single-agent environment for k (in a decentralized protocol in which each agent observes its own reward but not the reward of other agents).

Offset Decision-Estimation Coefficient. The results in this section are most naturally stated in terms of the *offset* variant of the DEC introduced in Foster et al. (2021)—specifically, the *regret* variant which restricts to $p = q$ (that is, exploration and exploitation are coupled). For an instance $\mathcal{M}$, reference model $\overline{M}$, and scale parameter $\gamma > 0$, we define

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}, \overline{M}) := \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[h^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi \sim p}[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)] \right\}. \tag{12}$$

We remark, via Foster et al. (2023), that this notion can be related to the constrained (PAC) DEC as follows.

Proposition E.1 (Foster et al. (2023)). *For all $\overline{M} \in \mathcal{M}^+$ and $\varepsilon > 0$,*

$$\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \inf_{\gamma > 0} \big\{ \mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}, \overline{M}) \vee 0 + \gamma \varepsilon^2 \big\}. \tag{13}$$

Proposition E.1 suffices to derive tight bounds on the constrained DEC for all of the examples we will consider. It is also possible to relate the two complexity measures in the opposite direction, but this can lead to loose results (Foster et al., 2023); this will not be necessary for our purposes.

E.1. Bounding the MA-DEC for convex decision spaces

Our first result considers a general class of instances in which agents' decision spaces Π_k satisfy a *convexity* property, formally stated as Assumption E.1.

Assumption E.1 (Convexity of decision spaces). *For each $k \in [K]$, there is a finite set $\mathcal{A}_k$ (called the* pure decision set*) so that $\Pi_k = \Delta(\mathcal{A}_k)$. Furthermore, the following holds:*

1. *Each $M \in \mathcal{M}$ is linear in π, i.e., for $\pi \in \Pi$, $M(\pi) = \mathbb{E}_{a_k \sim \pi_k \forall k}[M(a)]$, where we write $a = (a_1, \ldots, a_K)$.*

2. *There is a measurable function $\varphi : \mathcal{O} \to \mathcal{A}$ so that, for all $a \in \mathcal{A}$ and $M \in \mathcal{M}$, $\mathbb{P}_{o \sim M(a)}(\varphi(o) = a) = 1$, i.e., $M(a)$ reveals a.*

This assumption is quite mild, and is satisfied whenever players 1) are allowed to randomize their actions, and 2) observe the resulting actions that are sampled at each round. In particular, this encompasses (structured) normal-form games with bandit feedback (see examples in Appendix G.3). To simplify notation, we will write $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ and $\mathcal{A}_{-k} = \prod_{k' \neq k} \mathcal{A}_{k'}$.

Our main result for this subsection, Theorem E.1, shows that for any $\overline{M} \in \Delta(\mathcal{M})$, we can bound the multi-agent DEC $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}, \overline{M})$ in terms of the single-agent DECs $\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, \overline{M}_k)$, of the K model classes $\widetilde{\mathcal{M}}_k$ and reference models $\overline{M}_k$.

Theorem E.1 (Restatement of Theorem 1.4). *Suppose that $\mathcal{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is an NE instance of the MA-DMSO framework satisfying Assumption E.1. Then for any $\gamma > 0$, it holds that*

$$\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}, \overline{M}) \leq \sum_{k=1}^{K} \sup_{\overline{M}_k \in \mathrm{co}(\widetilde{\mathcal{M}}_k)} \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, \overline{M}_k).$$

This result is quite intuitive: It shows that the complexity of centralized equilibrium computation is no larger than the complexity required for each agent to optimize their own reward in the face of a worst-case environment induced by the other players. It is proven using the following fixed-point argument: For a given agent k, if all other agents commit to a joint distribution, this induces a single-agent DMSO class $\widetilde{\mathcal{M}}_k$, and it is natural for agent k to play the strategy that minimizes the single-agent DEC for this class. This is not enough to bound the MA-DEC as-is, because we need to specify a strategy for all agents, but by applying Kakutani's fixed point theoerem, we show that it is possible for all K agents to simultaneously minimize their respective single-agent DECs with respect to the other agents' strategies. Furthermore, we remark that an immediate consequence of Theorem E.1 is that the same upper bound on $\mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathscr{M})$ holds also when $\mathscr{M}$ is a CCE or a CE instance, since Nash equilibria are always (coarse) correlated equilibria (see Appendix G.3).

As a concrete example, for the multi-armed bandit problem with A actions, we have $\mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathcal{M}) \leq O\left(\frac{A}{\gamma}\right)$ (Foster et al., 2021). Using Theorem E.1, it follows that if $\mathscr{M}$ is the class of K-player normal-form games with bandit feedback and A_k actions per player, then

$$\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathscr{M}, \overline{M}) \leq O(K) \cdot \frac{\sum_{k=1}^K A_k}{\gamma}.$$

Using Proposition E.1, we conclude that $\mathsf{dec}_\varepsilon(\mathscr{M}) \leq O\left(\varepsilon \cdot \sqrt{K \sum_{k=1}^K A_k}\right)$. We refer to Appendix G.3 for details, as well as additional examples, including structured normal-form games with linear or concave payoffs. For many of these examples, the application of Theorem E.1 leads to nearly tight bounds on $\mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathscr{M})$. However, this is not always true: In Proposition G.11 (Appendix G.3.5), we show that there are instances $\mathscr{M}$ for which $\mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\widetilde{\mathcal{M}}_k)$ is much larger than $\mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathscr{M})$.

E.2. Bounding the MA-DEC for Markov games

While Assumption E.1 is quite general, and holds for most standard normal-form game setups, a notable setting that it does not capture is that of *Markov games*, where the joint decision space Π consists of randomized non-stationary policies (formalized in Assumption E.2 below).[11] In this section, we provide an analogous result specialized to this general, non-convex setting.

Assumption E.2 (Markov game instance). *The instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ is such that for some $H \in \mathbb{N}$, finite state space $\mathcal{S}$, and finite joint action space $\mathcal{A} = \prod_{k=1}^K \mathcal{A}_k$, each model $M \in \mathcal{M}$ is a K-player, horizon-H Markov game with state space $\mathcal{S}$ and joint action space $\mathcal{A}$ (see Example 1.2). In addition, for each k, the class Π_k consists of non-stationary, randomized Markov policies, i.e.,*

$$\Pi_k = \{(\pi_{k,1}, \ldots, \pi_{k,H}) \mid \pi_{k,h} : \mathcal{S} \to \Delta(\mathcal{A}_k) \ \forall h \in [H]\}.$$

The finiteness of $\mathcal{S}$ and $\mathcal{A}$ in Assumption E.2 is made for technical reasons, so as to enable the application of fixed point theorems; our bounds in this section will not depend quantitatively on $|\mathcal{S}|$ or $|\mathcal{A}|$, and we anticipate that this assumption can be relaxed.

---

11. One might try to satisfy Assumption E.1 by convexifying each agent's decision space Π_k; however, in the setting of Markov games, this will lead the model classes $\widetilde{\mathcal{M}}_k$ defined in (11) to be prohibitively large, since the policies π_{-k} will now be *mixtures* of non-stationary Markov policies. In particular, the DEC of the induced model classes $\widetilde{\mathcal{M}}_k$ that result will in general scale with the DEC of the class of mixtures of MDPs, which is exponential even in the tabular setting (Foster et al., 2022b).

Under Assumption E.2, we provide the following analogue of Theorem E.1.

Theorem E.2. *There is a constant $C > 0$ so that the following holds. Suppose that $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is an NE instance of the MA-DMSO framework satisfying Assumption E.2. Then for any $\gamma > 0$, it holds that*

$$\sup_{\overline{M} \in \mathcal{M}} \mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathscr{M}, \overline{M}) \leq \frac{CKH \log H}{\gamma} + \sum_{k=1}^{K} \sup_{\overline{M}_k \in \widetilde{\mathcal{M}}_k} \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/(CKH \log H)}(\widetilde{\mathscr{M}}_k, \overline{M}_k).$$

As an example, for when $\mathcal{M}$ is a class of tabular MDPs with $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$, we have $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}) \leq \frac{\mathrm{poly}(S,A,H)}{\gamma}$ (Foster et al., 2021). Theorem E.2 then implies that for tabular Markov games with $|\mathcal{S}| \leq S$ and $|\mathcal{A}_k| \leq A$, we have $\sup_{\overline{M} \in \mathcal{M}} \mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathscr{M}, \overline{M}) \leq \frac{\mathrm{poly}(S,A,H,K)}{\gamma}$ and via (13),

$$\sup_{\overline{M} \in \mathcal{M}} \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M}) \leq \varepsilon \cdot \sqrt{\mathrm{poly}(S, A, H, K)}.$$

We remark that while Theorem E.1 allows for improper reference models $\overline{M} \in \mathrm{co}(\mathcal{M})$, Theorem E.2 is restricted to proper reference models $\overline{M} \in \mathcal{M}$, and hence is mainly useful in settings (such as tabular MGs) in which proper estimators are available. See Appendix G.3 examples, as well as further details.

Appendix F. MA-DMSO: On the Curse of Multiple Agents

A nuisance encountered frequently in the study of multi-agent reinforcement learning is poor scaling of sample complexity with respect to the number of agents K. In particular, algorithms which directly estimate the model $M^\star$ or agents' Q-value functions typically incur sample complexity exponential in K, due to the fact that both the model and agents' Q-value functions require at least $\exp(K)$ parameters to specify; this phenomenon has been called the *curse of multiple agents* (Jin et al., 2021b). In this section, we investigate the curse of multiple agents in the MA-DMSO framework through the lens of the Multi-Agent Decision-Estimation Coefficient.

We first remark that the upper bound on the minimax risk in terms of the DEC in our upper bound, Theorem D.1 (as well as the more general version, Theorem J.1), does indeed suffer from the curse of multiple agents: even for very simple model classes such as K-player normal-form games, the estimation error $\log|\mathcal{M}|$ in Theorem D.1 will scale exponentially in K (see examples in Appendix G.3 for details and discussion), and therefore the upper bound in Theorem D.1 will also scale exponentially in K, even though the MA-DEC is not itself exponential. Note that our lower bound (Theorem D.2) does *not* have exponential dependence on K, since (a) the DEC typically scales as $\mathsf{dec}_\varepsilon(\mathscr{M}) \asymp C_{\mathrm{prob}} \cdot \varepsilon$, where the problem-dependent constant C_{prob} depends only on the size of agents' individual action sets, thus avoiding scaling exponential in K, and (b) the bound of Theorem D.2 does not include any term involving model estimation error (in particular, it does not multiply the scale $\underline{\varepsilon}(T)$ at which the DEC is evaluated).[12]

---

12. We recall that even in the single-agent setting, the appearance of the estimation error term in the upper bound, but not in the lower bound, leads to a gap between them. Foster et al. (2023) emphasize that narrowing this gap is an important open problem.

Evading the curse of multiple agents. Celebrated results in multi-agent (bandit) learning imply that the curse of multiple agents is not necessary, at least for multi-player normal-form games with bandit feedback: if each player runs an adversarial bandit no-regret algorithm, then the empirical average of their joint action profiles over T time steps approaches a (coarse) correlated equilibrium for the game at a rate of $\mathrm{poly}(K, \max_k A_k)/\sqrt{T}$ (e.g., Rakhlin and Sridharan (2013)), where A_k is the number of actions for player k. Furthermore, a sequence of recent works has extended these results to the setting of Markov games (Jin et al., 2021b; Song et al., 2021; Mao and Basar, 2022).

It is natural to wonder if it is possible to capture these results, which avoid exponential scaling with K, through our framework and the Multi-Agent Decision-Estimation Coefficient. In light of the discussion above, this question translates to asking whether the $\log|\mathcal{M}|$ term in Theorem D.1 (more generally, the term $\mathbf{Est}_{\mathsf{H}}(T)$ in Theorem J.1, which can be controlled in terms of covering numbers), which results from estimation error, can be decreased. Note that in general, as observed in Foster et al. (2021), the estimation error term $\log|\mathcal{M}|$ appearing in Theorem D.1 cannot be removed completely, even in single-agent settings, but one might hope to replace it with a weaker quantity. One possible avenue, if possible, would be to replace $\log|\mathcal{M}|$ with $\log|\mathcal{F}_{\mathcal{M}}|$, where $\mathcal{F}_{\mathcal{M}}$ denotes the induced class of *value functions*; this approach was explored for the single-agent setting in (Foster et al., 2022a), where it leads to tighter guarantees for model-free reinforcement learning settings. However, this approach is insufficient for the purpose of avoiding the curse of multiple agents, since (an ε-cover of) the value function class $\mathcal{F}$ typically has size whose logarithm scales exponentially in K, even for normal-form games with bandit feedback (Example 1.1).

In light of this discussion, perhaps most promising approach for evading the curse of multiple agents is to aim for bounds that are analogous to Theorem D.1, but replace the factor $\log|\mathcal{M}|$ with the logarithm of the size of the agents' *decision sets*. Indeed, the logarithm of the size of the joint (pure) decision set typically does not scale exponentially in K. For instance, for K-player normal-form games in which each player has A actions, the number of pure action profiles is A^K, so its logarithm is only *linear* in K; equivalently, one can look for bounds which scale as the sum of the logarithms of the agents' individual decision sets. In the single-agent DMSO setting, Foster et al. (2021, 2022b) indeed obtain bounds that scale with $\log|\Pi|$, as opposed to $\log|\mathcal{M}|$. There is a cost to pay for this improvement, however: the upper bounds of Foster et al. (2021, 2022b) that replace $\log|\mathcal{M}|$ with $\log|\Pi|$ depend on the DEC of the *convex hull of $\mathcal{M}$*, as opposed to the DEC of $\mathcal{M}$ itself.

Our upper bound. In Theorem F.1 below, we provide an upper bound that replaces the factor $\log|\mathcal{M}|$ appearing in Theorem D.1 with $\max_k \log|\Pi'_k|$, at the cost of scaling with the MA-DEC for a convexified version of the instance $\mathscr{M}$. The quantity $\max_k \log|\Pi'_k|$ is equal to $\max_k \log(|\Sigma_k|+1)$ in the special case of CCE instances (Definition 1.2), but is also small for CE instances (Definition G.1), as well as the following more general notion of correlated equilibrium, which we refer to as a "generalized correlated equilibrium".

Assumption F.1 (Generalized correlated equilibrium). *We say that an MA-DMSO instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ satisfies the* generalized correlated equilibrium *assumption if the following holds: we have $\Pi = \Delta(\Sigma_1 \times \cdots \times \Sigma_K)$, for finite sets $\Sigma_1, \ldots, \Sigma_K$, called* pure decision sets. *Furthermore, writing $\Sigma := \Sigma_1 \times \cdots \times \Sigma_K$, the instance $\mathscr{M}$ satisfies:*

1. *Each $M \in \mathcal{M}$ is linear in π, i.e., for $\pi \in \Pi$, $M(\pi) = \mathbb{E}_{\sigma \sim \pi}[M(\sigma)]$.*

2. *The deviation functions U_k respect linearity in the sense that for all $k \in [K]$, $M \in \mathcal{M}$, and $\pi \in \Pi, \pi'_k \in \Pi'_k$, we have $f_k^M(U_k(\pi'_k, \pi)) = \mathbb{E}_{\sigma \sim \pi}[f_k^M(U_k(\pi'_k, \sigma))]$.*

It is straightforward to check that both CCE instances (Definition 1.2) and CE instances (Definition G.1) satisfy Assumption F.1 as long as the pure decision sets Σ_k are all finite.

To state our result, for an instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ of the MA-DMSO framework, we define the *convex hull* of the instance $\mathscr{M}$ to be the instance $\mathrm{co}(\mathscr{M}) := (\mathrm{co}(\mathcal{M}), \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$. We with the results in the previous section, our guarantees are most naturally stated in terms of the regret variant of the MA-DEC (r-dec$_\gamma^{\mathrm{o}}$; cf. (12)).

Theorem F.1. *Suppose that $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is an MA-DMSO instance satisfying Assumption F.1. Then, for any $T \in \mathbb{N}$ and $\delta \in (0, 1)$, there exists an algorithm (*MAExO*; Algorithm 1 in Appendix L) which produces $\widehat{\pi} \in \Pi$ such that with probability at least $1 - \delta$,*

$$\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \leq O(K) \cdot \inf_{\gamma > 0} \left\{ \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathscr{M})) + \frac{\gamma}{T} \cdot \log\left(\frac{K \cdot \max_k |\Pi'_k|}{\delta} \right) \right\}.$$

We view this result as extending guarantees that replace $\log|\mathcal{M}|$ by $\log|\Pi|$ in the single-agent setting (Foster et al., 2021, 2022b); as with those prior results, the cost is that the DEC is applied to the convex hull of the instance. For the problem of computing CCE in normal form games with K players and A actions per player, we have $\mathsf{dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathscr{M})) \lesssim \frac{A}{\gamma}$ and $\max_k \log|\Pi'_k| = \log(A)$, so this result gives

$$\mathbf{Risk}(T) \lesssim \sqrt{\frac{\mathrm{poly}(K) \cdot A}{T}};$$

see Appendix G.3 for details and further examples. Theorem F.1 shows that it is possible to avoid the curse of multiple agents for convex classes, and leads to tight guarantees for structured classes of normal-form games with bandit feedback, such as games with linear or convex payoffs. In general though, it does not lead to tight guarantees non-convex classes such as Markov games. We prove the result by adapting the powerful *exploration-by-optimization* algorithm from the single-agent setting (Lattimore, 2022; Foster et al., 2022b) in a way that exploits the unique feedback structure of the multi-agent setting. One might wonder how the guarantee of Theorem F.1 compares to what one would obtain by having each agent k run the (single-agent) exploration-by-optimization algorithm of Foster et al. (2022b) separately (applied to the model class $\widetilde{\mathcal{M}}_k$ defined in (11)) and using the resulting regret bound of Foster et al. (2022b) for each agent to obtain an approximate CCE. As we show in Proposition G.11, the guarantee of Theorem F.1 can be arbitrarily better than this alternative approach, since it involves the multi-agent DEC, r-dec$_\gamma^{\mathrm{o}}(\mathrm{co}(\mathscr{M}))$, which can be arbitrarily smaller than the DEC for the single-agent classes, r-dec$_\gamma^{\mathrm{o}}(\mathrm{co}(\widetilde{\mathcal{M}}_k))$.

Extending the result to infinite decision sets. We next explain how to extend the guarantee of Theorem F.1 to the setting where the pure decision sets Σ_k and deviation sets Π'_k are not finite. We will focus on CCE instances: consider a MA-DMSO instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ satisfying Assumption F.1. Consider subsets $\widetilde{\Sigma}_k \subseteq \Sigma_k$ and $\widetilde{\Pi}'_k \subseteq \Pi'_k$ for each k, and write $\widetilde{\Pi} = \Delta(\widetilde{\Sigma}_1 \times \cdots \times \widetilde{\Sigma}_K) \subset \Pi$. (As an example, if $\mathscr{M}$ is a CCE instance, we will often take $\widetilde{\Pi}'_k = \widetilde{\Sigma}_k \cup \{\perp\}$.) It is straightforward to see that the instance $\widetilde{\mathscr{M}} = (\mathcal{M}, \widetilde{\Pi}, \mathcal{O}, \{\widetilde{\Pi}'_k\}_k, \{U_k\}_k)$ satisfies Assumption F.1 (with pure decision sets $\widetilde{\Sigma}_k$). We now define a sense in which the instance $\widetilde{\mathscr{M}}$ is a good cover for $\mathscr{M}$.

Definition F.1. *Let $\mathscr{M}, \widetilde{\mathscr{M}}$ be defined as above. For $\varepsilon \geq 0$, we say that that $\widetilde{\mathscr{M}}$ is an ε-decision space cover for $\mathscr{M}$ if*

$$\forall M \in \mathcal{M}, \quad \forall k \in [K], \quad \forall \widetilde{\pi} \in \widetilde{\Pi}, \quad \exists \widetilde{\pi}'_k \in \widetilde{\Pi}'_k \quad s.t. \quad \max_{\pi'_k \in \Pi'_k} f_k^M(U_k(\pi'_k, \widetilde{\pi})) - f_k^M(U_k(\widetilde{\pi}'_k, \widetilde{\pi})) \leq \frac{\varepsilon}{K}.$$

We let $\mathcal{N}_\Pi(\mathscr{M}, \varepsilon) := \max_{k \in [K]} |\widetilde{\Pi}'_k|$ denote the size of the largest deviation set in the smallest such cover, and define, for $T \in \mathbb{N}$,

$$\mathsf{est}_\Pi(\mathscr{M}, T) = \inf_{\varepsilon \geq 0} \left\{ \log \mathcal{N}_\Pi(\mathscr{M}, \varepsilon) + \varepsilon T \right\}.$$

Let $\widetilde{\mathscr{M}} = (\widetilde{\mathcal{M}}, \widetilde{\Pi}, \mathcal{O}, \{\widetilde{\Pi}'_k\}_k, \{U_k\}_k)$ be an ε-decision space cover for $\mathscr{M}$. Note that, for any $\widehat{\pi} \in \widetilde{\Pi}$, it follows from Definition F.1 that

$$h^M(\widehat{\pi}) = \sum_{k=1}^K \max_{\pi'_k \in \Pi'_k} f_k^M(U_k(\pi'_k, \widehat{\pi})) - f_k^M(\widehat{\pi}) \leq \sum_{k=1}^K \max_{\widetilde{\pi}'_k \in \widetilde{\Pi}_k} f_k^M(U_k(\widetilde{\pi}'_k, \widehat{\pi})) - f_k^M(\widehat{\pi}) + \varepsilon.$$

Therefore, applying the algorithm of Theorem F.1 to an appropriate decision space cover for the instance $\mathscr{M}$ (for an appropriate choice of ε), we get the following result as an immediate corollary:

Corollary F.1. Suppose that $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is a MA-DMSO instance satisfying Assumption F.1. Then, for any $T \in \mathbb{N}$ and $\delta \in (0, 1)$, there exists an algorithm which produces $\widehat{\pi} \in \Pi$ such that with probability at least $1 - \delta$,

$$\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \leq O(K) \cdot \inf_{\gamma > 0} \left\{ \mathsf{r\text{-}dec}_\gamma^\mathsf{o}(\mathrm{co}(\mathscr{M})) + \frac{\gamma}{T} \cdot (\mathsf{est}_\Pi(\mathscr{M}, T) + \log(K/\delta)) \right\}.$$

Lower bounds for Nash equilibrium instances. Theorem F.1 relies on the assumption that $\mathscr{M}$ is a generalized correlated equilibrium instance (Assumption F.1). To close the section, we complement this result by showing that it is not possible to achieve analogous guarantees for Nash equilibria. First, in Proposition F.1 we show such an impossibility result for K-player NE instances: We give an instance for which the upper bound in Theorem F.1 is polynomial in K, yet the minimax risk is exponential in K.

Proposition F.1. *There is a constant $c_0 > 0$ so that the following holds. For any $K \in \mathbb{N}$, there is a K-player NE instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ so that:*

1. *$\max_k |\Pi'_k| = 2$.*

2. *For all $\gamma > 0$, $\mathsf{r\text{-}dec}_\gamma^\mathsf{o}(\mathrm{co}(\mathscr{M})) \leq O(K/\gamma)$.*

3. *There is no algorithm that adaptively draws $2^{o(K)}$ samples and outputs a policy with expected risk at most $c_0 \cdot K$.*

For the instance $\mathscr{M}$ in Proposition F.1, we have $|\Pi'_k| = O(1)$ so a bound of the form in Theorem F.1 would imply that $\widetilde{O}(\mathrm{poly}(K)/\epsilon^2)$ samples suffice to learn an ϵ-approximate Nash equilibrium; the lower bound on sample complexity of $2^{\Omega(K)}$ from Proposition F.1 rules this out. The proof of Proposition F.1 follows directly from well-known lower bounds on the query complexity of K-player Nash equilibria (Rubinstein, 2016; Babichenko, 2016; Chen et al., 2017).

For our last result Theorem F.2, we go even further, and show that the impossibility of proving any variant of Theorem F.1 for NE instances persists even in the case when $K = 2$ and the game is zero-sum.

Theorem F.2. *There is a constant $C_0 > 0$ so that the following holds. Fix any $N \in \mathbb{N}$ with $N \geq C_0$ and $\epsilon \in (1/N, 1)$. There is a two-player zero-sum NE instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ such that the following holds:*

1. *$\max\{|\Pi'_1|, |\Pi'_2|\} \leq |\Pi| \leq C_0 \cdot N^2/\epsilon^2$.*

2. *For all $\gamma \geq C_0$, $\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathrm{co}(\mathscr{M})) \leq \epsilon$.*

3. *There is no algorithm that adaptively draws $\sqrt{N}/C_0$ samples and outputs a policy with expected risk at most $1/C_0$.*

Observe that for the instance $\mathscr{M}$ in Theorem F.2, we have $\log|\Pi| \lesssim \log(N/\epsilon)$, so a bound of the form in Theorem F.1 would imply that roughly $\frac{\log(N/\epsilon)}{\epsilon}$ samples suffice to learn an ϵ-approximate equilibrium. The lower bound on sample complexity in Theorem F.2, which shows that $\Omega(\sqrt{N})$ samples are required, thus rules out a guarantee of this type in a fairly strong sense.

We remark that the instance $\mathscr{M}$ constructed in Theorem F.2, while an NE instance per Definition 1.1, does not correspond to the standard notion of mixed Nash equilibrium in normal-form games (see the discussion following Definition 1.1). Since the marginals of coarse correlated equilibria in two-player zero-sum games constitute mixed Nash equilibria, Theorem F.1 rules out a strengthening of Theorem F.2 which constructs an NE instance corresponding to the standard notion of mixed Nash equilibrium.

The proof of Theorem F.2 is significantly more challenging (given prior work) than that of Proposition F.1. It uses the classical support estimation problem (e.g., Paninski (2008); Canonne (2020)) to construct an instance for which the DEC is small but the minimax risk is large. This idea is natural, because the support estimation problem has large model-estimation error, and the upper bound of Theorem J.1, which involves the model estimation error, must be respected by the instance $\mathscr{M}$. Using the support estimation problem as a building block, we construct a class of two-player zero-sum games, which bears some resemblance to the construction used in the proof of Theorem C.2. However, the construction in the latter result does not ensure that $\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathrm{co}(\mathscr{M}))$ remains small, necessitating a more sophisticated approach. To ensure that $\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathrm{co}(\mathscr{M}))$ is small while maintaining a lower bound on minimax risk, we need to embed a few additional components in the construction, namely the composition of a Reed-Solomon code and a randomness extractor. We refer the reader to Appendix M for further details.

Part II
Examples

Appendix G. MA-DMSO: Examples of instances

In this section of the appendix, we give examples of instances for the MA-DMSO framework, and apply our results to derive upper and lower bounds on the minimax risk.

- In Appendix G.1 we give additional examples equilibria that can be captured in the MA-DMSO framework, focusing on correlated equilibria and variants.

- In Appendix G.2 we give detailed examples of MA-DMSO instances, including normal-form games with linear or concave payoffs (Appendix G.2.1) and Markov games (Appendix G.2.2).

- Finally, in Appendix G.3, we give bounds on the Multi-Agent Decision-Estimation Coefficient and minimax risk for variance instances, including finite-action normal-form games (Appendix G.3.1), structured normal-form games (Appendix G.3.2, Appendix G.3.3), and tabular Markov games (Appendix G.3.4). In addition, in Appendix G.3.5, we give an instance which shows that the multi-agent to single-agent reduction in Theorem E.1 can be loose in general.

G.1. Additional examples of equilibria

Definition G.1 below shows how we can use the MA-DMSO framework to capture the problem of (normal-form) *correlated equilibrium* computation in games. The definition is similar to that of CCE instances (Definition 1.2), except players' deviation sets consist of mappings from their pure decision set to itself; these mappings describe how the player deviates as a function of their pure decision.

Definition G.1 (Correlated equilibrium instance). *We say that an MA-DMSO instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ is a* correlated equilibrium (CE) instance *if the following holds:*

1. *For some finite sets $\Sigma_1, \ldots, \Sigma_K$ (called* pure decisions*), we have $\Pi = \Delta(\Sigma_1 \times \cdots \times \Sigma_K)$. We write $\Sigma = \Sigma_1 \times \cdots \times \Sigma_K$.*

2. *For each $\pi \in \Pi$ and $M \in \mathcal{M}$, it holds that $M(\pi) = \mathbb{E}_{\sigma \sim \pi}[M(\sigma)]$.*

3. *For $k \in [K]$, we have $\Pi'_k = \Sigma_k^{\Sigma_k}$, i.e., Π'_k is the set of functions $\phi : \Sigma_k \to \Sigma_k$.*

4. *For each $k \in [K]$, $\pi \in \Pi$, and $\phi \in \Pi'_k$, $U_k(\phi, \pi) \in \Delta(\Sigma)$ is the distribution whose probability mass function is given as follows:*

$$\forall \sigma \in \Sigma, \quad U_k(\phi, \pi)(\sigma) = \pi \left(\left\{ (\sigma'_k, \sigma_{-k}) \in \Sigma \; : \; \phi(\sigma'_k) = \sigma_k \right\} \right).$$

In words, $U_k(\phi, \pi)$ is the distribution of $(\phi(\sigma_k), \sigma_{-k})$, for $\sigma \sim \pi$.

Our next example considers notions of equilibria specialized to Markov games. Recall that Definitions 1.2 and G.1 describe instances that capture the notions of (coarse) correlated equilibria in *normal-form games*, in which the pure actions belong to $\Sigma = \Sigma_1 \times \cdots \times \Sigma_K$. In the setting of Markov games, often a slightly different notion of (coarse) correlated equilibrium is used, whch we show is captured by Example G.1 below.

Example G.1 (Markov (coarse) correlated equilibria in Markov games). In Example 1.2, We will show how to capture the problem of computing *Markov coarse correlated equilibria (CCE)* and *Markov correlated equilibria (CE)* (e.g., Bai et al. (2020); Liu et al. (2021); Daskalakis et al. (2022)) in the MA-DMSO framework, generalizing the notion of Markov Has equilibrium from Example 1.2. As in Example 1.2, we assume that the class $\mathcal{M}$ consists of finite-horizon Markov games with horizon $H \in \mathbb{N}$, state spaces $\mathcal{S}_h$ for $h \in [H]$, action spaces $\mathcal{A}_k$ for $k \in [K]$, and distribution $d_1 \in \Delta(\mathcal{S}_1)$, all of which are identical across all models in the model class. The pure observation space $\mathcal{O}_\circ$ consists of trajectories, and the reward space is $\mathcal{R} = [0, 1]$. For both Markov CE and Markov CCE, the joint decision space is the set Π of *Markov correlated policies*, namely policies $\pi = (\pi_1, \ldots, \pi_H)$, where each $\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$ specifies a mapping from states to

joint distributions over actions. For a model M and a joint decision $\pi \in \Pi$, an observation (trajectory) $o = \{(s_h, (a_{1,h}, \ldots, a_{K,h}), (r_{1,h}, \ldots, r_{K,h}))\}_{h \in [H]}$ is drawn as follows: first, $s_1 \sim d_1$, and then for $h \in [H]$:

- $(a_{1,h}, \ldots, a_{K,h}) \sim \pi_h(s_h)$ and $r_{k,h} \sim R_k^M(s_h, (a_{1,h}, \ldots, a_{K,h}))$.

- $s_{h+1} \sim P_h^M(\cdot | s_h, (a_{1,h}, \ldots, a_{K,h}))$.

It remains to specify the deviation sets Π'_k and switching functions U_k:

- For the case of Markov CCE, for each $k \in [K]$, the deviation set Π'_k is the set of deterministic Markov policies for player k, which take the form $\pi'_k = (\pi'_{k,1}, \ldots, \pi'_{k,H})$, where $\pi'_{k,h} : \mathcal{S}_h \to \mathcal{A}_k$. For a joint policy $\pi \in \Pi$, $U_k(\pi'_k, \pi) \in \Pi$ is the Markov correlated policy where player k plays according to $\pi'_{k,h}$ at each state and all other players play according to π. In particular, denoting $\widetilde{\pi} := U_k(\pi'_k, \pi)$, we have that $\widetilde{\pi}_h(s_h) = \pi'_{k,h}(s_h) \times \pi_{-k,h}(s_h)$, where $\pi_{-k,h}(s_h)$ denotes the marginal of $\pi_h(s_h)$ on the actions of all players but k. Summarizing, for the MA-DMSO instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$, we have that $\widehat{\pi} \in \Pi$, $h^{M^\star}(\widehat{\pi}) = 0$ if and only if $\widehat{\pi}$ is a *Markov CCE* of $M^\star$.

- For the case of Markov CE, for each $k \in [K]$, the deviation set Π'_k is simply the set of tuples $\phi = (\phi_{k,h,s})_{h \in [H], s \in \mathcal{S}_h}$, where each $\phi_{k,h,s} : \mathcal{A}_k \to \mathcal{A}_k$ is a function from $\mathcal{A}_k$ to itself. For a joint policy $\pi \in \Pi$, $U_k(\phi, \pi)$ is the Markov correlated policy $\widetilde{\pi}$ defined as follows: the joint action distribution of $\widetilde{\pi}$ at step h and state $s_h \in \mathcal{S}_h$ is the distribution given by:

$$\widetilde{\pi}_h(s)(a) = \pi_h(s)(\{(a'_k, a_{-k}) \in \mathcal{A} \; : \; \phi(a'_k) = a_k\}),$$

for joint actions $a \in \mathcal{A}$. In words, $\widetilde{\pi}_h(s)$ is the distribution of $(\phi(a'_k), a_{-k})$, for $a \sim \pi_h(s)$. Summarizing, for the MA-DMSO instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$, we have that $\widehat{\pi} \in \Pi$, $h^{M^\star}(\widehat{\pi}) = 0$ if and only if $\widehat{\pi}$ is a *Markov CE* of $M^\star$.

Note that the instances constructed above are not special cases of the CCE or CE instances (Definitions 1.2 and G.1) we consider for normal-form games. This is because the notions of Markov (C)CE discussed above are more restrictive, forcing the joint decision $\widehat{\pi}$ to be a (joint) Markov policy, as opposed to an arbitrary distribution over joint policies. Nevertheless, as Example G.1 shows, the MA-DMSO framework is sufficiently general to capture all of these notions of equilibria.   ◁

G.2. Additional examples of instances

In this section, we give additional examples of instances that capture standard equilibrium learning problems found in the literature. We begin by describing examples of structured normal-form games in Appendix G.2.1, and then consider multi-agent reinforcement learning problems in Appendix G.2.2.

G.2.1. INSTANCES FOR BANDITS

In this section, we describe several instances of structured normal-form games, which may be thought of as multi-agent generalization of structured bandit problem found in the single-agent setting. For each example we consider, the models will have the following common structure (paralleling that of Example 1.1).

- Each agent $k \in [K]$ will have a set $\mathcal{A}_k$, referred to as its *pure action set*, and the joint policy space Π will be a subset of $\Delta(\mathcal{A}) = \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$ which contains all singleton distributions $\mathbb{I}_a$.

- We will take $\mathcal{R} := [-1, 1]$ as the reward space and $\mathcal{O}_\circ := \mathcal{A}$ as the pure observation space.

- Let a class of mean reward functions $\mathcal{F} \subseteq (\mathcal{A} \to \mathcal{R}^K)$ be given. We define the model class $\mathcal{M}_\mathcal{F}$ as the set of models $M : \Pi \to \Delta(\mathcal{R}^K \times \mathcal{O}_\circ)$ for which there is some $(f_1, \ldots, f_K) \in \mathcal{F}$ so that: (a) for all singleton distributions $\mathbb{I}_a \in \Pi$, the distribution of $(r_1, \ldots, r_K, o_\circ) \sim M(\mathbb{I}_a)$ satisfies $o_\circ = a$ a.s. and $\mathbb{E}^{M, \mathbb{I}_a}[r_k] = f^M(\mathbb{I}_a) = f_k(a)$, and (b) for all $\pi \in \Pi$, $M(\pi) = \mathbb{E}_{a \sim \pi}[M(\mathbb{I}_a)]$.

In words, $\mathcal{M}_\mathcal{F}$ consists of models M where (i) value functions $f_k^M(\cdot)$ are given by some element of $\mathcal{F}$, and (ii) observations reveal the action played (via the pure observation).

First, in Example G.2, we consider a normal-form game with linearly structured rewards, generalizing the single-agent linear bandit problem (Dani et al., 2007; Abernethy et al., 2008; Bubeck et al., 2012). This example generalizes Example 1.1, which can be thought of as the special case where each player's action set is the set of standard unit vectors.

Example G.2 (Normal-form games with linear rewards). Fix $K \in \mathbb{N}$; for each player $k \in [K]$, $\mathcal{A}_k \subset \mathbb{R}^{d_k}$ for some $d_k \in \mathbb{N}$. Write $d = d_1 d_2 \cdots d_K$. Suppose that $\Theta_1, \ldots, \Theta_K \subset \mathbb{R}^d$ are convex sets so that $|\langle a_1 \otimes \cdots \otimes a_K, \theta_k \rangle| \leq 1$ for all $a_1 \in \mathcal{A}_1, \ldots, a_K \in \mathcal{A}_K$, $k \in [K]$, and $\theta_k \in \Theta_k$. Define $\mathcal{F} \subset (\mathcal{A} \to \mathbb{R}^K)$ by $\mathcal{F} = \{(a_1, \ldots, a_K) \mapsto (\langle a_1 \otimes \cdots \otimes a_K, \theta_k \rangle)_{k \in [K]} : \theta_1 \in \Theta_1, \ldots, \theta_K \in \Theta_K\}$. We can now consider the instances corresponding to finding Nash equilibria, CE, and CCE for the class of games whose payoffs are given by functions in $\mathcal{F}$:

- We first treat Nash equilibria: suppose we set $\Pi_k = \Delta(\mathcal{A}_k)$ for each $k \in [K]$ and $\Pi = \Pi_1 \times \cdots \times \Pi_k$, and define Π_k', U_k as in Definition 1.1. We define $\mathcal{M} = \mathcal{M}_\mathcal{F}$. Then the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ captures the problem of finding Nash equilibria in an unknown linear bandit game.

- Next we treat (C)CE: we set $\Pi = \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$ and define Π_k', U_k as in Definition G.1 (respectively, Definition 1.2). We define $\mathcal{M} = \mathcal{M}_\mathcal{F}$, so that the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ captures the problem of finding (coarse) correlated equilibria in an unknown linear bandit game.

$\triangleleft$

Next, Example G.3 treats the setting of *concave games* (with bandit feedback), which has received extensive attention in the game theory literature (Rosen, 1965; Even-Dar et al., 2009), as well as machine learning (Bravo et al., 2018; Maheshwari et al., 2022; Lin et al., 2021). It can also be viewed as a generalization of the problem of single-player *concave bandits* (Kleinberg, 2004; Flaxman et al., 2005; Bubeck et al., 2017; Lattimore, 2020).[13]

---

13. Often referred to as *convex bandits*, or *zeroth-order convex optimization*, since it is typically phrased in the form of loss minimization, whereas we consider reward maximization.

Example G.3 (Concave games). Given $K \in \mathbb{N}$, for each $k \in [K]$, let $d_k \in \mathbb{N}$ and $\mathcal{A}_k \subset \mathbb{R}^{d_k}$ be a convex and compact subset with nonempty interior. Set $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_K \subset \mathbb{R}^d$, where $d = d_1 + \cdots + d_K$. Define $\mathcal{F} \subset (\mathcal{A} \to \mathbb{R}^K)$ by

$$\mathcal{F} = \left\{ f : \mathcal{A} \to [0,1]^K \mid \forall k \in [K], \ \forall a_{-k} \in \mathcal{A}_{-k}, \ \ \mathcal{A}_k \ni a_k \mapsto f_k(a_k, a_{-k}) \text{ is concave and 1-Lipschitz} \right\}.$$

Above, 1-Lipschitzness is with respect to the ℓ_2 norm. We consider the following Nash and CCE instances:

- We first consider Nash equilibria: define Π'_k, U_k as in Definition 1.1, and set $\Pi = \mathcal{A}$, $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$. Then the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ captures the problem of finding Nash equilibria in concave games, a classical problem (Rosen, 1965). In the two-player zero-sum case (namely, when $K = 2$ and $f_1(a) + f_2(a) = 0$ for all $a \in \Pi$), the problem of bandit feedback which we cover has received extensive attention (Bravo et al., 2018; Maheshwari et al., 2022; Lin et al., 2021).

- We next consider coarse correlated equilibria. Define $\Pi := \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$, namely the space of Borel measures on the compact set $\mathcal{A}_1 \times \cdots \times \mathcal{A}_K \subset \mathbb{R}^d$, and set $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$. Furthermore define Π'_k, U_k as in Definition 1.2. Then the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ captures the problem of finding coarse correlated equilibria in concave games; this has received less attention than Nash equilibria in concave games.

Since the action sets $\mathcal{A}_k$ are infinite in this setting, it is not particularly natural to define a CE instance in the sense of Definition G.1. ◁

G.2.2. INSTANCES FOR MULTI-AGENT REINFORCEMENT LEARNING

We now give concrete examples of Markov game classes $\mathcal{M}$. The first example considers the special case of the instances for computing Markov Nash equilibria and Markov (coarse) correlated equilibria described in Examples 1.2 and G.1 in which the Markov game under consider is *tabular* (i.e., has finite states and actions).

Example G.4 (Equilibria in tabular Markov games). Fix parameters $K, H \in \mathbb{N}$ representing the number of players and the horizon, finite action spaces $\mathcal{A}_k$ (of size $A_k \in \mathbb{N}$) for each player $k \in [K]$, and finite state spaces $\mathcal{S}_h$ (each of size $S \in \mathbb{N}$) at each step $h \in [H]$. The instances for each of the three types of equilibria (Nash, CE, CCE) share the same observation space $\mathcal{O}$: in particular, their pure observation space is $\mathcal{O}_\circ$, the space of all possible H-step trajectories over the state and action spaces $\mathcal{S}_1, \ldots, \mathcal{S}_H$ and $\mathcal{A}$, and the reward space is $\mathcal{R} = [0,1]$.

We refer to the *tabular setting* as the model class $\mathcal{M}$ parametrized by all possible K-player Markov games with horizon H, state spaces $\mathcal{S}_h$, and action spaces $\mathcal{A}_k$, so that the sum of each player's rewards is bounded in $[0,1]$ on any positive-probability trajectory.[14] Then for the deviation and switching functions Π'_k, U_k as described in Example 1.2, the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ captures the problem of computing Markov Nash equilibrium in an unknown tabular Markov game, and for Π'_k, U_k as described in Example G.1 corresponding to the notions of Markov CCE or Markov CE, respectively, the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ captures the problem of computing Markov CCE or Markov CE, respectively, in an unknown tabular Markov game. ◁

---

14. This assumption allows us to take $\mathcal{R} = [0,1]$.

A key question in (multi-agent) online reinforcement learning is to understand what structural properties of the model class $\mathcal{M}$ permit efficient learnability. In the simplest case (known as the *tabular* case), the state and action spaces $\mathcal{S}_h, \mathcal{A}$ are all finite, and $\mathcal{M}$ consists of all models specified by arbitrary transitions P_h^M and reward distributions $R_{k,h}^M$ with uniformly bounded support. By restricting $\mathcal{M}$, our formulation also captures a more complex settings that incorporate function approximation (Chen et al., 2022b; Li et al., 2022; Xie et al., 2020; Jin et al., 2022; Huang et al., 2021; Zhan et al., 2022; Liu et al., 2022). In what follows, we give one such example.

Example G.5 (Equilibria in linear mixture Markov games). Fix parameters $K, H \in \mathbb{N}$ representing the number of players and the horizon, finite action spaces $\mathcal{A}_k$ for each $k \in [K]$, and finite state spaces $\mathcal{S}_h$ for each $h \in [H]$. For a *dimension* parameter $d \in \mathbb{N}$, we are given mappings $\phi_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, $\psi_{k,h} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that for all $h \in [H]$ and $k \in [K]$

$$\sum_{s_{h+1} \in \mathcal{S}_{h+1}} \phi_h(s_{h+1}|s_h, a) = \mathbf{1} \in \mathbb{R}^d, \quad \text{and} \quad \|\psi_{k,h}(s_h, a)\|_2 \leq 1$$

for all $s_h \in \mathcal{S}_h, a \in \mathcal{A}, s_{h+1} \in \mathcal{S}_{h+1}$.[15] The instances we construct have pure observation space $\mathcal{O}_\circ$ given by the set of all possible H-step trajectories over the action and state spaces $\mathcal{A}$ and $\mathcal{S}_h$, and have reward space $\mathcal{R} = [0, 1]$.

For some $B \in \mathbb{N}$, the set of *linear mixture Markov games* is the model class $\mathcal{M}$ consisting of all K-player Markov games M with horizon H, state spaces $\mathcal{S}_h$, and action spaces $\mathcal{A}_k$, for which there are vectors $\theta_h^M \in \mathbb{R}^d$ satisfying $\|\theta_h^M\|_2 \leq B$ and

$$P_h^M(s_{h+1}|s_h, a) = \langle \theta_h^M, \phi_h(s_{h+1}|s_h, a) \rangle, \qquad R_{k,h}^M(s_h, a) = \langle \theta_h^M, \psi_{k,h}(s_h, a) \rangle$$

for all $h \in [H], k \in [K], s_h \in \mathcal{S}_h, a \in \mathcal{A}, s_{h+1} \in \mathcal{S}_{h+1}$, and for which under any positive-probability trajectory, $\sum_{h=1}^H r_{k,h} \in [0, 1]$.

For the deviation and switching functions Π_k', U_k as described in Example 1.2, the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ captures the problem of computing Markov Nash equilibrium in an unknown linear mixture Markov game, and for Π_k', U_k as described in Example G.1 corresponding to the notions of Markov CCE or Markov CE, respectively, the instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi_k'\}_k, \{U_k\}_k)$ captures the problem of computing Markov CCE or Markov CE, respectively, in an unknown linear mixture Markov game.

$\triangleleft$

G.3. Computing bounds on the DEC and minimax risk of multi-agent instances

In this section, we apply our results from Appendices D to F to (a) give bounds on the DEC of various MA-DMSO instances, and (b) use these bounds on the DEC to derive bounds on the minimax risk for learning equilibria in multi-agent interactive decision making.

G.3.1. NORMAL-FORM GAMES WITH FINITE ACTION SPACES

We begin with perhaps the simplest example: finite-action normal-form games with bandit feedback. We consider Nash, CE, and CCE instances, as described in Example 1.1. Let us fix $K \in \mathbb{N}$ along with action sets $\mathcal{A}_1, \ldots, \mathcal{A}_K$ for each of the K players, with joint action set $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$.

---

15. The values of $\phi_H(s_{h+1}|s_h, a)$ will not matter, so we may take $\mathcal{S}_{H+1}$ to be, e.g., the set consisting of single state.

We write $A_k := |\mathcal{A}_k|$ for $k \in [K]$. Let $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ denote the NE, CE, and CCE instances, respectively, constructed in Example 1.1. In this section, we bound the DEC of these instances; we begin with an upper bound on the offset DEC, which immediately yields an upper bound on the constrained DEC via Proposition E.1.

Proposition G.1. *For any $\gamma > 0$, the instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ defined above satisfy*

$$\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \leq \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}) \leq \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}) \leq \frac{K \cdot \sum_{k=1}^{K} A_k}{\gamma}.$$

Proof of Proposition G.1. Note that the instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ share the same observation space $\mathcal{O}$, i.e., they have pure observation space $\mathcal{O}_{\circ} = \mathcal{A}$ and reward space $\mathcal{R} = [0,1]$.[16] Thus, let us write $\mathscr{M}^{\mathrm{NE}} = (\mathcal{M}, \Pi^{\mathrm{NE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{NE}}\}_k, \{U_k^{\mathrm{NE}}\}_k)$, $\mathscr{M}^{\mathrm{CE}} = (\mathcal{M}, \Pi^{\mathrm{CE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CE}}\}_k, \{U_k^{\mathrm{CE}}\}_k)$, and $\mathscr{M}^{\mathrm{CCE}} = (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$. To distinguish between the three different settings, we augment the functions $f^M(\cdot)$ and $h^M(\cdot)$ with the superscripts NE/CE/CCE. For example, for the instance $\mathscr{M}^{\mathrm{NE}}$, we have, for $M \in \mathcal{M}, \pi \in \Pi^{\mathrm{NE}}$,

$$f_k^{M,\mathrm{NE}}(\pi) := \mathbb{E}^{M,\pi}[r_k], \quad \text{and} \quad h^{M,\mathrm{NE}}(\pi) = \sum_{k=1}^{K} \max_{\pi'_k \in (\Pi'_k)^{\mathrm{NE}}} f_k^{M,\mathrm{NE}}(U_k^{\mathrm{NE}}(\pi'_k, \pi)) - f_k^{M,\mathrm{NE}}(\pi).$$

The functions $h^{M,\mathrm{CE}} : \Pi^{\mathrm{CE}} \to \mathbb{R}$ and $h^{M,\mathrm{CCE}} : \Pi^{\mathrm{CCE}} \to \mathbb{R}$ are defined analogously.

It holds that $\Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$; furthermore, for any $M \in \mathcal{M}$ and $\pi \in \Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$, we have that $h^{M,\mathrm{CCE}}(\pi) \leq h^{M,\mathrm{CE}}(\pi)$. It immediately follows that $\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \leq \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}})$. Next, note that $\Pi^{\mathrm{NE}} \subset \Pi^{\mathrm{CE}}$, and for any $\pi \in \Pi^{\mathrm{NE}}$ and $M \in \mathcal{M}$, we have that $h^{M,\mathrm{NE}}(\pi) = h^{M,\mathrm{CE}}(\pi)$. Hence, for $\overline{M} \in \mathrm{co}(\mathcal{M})$,

$$\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}, \overline{M}) = \inf_{p \in \Delta(\Pi^{\mathrm{NE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{NE}}(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right]$$

$$\geq \inf_{p \in \Delta(\Pi^{\mathrm{CE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{CE}}(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right] = \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}, \overline{M}).$$

This establishes that

$$\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \leq \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}) \leq \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}).$$

It remains to upper bound $\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}})$. For $k \in [K]$, we write $\Pi_k := \Delta(\mathcal{A}_k)$ and $\Pi_{-k} := \prod_{k' \neq k} \Pi_{k'}$. For each $k \in [K]$, define the model class $\widetilde{\mathcal{M}}_k \subset (\Pi_k \to \Delta(\mathcal{R} \times \mathcal{O}_{\circ}))$ as in Eq. (11); in particular:

$$\widetilde{\mathcal{M}}_k = \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ \pi_{-k} \in \Pi_{-k}, M \in \mathcal{M}\}.$$

Next define the model class $\mathcal{M}'_k \subset (\mathcal{A}_k \to \Delta(\mathcal{R} \times \{\perp\}))$ by

$$\mathcal{M}'_k = \{M \ : \ M(a_k) \in \Delta(\mathcal{R} \times \{\perp\}) \ \forall a_k \in \mathcal{A}_k\},$$

---

16. Technically, the model class for the instance $\mathscr{M}^{\mathrm{NE}}$ only acts on product distributions in $\Pi^{\mathrm{NE}} = \Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_K)$, as opposed to $\Pi^{\mathrm{CCE}} = \Pi^{\mathrm{CE}} = \Delta(\mathcal{A}) \supset \Pi^{\mathrm{NE}}$; we will formally interpret the domain of $\mathcal{M}$ for the instance $\mathscr{M}^{\mathrm{NE}}$ as Π^{NE} to avoid cluttering notation.

i.e., $M(a_k)$ is allowed to be an arbitrary distribution over $\mathcal{R} \times \{\perp\}$ for each a_k. Proposition 5.2 of Foster et al. (2021) shows that $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}'_k) \leq \frac{A_k}{\gamma}$. Next, fix $\overline{M} \in \mathrm{co}(\widetilde{\mathcal{M}}_k)$, and let $\overline{M}' \in \mathrm{co}(\mathcal{M}'_k)$ be the unique model so that the reward $r \sim \overline{M}'(a_k)$ is distributed identically to the reward $r \sim \overline{M}(a_k)$ for all $a_k \in \mathcal{A}_k$. Then we have

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\widetilde{\mathcal{M}}_k, \overline{M})$$

$$= \inf_{p \in \Delta(\Pi_k)} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p} \left[\max_{\pi'_k \in \Pi_k} f_k^{M,\mathrm{NE}}(\pi'_k, \pi_{-k}) - f_k^{M,\mathrm{NE}}(\pi_k, \pi_{-k}) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k)\big) \right]$$

$$\leq \inf_{p \in \Delta(\mathcal{A}_k)} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{a_k \sim p} \left[\max_{a'_k \in \mathcal{A}_k} f_k^{M,\mathrm{NE}}(a'_k, \pi_{-k}) - f_k^{M,\mathrm{NE}}(a_k, \pi_{-k}) - \gamma \cdot D_{\mathsf{H}}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k)\big) \right]$$

$$\leq \inf_{p \in \Delta(\mathcal{A}_k)} \sup_{M' \in \mathcal{M}'_k} \mathbb{E}_{a_k \sim p} \left[\max_{a'_k \in \mathcal{A}_k} f_k^{M'}(a'_k) - f_k^{M'}(a_k) - \gamma \cdot D_{\mathsf{H}}^2\big(M'(a_k), \overline{M}'(a_k)\big) \right] = \mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}'_k, \overline{M}'),$$

where the first inequality follows since $\mathcal{A}_k \subset \Pi_k$ (by identifying each action $a_k \in \mathcal{A}_k$ with its indicator distribution $\mathbb{I}_{a_k} \in \Pi_k$), and the second inequality follows since for any $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$, there is a model $M' \in \mathcal{M}'_k$ so that for all $a_k \in \mathcal{A}_k$, the distribution of the reward $r \sim M(a_k, \pi_{-k})$ is identical to the distribution of $r \sim M'(a_k)$. Note that in the display above we have associated actions $a_k \in \mathcal{A}_k$ with their singleton distribution $\mathbb{I}_{a_k} \in \Pi_k$, per our convention. It follows that $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\widetilde{\mathcal{M}}_k) \leq \mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}'_k)$ for all $\gamma > 0$. Finally, by Theorem E.1 applied to the instance $\mathcal{M}^{\mathrm{NE}}$, we have that

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}^{\mathrm{NE}}) \leq \sum_{k=1}^{K} \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k) \leq \sum_{k=1}^{K} \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/K}(\mathcal{M}'_k) \leq \frac{K \cdot \sum_{k=1}^{K} A_k}{\gamma}.$$

Note that our application of Theorem E.1 is valid since Assumption E.1 is satisfied by the definition of $\mathcal{M}^{\mathrm{NE}}$ in Example 1.1 (in particular, our assumption that $M(\mathbb{I}_a) \in \Delta(\mathcal{R}^K) \times \{\mathbb{I}_a\}$, i.e., that M reveals a, satisfies the second point of Assumption E.1). ∎

Using Proposition G.1, we now bound the minimax rates for the instances $\mathcal{M}^{\mathrm{NE}}, \mathcal{M}^{\mathrm{CE}}, \mathcal{M}^{\mathrm{CCE}}$. To simplify matters slightly, we consider slightly simplified special cases of these instances in which the model class is constrained to models which output rewards according to the Bernoulli distribution (i.e., the rewards are $\{0, 1\}$-valued).[17] Furthermore, we assume for simplicity that $A_k \geq 2$ for all k. We denote the corresponding MA-DMSO instances with Bernoulli rewards by $\mathcal{M}_0^{\mathrm{NE}}, \mathcal{M}_0^{\mathrm{CE}}, \mathcal{M}_0^{\mathrm{CCE}}$. First, we bound the minimax rate for $\mathcal{M}_0^{\mathrm{NE}}$:

Proposition G.2. *There is an algorithm for the instance $\mathcal{M}_0^{\mathrm{NE}}$ which guarantees that with probability at least $1 - \delta$, $\mathbf{Risk}(T) \leq \sqrt{(\max_k A_k) \cdot A \cdot T^{-1}} \cdot \mathrm{polylog}(T, A, \delta^{-1})$, where $A = A_1 A_2 \cdots A_K$.*

It is evident that the same upper bound on risk for $\mathcal{M}_0^{\mathrm{NE}}$ in Proposition G.2 applies to $\mathcal{M}_0^{\mathrm{CE}}, \mathcal{M}_0^{\mathrm{CCE}}$ since for any decision $\widehat{\pi} \in \Pi^{\mathrm{NE}} \subset \Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$, we have $h^{M,\mathrm{CCE}}(\widehat{\pi}) \leq h^{M,\mathrm{CE}}(\widehat{\pi}) \leq h^{M,\mathrm{NE}}(\widehat{\pi})$ (recall the definition of $h^{M,\mathrm{NE}}, h^{M,\mathrm{CE}}, h^{M,\mathrm{CCE}}$ in the proof of Proposition G.1).

---

17. This restriction of the model class is essentially without loss of generality: given any model class with general reward distributions in $[0, 1]$, we can simulate samples from a model class with the same value functions $f_k^M(\cdot)$ and Bernoulli reward distributions by, upon receiving rewards $(r_1, \ldots, r_K) \sim M(\pi)$, replacing each r_k with a sample $r'_k \sim \mathrm{Ber}(r_k)$.

Proof of Proposition G.2. The combination of Proposition G.1 and Proposition E.1 yields that $\mathrm{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{NE}}) \leq \varepsilon \cdot 2\sqrt{K \cdot \sum_{k=1}^K A_k}$. Since, $|\mathcal{O}| \leq 2^K A$ (as rewards are assumed to be Bernoulli), the class $\mathcal{M}$ satisfies Assumption J.2 with $B = 2^K A$, and therefore Proposition J.1 gives that $\mathbf{Est}_{\mathsf{H}}(T, \delta) = O(\mathrm{est}(\mathcal{M}, T) + \log \delta^{-1}) \cdot K^2 \cdot \log^2(\max_k A_k)$. Finally, by discretizing the reward means into multiples of ε^2, we see that $\mathcal{N}(\mathcal{M}, \varepsilon) \leq (1/\varepsilon^2)^{AK}$, which implies that $\mathrm{est}(\mathcal{M}, T) \leq O(AK \cdot \log(T))$. Therefore, Theorem J.1 combined with Theorem C.1 gives that there is an algorithm with

$$\mathbf{Risk}(T) \leq \sqrt{K(A_1 + \cdots + A_K)} \cdot \sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(T, \delta)}{T}} \cdot \mathrm{polylog}(T, 1/\delta) \leq \sqrt{\frac{\max_k A_k \cdot A}{T}} \cdot \mathrm{polylog}(T, 1/\delta, A).$$

∎

Note that the upper bound of Proposition G.2 suffers from the curse of multiple agents: the number of joint action profiles A is exponential in the number of agents K. It is a well-known result that such exponential dependence on K is necessary for learning (e.g., Rubinstein (2016); see Proposition F.1), while it is not necessary for learning (coarse) correlated equilibria. We next show that our results in Appendix F allow us to recover this improved (polynomial) bound for (coarse) correlated equilibria:

Proposition G.3. *Fix any $T \in \mathbb{N}, \delta \in (0, 1)$. There is an algorithm for the instance $\mathscr{M}_0^{\mathrm{CE}}$ which produces $\widehat{\pi} \in \Pi^{\mathrm{CE}}$ such that with probability at least $1 - \delta$,*

$$\mathbf{Risk}(T) \leq \sqrt{\frac{K^4 \max_k A_k^2}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).$$

Furthermore, there is an algorithm for the instance $\mathscr{M}_0^{\mathrm{CCE}}$ which produces $\widehat{\pi} \in \Pi^{\mathrm{CCE}}$ such that with probability at least $1 - \delta$,

$$\mathbf{Risk}(T) \leq \sqrt{\frac{K^3 \sum_{k=1}^K A_k}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).$$

Proof of Proposition G.3. The statement of the proposition is an immediate consequence of Theorem F.1. For the instance $\mathscr{M}_0^{\mathrm{CCE}}$, we have that $\mathsf{r\text{-}dec}_\gamma^{\mathsf{o}}(\mathrm{co}(\mathscr{M}_0^{\mathrm{CCE}})) \leq \mathsf{r\text{-}dec}_\gamma^{\mathsf{o}}(\mathrm{co}(\mathscr{M}^{\mathrm{CCE}})) = \mathsf{r\text{-}dec}_\gamma^{\mathsf{o}}(\mathscr{M}^{\mathrm{CCE}}) \leq \frac{K \sum_{k=1}^K A_k}{\gamma}$, where we have used that the model class $\mathcal{M}$ is convex and Proposition G.1. Therefore, Theorem F.1 gives that there is an algorithm achieving

$$\mathbf{Risk}(T) \leq O\left(K \cdot \inf_{\gamma > 0} \left\{ \frac{K \sum_{k=1}^K A_k}{\gamma} + \frac{\gamma}{T} \cdot \log\left(\frac{K \cdot \max_k A_k}{\delta}\right) \right\}\right)$$

$$\leq \sqrt{\frac{K^3 \sum_{k=1}^K A_k}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).$$

Next, for the instance $\mathscr{M}_0^{\mathrm{CE}}$, the same upper bound on the DEC of $\mathrm{co}(\mathscr{M}_0^{\mathrm{CE}})$ holds, but the deviation sets are larger: we have $\max_k |\Pi_k'| = \max_k |A_k^{A_k}|$, and so Theorem F.1 gives

$$\mathbf{Risk}(T) \leq O\left(K \cdot \inf_{\gamma > 0} \left\{ \frac{K \sum_{k=1}^K A_k}{\gamma} + \frac{\gamma}{T} \cdot \log\left(\frac{K \cdot \max_k A_k^{A_k}}{\delta}\right) \right\}\right)$$

$$\leq \sqrt{\frac{K^4 \max_k A_k^2}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).$$

Lower bounds. Next we discuss lower bounds for the instances $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{NE}}$. It is straightforward to see that each of them embeds an instance of single-player $\max_k A_k$-armed bandits, by restricting the model class $\mathcal{M}$ to models for which the reward distribution depends only on the action taken by any single player k. It then follows from the proof of Proposition 5.3 of Foster et al. (2021) that $\mathsf{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{NE}}) \geq \mathsf{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CE}}) \geq \mathsf{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CCE}}) \geq \Omega(\varepsilon\sqrt{\max_k A_k})$ for $\varepsilon > 0$; in fact, these lower bounds are obtained by subclasses of $\mathcal{M}$ which have $C(T) = \log(T \wedge V(\mathcal{M})) = O(1)$. Therefore, Theorem D.2 (with $\underline{\varepsilon}(T) = \frac{c\sqrt{\max_k A_k}}{KT}$, for sufficiently small $c > 0$) together with Theorem C.1 gives that for any of the instances $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{NE}}$, and any algorithm, there is a model for which $\mathbb{E}[\mathbf{Risk}(T)] \geq \Omega(\max_k A_k/(KT))$ under any of these three instances.

For the instance $\mathscr{M}_0^{\mathrm{CCE}}$, in the learnable regime $T > \max_k A_k$, this lower bound is off from the upper bound of Proposition G.3 by a factor of $\sqrt{T/\max_k A_k} \cdot \mathrm{poly}(K, \max_k \log A_k, \log T)$; for $\mathscr{M}_0^{\mathrm{CE}}$, the gap increases to $\sqrt{T/\max_k A_k} \cdot \max_k \sqrt{A_k} \cdot \mathrm{poly}(K, \max_k \log A_k, \log T)$, and for $\mathscr{M}_0^{\mathrm{NE}}$, the gap increases further to $\sqrt{T/\max_k A_k} \cdot \sqrt{A} \cdot \mathrm{polylog}(T, A)$. In all these cases, the factor of $\sqrt{T}$ in the gap is due to the impossibility results discussed in Appendix D.2.2, and the remaining terms are due to model estimation error appearing in the upper bound but not the lower bound. In particular (up to a $O(K)$ factor), there is no gap in the upper and lower bounds we have computed on the MA-DEC for these instances.

G.3.2. NORMAL-FORM GAMES WITH LINEAR PAYOFFS

In this section we bound the DEC and minimax regret for the linearly structured normal-form game instances defined in Example G.2. In particular, fix action sets $\mathcal{A}_k \subset \mathbb{R}^{d_k}$ for each $k \in [K]$, as well as convex sets $\Theta_1, \ldots, \Theta_K \subset \mathbb{R}^d$ (with $d = d_1 \cdots d_K$) so that $|\langle a_1 \otimes \cdots \otimes a_K, \theta_k \rangle| \leq 1$ for all $(a_1, \ldots, a_K) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$, $k \in [K]$, $\theta_k \in \Theta_k$. Let $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ denote the NE, CE, and CCE instances constructed given the sets $\mathcal{A}_k, \Theta_k$ as in Example G.2. The below proposition bounds the (regret) offset DEC of these instances:

Proposition G.4. *For any $\gamma > 0$, the instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ defined above satisfy*

$$\mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathscr{M}^{\mathrm{CCE}}) \leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathscr{M}^{\mathrm{CE}}) \leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathscr{M}^{\mathrm{NE}}) \leq \frac{K \cdot \sum_{k=1}^K d_k}{\gamma}.$$

Proof of Proposition G.4. The proof is essentially identical to that of Proposition G.1, except that each induced model class $\widetilde{\mathcal{M}}_k$ can be viewed as a single-agent linear bandit problem in d dimensions, allowing us to use Proposition 6.1 of Foster et al. (2021) to bound the DEC for (single-player) linear bandits. ∎

Using Proposition G.4, we now bound the minimax rates for the instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$. As in the previous subsection, to simplify matters, we restrict the instances so that the model class is constrained to models which output (random) rewards that take values in $\{-1, 1\}$ (recall that, for the linear bandit instances defined in Example G.2, $f_k^M(a) \in [-1, 1]$ for all $M \in \mathcal{M}, a \in \mathcal{A}$). Furthermore, we assume that for each k, $d_k \geq 2$ and all $\theta_k \in \Theta_k$ satisfy $\|\theta_k\|_2 \leq D$ and all $a_k \in \mathcal{A}_k$ satisfy $\|a_k\|_2 \leq D$ for some $D > 0$. It follows that $\|a_1 \otimes \cdots \otimes a_K\|_2 \leq D^K$ for all $a_1 \in \mathcal{A}_1, \ldots, a_k \in \mathcal{A}_k$; our bounds depend only logarithmically on D. We denote the corresponding MA-DMSO instances with $\{-1, 1\}$-valued rewards by $\mathscr{M}_0^{\mathrm{NE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{CCE}}$. First, we bound the minimax rate for $\mathscr{M}_0^{\mathrm{NE}}$:

Proposition G.5. *For any $T \in \mathbb{N}, \delta \in (0,1)$, there is an algorithm for the instance $\mathscr{M}_0^{\mathrm{NE}}$ which guarantees that with probability at least $1 - \delta$, $\mathbf{Risk}(T) \leq \sqrt{(\max_k d_k) \cdot d \cdot T^{-1}} \cdot \mathrm{polylog}(T, d, \delta^{-1})$, where $d = d_1 d_2 \cdots d_K$.*

Proof of Proposition G.5. Analogous to our notation for finite-action normal-form games, let us write $\mathscr{M}_0^{\mathrm{NE}} = (\mathcal{M}, \Pi^{\mathrm{NE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{NE}}\}_k, \{U_k^{\mathrm{NE}}\}_k)$. The combination of Proposition G.1 and Proposition E.1 yields that $\mathrm{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{NE}}) \leq 2\varepsilon \cdot \sqrt{K \cdot \sum_{k=1}^{K} d_k}$. For any $\pi \in \Pi^{\mathrm{NE}} \subset \Delta(\mathcal{A})$, the distribution on $\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ = \mathcal{R}^K \times \mathcal{A}$ defined by $\nu(\cdot | \pi) := \mathrm{Unif}(\{-1, 1\}^K) \times \pi$ verifies that $\mathcal{M}$ satisfies Assumption J.2 with $B = 2^K$, and therefore Proposition J.1 gives that $\mathbf{Est}_{\mathsf{H}}(T, \delta) = O(\mathrm{est}(\mathcal{M}, T) + \log \delta^{-1}) \cdot K^2$. Finally, note that a product of ε^2/D^K-covers of Θ_k, for $k \in [K]$, with respect to the Euclidean norm yields a ε-model class cover of $\mathcal{M}$ in the sense of Definition J.1. Since each Θ_k has a ε^2/D^K-cover of size $O(D^{K+1}/\varepsilon^2)^d$, it follows that $\mathcal{N}(\mathcal{M}, \varepsilon) \leq (D^{K+1}/\varepsilon^2)^{Kd}$, which implies that $\mathrm{est}(\mathcal{M}, T) \leq O(K^2 d \cdot \log(TD))$. Therefore, Theorem J.1 combined with Theorem C.1 gives that there is an algorithm with

$$\mathbf{Risk}(T) \leq \sqrt{K(d_1 + \cdots + d_K)} \cdot \sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(T, \delta)}{T}} \cdot \mathrm{polylog}(T, 1/\delta) \leq \sqrt{\frac{\max_k d_k \cdot d}{T}} \cdot \mathrm{polylog}(T, 1/\delta, d).$$

■

As in the case of finite-action normal-form games, the upper bound in Proposition G.5 (which also applies to the instances $\mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{CCE}}$) suffers from the curse of multiple agents. For the instance $\mathscr{M}_0^{\mathrm{CCE}}$, we obtain improved bounds with minimax risk scaling only polynomially with K by appealing to our results in Appendix F.

Proposition G.6. *For any $T \in \mathbb{N}, \delta \in (0, 1)$, there is an algorithm for the instance $\mathscr{M}_0^{\mathrm{CCE}}$ which guarantees that with probability at least $1 - \delta$,*

$$\mathbf{Risk}(T) \leq \sqrt{\frac{K^5 \cdot \max_k\{d_k\}}{T}} \cdot \log(KDT/\delta).$$

One might wonder whether a similar bound can be established for the instance $\mathscr{M}_0^{\mathrm{CE}}$. According to our definition of $\mathscr{M}_0^{\mathrm{CE}}$ (which is a CE instance per Definition G.1) we have $|\Pi'_k| = |\mathcal{A}_k|^{|\mathcal{A}_k|}$ for each k, meaning that the upper bound of Theorem F.1 would yield a risk bound with polynomial dependence on $|\mathcal{A}_k|$, which is unacceptable in the linear bandit setting since $\mathcal{A}_k$ is often taken to be exponentially large or infinite. Even if we were to attempt to use Corollary F.1 to decrease the size of the deviation sets, the only choice of deviation set that works generically is $\widetilde{\Pi}'_k := \widetilde{\Sigma}_k^{\widetilde{\Sigma}_k}$, which has logarithm scaling exponentially in the dimension d_k. A more promising avenue is to consider notions of equilibria between CCE and CE (sometimes known as Φ-equilibria), as in, e.g., Gordon et al. (2008); Anagnostides et al. (2022); Mansour et al. (2022); we leave this direction for future work.

Proof of Proposition G.6. The proposition follows as a consequence of Corollary F.1. Paralleling our notation for normal-form games, let us write $\mathscr{M}_0^{\mathrm{CCE}} = (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$. Let us write $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_K := \{a_1 \otimes \cdots \otimes a_K : a_1 \in \mathcal{A}_1, \ldots, a_K \in \mathcal{A}_K\}$. For each $k \in [K]$, there is an $\varepsilon/(KD^K)$-cover with respect to the ℓ_2-norm of $\mathcal{A}_k$ of size at most $O(KD^{K+1}/\varepsilon)^{d_k}$. Let us denote such a cover by $\widetilde{\mathcal{A}}_k \subseteq \mathcal{A}_k$. Let us write $\widetilde{\mathcal{A}} = \widetilde{\mathcal{A}}_1 \times \cdots \times \widetilde{\mathcal{A}}_K$, $\widetilde{\Pi}^{\mathrm{CCE}} = \Delta(\widetilde{\mathcal{A}})$, and $(\widetilde{\Pi}'_k)^{\mathrm{CCE}} := \widetilde{\mathcal{A}}_k \cup \{\perp\}$ for each $k \in [K]$. Consider any model $M \in \mathcal{M}$. Note that, for any $k \in [K]$,

and $a_k \in \mathcal{A}_k$, there is some $\widetilde{a}'_k \in \widetilde{\mathcal{A}}_k$ so that for all $\widetilde{a} \in \widetilde{\mathcal{A}}$,

$$
\begin{aligned}
|f_k^M(U_k(a_k, \widetilde{a})) - f_k^M(U_k(\widetilde{a}'_k, \widetilde{a}))| =& |\langle \widetilde{a}_1 \otimes \cdots \otimes a_k \otimes \cdots \otimes \widetilde{a}_K, \theta_k^M \rangle - \langle \widetilde{a}_1 \otimes \cdots \otimes \widetilde{a}'_k \otimes \cdots \otimes \widetilde{a}_K, \theta_k^M \rangle| \\
\leq& \|\theta_k^M\|_2 \cdot D^{K-1} \cdot \|\widetilde{a}_k - \widetilde{a}'_k\|_2 \leq \varepsilon/K,
\end{aligned}
$$

which in particular implies that the instance $\widetilde{\mathscr{M}}_0^{\mathrm{CCE}} := (\mathcal{M}, \widetilde{\Pi}^{\mathrm{CCE}}, \mathcal{O}, \{(\widetilde{\Pi}'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$ is a ε-decision space cover for $\mathscr{M}^{\mathrm{CCE}}$ (per Definition F.1). It therefore follows that $\mathrm{est}_\Pi(\mathscr{M}_0^{\mathrm{CCE}}, T) \leq K \cdot \max_k\{d_k\} \cdot \log(KDT)$. We have $\mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathscr{M}_0^{\mathrm{CCE}})) \leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathscr{M}^{\mathrm{CCE}})) = \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathscr{M}^{\mathrm{CCE}}) \leq \frac{K \cdot \sum_{k=1}^K d_k}{\gamma}$ by Proposition G.4 and convexity of the class $\mathcal{M}$, which follows since the sets Θ_k are convex. By Corollary F.1, we have that there is an algorithm with

$$
\begin{aligned}
\mathbf{Risk}(T) \leq& O(K) \cdot \inf_{\gamma > 0} \left\{ \frac{K \cdot \sum_{k=1}^K d_k}{\gamma} + \frac{\gamma}{T} \cdot K \cdot \max_k\{d_k\} \cdot \log(KDT/\delta) \right\} \\
\leq& \sqrt{\frac{K^5 \cdot \max_k\{d_k\}}{T}} \cdot \log(KDT/\delta).
\end{aligned}
$$

$\blacksquare$

Lower bounds. We now derive lower bounds for the instances $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{NE}}$ under the assumption that $\mathcal{A}_k, \Theta_k$ contain the unit ℓ_2 ball in their respective spaces.[18] It follows from the proof of Proposition 6.2 of Foster et al. (2021) that $\mathsf{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{NE}}) \geq \mathsf{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CE}}) \geq \mathsf{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CCE}}) \geq \Omega(\varepsilon \sqrt{\max_k d_k})$ for $\varepsilon > 0$, using the fact that each of these instances embeds an instance of single-player linear bandits in dimension $\max_k d_k$. Therefore, Theorem D.2 (with $\underline{\varepsilon}(T) = \frac{c\sqrt{\max_k d_k}}{KT}$, for sufficiently small $c > 0$) together with Theorem C.1 gives that for any of the instances $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{NE}}$, and any algorithm, there is a model for which $\mathbb{E}[\mathbf{Risk}(T)] \geq \Omega(\max_k d_k/(KT))$ under any of these three instances. Similar considerations apply to the gaps between the upper and lower bounds as discussed in Appendix G.3.1.

G.3.3. CONCAVE (BANDIT) GAMES

We now bound the DEC and minimax regret for the normal-form games wth concave rewards given in Example G.3. Fix sets $\mathcal{A}_k \subset \mathbb{R}^{d_k}$ and the class $\mathcal{F} \subset (\mathcal{A} \to \mathbb{R}^K)$ as described in Example G.3. We assume that $\|a\|_2 \leq D$ for all $a \in \mathcal{A}$, for some $D > 0$.; our bounds depend only logarithmically on D. Let $\mathscr{M}^{\mathrm{NE}} = (\mathcal{M}, \Pi^{\mathrm{NE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{NE}}\}_k, \{U_k^{\mathrm{NE}}\}_k)$ denote the NE instance constructed in Example G.3, and let $\mathscr{M}^{\mathrm{CCE}} = (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$ denote the CCE instance constructed in Example G.3.[19] The below proposition bounds the (regret) offset DEC of these instances:

Proposition G.7. *For any $\gamma > 0$, the instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CCE}}$ defined above satisfy*

$$
\mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathscr{M}^{\mathrm{CCE}}) \leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathscr{M}^{\mathrm{NE}}) \leq \frac{K \cdot \sum_{k=1}^K d_k^4}{\gamma} \cdot \mathrm{polylog}\left(\max_k\{d_k\}, D, \gamma\right).
$$

---

18. Analogous lower bounds can be obtained under alternative action and parameter sets; for instance, if $\mathcal{A}_k$ each contains the ℓ_1 ball and Θ_k each contains the ℓ_∞ ball, then we can embed the normal-form game setting from the previous subsection.

19. As we have done previously, we use the model class $\mathcal{M}$ for both instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CCE}}$, where it is understood that models have domain appropriate for each instance.

Proof of Proposition G.7. The fact that $\mathsf{r\text{-}dec}^{\mathsf{o}}_\gamma(\mathscr{M}^{\mathrm{CCE}}) \leq \mathsf{r\text{-}dec}^{\mathsf{o}}_\gamma(\mathscr{M}^{\mathrm{NE}})$ follows from the fact that Π^{NE} may be identified as a subset of Π^{CCE} (namely, Π^{NE} consists of singleton distributions in Π^{CCE}), in a similar manner to the proof of Proposition G.1. To prove the second upper bound, we will use Theorem E.1 applied to the instance $\mathscr{M}^{\mathrm{NE}}$, which gives that $\mathsf{r\text{-}dec}^{\mathsf{o}}_\gamma(\mathscr{M}^{\mathrm{NE}}) \leq \sum_{k=1}^K \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k)$, for $\widetilde{\mathcal{M}}_k$ defined as in (11). In turn, to bound the DEC of $\widetilde{\mathcal{M}}_k$, we define the model class $\mathcal{M}'_k \subset (\mathcal{A}_k \to \Delta(\mathcal{R} \times \{\bot\}))$, by $\mathcal{M}'_k = \{M \ : \ f^M(\cdot) \text{ is concave}\}$. Since, for any $k \in [K]$, $M \in \mathcal{M}^{\mathrm{NE}}$, $a_{-k} \in \mathcal{A}_{-k}$, there is a model $M'_k \in \mathcal{M}'_k$ so that, for all $a_k \in \mathcal{A}_k$, the distribution of $r \sim M'_k(a_k)$ is the same as the distribution of $r_k \sim M(a_k, a_{-k})$, it holds that $\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k) \leq \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma/K}(\mathcal{M}'_k)$. Finally, Proposition 6.3 of Foster et al. (2021) (which is a restatement of Theorem 3 of Lattimore (2020)) gives that, for all $\gamma' > 0$, $\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma'}(\mathcal{M}'_k) \leq \frac{d_k^4}{\gamma} \cdot \mathrm{polylog}(d_k, \mathrm{diam}(\mathcal{A}_k), \gamma)$, which yields that $\mathsf{r\text{-}dec}^{\mathsf{o}}_\gamma(\mathscr{M}^{\mathrm{NE}}) \leq \frac{K \sum_{k=1}^K d_k^4}{\gamma} \cdot \mathrm{polylog}\left(\max_k\{d_k\}, D, \gamma\right)$. ∎

We now turn our attention to bounding the minimax risk. The model classes $\mathcal{M}^{\mathrm{NE}}, \mathcal{M}^{\mathrm{CCE}}$ for our concave game instances are extremeley large: any cover of $\mathcal{M}^{\mathrm{NE}}$ or $\mathcal{M}^{\mathrm{CCE}}$ in the sense of Definition J.1 must have logarithm exponential in the dimensions d_k, so the model-based guarantee from Theorem J.1 is not particularly interesting, even in the case where K is small. Therefore, we turn directly to the policy-based guarantees given in Appendix F, and will prove a minimax risk upper bound for the instance $\mathscr{M}^{\mathrm{CCE}}$. It turns out that such an upper bound will immediately imply upper bounds for the instance $\mathscr{M}^{\mathrm{NE}}$, under the following assumption, specializing Even-Dar et al. (2009).

Assumption G.1 (Zero-sum socially concave)**.** *We say that a model class $\mathcal{M}$ is* zero-sum socially concave *if for all $M \in \mathcal{M}$, $k \in [K]$ and $a_k \in \mathcal{A}_k$, the mapping $\mathcal{A}_{-k} \ni a_{-k} \mapsto f_k^M(a_k, a_{-k})$ is a convex function and for all $a \in \mathcal{A}$, $\sum_{k=1}^K f_k^M(a) = 0$.*

In the special case that the model M is a two-player zero-sum concave game (i.e., $f_1^M + f_2^M \equiv 0$), zero-sum social concavity necessarily holds.

Proposition G.8. *Then for any $T \in \mathbb{N}, \delta \in (0, 1)$, there is an algorithm for the instance $\mathscr{M}^{\mathrm{CCE}}$ which guarantees that with probability at least $1 - \delta$,*

$$\mathbf{Risk}(T) \leq \frac{K^2 \cdot \max_k\{d_k^{2.5}\}}{\sqrt{T}} \cdot \mathrm{polylog}\left(D, T, \gamma, \max_k\{d_k\}, K, 1/\delta\right). \tag{14}$$

Suppose further that the model class $\mathcal{M}$ is zero-sum socially concave (i.e., it satisfies Assumption G.1). Then there is an algorithm for the instance $\mathscr{M}^{\mathrm{NE}}$ which guarantes the same upper bound on risk in Eq. (14) with probability at least $1 - \delta$.

Proof of Proposition G.8. For each $k \in [K]$, there is an ε-cover with respect to the ℓ_2-norm of $\mathcal{A}_k$ of size at most $O(D/\varepsilon)^{d_k}$. Let us denote such a cover by $\widetilde{\mathcal{A}}_k \subset \mathcal{A}_k$. Write $\widetilde{\mathcal{A}} := \widetilde{\mathcal{A}}_1 \times \cdots \times \widetilde{\mathcal{A}}_K$. Let $\widetilde{\Pi}^{\mathrm{CCE}} := \Delta(\mathcal{A})$, and $(\widetilde{\Pi}'_k)^{\mathrm{CCE}} := \widetilde{\mathcal{A}}_k \cup \{\bot\}$. Note that, for any $M \in \mathcal{M}$, $k \in [K]$, and $a_k \in \mathcal{A}_k$, there is some $\widetilde{a}'_k \in \widetilde{\mathcal{A}}_k$ so that for all $\widetilde{a} \in \widetilde{\mathcal{A}}$,

$$|f_k^M(U_k(a_k, \widetilde{a})) - f_k^M(U_k(\widetilde{a}'_k, \widetilde{a}))| = |f_k^M(a_k, \widetilde{a}_{-k}) - f_k^M(\widetilde{a}'_k, \widetilde{a}_{-k})| \leq \|a_k - \widetilde{a}'_k\|_2 \leq \varepsilon,$$

where the first inequality uses 1-Lipschitzness of $f_k^M(\cdot)$. Hence, the CCE instance $\widetilde{\mathscr{M}}^{\mathrm{CCE}} := (\mathcal{M}, \widetilde{\Pi}^{\mathrm{CCE}}, \mathcal{O}, \{(\widetilde{\Pi}'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$ is an ε-decision space cover for $\mathscr{M}^{\mathrm{CCE}}$ (per Definition F.1). It follows that $\mathsf{est}_\Pi(\mathscr{M}^{\mathrm{CCE}}, T) \leq \max_k\{d_k\} \cdot \log(DT)$.

Next, we have $\text{r-dec}_\gamma^{\text{o}}(\text{co}(\mathscr{M}^{\text{CCE}})) = \text{r-dec}_\gamma^{\text{o}}(\mathscr{M}^{\text{CCE}}) \leq \frac{K \cdot \sum_{k=1}^{K} d_k^4}{\gamma} \cdot \text{polylog}(\max_k\{d_k\}, D, \gamma)$ by Proposition G.7 and convexity of the class $\mathcal{M}$ (which follows since the convex combination of concave and 1-Lipschitz functions is concave and 1-Lipschitz). By Corollary F.1, for any $T \in \mathbb{N}$ and $\delta > 0$, there is an algorithm which outputs $\widehat{\pi} \in \Pi^{\text{CCE}}$ so that with probability at least $1 - \delta$,

$$
\begin{aligned}
\mathbf{Risk}(T) \leq & K \cdot \inf_{\gamma > 0} \left\{ \frac{K \cdot \sum_{k=1}^{K} d_k^4}{\gamma} + \frac{\gamma}{T} \cdot \max_k\{d_k\} \right\} \cdot \text{polylog}\left(D, T, \gamma, \max_k\{d_k\}, K, 1/\delta \right) \\
\leq & \frac{K^2 \cdot \max_k\{d_k^{2.5}\}}{\sqrt{T}} \cdot \text{polylog}\left(D, T, \gamma, \max_k\{d_k\}, K, 1/\delta \right).
\end{aligned}
$$

Next we prove the upper bound for $\mathscr{M}^{\text{NE}}$. As we have done previously in this section, for $\widehat{\pi} \in \Pi^{\text{CCE}}$ and $M \in \mathcal{M}$, we write $h^{M,\text{CCE}}(\widehat{\pi})$ to denote the suboptimality of $\widehat{\pi}$ with respect to the instance $\mathscr{M}^{\text{CCE}}$, and for $\widehat{\pi} \in \Pi^{\text{NE}}$, we write $h^{M,\text{NE}}(\widehat{\pi})$ to denote the suboptimality of $\widehat{\pi}$ with respect to the instance $\mathscr{M}^{\text{NE}}$. Given $\widehat{\pi} \in \Pi^{\text{CCE}}$, define $\widehat{a} := \mathbb{E}_{a \sim \widehat{\pi}}[a] \in \mathcal{A}$. For each $k \in [K]$ and $M \in \mathcal{M}$, we have that

$$
\begin{aligned}
h^{M,\text{NE}}(\widehat{a}) &= \sum_{k=1}^{K} \max_{a_k' \in \mathcal{A}_k} f_k^M(a_k', \widehat{a}_{-k}) - f_k^M(\widehat{a}) \\
&\leq \sum_{k=1}^{K} \max_{a_k' \in \mathcal{A}_k} \mathbb{E}_{a_{-k} \sim \widehat{\pi}}[f_k^M(a_k', a_{-k})] - f_k^M(\widehat{a}) \\
&= \sum_{k=1}^{K} \max_{a_k' \in \mathcal{A}_k} \mathbb{E}_{a_{-k} \sim \widehat{\pi}}[f_k^M(a_k', a_{-k})] - \mathbb{E}_{a \sim \widehat{\pi}}[f_k^M(a)] = h^{M,\text{CCE}}(\widehat{\pi}),
\end{aligned}
$$

where the first inequality follows from social concavity and the second equality follows from the fact that $\sum_{k=1}^{K} f_k^M(\widehat{a}) = 0 = \sum_{k=1}^{K} \mathbb{E}_{a \sim \widehat{\pi}}[f_k^M(a)]$. Thus, given a decision $\widehat{\pi}$ output by our algorithm for the instance $\mathscr{M}^{\text{CCE}}$, we may simply output $\widehat{a} = \mathbb{E}_{a \sim \widehat{\pi}}[a]$, which yields the same upper bound on risk. ∎

Lower bounds. Assume that $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ contains the unit ℓ_2-ball. Then the instances $\mathscr{M}^{\text{CCE}}, \mathscr{M}^{\text{NE}}$ each embed a single-player linear bandit instance with dimension $\max_k d_k$ (namely, by taking the subclass of $\mathcal{F}$ to consist of linear functions in a_k only), and so the lower bounds from Appendix G.3.2 give $\text{dec}_\varepsilon(\mathscr{M}^{\text{NE}}) \geq \text{dec}_\varepsilon(\mathscr{M}^{\text{CCE}}) \geq \Omega(\varepsilon \sqrt{\max_k d_k})$ and a minimax risk lower bound of $\Omega(\max_k d_k/(KT))$. In this setting, even the DEC lower bound (in the single-agent setting) is off from the upper bound implied by Proposition G.7 and Proposition E.1 (Foster et al., 2021).

G.3.4. TABULAR MARKOV GAMES

We next give bounds on minimax risk for the instances corresponding to Markov Nash equilibria, Markov CE, and Markov CCE in tabular Markov games, as described in Example G.4. Given $H \in \mathbb{N}$, state spaces $\mathcal{S}_h$ each of size S, action spaces $\mathcal{A}_k$ of size $A_k := \mathcal{A}_k$, and an initial distribution $d^1 \in \Delta(\mathcal{S}_1)$, let $\mathscr{M}^{\text{NE}} = (\mathcal{M}, \Pi^{\text{NE}}, \mathcal{O}, \{(\Pi_k')^{\text{NE}}\}_k, \{U_k^{\text{NE}}\}_k)$, $\mathscr{M}^{\text{CE}} = (\mathcal{M}, \Pi^{\text{CE}}, \mathcal{O}, \{(\Pi_k')^{\text{CE}}\}_k, \{U_k^{\text{CE}}\}_k)$, and $\mathscr{M}^{\text{CCE}} = (\mathcal{M}, \Pi^{\text{CCE}}, \mathcal{O}, \{(\Pi_k')^{\text{CCE}}\}_k, \{U_k^{\text{CCE}}\}_k)$ be the MA-DMSO instances corresponding to Markov Nash equilibria, Markov CE, and Markov CCE as defined in Example G.4. Technically, the

model class $\mathcal{M}$ for $\mathscr{M}^{\mathrm{NE}}$ acts on policies in Π^{NE}, whereas the model class $\mathcal{M}$ for $\mathscr{M}^{\mathrm{CE}}$ and $\mathscr{M}^{\mathrm{CCE}}$ acts on policies in $\Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}} \neq \Pi^{\mathrm{NE}}$; we will write the model class for each instance as $\mathcal{M}$ and formally interpret its domain as the appropriate decision space, to avoid cluttering notation.

In Proposition G.9 below, we begin with an upper bound on their offset DEC, which immediately yields an upper bound on the constrained DEC via Proposition E.1.

Proposition G.9. *For any $\gamma > 0$, and any $\overline{M} \in \mathcal{M}$, the instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ defined above satisfy*

$$\mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{CCE}}, \overline{M}) \leq \mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{CE}}, \overline{M}) \leq \mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{NE}}, \overline{M}) \leq \frac{27 K H^3 \log(H) S \sum_{k=1}^{K} A_k}{\gamma}.$$

Proof of Proposition G.9. As in the proof of Proposition G.1, we augment the functions $f^M(\cdot)$ and $h^M(\cdot)$ with the superscripts NE/CE/CCE to distinguish between the value functions for models in the three different instances. For example, for the instance $\mathscr{M}^{\mathrm{NE}}$, we have, for $M \in \mathcal{M}, \pi \in \Pi^{\mathrm{NE}}$,

$$f_k^{M,\mathrm{NE}}(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^{H} r_{k,h}\right], \qquad h^{M,\mathrm{NE}}(\pi) = \sum_{k=1}^{K} \max_{\pi_k' \in (\Pi_k')^{\mathrm{NE}}} f_k^{M,\mathrm{NE}}(U_k^{\mathrm{NE}}(\pi_k', \pi)) - f_k^{M,\mathrm{NE}}(\pi).$$

The functions $h^{M,\mathrm{CE}} : \Pi^{\mathrm{CE}} \to \mathbb{R}$ and $h^{M,\mathrm{CCE}} : \Pi^{\mathrm{CCE}} \to \mathbb{R}$ are defined similarly.

We have $\Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$; furthermore, for any $M \in \mathcal{M}$ and $\pi \in \Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$, we have that $h^{M,\mathrm{CCE}}(\pi) \leq h^{M,\mathrm{CE}}(\pi)$. Thus $\mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{CCE}}, \overline{M}) \leq \mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{CE}}, \overline{M})$. Next, note that we may identify Π^{NE} as a subset of Π^{CE} as follows: for $\pi = (\pi_1, \ldots, \pi_K) \in \Pi^{\mathrm{NE}}$, we associate it to the joint Markov policy $\widetilde{\pi} = (\widetilde{\pi}_1, \ldots, \widetilde{\pi}_H) \in \Pi^{\mathrm{CE}}$ where $\widetilde{\pi}_h(s_h)$ is the product distribution $\widetilde{\pi}_h(s_h) := \pi_{1,h}(s_h) \times \cdots \times \pi_{K,h}(s_h)$. It is straightforward to see that, for such π and any model $M \in \mathcal{M}$, the distributions of $M(\pi)$ and $M(\widetilde{\pi})$ are identical. Accordingly, with slight abuse of notation, for $\pi \in \Pi^{\mathrm{NE}}$, we denote its corresponding policy in Π^{CE} as π as well. Thus we have $h^{M,\mathrm{NE}}(\pi) = h^{M,\mathrm{CE}}(\pi)$, and for any $\overline{M} \in \mathcal{M}$, we have

$$
\begin{aligned}
\mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{NE}}, \overline{M}) &= \inf_{p \in \Delta(\Pi^{\mathrm{NE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{NE}}(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right] \\
&\geq \inf_{p \in \Delta(\Pi^{\mathrm{CE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{CE}}(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right] = \mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{CE}}, \overline{M}).
\end{aligned}
$$

It remains to upper bound $\mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathscr{M}^{\mathrm{NE}})$. For $k \in [K]$, let Π_k be the class of randomized Markov policies of player k (so that $\Pi^{\mathrm{NE}} = \Pi_1 \times \cdots \times \Pi_K$). For each $k \in [K]$, define the model class $\widetilde{\mathcal{M}}_k \subset (\Pi_k \to \Delta(\mathcal{R} \times \mathcal{O}_{\circ}))$ as in (11):

$$\widetilde{\mathcal{M}}_k = \left\{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \; : \; \pi_{-k} \in \Pi_{-k}, \; M \in \mathcal{M}\right\}.$$

Define $\mathcal{M}_k'$ to be the model class consisting of all horizon-H Markov decision processes with action set $\mathcal{A}_k$ and state spaces $\mathcal{S}_1, \ldots, \mathcal{S}_H$, and so that the sum of rewards under any trajectory that occurs with positive probability is bounded in $[0, 1]$. Formally, the pure observation space of $\mathcal{M}_k'$ is the space $\mathcal{O}_{\circ}'$ of trajectories $\{(s_h, a_{k,h}, r_{k,h})\}_{h \in [H]}$, with $s_h \in \mathcal{S}_h, a_{k,h} \in \mathcal{A}_k, r_{k,h} \in \mathbb{R}$, its reward space is $\mathcal{R} = [0, 1]$, and its decision space is Π_k. Thus $\mathcal{M}_k' \subset (\Pi_k \to \Delta(\mathcal{R} \times \mathcal{O}_{\circ}'))$. Proposition 5.4 of Foster et al. (2021) shows that for all $\overline{M} \in \mathcal{M}_k'$, $\mathsf{r}\text{-}\mathsf{dec}_{\gamma}^{\mathrm{o}}(\mathcal{M}_k', \overline{M}) \leq 26 \frac{H^2 S A_k}{\gamma}$.

Next, fix $\widetilde{M} \in \widetilde{\mathcal{M}}_k$. By definition of $\widetilde{\mathcal{M}}_k$, we can find $\overline{M} \in \mathcal{M}$ and $\overline{\pi}_{-k} \in \Pi_{-k}$ so that $\widetilde{M}(\pi_k) = \overline{M}|_k(\pi_k, \overline{\pi}_{-k})$ for all $\pi_k \in \Pi_k$. Let $\overline{M}' \in \mathcal{M}_k'$ be the unique model so that for all $\pi_k \in \Pi_k$,

the marginal distribution of $\{(s_h, a_{k,h}, r_{k,h})\}_{h \in [H]}$ for a trajectory drawn from $\widetilde{M}(\pi_k)$ is identical to the distribution of the pure observation drawn from $\overline{M}'(\pi_k)$. Such a model exists, since for each state $s_h \in \mathcal{S}_h$ and action $a_{k,h} \in \mathcal{A}_k$, the transition distribution $P_h^{\overline{M}'}(\cdot|s_h, a_{k,h}) \in \Delta(\mathcal{S}_{h+1})$ is defined as $\mathbb{E}_{a_{k',h} \sim \overline{\pi}_{k',h}(s_h) \forall k' \neq k}[P_h^{\overline{M}}(\cdot|s_h, (a_{k,h}, a_{-k,h}))]$ and the reward distribution $R_{k,h}^{\overline{M}'}(s_h, a_{k,h}) \in \Delta(\mathbb{R})$ is defined as $\mathbb{E}_{a_{k',h} \sim \overline{\pi}_{k',h}(s_h) \forall k' \neq k}[R_h^{\overline{M}}(s_h, (a_{k,h}, a_{-k,h}))]$. We now compute

$$\text{r-dec}_\gamma^o(\widetilde{\mathcal{M}}_k, \widetilde{M}) \tag{15}$$

$$= \inf_{p \in \Delta(\Pi_k)} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p} \left[\max_{\pi_k' \in \Pi_k} f_k^{M,\text{NE}}(\pi_k', \pi_{-k}) - f_k^{M,\text{NE}}(\pi) - \gamma \cdot D_{\mathsf{H}}^2\Big(M(\pi_k, \pi_{-k}), \widetilde{M}(\pi_k)\Big) \right]$$

$$\leq \inf_{p \in \Delta(\Pi_k)} \sup_{M' \in \mathcal{M}_k'} \mathbb{E}_{\pi_k \sim p} \left[\max_{\pi_k' \in \Pi_k} f_k^{M'}(\pi_k') - f_k^{M'}(\pi_k) - \gamma \cdot D_{\mathsf{H}}^2\big(M'(\pi_k), \overline{M}'(\pi_k)\big) \right]$$

$$= \text{r-dec}_\gamma^o(\mathcal{M}_k', \overline{M}'),$$

where the inequality follows since, via the same argument used to construct $\overline{M}'$, for any $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$, there is some $M' \in \mathcal{M}_k'$ so that for any $\pi_k \in \Pi_k$, the marginal distribution of $\{(s_h, a_{k,h}, r_{k,h})\}_{h \in [H]}$ for a trajectory drawn from $M(\pi_k, \pi_{-k})$ is the same as the distribution of a trajectory drawn from $M'(\pi_k)$. In addition, we have applied the data processing inequality for the Hellinger distance to conclude that $D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \widetilde{M}(\pi_k)\big)$ is an upper bound for the squared Hellinger distance between the marginal distributions of $\{(s_h, a_{k,h}, r_{k,h})\}_{h \in [H]}$ under $M(\pi_k, \pi_{-k})$ and $\widetilde{M}(\pi_k)$. Finally, by Theorem E.2 applied to the instance $\mathcal{M}^{\text{NE}}$, we have that

$$\sup_{\overline{M} \in \mathcal{M}} \text{r-dec}_\gamma^o(\mathcal{M}^{\text{NE}}, \overline{M}) \leq \frac{CKH \log H}{\gamma} + \sum_{k=1}^K \sup_{\widetilde{M}_k \in \widetilde{\mathcal{M}}_k} \text{r-dec}_{\gamma/(CKH \log H)}^o(\widetilde{\mathcal{M}}_k, \widetilde{M}_k)$$

$$\leq \frac{CKH \log H}{\gamma} + \sum_{k=1}^K \sup_{\overline{M}_k' \in \mathcal{M}_k'} \text{r-dec}_{\gamma/(CKH \log H)}^o(\mathcal{M}_k', \overline{M}_k')$$

$$\leq \frac{CKH \log H}{\gamma} + \sum_{k=1}^K 26H^2 S A_k \cdot \frac{CKH \log H}{\gamma}$$

$$\leq \frac{27KH^3 \log(H) S \sum_{k=1}^K A_k}{\gamma}.$$

■

Using Proposition G.9, we now bound the minimax rates for the instances $\mathcal{M}^{\text{NE}}, \mathcal{M}^{\text{CE}}, \mathcal{M}^{\text{CCE}}$. To simplify matters, we assume that reward distributions are known. Formally, we fix some functions $R_{k,h}^\star : \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{R})$ (for $k \in [K], h \in [H]$) and restrict the model class $\mathcal{M}$ to models $M \in \mathcal{M}$ for which $R_{k,h}^\star(s_h, a) \equiv R_{k,h}^M(s_h, a) \in \Delta([0, 1/H])$ for all $M \in \mathcal{M}$. We also assume that $A_k \geq 2$ for all k. With the functions $R_{k,h}^\star$ fixed, let us denote the resulting instances by $\mathcal{M}_0^{\text{NE}}, \mathcal{M}_0^{\text{CE}}, \mathcal{M}_0^{\text{CCE}}$.[20]

---

20. Essentially the same argument in Proposition G.10 allows us to upper bound the minimax risk for the original instances $\mathcal{M}^{\text{NE}}, \mathcal{M}^{\text{CE}}, \mathcal{M}^{\text{CCE}}$, for which rewards are not known, but doing so requires a renormalization argument (and the loss of a factor of H) to ensure rewards are always bounded in $[0, 1]$, which we omit for brevity.

Proposition G.10. *There is an algorithm for each of the instances $\mathscr{M}_0^{\mathrm{NE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{CCE}}$ which guarantees that with probability at least $1 - \delta$, $\mathbf{Risk}(T) \leq \sqrt{\frac{\max_k A_k \cdot AS^3H^4}{T}} \cdot \mathrm{polylog}(T, \delta^{-1}, A, S, H).$*

Proof of Proposition G.10. Note that it suffices to bound the minimax risk for the instance $\mathscr{M}_0^{\mathrm{NE}}$, since for any $\widehat{\pi} \in \Pi^{\mathrm{NE}} \subset \Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$, we have that $h^{M,\mathrm{CCE}}(\widehat{\pi}) \leq h^{M,\mathrm{CE}}(\widehat{\pi}) \leq h^{M,\mathrm{NE}}(\widehat{\pi})$. (Recall the definition of $h^{M,\mathrm{CCE}}, h^{M,\mathrm{CE}}, h^{M,\mathrm{NE}}$ in the proof of Proposition G.9.) The combination of Proposition G.9 and Proposition E.1 yields that, for any $\overline{M} \in \mathcal{M}$,

$$\mathsf{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{NE}}, \overline{M}) \leq O\left(\varepsilon \cdot \sqrt{KH^3 \log(H)S \sum_{k=1}^{K} A_k}\right).$$

Because of the constraint that $\overline{M} \in \mathcal{M}$ in the DEC upper bound, we need a *proper* estimation algorithm, i.e., one with $\widehat{\mathcal{M}} = \mathcal{M}$ (in the context of Assumption J.1). To do so, we use the approach of *layer-wise* estimators from Foster et al. (2021). Note that the model class $\mathcal{M}$ has the product structure $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_H$, where each $\mathcal{M}_h$ is the set of transition kernels $\mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$, which is a convex set, thus satisfying Assumption 7.2 of Foster et al. (2021). Furthermore, by gridding the transition densities into multiples of ε^2, we have that $\mathcal{N}(\mathcal{M}_h, \varepsilon) \leq (1/\varepsilon^2)^{S^2A}$, and therefore, by Proposition 7.1 and Lemma A.16 of Foster et al. (2021), there is an estimation algorithm $\mathbf{Alg}_{\mathsf{Est}}$ with $\widehat{\mathcal{M}} = \mathcal{M}$ and which has estimation error $\mathbf{Est}_{\mathsf{H}}(T, \delta) \leq O(S^2AH) \cdot \mathrm{polylog}(S, H, \delta^{-1}, T)$. Therefore, Theorem J.1 combined with Theorem C.1 gives that there is an algorithm with

$$\begin{aligned}\mathbf{Risk}(T) \leq &\sqrt{KH^3 \log(H)S \sum_{k=1}^{K} A_k} \cdot \sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(T, \delta)}{T}} \cdot \mathrm{polylog}(T, 1/\delta) \\ \leq &\sqrt{\frac{\max_k A_k \cdot AS^3H^4}{T}} \cdot \mathrm{polylog}(T, \delta^{-1}, A, S, H).\end{aligned}$$

∎

Lower bounds. It is straightforward to see that for any k, each of the instances $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ embeds an instance corresponding the class of single-player MDPs on state spaces $\mathcal{S}_h$, action space $\mathcal{A}_k$, and horizon H: in particular, take the subclass of $\mathcal{M}$ whose transitions and rewards only depend on player k's action at each step. Then it follows from the proof of Proposition 5.8 of Foster et al. (2021) that $\mathsf{dec}_\varepsilon(\mathscr{M}^{\mathrm{NE}}) \geq \mathsf{dec}_\varepsilon(\mathscr{M}^{\mathrm{CE}}) \geq \mathsf{dec}_\varepsilon(\mathscr{M}^{\mathrm{CCE}}) \geq \Omega(\varepsilon\sqrt{SH \cdot \max_k A_k})$. Therefore, Theorem D.2 (with $\underline{\varepsilon}(T) = \frac{c\sqrt{SH \cdot \max_k A_k}}{KT \log T}$, for sufficiently small $c > 0$) together with Theorem C.1 gives that for any of the instances $\mathscr{M}^{\mathrm{CCE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{NE}}$, and any algorithm, there is a model for which $\mathbb{E}[\mathbf{Risk}(T)] \geq \widetilde{\Omega}(SH \cdot \max_k A_k/(KT))$.

G.3.5. A SEPARATION BETWEEN MULTI-AGENT DEC AND SINGLE-AGENT DEC

In the previous subsections, we bounded the multi-agent DEC, and thereby the minimax risk (via an application of Theorem J.1 and Theorem C.1), for several bandit problems. In all cases, our upper bound on the multi-agent DEC (for CCE, CE, and Nash instances) followed via an application of Theorem E.1 to upper bound the multi-agent DEC by the single-agent DEC of the model classes $\widetilde{\mathcal{M}}_k$ defined in Eq. (11). The next (straightforward) proposition shows that this approach is not tight in

general, indicating that the multi-agent DEC represents a fundamental complexity measure that is distinct from existing ones.

Proposition G.11. *For any $K, A \in \mathbb{N}$, there is a K-player MA-DMSO NE instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ so that $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathscr{M}) = 0$ but $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\widetilde{\mathcal{M}}_k) \geq \Omega(A/\gamma)$ for all $\gamma > 0$, where $\widetilde{\mathcal{M}}_k$ are defined as in (11).*

Proof of Proposition G.11. Fix $K, A \in \mathbb{N}$, and set $\Pi_k = \{0, 1, \ldots, A\}$ for each k. Let $\mathcal{R} := [-1, 1]$ and $\mathcal{O}_\circ := \mathcal{A}$. Define $\mathcal{F} \subseteq (\Pi \to \mathcal{R}^K)$ to be the class of all tuples $(f_1, \ldots, f_K)$ with $f_k : \Pi \to \mathcal{R}^K$ with the property that for all $\pi \in \Pi$, if there is any k so that $\pi_k = 0$, then $f_{k'}(\pi) = 0$ for all $k' \in [K]$. Set $\mathcal{M} := \mathcal{M}_\mathcal{F}$, and define Π'_k, U_k as in Definition 1.1.

Define $\pi_0 = (0, \ldots, 0)$. Since $h^M(\pi_0) = 0$ for all $M \in \mathcal{M}$, it follows that $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathscr{M}) = 0$. On the other hand, it is straightforward to see that each class $\widetilde{\mathcal{M}}_k$ embeds a standard multi-armed bandit instance with A arms, meaning that by Proposition 5.3 of Foster et al. (2021), we have that $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\widetilde{\mathcal{M}}_k) \geq \Omega(A/\gamma)$ for all $\gamma > 0$. ∎

Part III
Proofs

Appendix H. Technical tools

H.1. Information theory

In this section we collect several technical lemmas which are used in our proofs.

Lemma H.1. *Let $(\mathcal{X}, \mathscr{X})$, $(\mathcal{I}, \mathscr{I})$ be measure spaces. Suppose that for each $i \in \mathcal{I}$, there are distributions $P_i, P'_i \in \Delta(\mathcal{X})$, and $Q \in \Delta(\mathcal{I})$. Suppose further that there is a measurable function $\varphi : \mathcal{X} \to \mathcal{I}$ so that, for each $i \in \mathcal{I}$, $\mathbb{P}_{x \sim P_i}(\varphi(x) = i) = \mathbb{P}_{x \sim P'_i}(\varphi(x) = i) = 1$. Then for any f-divergence $D(\cdot \parallel \cdot)$, it holds that*

$$D\big(\mathbb{E}_{i \sim Q}[P_i] \parallel \mathbb{E}_{i \sim Q}[P'_i]\big) = \mathbb{E}_{i \sim Q}[D\big(P_i \parallel P'_i\big)].$$

Proof of Lemma H.1. That $D(\mathbb{E}_{i \sim Q}[P_i] \parallel \mathbb{E}_{i \sim Q}[P'_i]) \leq \mathbb{E}_{i \sim Q}[D(P_i \parallel P'_i)]$ follows from convexity of $D(\cdot \parallel \cdot)$. To establish the opposite direction, our assumption on the function φ together with the data processing inequality yields

$$D\big(\mathbb{E}_{i \sim Q}[P_i] \parallel \mathbb{E}_{i \sim Q}[P'_i]\big) \geq D\big(\mathbb{E}_{i \sim Q}[\mathbb{I}_i \times P_i] \parallel \mathbb{E}_{i \sim Q}[\mathbb{I}_i \times P'_i]\big) = \mathbb{E}_{i \sim Q}[D\big(P_i \parallel P'_i\big)],$$

where the final inequality follows from, e.g., Polyanskiy and Wu (2014). ∎

Lemma H.2 (e.g., Polyanskiy and Wu (2014))**.** *Let $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$ be measure spaces, and let $\mathcal{X} \times \mathcal{Y}$ be equipped with the product sigma-algebra $\mathscr{X} \otimes \mathscr{Y}$. Let (x, y) be a pair of random variables on $\mathcal{X} \times \mathcal{Y}$, distributed according to some distribution $\mathbb{P}_{x,y}$. For any f-divergence $D(\cdot \parallel \cdot)$, it holds that*

$$\mathbb{E}_{x \sim \mathbb{P}_x}\left[D\big(\mathbb{P}_{y|x} \parallel \mathbb{P}_y\big)\right] = \mathbb{E}_{y \sim \mathbb{P}_y}\left[D\big(\mathbb{P}_{x|y} \parallel \mathbb{P}_x\big)\right].$$

Lemma H.3 (Lemma B.5 of Foster et al. (2022b)). *Let $\mathbb{P}, \mathbb{Q}$ be probability distributions on a measure space $(\mathcal{X}, \mathscr{X})$. For any $\alpha \geq 1$, let $\mathcal{G}_\alpha := \{g : \mathcal{X} \to \mathbb{R} \ : \ \|g\|_\infty \leq \alpha\}$. Then*

$$\frac{1}{2}D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) \leq \sup_{g \in \mathcal{G}_\alpha} \left\{ 1 - \mathbb{E}_{\mathbb{P}}[e^g] \cdot \mathbb{E}_{\mathbb{Q}}[e^{-g}] \right\} + 4e^{-\alpha}.$$

Lemma H.4 (e.g., Foster et al. (2022b)). *Consider measure spaces $(\mathcal{X}, \mathscr{X})$, $(\mathcal{Y}, \mathscr{Y})$, and let (x, y) be a pair of random variables distributed according to some distribution $\mathbb{P}_{x,y}$ on $(\mathcal{X} \times \mathcal{Y}, \mathscr{X} \otimes \mathscr{Y})$. Then*

$$\mathbb{E}_{x \sim \mathbb{P}_x}\left[D_{\mathsf{H}}^2\left(\mathbb{P}_{y|x}, \mathbb{P}_y\right) \right] \leq 4 \cdot \inf_{\mathbb{Q} \in \Delta(\mathcal{Y})} \mathbb{E}_{x \sim \mathbb{P}_x}\left[D_{\mathsf{H}}^2\left(\mathbb{P}_{y|x}, \mathbb{Q}\right) \right].$$

Proof of Lemma H.4. Consider any $\mathbb{Q} \in \Delta(\mathcal{Y})$. Using the fact that the Hellinger distance satisfies the triangle inequality, we have

$$\begin{aligned}
\mathbb{E}_{x \sim \mathbb{P}_x}\left[D_{\mathsf{H}}^2\left(\mathbb{P}_{y|x}, \mathbb{P}_y\right) \right] \leq & \mathbb{E}_{x \sim \mathbb{P}_x}\left[2 \cdot D_{\mathsf{H}}^2\left(\mathbb{P}_{y|x}, \mathbb{Q}\right) + 2 \cdot D_{\mathsf{H}}^2(\mathbb{Q}, \mathbb{P}_y) \right] \\
\leq & 2 \cdot \mathbb{E}_{x \sim \mathbb{P}_x}\left[D_{\mathsf{H}}^2\left(\mathbb{P}_{y|x}, \mathbb{Q}\right) \right] + 2 \cdot D_{\mathsf{H}}^2\left(\mathbb{Q}, \mathbb{E}_{x \sim \mathbb{P}_x}[\mathbb{P}_{y|x}]\right) \\
\leq & 4 \cdot \mathbb{E}_{x \sim \mathbb{P}_x}\left[D_{\mathsf{H}}^2\left(\mathbb{P}_{y|x}, \mathbb{Q}\right) \right],
\end{aligned}$$

where the final inequality follows from convexity of the squared Hellinger distance. ∎

Lemma H.5 (Donsker-Varadhan; see Polyanskiy and Wu (2014)). *Let $(\mathcal{X}, \mathscr{X})$ be a measure space, and let $\mathbb{P}, \mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathscr{X})$. Then*

$$D_{\mathsf{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \sup_{h:\mathcal{X} \to \mathbb{R}} \left\{ \mathbb{E}_{X \sim \mathbb{P}}[h(X)] - \log \mathbb{E}_{X \sim \mathbb{Q}}[\exp(h(X))] \right\},$$

where the supremum is over all (measurable) functions $h : \mathcal{X} \to \mathbb{R}$ satisfying $\mathbb{E}_{X \sim \mathbb{Q}}[\exp(h(X))] < \infty$.

Lemma H.6. *Let $\mathbb{P}, \mathbb{Q}$ be probability measures on some probability space $(\Omega, \mathscr{F})$. Consider some event $\mathcal{E} \in \mathscr{F}$ so that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$, for some $\delta \in (0, 1)$. Suppose also that for all events $\mathcal{F} \in \mathscr{F}$, we have $\mathbb{P}(\mathcal{E} \cap \mathcal{F}) = \mathbb{Q}(\mathcal{E} \cap \mathcal{F})$. Then $D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) \leq \delta$.*

Proof of Lemma H.6. Choosing $\mathcal{E}' = \Omega$ gives $\mathbb{Q}(\mathcal{E}) = \mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Then for any event $\mathcal{F} \in \mathscr{F}$, we have

$$\begin{aligned}
|\mathbb{P}(\mathcal{F}) - \mathbb{Q}(\mathcal{F})| \leq & |\mathbb{P}(\mathcal{F} \cap \mathcal{E}) - \mathbb{Q}(\mathcal{F} \cap \mathcal{E})| + |\mathbb{P}(\mathcal{F} \cap \mathcal{E}^{\mathsf{c}}) - \mathbb{Q}(\mathcal{F} \cap \mathcal{E}^{\mathsf{c}})| \\
= & |\mathbb{P}(\mathcal{F} \cap \mathcal{E}^{\mathsf{c}}) - \mathbb{Q}(\mathcal{F} \cap \mathcal{E}^{\mathsf{c}})| \leq \delta.
\end{aligned}$$

∎

H.2. Concentration inequalities

Lemma H.7 (Lemma A.4 of Foster et al. (2021)). *Let $(X_t)_{t \in [T]}$ be any sequence of real-valued random variables adapted to a filtration $\mathscr{F}^t$. Then with probability at least $1 - \delta$,*

$$\sum_{t=1}^T X_t \leq \sum_{t=1}^T \log\left(\mathbb{E}\left[e^{X_t} | \mathscr{F}^{t-1} \right] \right) + \log(1/\delta).$$

H.3. Topological lemmas

The below lemma is a special case of the Berge maximum theorem.

Lemma H.8. *Let $\mathcal{U}, \mathcal{V}$ be compact subsets of Euclidean space, and consider any continuous function $G : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$. Define $\mathcal{C} : \mathcal{V} \to \mathcal{P}(\mathcal{U})$ by $\mathcal{C}(v) := \arg\min_{u \in \mathcal{U}}\{G(u, v)\}$. Then $\mathcal{C}$ is upper hemicontinuous.*

Proof of Lemma H.8. Consider any sequences $u_n \to u \in \mathcal{U}$, $v_n \to v \in \mathcal{V}$ so that $u_n \in \mathcal{C}(v_n)$ for all n. We wish to show that $u \in \mathcal{C}(v)$, i.e., $G(u, v) \le G(u', v)$ for all $u' \in \mathcal{U}$ (which suffices to prove upper hemicontinuity by compactness of $\mathcal{V}$; see (Beer, 1993, Lemma 6.2.6)). To do so, fix any $u' \in \mathcal{U}$ and $\epsilon > 0$. There exists N so that for $n \ge N$, we have $|G(u_n, v_n) - G(u, v)| \le \epsilon$ and $|G(u', v_n) - G(u', v)| \le \epsilon$, by continuity of G. Then

$$G(u, v) \le G(u_n, v_n) + \epsilon \le G(u', v_n) + \epsilon \le G(u', v) + 2\epsilon,$$

and by taking $\epsilon \to 0$ we get that $G(u, v) \le G(u', v)$. ∎

The next lemma is a straightforward consequence of Kakutani's fixed point theorem. In its statement, we write $\mathcal{X}_{-k} := \prod_{k' \ne k} \mathcal{X}_{k'}$ and $\mathcal{X} = \prod_{k \in [K]} \mathcal{X}_k$.

Lemma H.9. *Suppose that $\mathcal{X}_1, \ldots, \mathcal{X}_K$ are nonempty, compact, and convex subsets of Euclidean space. Suppose that for each $k \in [K]$ we are given an upper hemicontinuous function $F_k : \mathcal{X} \to \mathcal{P}(\mathcal{X}_k)$ so that, for all $x \in \mathcal{X}$, $F_k(x)$ is nonempty, closed, and convex. Then there is some $x \in \mathcal{X}$ so that*

$$x \in F_1(x) \times \cdots \times F_K(x).$$

Proof of Lemma H.9. Define $F : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ by $F(x) := F_1(x) \times \cdots \times F_K(x)$. It is evident that for each $x \in \mathcal{X}$, $F(x)$ is nonempty, closed, and convex. Furthermore, we claim that F is upper hemicontinuous. To see this, consider any sequences $x_n \to x$ and $y_n \to y$ so that $y_n \in F(x_n)$ for each $n \in \mathbb{N}$. Writing $y_n = (y_{n,1}, \ldots, y_{n,K})$ and $y = (y_1, \ldots, y_K)$, by the product structure of $F(x_n)$, we have that $y_{n,k} \in F_k(x_n)$ for each $k \in [K]$. By upper hemicontinuity of F_k and the fact that $y_{n,k} \to y_k$, it holds that $y_k \in F_k(x)$. Thus $y \in F(x)$. By Kakutani's fixed point theorem (Osborne and Rubinstein, 1994, Lemma 20.1), it holds that F has a fixed point, namely some $x \in \mathcal{X}$ so that $x \in F(x)$. ∎

H.4. Minimax theorem

Theorem H.1 (Sion's minimax theorem). *Let $\mathcal{X}, \mathcal{Y}$ be convex subsets of topological vector spaces, with $\mathcal{X}$ compact. Let $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a function such that (a) the mapping $y \mapsto F(x, y)$ is concave and upper semicontinuous for all $x \in \mathcal{X}$, and (b) the mapping $x \mapsto F(x, y)$ is convex and lower semicontinuous for all $y \in \mathcal{Y}$. Then*

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} F(x, y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} F(x, y).$$

Appendix I. Proofs for Appendix C

Proof of Theorem C.1. Consider an instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}_\circ, \mathcal{R}, \{\Pi'_k\}_k, \{U_k\}_k)$ of the MA-DMSO framework.

For all models M and decisions $\pi \in \Pi$, define[21]

$$\widetilde{f}^M(\pi) := K - h^M(\pi) = K - \sum_{k=1}^{K} \sup_{\pi'_k \in \Pi'_k} \left\{ f_k^M(U_k(\pi'_k, \pi)) - f_k^M(\pi) \right\}.$$

Now fix any $M \in \mathcal{M}$. By Assumption 1.2, there is some $\pi^\star \in \Pi$ so that $h^M(\pi^\star) = 0$. By Assumption 1.3, it holds that $h^M(\pi) \geq 0$ for all $\pi \in \Pi$. Then

$$\sup_{\pi' \in \Pi} \widetilde{f}^M(\pi') - \widetilde{f}^M(\pi) = K - (K - h^M(\pi)) = h^M(\pi). \tag{16}$$

Note that the instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{\widetilde{f}^M\}_M)$ is well-defined since models in M are probability kernels $M : \Pi \to \mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}^K$, and the observation space in the instance $\mathscr{H}$ is by definition $\mathcal{O}$. Thus, the first claimed point is follows from (16) since for any $\overline{M}$, we have:

$$\mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M}) = \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{q,\varepsilon}(\overline{M})} \mathbb{E}_{\pi \sim p}[h^M(\pi)] \tag{17}$$

$$= \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{q,\varepsilon}(\overline{M})} \mathbb{E}_{\pi \sim p} \left[\sup_{\pi' \in \Pi} \widetilde{f}^M(\pi') - \widetilde{f}^M(\pi) \right] = \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}).$$

Finally, we note that since the decision and (full) observation spaces of $\mathscr{H}, \mathscr{M}$ are identical, the space of algorithms (p, q) and distributions $\mathbb{P}^{M^\star,(p,q)}$ are identical in the two frameworks. It follows from the definitions of $\mathfrak{M}(\mathscr{H}, T)$ and $\mathfrak{M}(\mathscr{M}, T)$ that they are equal. ∎

Proof of Theorem C.2. Consider an instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ and some $V \in \mathbb{N}$. We first specify the instance $\mathscr{M}$ by defining each of its components:

- Define $\Sigma_1 = \Pi$ and $\Sigma_2 = \{0, 1, \ldots, V\}, \widetilde{\Pi}_k = \Delta(\Sigma_k)$ for $k \in \{1, 2\}$, and $\widetilde{\Pi} := \widetilde{\Pi}_1 \times \widetilde{\Pi}_2$.

- We define Π'_k, U_k for $k \in [2]$ in the standard fashion for NE instances, per Definition 1.1; in particular, set $\Pi'_k = \widetilde{\Pi}_k = \Delta(\Sigma_k)$ for each k and $U_k(\pi'_k, \pi) = (\pi'_k, \pi_{-k})$.

- Define $\mathcal{O}_\circ := \mathcal{O} \cup \{\perp\}, \mathcal{R} = [-1, 1]$, and set $\widetilde{\mathcal{O}} = \mathcal{O}_\circ \times \mathcal{R}^2$.

- The model class $\widetilde{\mathcal{M}}$ is indexed by tuples $(M, v) \in \mathcal{M} \times \{1, 2, \ldots, V\} = \mathcal{M} \times [V]$. In particular, for each such tuple (M, v), we have a model $\widetilde{M}_{M,v} \in \widetilde{\mathcal{M}}$, which is defined as explained below. As the instance $\mathscr{M}$ we are constructing corresponds to that of computing mixed Nash equilibria in a game whose pure action sets are Σ_1, Σ_2, we call elements of $\Sigma_1 \times \Sigma_2$ *pure decisions*.

  - For pure decisions of the form $(\sigma_1, 0) \in \Sigma_1 \times \Sigma_2$, the distribution of $(o_\circ, r_1, r_2) \sim \widetilde{M}_{M,v}((\sigma_1, 0))$ is given by:

    $$o_\circ \sim M(\sigma_1) \in \mathcal{O} \subset \mathcal{O}_\circ, \qquad r_1 = r_2 = 0.$$

---

21. The addition of K in the definition of $\widetilde{f}^M(\pi)$ is for convenience, so as to ensure that if $h^M(\pi) \in [0, K]$ for all M, π, then the same holds for $\widetilde{f}^M(\pi)$.

- For pure decisions of the form $(\sigma_1, i) \in \Sigma_2 \times \Sigma_2$ with $i > 0$, the distribution of $(o_\circ, r_1, r_2) \sim \widetilde{M}_{M,v}((\sigma_1, i))$ is given by:

$$o_\circ = \perp, \qquad r_2 = -r_1 = \begin{cases} -1 & : i \neq v \\ g^M(\sigma_1) & : i = v. \end{cases}$$

- For general decisions $\pi \in \widetilde{\Pi}$, we can write $\pi = \pi_1 \times \pi_2$ for $\pi_k \in \Delta(\Sigma_k)$ for $k \in [2]$. Then the distribution $\widetilde{M}_{M,v}(\pi)$ is the distribution of $\widetilde{M}_{M,v}(\sigma)$, for $\sigma = (\sigma_1, \sigma_2)$ is distributed as: $\sigma_k \sim \pi_k$ for $k \in [2]$.

For reference later in the proof, we state a basic technical lemma, which is an immediate consequence of the construction of $\widetilde{\mathcal{M}}$.

Lemma I.1. *For any $\pi = \pi_1 \times \pi_2 \in \widetilde{\Pi}$, and any $M' = \widetilde{M}_{M,v} \in \widetilde{\mathcal{M}}$, it holds that*

$$h^{M'}(\pi) \geq \pi_2(\Sigma_2 \backslash \{v\}) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] + \pi_2(\Sigma_2 \backslash \{0, v\}).$$

Proof of Lemma I.1. By considering the deviation $\pi_2' = v$, we have

$$\begin{aligned} h^{M'}(\pi) &\geq \sup_{\pi_2' \in \Pi_2'} \{f_2^{M'}(U_2(\pi_2', \pi)) - f_2^{M'}(\pi)\} \\ &\geq f_2^{M'}(\pi_1 \times \mathbb{I}_v) - f_2^{M'}(\pi_1 \times \pi_2) \\ &= \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] + \pi_2(\Sigma_2 \backslash \{0, v\}) - \pi_2(v) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] \\ &= \pi_2(\Sigma_2 \backslash \{v\}) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] + \pi_2(\Sigma_2 \backslash \{0, v\}). \end{aligned}$$

∎

Bounding $\mathfrak{M}(\mathscr{H}, T)$ by $\mathfrak{M}(\mathscr{M}, T)$. Consider any algorithm $(\widetilde{p}, \widetilde{q})$ which achieves $\mathfrak{M}(\mathscr{M}, T)$. We have $\widetilde{p} : \prod_{t=1}^{T}(\widetilde{\Pi} \times \widetilde{\mathcal{O}}) \to \Delta(\widetilde{\Pi})$ and $\widetilde{q} = (\widetilde{q}^1, \ldots, \widetilde{q}^T)$, with each $\widetilde{q}^t : \prod_{i=1}^{t-1}(\widetilde{\Pi} \times \widetilde{\mathcal{O}}) \to \Delta(\widetilde{\Pi})$ (we refer to Appendix B for background on how algorithms in the MA-DMSO framework and HR-DMSO framework are formalized).

Given $(\widetilde{p}, \widetilde{q})$, we define an algorithm (p, q) for the instance $\mathscr{H}$ as follows. For any model $M \in \mathcal{M}$, the algorithm attempts to simulate the interaction of $(\widetilde{p}, \widetilde{q})$ with $\widetilde{M}_{M,v}$ by only interacting with $M \in \mathcal{M}$. The algorithm will store internal state, denoted by $(\widetilde{\pi}^i, (\widetilde{o}^i, r_1^i, r_2^i))$, for each $i \in [T]$, which store the "simulated" decisions and observations taken with respect to $\widetilde{M}_{M,v}$. As a result of this internal state, our description below does not explicitly identify the probability kernels $p(\cdot|\cdot)$, $q^t(\cdot|\cdot)$. Since these kernels take as input the entire history, there exist kernels $p(\cdot|\cdot), q^t(\cdot|\cdot)$ which produce exactly the same distribution over trajectories as the below algorithm, but writing them down explicitly is somewhat cumbersome.

In particular, the distributions q^t (for $t \in [T]$) and p are defined (implicitly) as follows:

1. For $t = 1, 2, \ldots, T$:

   (a) Draw $\widetilde{\pi}^t \sim \widetilde{q}^t(\cdot|(\widetilde{\pi}^1, (\widetilde{o}^1, r_1^1, r_2^1)), \ldots, (\widetilde{\pi}^{t-1}, (\widetilde{o}^{t-1}, r_1^{t-1}, r_2^{t-1})))$, so that $\pi^t \in \widetilde{\Pi}$.

   (b) Draw $(\sigma_1^t, \sigma_2^t) \sim \widetilde{\pi}^t$.

(c) The distribution q^t is defined (implicitly) by taking the decision $\sigma_1^t \in \Sigma_1 = \Pi$.

(d) For use in choosing future decisions: as a function of the observation o^t received after σ_1^t is played, define

$$(\widetilde{o}^t, r_1^t, r_2^t) = \begin{cases} (o^t, 0, 0) & : \sigma_2^t = 0 \\ (\bot, 1, -1) & : \sigma_2^t > 0. \end{cases}$$

2. Finally, the distribution p is defined as the distribution of $\widehat{\sigma}_1$, where $\widehat{\pi} \sim \widetilde{p}(\cdot | (\widetilde{\pi}^1, (\widetilde{o}^1, r_1^t, r_2^t)), \ldots, (\widetilde{\pi}^T, (\widetilde{o}^T, r_1^T, r_2^T)))$ and $\widehat{\sigma}_1 \sim \widehat{\pi}_1$.

To analyze this algorithm, for each $M \in \mathcal{M}$, we introduce a model $\widetilde{M}_{M,0}$ which is defined identically to $\widetilde{M}_{M,v}$ for any $v \in [V]$ except that $\widetilde{M}_{M,0}((\sigma_1, i))$ outputs $(\bot, 1, -1)$ a.s. for any $\sigma_1 \in \Sigma_1, i \in [V]$. It is straightforward to see that if there is some underlying model $M \in \mathcal{M}$ so that $o^t \sim M(\sigma_1^t)$ when the algorithm (p, q) defined above is used, then the distribution of $\{(\widetilde{\pi}^t, (\widetilde{o}^t, r_1^t, r_2^t))\}_{t=1}^T$ defined above is exactly the distribution of the history under $\mathbb{P}^{\widetilde{M}_{M,0},(\widetilde{p},\widetilde{q})}$. We next appeal to the following claim, which states that we can pass from this distribution to the distribution $\mathbb{P}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}$ for some $v^\star \in [V]$:

Lemma I.2. *There is an absolute constant $C > 0$ so that for any choice of algorithm $(\widetilde{p}, \widetilde{q})$ and model $M \in \mathcal{M}$, there exists $v^\star \in [V]$ so that:*

1. $D_{\mathsf{TV}}\left(\mathbb{P}^{\widetilde{M}_{M,0},(\widetilde{p},\widetilde{q})}, \mathbb{P}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\right) \leq C\sqrt{T\log(T)/V}.$

2. $\mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v^\star)] \leq C\sqrt{T\log(T)/V}.$

The proof of Lemma I.2 is provided following in the sequel. Let $\delta := C\sqrt{T\log(T)/V}$, where C is the constant from Lemma I.2. If $\delta > 1$, then it is immediate that $\mathfrak{M}(\mathcal{H}, T) \leq \mathfrak{M}(\mathcal{M}, T) + O(\delta)$, so we may assume henceforth that $\delta \leq 1$. We then have:

$$\begin{aligned}
\mathbb{E}^{M,(p,q)}\mathbb{E}_{\widehat{\sigma}_1\sim p}[g^M(\widehat{\sigma}_1)] &= \mathbb{E}^{\widetilde{M}_{M,0},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}\mathbb{E}_{\widehat{\sigma}_1\sim\widehat{\pi}_1}[g^M(\widehat{\sigma}_1)] \\
&\leq \mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}\mathbb{E}_{\widehat{\sigma}_1\sim\widehat{\pi}_1}[g^M(\widehat{\sigma}_1)] + \delta \\
&\leq \mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}\left[\min\left\{\frac{h^{\widetilde{M}_{M,v^\star}}(\widehat{\pi})}{1 - \widehat{\pi}_2(v^\star)}, 1\right\}\right] + \delta \\
&\leq (1 + \sqrt{\delta}) \cdot \mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[h^{\widetilde{M}_{M,v^\star}}(\widehat{\pi})] + 2\sqrt{\delta} \\
&\leq \mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[h^{\widetilde{M}_{M,v^\star}}(\widehat{\pi})] + 6\sqrt{\delta},
\end{aligned}$$

where the first inequality follows from the first point of Lemma I.2, the second inequality follows from Lemma I.1, the second-to-last inequality uses the second point of Lemma I.2 together with Markov's inequality to conclude that $\mathbb{P}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}(\widehat{\pi}_2(v^\star) \geq \sqrt{\delta}) < \sqrt{\delta}$, and the final inequality uses that $h^{M'}(\pi) \leq 4$ for all $M' \in \widetilde{\mathcal{M}}, \pi \in \widehat{\Pi}$.

Taking a supremum over all models $M \in \mathcal{M}$, we conclude that

$$\sup_{M \in \mathcal{M}} \mathbb{E}^{M,(p,q)}\mathbb{E}_{\widehat{\sigma}_1\sim p}[g^M(\widehat{\sigma}_1)] \leq \sup_{\widetilde{M} \in \widetilde{\mathcal{M}}} \mathbb{E}^{\widetilde{M},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[h^{\widetilde{M}}(\widehat{\pi})] + 6\sqrt{\delta}.$$

Bounding $\mathfrak{M}(\mathcal{M}, T)$ by $\mathfrak{M}(\mathcal{H}, T)$. Consider any algorithm (p, q) which achieves $\mathfrak{M}(\mathcal{H}, T)$. We have $p : \prod_{t=1}^{T}(\Pi \times \mathcal{O}) \to \Delta(\Pi)$, and $q = (q^1, \ldots, q^T)$, with each $q^t : \prod_{i=1}^{t-1}(\Pi \times \mathcal{O}) \to \Delta(\Pi)$.

We define an algorithm $(\widetilde{p}, \widetilde{q})$ for the instance $\mathcal{M}$ as follows. Given $\pi^t \in \widetilde{\mathcal{O}}, o^t \in \mathcal{O}_\circ, r_1^t, r_2^t \in \mathcal{R}$ for each $t \in [T]$, we define

$$\widetilde{p}(\cdot | (\pi^1, (o^1, r_1^1, r_2^1)), \ldots, (\pi^T, (o^T, r_1^T, r_2^T))) \in \Delta(\widetilde{\Pi})$$

to be the distribution obtained by sampling $(\sigma_1^t, \sigma_2^t) \sim \pi^t$ for each t, and taking the pure decision $(\widehat{\sigma}, 0)$, where $\widehat{\sigma}$ is distributed according to $p(\cdot | (\sigma_1^1, o^1), \ldots, (\sigma_1^T, o^T))$. Similarly, define

$$\widetilde{q}^t(\cdot | (\pi^1, (o^1, r_1^1, r_2^1)), \ldots, (\pi^{t-1}, (o^{t-1}, r_1^{t-1}, r_2^{t-1}))) \in \Delta(\widetilde{\Pi})$$

to be the distribution obtained by sampling $(\sigma_1^i, \sigma_2^i) \sim \pi^i$ for each $i < t$, and taking the pure decision $(\sigma_1^t, 0)$, where σ_1^t is distributed according to $q^t(\cdot | (\sigma_1^1, o^1), \ldots, (\sigma_1^{t-1}, o^{t-1}))$. Since each $\widetilde{q}^t$ is supported only on (pure) decisions in $\Sigma_1 \times \{0\}$, for any model $M \in \mathcal{M}$ and any $v \in [V]$, letting $M' = \widetilde{M}_{M,v}$, the distribution of $\{(\sigma_1^t, o^t)\}_{t=1}^T$ under $\mathbb{P}^{M', (\widetilde{p}, \widetilde{q})}$ is the same as the distribution of $\{(\pi^t, o^t)\}_{t=1}^T$ under $\mathbb{P}^{M, (p, q)}$. Thus, we have

$$\mathbb{E}^{M, (p, q)} \mathbb{E}_{\widehat{\pi} \sim p}[g^M(\widehat{\pi})] = \mathbb{E}^{M', (\widetilde{p}, \widetilde{q})} \mathbb{E}_{(\widehat{\sigma}_1, 0) \sim \widetilde{p}}[g^M(\widehat{\sigma}_1)]$$

$$= \mathbb{E}^{M', (\widetilde{p}, \widetilde{q})} \mathbb{E}_{(\widehat{\sigma}_1, 0) \sim \widetilde{p}}\left[\sup_{\pi_2' \in \Pi_2'} \{f_2^{M'}(U_2(\pi_2', (\widehat{\sigma}_1, 0))) - f_2^{M'}((\widehat{\sigma}_1, 0))\}\right]$$

$$= \mathbb{E}^{M', (\widetilde{p}, \widetilde{q})} \mathbb{E}_{(\widehat{\sigma}_1, 0) \sim \widetilde{p}}\left[h^{M'}((\widehat{\sigma}_1, 0))\right],$$

where above we have shortened $p = p(\cdot | (\pi^1, o^1), \ldots, (\pi^T, o^T))$ to denote the random variable under $\mathbb{P}^{M, (p, q)}$ and $\widetilde{p} = \widetilde{p}(\cdot | (\pi^1, (o^1, r_1^1, r_2^1)), \ldots, (\pi^T, (o^T, r_1^T, r_2^T)))$ to denote the random variable under $\mathbb{P}^{M', (\widetilde{p}, \widetilde{q})}$. This establishes that $\mathfrak{M}(\mathcal{H}, T) \leq \mathfrak{M}(\mathcal{M}, T)$.

Bounding $\mathsf{dec}_\varepsilon(\mathcal{H})$ by $\mathsf{dec}_{\varepsilon'}(\mathcal{M})$. Consider any reference model $\overline{M} \in \mathrm{co}(\mathcal{M})$. Given $\varepsilon > 0$, set $\varepsilon' := \varepsilon + \sqrt{6/V}$. We will upper bound $\mathsf{dec}_\varepsilon(\mathcal{H}, \overline{M})$ by $\mathsf{dec}_{\varepsilon'}(\mathcal{M}, \widetilde{M})$ for some $\widetilde{M} \in \mathrm{co}(\widetilde{\mathcal{M}})$. For some distribution $\nu \in \Delta(\mathcal{M})$, we can write $\overline{M}(\pi) = \mathbb{E}_{M \sim \nu}[M(\pi)]$ for all $\pi \in \Pi$. Define $\mu := \nu \times \mathrm{Unif}([V]) \in \Delta(\mathcal{M} \times [V])$, and $\widetilde{M}(\pi) := \mathbb{E}_{(M, v) \sim \mu}[\widetilde{M}_{M,v}(\pi)]$ for all $\pi \in \Pi$. Choose some $\widetilde{p}, \widetilde{q} \in \Delta(\widetilde{\Pi})$ so that

$$\mathsf{dec}_{\varepsilon'}(\mathcal{M}, \widetilde{M}) = \sup_{M' \in \widetilde{\mathcal{M}}} \left\{\mathbb{E}_{\pi \sim \widetilde{p}}[h^{M'}(\pi)] \mid \mathbb{E}_{\pi \sim \widetilde{q}}\left[D_{\mathsf{H}}^2\left(M'(\pi), \widetilde{M}(\pi)\right)\right] \leq (\varepsilon')^2\right\}.$$

Define $p \in \Delta(\Pi)$ to be the distribution of σ_1 where $\pi = \pi_1 \times \pi_2 \sim \widetilde{p}$ and $\sigma_1 \sim \pi_1$. Similarly define $q \in \Delta(\Pi)$ to be the distribution of σ_1 where $\pi = \pi_1 \times \pi_2 \sim \widetilde{q}$ and $\sigma_1 \sim \pi_1$. Now choose $v^\star \in [V]$ as follows:

$$v^\star := \arg\min_{v \in [V]} \left\{\mathbb{E}_{\pi \sim \widetilde{p}}[\pi_2(v)] + \mathbb{E}_{\pi \sim \widetilde{q}}[\pi_2(v)]\right\},$$

where we have used the convention that $\pi = \pi_1 \times \pi_2$ above. Then we have

$$\mathbb{E}_{\pi \sim \widetilde{p}}[\pi_2(v^\star)] + \mathbb{E}_{\pi \sim \widetilde{q}}[\pi_2(v^\star)] \leq \frac{2}{V}.$$

Consider any model $M \in \mathcal{M}$, and let $M' := \widetilde{M}_{M,v^\star}$. We now compute

$$
\begin{aligned}
\mathbb{E}_{\sigma_1 \sim p}[g^M(\sigma_1)] =& \mathbb{E}_{\pi \sim \widetilde{p}} \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] \\
\leq& \mathbb{E}_{\pi \sim \widetilde{p}} \left[\min \left\{ \frac{h^{M'}(\pi) - \pi_2(\Sigma_2 \backslash \{0, v^\star\})}{\pi_2(\Sigma_2 \backslash \{v^\star\})}, \ 1 \right\} \right] \\
\leq& 2V^{-1/2} + \mathbb{E}_{\pi \sim \widetilde{p}} \left[\frac{h^{M'}(\pi)}{1 - V^{-1/2}} \right] \\
\leq& 2V^{-1/2} + (1 + 2V^{-1/2}) \cdot \mathbb{E}_{\pi \sim \widetilde{p}}[h^{M'}(\pi)],
\end{aligned}
$$

where the first inequality uses Lemma I.1 and the second-to-last inequality uses Markov's inequality to conclude that $\mathbb{P}_{\pi \sim \widetilde{p}}[\pi_2(v^\star) > V^{-1/2}] \leq 2V^{-1/2}$. Furthermore, we have

$$
\begin{aligned}
\mathbb{E}_{\pi \sim \widetilde{q}} \left[D_{\mathsf{H}}^2 \Big(M'(\pi), \widetilde{M}(\pi) \Big) \right] \leq& \mathbb{E}_{\pi \sim \widetilde{q}} \left[\mathbb{E}_{(\sigma_1, \sigma_2) \sim \pi} \left[D_{\mathsf{H}}^2 \Big(M'((\sigma_1, \sigma_2)), \widetilde{M}((\sigma_1, \sigma_2)) \Big) \right] \right] \\
\leq& \mathbb{E}_{\pi \sim \widetilde{q}} \left[\pi_2(0) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1} \left[D_{\mathsf{H}}^2 \Big(M'((\sigma_1, 0)), \widetilde{M}((\sigma_1, 0)) \Big) \right] \right] \\
&+ \mathbb{E}_{\pi \sim \widetilde{q}} \left[\pi_2(v^\star) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1} \left[D_{\mathsf{H}}^2 \Big(M'((\sigma_1, v^\star)), \widetilde{M}((\sigma_1, v^\star)) \Big) \right] \right] \\
&+ \mathbb{E}_{\pi \sim \widetilde{q}} \left[\pi_2(\Sigma \backslash \{0, v^\star\}) \cdot D_{\mathsf{H}}^2(\mathrm{Ber}(0), \mathrm{Ber}(1/V)) \right] \\
\leq& \mathbb{E}_{\sigma_1 \sim q} \left[D_{\mathsf{H}}^2 \Big(M'((\sigma_1, 0)), \widetilde{M}((\sigma_1, 0)) \Big) \right] + 2 \cdot \mathbb{E}_{\pi \sim \widetilde{q}}[\pi_2(v^\star)] + 2/V \\
\leq& \mathbb{E}_{\sigma_1 \sim q} \left[D_{\mathsf{H}}^2 \big(M(\pi), \overline{M}(\pi) \big) \right] + 6/V,
\end{aligned}
$$

where the first equality uses convexity of the squared hellinger distance, the second inequality uses that $\widetilde{M}$ is a mixture of $\widetilde{M}_{M,v}$ with $v \sim \mathrm{Unif}([V])$, and the third inequality uses that $D_{\mathsf{H}}^2(\mathrm{Ber}(0), \mathrm{Ber}(1/V)) \leq 2 \cdot D_{\mathsf{TV}}(\mathrm{Ber}(0), \mathrm{Ber}(1/V)) = 2/V$. Thus, it follows that

$$
\begin{aligned}
\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) \leq& \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\sigma_1 \sim p}[g^M(\pi)] \mid \mathbb{E}_{\sigma_1 \sim q} \left[D_{\mathsf{H}}^2 \big(M(\sigma_1), \overline{M}(\sigma_1) \big) \right] \leq \varepsilon^2 \right\} \\
\leq& \sup_{M' \in \widetilde{\mathcal{M}}} \left\{ 2V^{-1/2} + (1 + 2V^{-1/2}) \cdot \mathbb{E}_{\pi \sim \widetilde{p}}[h^{M'}(\pi)] \mid \mathbb{E}_{\pi \sim \widetilde{q}} \left[D_{\mathsf{H}}^2 \Big(M'(\pi), \widetilde{M}(\pi) \Big) \right] \leq \varepsilon^2 + 6/V \right\} \\
\leq& 2V^{-1/2} + (1 + 2V^{-1/2}) \cdot \mathsf{dec}_{\varepsilon + (6/V)^{-1/2}}(\mathscr{M}, \widetilde{M}) \\
\leq& 6V^{-1/2} + \mathsf{dec}_{\varepsilon + (6/V)^{-1/2}}(\mathscr{M}, \widetilde{M}).
\end{aligned}
$$

Bounding $\mathsf{dec}_\varepsilon(\mathscr{M})$ by $\mathsf{dec}_\varepsilon(\mathscr{H})$. Next consider any reference model $\widetilde{M} \in \mathrm{co}(\widetilde{\mathcal{M}})$. We will upper bound $\mathsf{dec}_\varepsilon(\mathscr{M}, \widetilde{M})$ by $\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M})$ for some $\overline{M} \in \mathrm{co}(\mathcal{M})$. For some distribution $\mu \in \Delta(\mathcal{M} \times [V])$, we have $\widetilde{M}(\pi) = \mathbb{E}_{(M,v) \sim \mu}[\widetilde{M}_{M,v}(\pi)]$ for all $\pi \in \widetilde{\Pi}$. Define $\overline{M}$ by letting $\nu \in \Delta(\mathcal{M})$ to be the marginal of μ over $\mathcal{M}$, and then: $\overline{M}(\pi) := \mathbb{E}_{M \sim \nu}[M(\pi)]$ for each $\pi \in \Pi$. Choose some $p, q \in \Delta(\Pi)$ so that

$$
\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) = \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] \mid \mathbb{E}_{\pi \sim q}[D_{\mathsf{H}}^2 \big(M(\pi), \overline{M}(\pi) \big)] \leq \varepsilon^2 \right\}.
$$

Let $\widetilde{p} \in \Delta(\widetilde{\Pi})$ be the distribution of $(\sigma_1, 0)$ where $\sigma_1 \sim p$, and $\widetilde{q} \in \Delta(\widetilde{\Pi})$ be the distribution of $(\sigma_1, 0)$ where $\sigma_1 \sim q$. By definition of the models $\widetilde{M}_{M,v}$, for any $M' = \widetilde{M}_{M,v} \in \widetilde{\mathcal{M}}$, we have:

$$
\begin{aligned}
\mathbb{E}_{\pi \sim \widetilde{p}}[h^{M'}(\pi)] &= \mathbb{E}_{\pi \sim \widetilde{p}} \left[\sup_{\pi'_1 \in \Pi'_1} \{ f_1^{M'}(U_1(\pi'_1, \pi)) - f_1^{M'}(\pi) \} + \sup_{\pi'_2 \in \Pi'_2} \{ f_2^{M'}(U_2(\pi'_2, \pi)) - f_2^{M'}(\pi) \} \right] \\
&= \mathbb{E}_{\pi \sim \widetilde{p}} \left[\sup_{\pi'_2 \in \Pi'_2} \{ f_2^{M'}(U_2(\pi'_2, \pi)) \} \right] \\
&= \mathbb{E}_{\sigma_1 \sim p}[g^M(\sigma_1)].
\end{aligned}
$$

Furthermore, since $\widetilde{q}$ is supported entirely on $\Sigma_1 \times \{0\}$ and all models in $\widetilde{\mathcal{M}}$ have $r_1 = r_2 = 0$ a.s. under such policies, it holds that

$$
\mathbb{E}_{\pi \sim \widetilde{q}} \left[D_{\mathsf{H}}^2 \left(M'(\pi), \widetilde{M}(\pi) \right) \right] = \mathbb{E}_{\sigma_1 \sim q} \left[D_{\mathsf{H}}^2 \left(M(\pi), \overline{M}(\pi) \right) \right],
$$

which certifies that

$$
\mathsf{dec}_\varepsilon(\mathscr{M}, \widetilde{M}) \leq \sup_{M' \in \widetilde{\mathcal{M}}} \left\{ \mathbb{E}_{\pi \sim \widetilde{p}}[h^{M'}(\pi)] \mid \mathbb{E}_{\pi \sim \widetilde{q}} \left[D_{\mathsf{H}}^2 \left(M'(\pi), \widetilde{M}(\pi) \right) \right] \leq \varepsilon^2 \right\} \leq \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}).
$$

for each $\pi \in \Pi$, $\overline{M}(\pi)$ to be the distribution of $o_\circ$ when $(o_\circ, r_1, r_2) \sim \widetilde{M}_{M,v}((\pi, 0))$ and $(M, v) \sim \mu$. ∎

Proof of Lemma I.2. We denote a history drawn according to any of the distributions $\widetilde{M}_{M,v}$ (for $v \geq 0$) by $\{(\widetilde{\pi}^t, (\widetilde{o}^t, r_1^t, r_2^t))\}_{t=1}^T$. Furthermore, we abbreviate $\widetilde{q}^t = \widetilde{q}^t(\cdot | (\widetilde{\pi}^1, (\widetilde{o}^1, r_1^t, r_2^t)), \ldots, (\widetilde{\pi}^{t-1}, (\widetilde{o}^{t-1}, r_1^{t-1}, r_2^{t-1})))$ and $\widetilde{p} = \widetilde{p}(\cdot | (\widetilde{\pi}^1, (\widetilde{o}^1, r_1^t, r_2^t)), \ldots, (\widetilde{\pi}^{T-1}, (\widetilde{o}^{T-1}, r_1^{T-1}, r_2^{T-1}))))$. Define $M' := \widetilde{M}_{M,0}$ and choose

$$
v^\star := \arg\min_{v \in [V]} \left\{ \sum_{t=1}^T \mathbb{E}^{M', (\widetilde{p}, \widetilde{q})} \mathbb{E}_{\widetilde{\pi}^t \sim \widetilde{q}^t}[\widetilde{\pi}_2^t(v)] + T \cdot \mathbb{E}^{M', (\widetilde{p}, \widetilde{q})} \mathbb{E}_{\widehat{\pi} \sim \widetilde{p}}[\widehat{\pi}_2(v)] \right\}.
$$

Then the choice of $v^\star$ together with the fact that all $\pi \in \widetilde{\Pi}$ satisfy $\sum_{v=1}^V \pi_2(v) \leq 1$ ensures that

$$
\sum_{t=1}^T \mathbb{E}^{M', (\widetilde{p}, \widetilde{q})} \mathbb{E}_{\widetilde{\pi}^t \sim \widetilde{q}^t}[\widetilde{\pi}_2^t(v^\star)] + T \cdot \mathbb{E}^{M', (\widetilde{p}, \widetilde{q})} \mathbb{E}_{\widehat{\pi} \sim \widetilde{p}}[\widehat{\pi}_2(v^\star)] \leq \frac{2T}{V}. \tag{18}
$$

Write $M'' = \widetilde{M}_{M,v^\star}$. Next, using (Foster et al., 2021, Lemma A.13),[22] we have:

$$
\begin{aligned}
D_{\mathsf{H}}^2\Big(\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}, \mathbb{P}^{M'',(\widetilde{p},\widetilde{q})}\Big) =&\, O(\log T) \cdot \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\left[\sum_{t=1}^{T} D_{\mathsf{H}}^2\big(M'(\widetilde{\pi}^t), M''(\widetilde{\pi}^t)\big)\right]\\
\leq&\, O(\log T) \cdot \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\left[\sum_{t=1}^{T} \mathbb{E}_{(\sigma_1^t,\sigma_2^t)\sim\widetilde{\pi}^t}[D_{\mathsf{H}}^2\big(M'((\sigma_1^t,\sigma_2^t)), M''((\sigma_1^t,\sigma_2^t)))]\right]\\
\leq&\, O(\log T) \cdot \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\left[\sum_{t=1}^{T} \mathbb{E}_{(\sigma_1^t,\sigma_2^t)\sim\widetilde{\pi}^t}[2\cdot\mathbb{I}\{\sigma_2^t=v^\star\}]\right]\\
\leq&\, O\left(\frac{T\log T}{V}\right),
\end{aligned}
$$

where the final inequality uses (18). Since total variation distance is bounded above by Hellinger distance, it follows that $D_{\mathsf{TV}}\big(\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}, \mathbb{P}^{M'',(\widetilde{p},\widetilde{q})}\big) \leq C\sqrt{T\log(T)/V}$ for some constant $C > 0$. Using this fact together with (18), we see that

$$
\mathbb{E}^{M'',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v^\star)] \leq \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v^\star)] + C\sqrt{T\log(T)/V} \leq 2T/V + C\sqrt{T\log(T)/V},
$$

where the second inequality above uses (18). ∎

Appendix J. Proofs for Appendix D

J.1. Proofs from Appendix D.1

J.1.1. FURTHER DETAILS FOR UPPER BOUND

The upper bound from Theorem D.1 is derived by appealing to the $\mathsf{E2D}^+$ for PAC algorithm from Foster et al. (2023). In what follows, we give some background on the algorithm, as well as a more general upper bound. In brief, the $\mathsf{E2D}^+$ for PAC algorithm proceeds as follows: The algorithm uses an *online estimation oracle*, denoted by $\mathbf{Alg}_{\mathsf{Est}}$ (defined formally in Assumption J.1), which is given as input a model class $\mathcal{M}$ and attempts to estimate the true model $M^\star \in \mathcal{M}$ given data obtained from playing various decisions under $M^\star$. To generate each successive datapoint at iteration t, which will be fed to the estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$, the $\mathsf{E2D}^+$ for PAC algorithm solves the minimax problem in (5) to compute distributions p^t, q^t, where the model $\overline{M}$ is set to be the output of the estimation oracle from the previous iteration. Then, a decision π^t is sampled from q^t, and we observe the resulting observation $o^t \sim M^\star(\pi^t)$. The tuple (π^t, o^t) is then fed to the estimation oracle, which produces its next estimate $\widehat{M}^{t+1}$. The algorithm's output after T iterations is given by a sample from one of the distributions $p^{t^\star}$, where $t^\star \sim [T]$ is uniform. See Foster et al. (2023) for further background.

Assumption J.1 (Estimation oracle for $\mathcal{M}$)**.** *For each time $t \in [T]$, an online estimation oracle* $\mathbf{Alg}_{\mathsf{Est}}$ *for the class $\mathcal{M}$ takes as input $\mathfrak{H}^{t-1} = (\pi^1, o^1), \ldots, (\pi^{t-1}, o^{t-1})$ where $o^i \sim M^\star(\pi^i)$ and $\pi^i \sim q^i$, for arbitrary (adaptive) choices of the distributions $q^i \in \Delta(\Pi)$. Then, for some class*

---

22. In particular, we apply this lemma to the sequence $X_1, \ldots, X_{2T}$, where, for odd values of t we have $X_t = \widetilde{\pi}^t$, $X_{t+1} = (\widetilde{o}^t, r_1^t, r_2^t)$, and use that the conditional distribution of $\widetilde{\pi}^t$ given the history up to step $t-1$ is the same under the distributions $\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}$ and $\mathbb{P}^{M'',(\widetilde{p},\widetilde{q})}$ since the algorithm $(\widetilde{p},\widetilde{q})$ is the same.

$\widehat{\mathcal{M}} \subseteq \mathrm{co}(\mathcal{M})$, the oracle $\mathbf{Alg}_{\mathsf{Est}}$ returns an estimator $\widehat{M}^t \in \widehat{\mathcal{M}}$. We assume that if $M^\star \in \mathcal{M}$, the estimators produced by the algorithm satisfy

$$\mathbf{Est}_{\mathsf{H}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim q^t}\left[D_{\mathsf{H}}^2\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right] \leq \mathbf{Est}_{\mathsf{H}}(T, \delta),$$

with probability at least $1 - \delta$, where $\mathbf{Est}_{\mathsf{H}}(T, \delta)$ is a known upper bound.

For most estimation oracles, the class $\widehat{\mathcal{M}}$ in Assumption J.1 will be $\mathrm{co}(\mathcal{M})$, though in some cases it is possible to take it to be smaller (see Proposition G.10 for an example).

Theorem J.1 (Foster et al. (2023), Theorem 3.1; Upper bound for HR-DMSO). *Fix $\delta \in \left(0, \frac{1}{10}\right)$ and $T \in \mathbb{N}$, and consider any instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$. Suppose that Assumptions 1.4 and J.1 hold for the model class $\mathcal{M}$ and some class $\widehat{\mathcal{M}} \subseteq \mathrm{co}(\mathcal{M})$, and let $\overline{\mathbf{Est}}_{\mathsf{H}} := \mathbf{Est}_{\mathsf{H}}\left(\frac{2T}{\lceil \log 2/\delta \rceil}, \frac{\delta}{4\lceil \log 2/\delta \rceil}\right)$. Letting $\bar{\varepsilon}(T) := 8\sqrt{\frac{\lceil \log 2/\delta \rceil}{T} \cdot \overline{\mathbf{Est}}_{\mathsf{H}}}$, E2D$^+$ for PAC, with access to the oracle $\mathbf{Alg}_{\mathsf{Est}}$, guarantees that with probability at least $1 - \delta$,*

$$\mathbf{Risk}(T) \leq \sup_{\overline{M} \in \widehat{\mathcal{M}}} \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}, \overline{M}) \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}).$$

If further $f^M(\cdot) \in [0, R]$ for all $M \in \mathcal{M}$ and some $R > 0$, then the expected risk is bounded as $\mathbb{E}[\mathbf{Risk}(T)] \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) + R\delta$.

We remark that Theorem J.1 is only stated in Foster et al. (2023) for the case $\widehat{\mathcal{M}} = \mathrm{co}(\mathcal{M})$, but an inspection of the proof shows that the same guarantee holds for an arbitrary subclass $\widehat{\mathcal{M}} \subseteq \mathrm{co}(\mathcal{M})$ in which $\mathbf{Alg}_{\mathsf{Est}}$ produces its predictions (with no modifications to the proof being necessary).

Theorem D.1 follows from Theorem J.1 by noting that there exists an estimation oracle with $\mathbf{Est}_{\mathsf{H}}(T, \delta) \leq 2\log(|\mathcal{M}|/\delta)$ for finite classes (Foster et al., 2023).

Remark J.1 (Analogue for MA-DMSO). Using the transformation of Theorem C.1 (which does not change the model class of the instance, and therefore preserves estimation error guarantees), there is an analogue of Theorem J.1 for the multi-agent setting. In particular, for any instance $\mathscr{M}$ of MA-DMSO, under Assumptions 1.1 and J.1, there is an algorithm that ensures with probability $1 - \delta$, $\mathbf{Risk}(T) \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M})$.

Infinite model classes. As some of our applications in Appendix G involve infinite model classes $\mathcal{M}$, we next describe a simple way to bound the estimation error $\mathbf{Est}_{\mathsf{H}}(T, \delta)$ for such classes, following the approach in Foster et al. (2021).

Definition J.1 (Model class cover; Foster et al. (2021), Definition 3.2). *A model class $\mathcal{M}' \subseteq \mathcal{M}$ is an ε-cover for $\mathcal{M}$ if for all $M \in \mathcal{M}$, there is $M' \in \mathcal{M}'$ so that $\sup_{\pi \in \Pi} D_{\mathsf{H}}^2(M(\pi), M'(\pi)) \leq \varepsilon^2$. Let $\mathcal{N}(\mathcal{M}, \varepsilon)$ denote the size of the smallest such cover $\mathcal{M}'$, and define*

$$\mathsf{est}(\mathcal{M}, T) := \inf_{\varepsilon \geq 0}\left\{\log \mathcal{N}(\mathcal{M}, \varepsilon) + \varepsilon^2 T\right\}.$$

We will bound the estimation error $\mathbf{Est}_{\mathsf{H}}(T, \delta)$ for a model class in terms of the quantity $\mathsf{est}(\mathcal{M}, T)$; to do so, we need the following mild assumption.

Assumption J.2. *Suppose that there is a kernel ν from $(\Pi, \mathscr{P})$ to $(\mathcal{O}, \mathscr{O})$ so that $M(\pi) \ll \nu(\pi)$ for all $M \in \mathcal{M}, \pi \in \Pi$, and let $m^M(\cdot|\pi)$ denote the density of $M(\cdot|\pi)$ with respect to $\nu(\cdot|\pi)$. Furthermore, suppose there is a constant $B \geq e$ so that*

1. *$\nu(\mathcal{O}|\pi) \leq B$ for all $\pi \in \Pi$.*

2. *$\sup_{\pi \in \Pi} \sup_{o \in \mathcal{O}} m^M(o|\pi) \leq B$ for all $M \in \mathcal{M}$.*

Proposition J.1 below shows that the estimation error $\mathbf{Est}_{\mathsf{H}}(T, \delta)$ scales with $\log B$. This quantity is typically small: for instance, it is a constant for standard multi-armed bandit problems (e.g., Bernoulli bandits and Gaussian bandits), and is polylogarithmic in the size of the state and action spaces for reinforcement learning problems with finite state and action spaces.

Proposition J.1 (Lemma A.16 of Foster et al. (2021))**.** *Suppose Assumption J.2 holds. Fix $T \in \mathbb{N}, \delta \in (0, e^{-1})$, and write $b_T = \log(2B^2 T)$. Then there is an algorithm $\mathbf{Alg}_{\mathsf{Est}}$ that guarantees that, with probability $1 - \delta$, we have*

$$\mathbf{Est}_{\mathsf{H}}(T) \leq O(b_T \cdot \mathsf{est}(\mathcal{M}, T) + b_T^2 \log(\delta^{-1})),$$

i.e., we can take $\mathbf{Est}_{\mathsf{H}}(T, \delta) = C \cdot (b_T \cdot \mathsf{est}(\mathcal{M}, T) + b_T^2 \log(\delta^{-1}))$ for some universal constant C.

J.1.2. PROOF OF THEOREM D.2

Proof of Theorem D.2. Fix $T \in \mathbb{N}$ and an algorithm $(p, q) = \{q^t(\cdot|\cdot), p(\cdot|\cdot)\}_{t=1}^T$. For each model $M \in \mathcal{M}^+$, we use the abreviation $\mathbb{P}^M \equiv \mathbb{P}^{M,(p,q)}$, and write $\mathbb{E}^M$ for the corresponding expectation. We also define

$$p_M = \mathbb{E}^M[p(\cdot|\mathfrak{H}^T)], \qquad q_M = \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^T q^t(\cdot|\mathfrak{H}^{t-1})\right].$$

Choose $\underline{\varepsilon}(T)$ as in the theorem statement, and write $\varepsilon = \underline{\varepsilon}(T)$. Choose $\overline{M} \in \mathrm{co}(\mathcal{M})$ so that $\mathsf{dec}_\varepsilon(\mathcal{M}) = \mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M})$.[23] We will prove a lower bound on the expected risk in terms of $\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M})$. Define

$$M := \underset{M \in \mathcal{M}}{\arg\max}\left\{\mathbb{E}_{\pi \sim p_{\overline{M}}}[g^M(\pi)] \mid \mathbb{E}_{\pi \sim q_{\overline{M}}}[D_{\mathsf{H}}^2(M(\pi), \overline{M}(\pi))] \leq \varepsilon^2\right\},$$

where we recall that $C(T) := \log(T \wedge V(\mathcal{M}))$. Note that if the $\mathcal{H}_{q_{\overline{M}}, \varepsilon}(\overline{M}) = \varnothing$, then by definition $\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) = 0$ and the result follows. Thus, we may assume that $\mathcal{H}_{q_{\overline{M}}, \varepsilon}(\overline{M}) \neq \varnothing$, and hence the choice of M above is well-defined. Furthermore, the choice of M ensures that

$$\mathbb{E}_{\pi \sim p_{\overline{M}}}[g^M(\pi)] \geq \mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) = \mathsf{dec}_\varepsilon(\mathcal{M}). \tag{19}$$

By Lemma A.13 in Foster et al. (2021), we have[24]

$$D_{\mathsf{H}}^2\left(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\right) \leq C(T) \cdot T \cdot \mathbb{E}_{\pi \sim q_{\overline{M}}}[D_{\mathsf{H}}^2(M(\pi), \overline{M}(\pi))] \leq C(T) \cdot T \cdot \varepsilon^2.$$

---

23. If the supremum over $\overline{M}$ is not achievable, then we may apply the argument that follows for a sequence that achieves the supremum.

24. In order to apply this result, we need to ensure that for all measurable sets $\mathcal{E} \subseteq \mathcal{O}$ and all $\pi \in \Pi$, we have $\frac{\overline{M}(\mathcal{E}|\pi)}{M(\mathcal{E}|\pi)} \leq V(\mathcal{M})$. This follows from the definition of $V(\mathcal{M})$ in (7) and the fact that $\overline{M} \in \mathrm{co}(\mathcal{M})$.

Using the data processing inequality, it follows that

$$D_{\mathsf{H}}^2(p_M, p_{\overline{M}}) \le C(T) \cdot T \cdot \varepsilon^2 \le \frac{1}{8R} \cdot \mathsf{dec}_\varepsilon(\mathcal{M}), \tag{20}$$

where the second inequality follows from the choice of $\varepsilon = \underline{\varepsilon}(T)$.

Next, using Lemma A.11 in Foster et al. (2021) and the fact that $g^M(\pi) \in [0, R]$ for all π, we have

$$\mathbb{E}_{\pi \sim p_{\overline{M}}}[g^M(\pi)] \le 3 \cdot \mathbb{E}_{\pi \sim p_M}[g^M(\pi)] + 4R \cdot D_{\mathsf{H}}^2(p_M, p_{\overline{M}}).$$

Combining the above display with (19) and (20) and rearranging, we see that

$$\frac{1}{6} \cdot \mathsf{dec}_\varepsilon(\mathcal{M}) \le \mathbb{E}_{\pi \sim p_M}[g^M(\pi)] = \mathbb{E}^M \mathbb{E}_{\pi \sim p(\cdot | \mathfrak{H}^T)}[g^M(\pi)] = \mathbb{E}^M[\mathbf{Risk}(T)],$$

which gives the desired lower bound on expected risk. ∎

J.2. Proofs from Appendix D.2

J.2.1. PROOF OF PROPOSITION D.1

Proof of Proposition D.1. Define $\Delta := \frac{\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M})}{8 \cdot \bar{\varepsilon}(T)^2 \cdot C(T) \cdot T}$. If $\Delta \ge 1$, then we have $\underline{\varepsilon}(T) \ge \bar{\varepsilon}(T)$, so we may assume from here on that $\Delta < 1$. Choose

$$\alpha = \left\lceil \frac{\log 1/\Delta}{2 \log(C_{\mathrm{reg}}/c_{\mathrm{reg}})} \right\rceil \ge 1,$$

which in particular is the smallest positive integer so that $(C_{\mathrm{reg}}^2/c_{\mathrm{reg}}^2)^\alpha \ge 1/\Delta$. Such α is well-defined by our assumption that $C_{\mathrm{reg}} > c_{\mathrm{reg}}$ and since $1/\Delta > 1$. Applying Assumption D.1 to $\varepsilon = \bar{\varepsilon}(T) \cdot (c_{\mathrm{reg}}/C_{\mathrm{reg}})^j$ for $0 \le j < \alpha$, it follows that

$$\begin{aligned}
\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M}) &\le c_{\mathrm{reg}}^{2\alpha} \cdot \mathsf{dec}_{\bar{\varepsilon}(T)/C_{\mathrm{reg}}^\alpha}(\mathcal{M}) \\
&\le \Delta \cdot C_{\mathrm{reg}}^{2\alpha} \cdot \mathsf{dec}_{\bar{\varepsilon}(T)/C_{\mathrm{reg}}^\alpha}(\mathcal{M}) \\
&\le \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M}) \cdot \frac{\mathsf{dec}_{\bar{\varepsilon}(T)/C_{\mathrm{reg}}^\alpha}(\mathcal{M})}{8 \cdot (\bar{\varepsilon}(T)/C_{\mathrm{reg}}^\alpha)^2 \cdot C(T) \cdot T}.
\end{aligned}$$

Hence $\underline{\varepsilon}(T) \ge \bar{\varepsilon}(T)/C_{\mathrm{reg}}^\alpha$, and so

$$\mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \ge \mathsf{dec}_{\bar{\varepsilon}(T)/C_{\mathrm{reg}}^\alpha}(\mathcal{M}) \ge \frac{1}{c_{\mathrm{reg}}^{2\alpha}} \cdot \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M}).$$

The definition of α and $\bar{\varepsilon}(T)$ gives that

$$\left(\frac{C_{\mathrm{reg}}}{c_{\mathrm{reg}}} \right)^{2\alpha} \le \frac{C_{\mathrm{reg}}}{c_{\mathrm{reg}}} \cdot \frac{1}{\Delta} \le \frac{C_{\mathrm{reg}}}{c_{\mathrm{reg}}} \cdot \frac{8^3 \cdot \lceil \log 2/\delta \rceil \cdot \overline{\mathbf{Est}}_{\mathsf{H}} \cdot C(T)}{\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M})}.$$

Our definition of β ensures that $c_{\mathrm{reg}}^{2\alpha} \le ((C_{\mathrm{reg}}/c_{\mathrm{reg}})^{2\alpha})^\beta$, meaning that, for some constant C,

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M}) \le c_{\mathrm{reg}}^{2\alpha} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \le (C \cdot C_{\mathrm{reg}}/c_{\mathrm{reg}})^\beta \cdot \log^\beta 1/\delta \cdot \overline{\mathbf{Est}}_{\mathsf{H}}^\beta \cdot C(T)^\beta \cdot \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M})^{-\beta} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}),$$

and rearranging yields:

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M}) \leq \left(C \log 1/\delta \cdot \overline{\mathbf{Est}}_{\mathsf{H}} \cdot C(T) \cdot C_{\mathrm{reg}}/c_{\mathrm{reg}}\right)^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})^{\frac{1}{1+\beta}}.$$

∎

J.2.2. PROOF OF PROPOSITION D.2 AND PROPOSITION D.3

Proof of Proposition D.2.   We set $\Pi = [A]$ and $\mathcal{O} = \{0, 1\}$. For each $\delta > 0$, we define a model class $\mathcal{M}_\delta \subset (\Pi \to \Delta(\mathcal{O}))$, as follows: $\mathcal{M}_\delta = \{M_{\delta,a} : a \in [A]\}$, and define $M_{\delta,a}(\pi) = \mathrm{Ber}(1/2 + \delta \mathbb{I}\{\pi = a\})$. We now set $\mathcal{M} = \bigcup_{i=2}^L \mathcal{M}_{2^{-i}}$, from which it follows that $|\mathcal{M}| \leq LA$. Define $f^M(\pi) = \mathbb{E}^{M,\pi}[r]$, where $r \sim M(\pi)$. Finally set $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$. Note that the instance $\mathscr{H}$ is actually a standard (non-hidden reward) DMSO instance in the sense of Foster et al. (2021).

Since the model class $\mathcal{M}$ is a subclass of the class of all A-armed bandit problems, we have from Proposition 5.1 of Foster et al. (2021) and Proposition E.1 (which applies identically to HR-DMSO instances in addition to MA-DMSO instances) that $\mathsf{dec}_\varepsilon(\mathscr{H}) \leq O(\varepsilon \sqrt{A})$. Furthermore, we have $\mathfrak{M}(\mathscr{H}, T) \leq O(\sqrt{TA})$ (Audibert and Bubeck, 2009) (up to logarithmic factors, this bound is also a consequence of, e.g., Theorem D.1).

For each $\delta > 0$, write $\mathscr{H}_\delta := (\mathcal{M}_\delta, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$. Also write $\overline{M}_\delta = \frac{1}{A} \sum_{a=1}^A M_{\delta,a} \in \mathrm{co}(\mathcal{M}_\delta)$. Since for all $\pi \in \Pi$ and $a \in [A]$, $D_{\mathsf{H}}^2(\overline{M}_\delta(\pi), M_{\delta,a}(\pi)) \leq 4\delta^2$, it is straightforward to see that $\mathsf{dec}_{4\delta/\sqrt{A}}(\mathscr{H}_\delta) \geq \mathsf{dec}_{4\delta/\sqrt{A}}(\mathscr{H}_\delta, \overline{M}_\delta) \geq \Omega(\delta)$. Since increasing the size of the model class cannot decrease the DEC, it follows that, for all ε satisfying $1/\sqrt{A} > \varepsilon > 2^{-L}$, $\mathsf{dec}_\varepsilon(\mathscr{H}) \geq \Omega(\varepsilon \sqrt{A})$. Finally, since the rewards are observed in the instance $\mathscr{H}$, we can use Theorem 2.1 of Foster et al. (2023) to conclude that for A at least some sufficiently large constant, and $T \leq 2^{L/2}$, $\mathfrak{M}(\mathscr{H}, T) \geq \Omega(\sqrt{A/T})$. ∎

Proof of Proposition D.3.   Fix L to be larger than some universal constant (whose value will be specified below), and consider any value for a constant $C_{\mathrm{prob}} \geq 1$. We define the following instance $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_{M \in \mathcal{M}})$, with the individual components defined as follows:

- For $1 \leq \ell \leq L$, define $\alpha_\ell := 1/L$, $N_\ell = 2^\ell$, and $\delta_\ell = \frac{1}{(C_{\mathrm{prob}} N_\ell)^2}$.

- Let $\mathcal{V} := \prod_{\ell=1}^L [N_\ell]$, and set $\Pi := \mathcal{V}$.

- Let $\mathcal{O} = \prod_{\ell=1}^L ([N_\ell] \cup \{\perp_\ell\})$. For ease of notation we write $\mathcal{O}_\ell := [N_\ell] \cup \{\perp_\ell\}$.

- For $o_\ell \in \mathcal{O}_\ell$ and $v_\ell \in [N_\ell]$, define

$$P_{v_\ell}(o_\ell) := \begin{cases} 1 - \delta(N_\ell - 1) & : o_\ell = \perp_\ell \\ \delta_\ell & : o_\ell \in [N_\ell] \backslash \{v_\ell\} \\ 0 & : o_\ell = v_\ell. \end{cases}$$

Then $P_{v_\ell} \in \Delta(\mathcal{O}_\ell)$.

73

- The class $\mathcal{M}$ is indexed by tuples $v \in \mathcal{V}$; in particular, for each $v = (v_1, \ldots, v_\ell) \in \mathcal{V}$, there is a model M_v, defined as follows. For $\pi \in \Pi$, $M_v(\pi) \in \Delta(\mathcal{O})$ is the following distribution which does not depend on π: for $o = (o_1, \ldots, o_L) \in \mathcal{O}$,

$$M_v(\pi)(o) = \prod_{\ell=1}^{L} P_{v_\ell}(o_\ell).$$

Since the distribuiton $M_v(\pi)$ does not depend on π, we will often drop the argument π and simply write $M_v \in \Delta(\mathcal{O})$. Accordingly, the Hellinger distance between observation distributions of two models $M, M' \in \mathrm{co}(\mathcal{M})$ will be denoted by $D_{\mathsf{H}}^2(M, M')$.

- For all $v \in \mathcal{V}$ and $\pi = (\pi_1, \ldots, \pi_L) \in \Pi$, the value function $f : \Pi \to [0, 1]$ is defined as follows:

$$f^{M_v}(\pi) := \sum_{\ell=1}^{L} \alpha_\ell \cdot (1 - \mathbb{I}\{\pi_\ell = v_\ell\}).$$

For convenience we write $f_\ell^{M_v}(\pi) := (1 - \mathbb{I}\{\pi_\ell = v_\ell\})$, so that $f^{M_v}(\pi) = \sum_{\ell=1}^{L} \alpha_\ell \cdot f_\ell^{M_v}(\pi)$. It is clear that for all $v \in \mathcal{V}$ there is some π (namely, any π so that $\pi_\ell \neq v_\ell$ for all ℓ) for which $f^{M_v}(\pi) = 1$, meaning that $g^{M_v}(\pi) = 1 - f^{M_v}(\pi)$.

Upper bounding the minimax sample complexity. Fix some $T \in \mathbb{N}$; we next upper bound $\mathfrak{M}(\mathcal{H}, T)$. Since the distribution over observations for all models in the class $\mathcal{M}$ does not depend on the decision, to specify an algorithm (p, q) we need only to specify the distribution p, which is a mapping from T-tuples of observations to distributions over decisions. To define p, we first define mappings $p_\ell : \mathcal{O}_\ell^T \to [N_\ell]$, as follows:

$$p_\ell(o_{\ell,1}, \ldots, o_{\ell,T}) := \begin{cases} \mathrm{Unif}(\Pi) & : o_{\ell,1} = \cdots = o_{\ell,T} = \bot_\ell, \\ \mathbb{I}_{o_{\ell,t}} & : t := \arg\min\{s \in [T] \mid o_{\ell,s} \neq \bot_\ell\} \text{ exists.} \end{cases}$$

In particular, p_ℓ outputs the first index of an observation which is not $\bot_\ell$; if no such index exists, then p_ℓ outputs the uniform distribution over $[N_\ell]$. Now we define

$$p(o_1, \ldots, o_T) := (p_\ell(o_{\ell,1}, \ldots, o_{\ell,T}))_{\ell=1}^{L},$$

where we have written $o_t = (o_{1,t}, \ldots, o_{L,t})$ for each $T \in [T]$.

We now upper bound the risk of the algorithm p. We abbreviate the distribution over histories under a given model $M_v \in \mathcal{M}$ by $\mathbb{P}^{M_v}(\cdot)$, and write $\mathbb{E}^{M_v}[\cdot]$ for the corresponding expectation. For each $\ell \in [L]$, we have, for all $M_v \in \mathcal{M}$,

$$\mathbb{E}^{M_v}\mathbb{E}_{\pi \sim p(o_1, \ldots, o_T)}\left[1 - f_\ell^{M_v}(\pi)\right] \leq (1 - \delta_\ell(N_\ell - 1))^T \cdot \frac{1}{N_\ell},$$

since the probability that there is no $t \in [T]$ so that $o_{\ell,t} \neq \bot_\ell$ is $(1 - \delta_\ell(N_\ell - 1))^T$, and on the complement of this event (so that such t exists), $p_\ell(o_1, \ldots, o_T)$ puts all its mass on such $o_{\ell,t} \neq v_\ell$, so that $f_\ell^{M_v}(\pi) = 1$. Hence

$$\mathbb{E}^{M_v}\mathbb{E}_{\pi \sim p(o_1, \ldots, o_T)}[g^{M_v}(\pi)] \leq \sum_{\ell=1}^{L} \frac{\alpha_\ell}{N_\ell} \cdot (1 - \delta_\ell(N_\ell - 1))^T \leq \sum_{\ell=1}^{L} \frac{\alpha_\ell}{N_\ell} \cdot \left(1 - \frac{1}{2C_{\mathrm{prob}}^2 N_\ell}\right)^T.$$

74

Given $T \leq N_L$, choose $\ell_\star = \ell_\star(T) \in [L]$ as large as possible so that $T \geq 2\log(N_{\ell_\star}) \cdot 2C_{\text{prob}}^2 N_{\ell_\star}$, which gives

$$
\mathbb{E}^{M_v}\mathbb{E}_{\pi \sim p(o_1,\ldots,o_T)}[g^{M_v}(\pi)] \leq \sum_{\ell=1}^{\ell_\star} \frac{\alpha_\ell}{N_\ell} \cdot \exp\left(-\frac{2\log N_{\ell_\star}}{T}\right)^T + \sum_{\ell=\ell_\star+1}^{L} \frac{\alpha_\ell}{N_\ell}
$$

$$
\leq \sum_{\ell=1}^{\ell_\star} \frac{\alpha_\ell}{N_\ell} \cdot \frac{1}{N_{\ell_\star}^2} + \sum_{\ell=\ell_\star+1}^{L} \frac{\alpha_\ell}{N_\ell} \leq \frac{1}{N_{\ell_\star}} \leq \frac{8C_{\text{prob}}^2 \log T}{T},
$$

where the final inequality uses that our choice of $\ell_\star$ gives that $N_{\ell_\star} \geq \frac{T}{8C_{\text{prob}}^2 \log T}$,

Lower bounding the DEC. By the tensorization property of the squared Hellinger distance, we have, for any two models $M_v, M_u \in \mathcal{M}$,

$$
1 - \frac{1}{2}D_{\mathsf{H}}^2(M_u, M_v) = \prod_{\ell=1}^{L}\left(1 - \frac{1}{2}D_{\mathsf{H}}^2(P_{v_\ell}, P_{u_\ell})\right) = \prod_{\ell=1}^{L}\left(1 - \frac{1}{2}\cdot\mathbb{I}\{u_\ell \neq v_\ell\}\cdot 2\delta_\ell\right) \geq 1 - \sum_{\ell=1}^{L}\delta_\ell \cdot \mathbb{I}\{u_\ell \neq v_\ell\},
$$

which implies that $D_{\mathsf{H}}^2(M_u, M_v) \leq 2\sum_{\ell=1}^{L}\delta_\ell \cdot \mathbb{I}\{u_\ell \neq v_\ell\}$. Let $\overline{v} := (1, 1, \ldots, 1) \in \mathcal{V}$, and set $\overline{M} := M_{\overline{v}}$.

Now consider any $2 \geq \varepsilon \geq \sqrt{2\delta_L}$. Choose $\ell^\star = \ell^\star(\varepsilon)$ to be the smallest possible value of $\ell \in [L]$ so that $\varepsilon^2 \geq 2\delta_\ell$. For each $i \in [N_{\ell^\star}]$, define $\overline{v}^i \in \mathcal{V}$ by:

$$
\overline{v}_\ell^i = \begin{cases} \overline{v}_\ell & : \ell \neq \ell^\star \\ i & : \ell = \ell^\star, \end{cases}
$$

and write $M^i := M_{\overline{v}^i}$. Then for all $i \in [N_{\ell^\star}]$, we have $D_{\mathsf{H}}^2(\overline{M}, M^i) \leq 2\delta_{\ell^\star} \leq \varepsilon^2$. For any distribution $p \in \Delta(\Pi)$, there must be some $i^\star \in [N_{\ell^\star}]$ so that

$$
\mathbb{E}_{\pi \sim p}[1 - f_{\ell^\star}^{M^{i^\star}}(\pi)] = \mathbb{P}_{\pi \sim p}(\pi_{\ell^\star} = i^\star) \geq 1/N_{\ell^\star}.
$$

Therefore,

$$
\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \frac{\alpha_{\ell^\star}}{N_{\ell^\star}} = \alpha_{\ell^\star}C_{\text{prob}}\sqrt{\delta_{\ell^\star}} \geq \frac{\alpha_{\ell^\star}C_{\text{prob}}}{\sqrt{8}}\cdot\varepsilon = \frac{C_{\text{prob}}}{\sqrt{8}\cdot L}\cdot\varepsilon, \tag{21}
$$

where the final inequality uses that $\varepsilon^2 \leq 8\delta_\ell$ since $\delta_{\ell+1} = \delta_\ell/4$ for all $\ell < L$ and $\varepsilon \leq 2$.

Upper bounding the DEC. Next we upper bound $\mathsf{dec}_\varepsilon(\mathcal{M})$ for $\varepsilon \in (0, 2)$; while not necessary for lower bounding $\underline{\varepsilon}(T)$, an upper bound on the $\mathsf{dec}_\varepsilon(\mathcal{M})$ serves to ensure that the class $\mathcal{M}$ satisfies the regularity condition of Assumption D.1. This certifies that the instance $\mathscr{H}$ we construct satisfies the assumptions that we use to upper and lower bounding minimax risk in terms of the DEC.

Consider any $\overline{M} \in \mathsf{co}(\mathcal{M})$. We can write $\overline{M} = \mathbb{E}_{v \sim \mu}[M_v]$ for some distribution $\mu \in \Delta(\mathcal{V})$. For each $\ell \in [L]$, let $\mu_\ell \in \Delta([N_\ell])$ be the marginal of μ on $[N_\ell]$ (recall that $\mathcal{V} = \prod_{\ell=1}^{L}[N_\ell]$). Since $D_{\mathsf{H}}^2(P_{u_\ell}, P_{v_\ell}) = 2\delta_\ell$ for $u_\ell \neq v_\ell$, any two distinct values $v_\ell, v_\ell' \in [N_\ell]$ satisfying $D_{\mathsf{H}}^2(\mathbb{E}_{u_\ell \sim \mu_\ell}[P_{u_\ell}], P_{v_\ell}) \leq \varepsilon^2$ and $D_{\mathsf{H}}^2(\mathbb{E}_{u_\ell \sim \mu_\ell}[P_{u_\ell}], P_{v_\ell'}) \leq \varepsilon^2$ must in turn satisfy

$$
2\delta_\ell = D_{\mathsf{H}}^2(P_{v_\ell}, P_{v_\ell'}) \leq 2 \cdot D_{\mathsf{H}}^2(\mathbb{E}_{u_\ell \sim \mu_\ell}[P_{u_\ell}], P_{v_\ell}) + 2 \cdot D_{\mathsf{H}}^2(\mathbb{E}_{u_\ell \sim \mu_\ell}[P_{u_\ell}], P_{v_\ell'}) \leq 4\varepsilon^2. \tag{22}
$$

Now consider any $\varepsilon \in (\sqrt{\delta_L}, 2)$. Define $\bar{\ell} = \bar{\ell}(\varepsilon)$ to be the largest possible value of $\ell \in [L]$ so that $\varepsilon^2 < \delta_{\bar{\ell}}/2$. By (22) it follows that for all $\ell \leq \bar{\ell}$, there is at most a single value of $v_\ell \in [N_\ell]$ so that $D_{\mathsf{H}}^2(\mathbb{E}_{u_\ell \sim \mu_\ell}[P_{u_\ell}], P_v) \leq \varepsilon^2$. Denote this value of v_ℓ by $\bar{v}_\ell$ if such a v_ℓ exists; if not, choose $\bar{v}_\ell \in [N_\ell]$ arbitrarily.

By the data processing inequality, for any $v \in \mathcal{V}$ and each $\ell \in [L]$, it holds that $D_{\mathsf{H}}^2(\overline{M}, M_v) \geq D_{\mathsf{H}}^2(\mathbb{E}_{u_\ell \sim \mu_\ell}[P_{u_\ell}], P_{v_\ell})$. Thus, for each $M_v \in \mathcal{M}$ so that $D_{\mathsf{H}}^2(\overline{M}, M_v) \leq \varepsilon^2$, we must have that $v_\ell = \bar{v}_\ell$ for all $\ell \leq \bar{\ell}$.

Now choose any $v^\star \in \mathcal{V}$ so that $v_\ell^\star \neq \bar{v}_\ell$ for all $\ell \leq \bar{\ell}$, and define $p^\star \in \Delta(\Pi)$ as follows:

$$p^\star := \mathrm{Unif}\left(\left\{v \in \mathcal{V} \mid v_\ell = v_\ell^\star \ \forall \ell \leq \bar{\ell}\right\}\right).$$

We may now compute:

$$\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \sup_{M \in \mathcal{M}} \left\{\mathbb{E}_{\pi \sim p^\star}[g^M(\pi)] \mid D_{\mathsf{H}}^2(\overline{M}, M) \leq \varepsilon^2\right\} \leq \sum_{\ell=\bar{\ell}+1}^{L} \frac{\alpha_\ell}{N_\ell} \leq \frac{1}{N_{\bar{\ell}+1}} \leq 2C_{\mathrm{prob}}\varepsilon,$$

where the final inequality uses that $\varepsilon^2 \geq \delta_{\bar{\ell}}/8$ by definition of $\bar{\ell}$ and the fact that $\varepsilon \geq \sqrt{\delta_L}$.

Bounding $\underline{\varepsilon}(T)$. Consider any $T \leq N_L/L^3$, which ensures that (for sufficiently large L),

$$\frac{C_{\mathrm{prob}}}{8\sqrt{8} \cdot L \cdot C(T) \cdot T} \geq \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot N_L} = \sqrt{2\delta_L}, \tag{23}$$

where we have used that $C(T) \leq C_0 \cdot \log(T)$ for some universal constant C_0. Recall that $\underline{\varepsilon}(T)$ is defined to be as large as possible so that $\underline{\varepsilon}(T)^2 \cdot C(T) \cdot T \leq \frac{1}{8} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})$. Set

$$\varepsilon_0 := \frac{C_{\mathrm{prob}}}{8\sqrt{8} \cdot L \cdot C(T) \cdot T} \geq \Omega\left(\frac{C_{\mathrm{prob}}}{T \log(T) \cdot L}\right),$$

which, using (23) and (21), satisfies $\varepsilon_0^2 \cdot C(T) \cdot T \leq \frac{1}{8} \cdot \mathsf{dec}_{\varepsilon_0}(\mathcal{M})$, and thus $\underline{\varepsilon}(T) \geq \varepsilon_0$. ∎

J.2.3. PROOF OF THEOREM D.3

Proof of Theorem D.3. Given any $C_{\mathrm{prob}} \geq 1$, fix $N = \lceil\sqrt{T/C_{\mathrm{prob}}}\rceil$. For real numbers $\delta, \beta \in (0, 1)$, we will define instances $\mathscr{H}^{\delta,\beta}$ of the HR-DMSO framework. We will later choose $\mathscr{H}_1, \mathscr{H}_2$ to be such instances for certain choices of δ, β. For some model classes $\mathcal{M}^{\delta,\beta}$, each of size N, we will have, for all δ, β, $\mathscr{H}^{\delta,\beta} = (\mathcal{M}^{\delta,\beta}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$, i.e., the instances $\mathscr{H}^{\delta,\beta}$ share the same decision space, observation space, and value functions. We next define these components:

- $\Pi = [N]$ and $\mathcal{O} = [N] \cup \{\perp\}$.

- For all δ, β, we have $\mathcal{M}^{\delta,\beta} = \{M_1^{\delta,\beta}, \ldots, M_N^{\delta,\beta}\}$. For $i \in [N]$ and $\pi \in \Pi$, $M_i^{\delta,\beta}(\pi) \in \Delta(\mathcal{O})$ is the following distribution, which does not depend on π:

$$M_i^{\delta,\beta}(\pi)(j) = \begin{cases} 1 - \delta(N-1) - \beta & : j = \perp \\ \delta & : j \in [N]\backslash\{i\} \\ \beta & : j = i. \end{cases}$$

Since the distribution $M_i^{\delta,\beta}(\pi)$ does not depend on π, we will often drop the argument π and simply write $M_i^{\delta,\beta} \in \Delta(\mathcal{O})$.

- For all δ, β and for all $\pi \in \Pi$, $i \in [N]$, the value function $f^{M_i^{\delta,\beta}} : \Pi \to [0,1]$ is defined as follows:

$$f^{M_i^{\delta,\beta}}(\pi) := 1 - \mathbb{I}\{i = \pi\}.$$

Since the above value function does not depend on δ, β, we will simply write $f^i(\pi) := f^{M_i^{\delta,\beta}}(\pi)$ and $g^i(\pi) := g^{M_i^{\delta,\beta}}(\pi) = \mathbb{I}\{i = \pi\}$.

In the model $M_i^{\delta,\beta}$, all decisions except decision i are optimal. Furthermore, we will always have $\beta < \delta$, meaning that, under $M_i^{\delta,\beta}$, it is more likely to observe any given index in $[N]\backslash\{i\}$ than it is to observe i.

Upper bounding the minimax risk. Next, for $T \in \mathbb{N}$, we upper bound $\mathfrak{M}(\mathscr{H}^{\delta,\beta}, T)$. Since the distribution over observations for all models in the classes $\mathcal{M}^{\delta,\beta}$ does not depend on the decision, to specify an algorithm (p,q) we need only to specify the distribution p, which is a mapping from T-tuples of observations to distributions over decisions. Furthermore, to specify the distribution over histories under a given model $M_i^{\delta,\beta}$, we write $\mathbb{E}^{M_i^{\delta,\beta}}[\cdot]$. Now consider the algorithm p defined by:

$$p(o_1, \ldots, o_T) := \begin{cases} \mathrm{Unif}(\Pi) & : o_1 = \cdots = o_T = \bot \\ \mathbb{I}_{o_t} & : t := \arg\min\{s \in [T] \mid o_s \neq \bot\}, \quad \text{exists.} \end{cases} \tag{24}$$

In particular, p outputs the index of the first observation which is not $\bot$; if no such index exists, then p outputs the uniform distribution over decisions. To upper bound the expected risk of p, note that, for any model $M_i^{\delta,\beta}$, we have

$$\mathbb{E}^{M_i^{\delta,\beta}}\mathbb{E}_{\pi \sim p(o_1,\ldots,o_T)}[g^i(\pi)] \leq \frac{(1 - \delta(N-1) - \beta)^T}{N} + \frac{\beta}{\beta + \delta(N-1)} \leq \frac{(1 - \delta(N-1))^T}{N} + \frac{\beta}{\delta(N-1)}, \tag{25}$$

where the first term on the right-hand side accounts for the case that $o_1 = \cdots = o_T = \bot$, and the second term gives the probability that, given that there exists s such that $o_s \neq \bot$, the index t of the first such observation satisfies $o_t = i$.

Lower bounding the minimax risk. We next lower bound the minimax risk for the instances $\mathscr{H}^{\delta,\beta}$ in the following lemma; the proof is provided at the end of the section.

Lemma J.1. *Fix any real numbers $C \geq 1$ and $\epsilon \in (0,1)$, suppose $\delta \leq 1/N$, and write $\beta = \delta/C$. The minimax risk for the instance $\mathscr{H}^{\delta,\beta}$ is bounded below as follows: for $S \leq 1/\delta^{1-\epsilon}$, $\mathfrak{M}(\mathscr{H}^{\delta,\beta}, S) \geq \frac{1}{2NC^{2/\epsilon}}$.*

Computing D_ϕ. It is now straightforward to compute the D_ϕ-divergence between any two models in $\mathcal{M}^{\delta,\beta}$. In particular, for $i \in [N]$, we have:

$$D_\phi(M_i^{\delta,\beta} \parallel M_j^{\delta,\beta}) = \begin{cases} 0 & : i = j \\ \beta \cdot \phi(\delta/\beta) + \delta \cdot \phi(\beta/\delta) & : i \neq j. \end{cases} \tag{26}$$

77

Choosing δ, β. Let $\epsilon \in (0,1)$ and $T \in \mathbb{N}$ be as in the theorem statement. Since ϕ is assumed to be (α, β)-bounded, we have that

$$\frac{\phi(N^{\epsilon/\alpha}) \cdot N^{-\epsilon/\alpha} + \phi(N^{-\epsilon/\alpha})}{\phi(2)/2 + \phi(1/2)} \leq \frac{\beta \cdot N^\epsilon}{\phi(2)/2 + \phi(1/2)} \leq \beta' \cdot N^\epsilon, \tag{27}$$

where we have written $\beta' := \max\left\{1, \frac{\beta}{\phi(2)/2 + \phi(1/2)}\right\}$.

For some constant $C_0 > 0$ to be specified below, we choose

$$\delta_1 = \frac{C_0 \ln T}{(N-1)T}, \quad \beta_1 = \frac{\delta_1}{N^{\epsilon/\alpha}}; \qquad \delta_2 = \frac{\phi(N^{\epsilon/\alpha}) \cdot N^{-\epsilon/\alpha} + \phi(N^{-\epsilon/\alpha})}{\phi(2)/2 + \phi(1/2)} \cdot \delta_1, \quad \beta_2 = \frac{\delta_2}{2}. \tag{28}$$

The choices of $\delta_1, \beta_1, \delta_2, \beta_2$ ensure that

$$\beta_1 \cdot \phi(\delta_1/\beta_1) + \delta_1 \cdot \phi(\beta_1/\delta_1) = \beta_2 \cdot \phi(\delta_2/\beta_2) + \delta_2 \cdot \phi(\beta_2/\delta_2), \tag{29}$$

which, together with (26), ensures that for all $i,j \in [N]$, $D_\phi(M_i^{\delta_1,\beta_1} \parallel M_j^{\delta_1,\beta_1}) = D_\phi(M_i^{\delta_2,\beta_2} \parallel M_j^{\delta_2,\beta_2})$.

Wrapping up. We set $\mathscr{H}_1 = \mathscr{H}^{\delta_1,\beta_1}$ and $\mathscr{H}_2 = \mathscr{H}^{\delta_2,\beta_2}$, and correspondingly set $\mathcal{M}_1 = \mathcal{M}^{\delta_1,\beta_1}$ and $\mathcal{M}_2 = \mathcal{M}^{\delta_2,\beta_2}$. We define the one-to-one mapping $\mathscr{E} : \mathcal{M}_1 \to \mathcal{M}_2$ by the mapping that sends $M_i^{\delta_1,\beta_1} \mapsto M_i^{\delta_2,\beta_2}$ for all $i \in [N]$. It is clear that these definitions satisfy Item 1 and Item 2 of the proposition statement.

From (25), the expected risk of p against a worst-case model in $\mathcal{M}^{\delta_1,\beta_1}$ is bounded above as follows:

$$\sup_{M^\star \in \mathcal{M}^{\delta_1,\beta_1}} \mathbb{E}^{M^\star, p}[\mathbf{Risk}(T)] \leq \left(1 - \frac{C_0 \ln T}{T}\right)^T + \frac{\beta_1}{\delta_1 \cdot (N-1)}$$

$$\leq \exp\left(-\frac{C_0 \ln T}{T}\right)^T + \frac{2}{N^{1+\epsilon/\alpha}}$$

$$\leq T^{-1} + 2 \cdot \left(\frac{C_{\text{prob}}}{T}\right)^{1/2 + \epsilon/(2\alpha)},$$

where the final inequality holds as long as we choose $C_0 = 1$; recall that $N := \lceil \sqrt{T/C_{\text{prob}}} \rceil$. The above display establishes the upper bound of Item 3.

Next, for the lower bound, recall that (27) gives that $\delta_2 \leq \beta' \cdot N^\epsilon \cdot \delta_1$, so

$$\frac{1}{\delta_2^{1-\epsilon}} \geq \frac{1}{(\beta' N^\epsilon \cdot \delta_1)^{1-\epsilon}} \geq \frac{(N^{1-\epsilon}T)^{1-\epsilon}}{2\beta' C_0 \ln T} \geq \frac{N^{1-4\epsilon} \cdot T}{2\beta' C_0 C_{\text{prob}}^\epsilon \ln T} \geq \frac{T^{3/2-2\epsilon}}{2\beta' C_0 C_{\text{prob}}^{1/2+\epsilon} \ln T},$$

where the second-to-last inequality uses that $T^\epsilon \leq N^{2\epsilon} \cdot C_{\text{prob}}^\epsilon$. Thus, from Lemma J.1 with $(\delta, \beta) = (\delta_2, \beta_2)$ (so that $C = 2$), we have that for all $T' \leq \frac{T^{3/2-2\epsilon}}{2\beta' C_0 C_{\text{prob}}^{1/2+\epsilon} \ln T}$,

$$\mathfrak{M}(\mathscr{H}^{\delta_2,\beta_2}) \geq \frac{1}{2^{1+2/\epsilon}} \cdot \frac{1}{N} \geq \frac{1}{2^{2+2/\epsilon}} \cdot \sqrt{\frac{T}{C_{\text{prob}}}}.$$

Thus, taking $C_\phi = 2\beta' C_0$, the above inequality verifies the lower bound of Item 3.

∎

Proof of Lemma J.1. Consider any algorithm $p : \mathcal{O}^S \to \Pi$. Note that the distributions of $M_1^{\delta,\delta}, \ldots, M_N^{\delta,\delta}$ are all identical. Thus, there is some $i \in [N]$ so that

$$\mathbb{E}^{M_i^{\delta,\delta}} \mathbb{E}_{\pi \sim p(o_1,\ldots,o_S)}[\mathbb{I}\{\pi = i\}] = \mathbb{E}^{M_i^{\delta,\delta}} \mathbb{E}_{\pi \sim p(o_1,\ldots,o_S)}[g^i(\pi)] \geq 1/N.$$

For $\lambda \geq 0$, define

$$\mathcal{S}_\lambda := \left\{ (o_1, \ldots, o_S) \in \mathcal{O}^S \ : \ |\{t \in [S] \ : \ o_t = i\}| > \lambda \right\}.$$

Then the probability that $(o_1, \ldots, o_S) \in \mathcal{S}_\lambda$ is bounded above as follows:

$$\mathbb{P}^{M_i^{\delta,\delta}} ((o_1, \ldots, o_S) \in \mathcal{S}_\lambda) \leq \binom{S}{\lambda} \cdot \delta^\lambda \leq (S\delta)^\lambda \leq \delta^{\epsilon\lambda}.$$

Choosing $\lambda = 2/\epsilon$ yields $\delta^{\epsilon\lambda} = \delta^2 \leq 1/N^2$, meaning that

$$\mathbb{E}^{M_i^{\delta,\delta}} \mathbb{E}_{\pi \sim p(o_1,\ldots,o_S)} \left[\mathbb{I}\{(o_1, \ldots, o_S) \notin \mathcal{S}_{2/\epsilon}\} \cdot \mathbb{I}\{\pi = i\} \right] \geq 1/N - 1/N^2 \geq 1/(2N).$$

For any $(o_1, \ldots, o_S) \notin \mathcal{S}_\lambda$, we have that

$$\frac{M_i^{\delta,\beta}((o_1, \ldots, o_S))}{M_i^{\delta,\delta}((o_1, \ldots, o_S))} \geq (\beta/\delta)^\lambda \leq 1/C^\lambda.$$

Thus,

$$
\begin{aligned}
\mathbb{E}^{M_i^{\delta,\beta}} \mathbb{E}_{\pi \sim p(o_1,\ldots,o_S)}[g^i(\pi)] &\geq \mathbb{E}^{M_i^{\delta,\beta}} \mathbb{E}_{\pi \sim p(o_1,\ldots,o_S)}[\mathbb{I}\{\pi = i\} \cdot \mathbb{I}\{(o_1, \ldots, o_S) \notin \mathcal{S}_{2/\epsilon}\}] \\
&\geq \mathbb{E}^{M_i^{\delta,\delta}} \mathbb{E}_{\pi \sim p(o_1,\ldots,o_S)}[\mathbb{I}\{\pi = i\} \cdot \mathbb{I}\{(o_1, \ldots, o_S) \notin \mathcal{S}_{2/\epsilon}\}] \cdot 1/C^{2/\epsilon} \\
&\geq \frac{1}{2NC^{2/\epsilon}}.
\end{aligned}
$$

∎

J.3. Proofs from Appendix D.3

Proof of Theorem D.4. Given $C_{\text{prob}} \geq 1$, fix $N = \lceil \sqrt{T/C_{\text{prob}}} \rceil$. Recall the definition of the instances $\mathcal{H}^{\delta,\beta} = (\mathcal{M}^{\delta,\beta}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ (for $\delta, \beta \in (0,1)$) of the HR-DMSO framework defined in the proof of Theorem D.3, where we have $\Pi = [N]$ and $\mathcal{O} = [N] \cup \{\perp\}$. For each δ, β, we now define $\mathscr{M}^{\delta,\beta} = (\widetilde{\mathcal{M}}^{\delta,\beta}, \widetilde{\Pi}, \widetilde{\mathcal{O}}, \{\Pi'_k\}_k, \{U_k\}_k)$ to be the instance of the (2-player) MA-DMSO framework constructed given the instance $\mathcal{H}^{\delta,\beta}$ per the construction in the proof of Theorem C.2 with a value of V to be specified below. In particular, $\widetilde{\Pi}, \widetilde{\mathcal{O}}, \Pi'_k, U_k$ do not depend on δ, β. For clarity, we explicitly write out the definition of the components of $\mathscr{M}^{\delta,\beta}$ in terms of the components of $\mathcal{H}^{\delta,\beta}$:

- Define $\Sigma_1 = \Pi = [N]$ and $\Sigma_2 = \{0, 1, \ldots, V\}$, $\widetilde{\Pi}_k = \Delta(\Sigma_k)$ for $k \in \{1, 2\}$, and $\widetilde{\Pi} = \widetilde{\Pi}_1 \times \widetilde{\Pi}_2$.

- Define Π'_k, U_k for $k \in [2]$ so that $\mathscr{M}^{\delta,\beta}$ is an NE instance (Definition 1.1); in particular, $\Pi'_k = \widetilde{\Pi}_k$ for each k and $U_k(\pi'_k, \pi) = (\pi_k, \pi_{-k})$.

- Define the pure observation space to be $\mathcal{O}_\circ := \mathcal{O} \cup \{\widetilde{\bot}\}$, the reward space to be $\mathcal{R} = [-1, 1]$, and the full observation space to $\widetilde{\mathcal{O}} := \mathcal{O}_\circ \times \mathbb{R}^2$.

- The model class $\widetilde{\mathcal{M}}^{\delta,\beta}$ is indexed by tuples $(M, v) \in \mathcal{M}^{\delta,\beta} \times \{1, 2, \ldots, V\} = \mathcal{M}^{\delta,\beta} \times [V]$. (Thus $|\widetilde{\mathcal{M}}^{\delta,\beta}| = NV$.) In particular, for each such tuple (M, v), we have a model $\widetilde{M}_{M,v}$, which is defined as follows:

  - For pure decisions of the form $(\sigma_1, 0) \in \Sigma_1 \times \Sigma_2$ the distribution of $(o_\circ, r_1, r_2) \sim \widetilde{M}_{M,v}((\sigma_1, 0))$ is given by:

    $$o_\circ \sim M(\sigma_1) \in \mathcal{O} \subset \mathcal{O}_\circ, \qquad r_1 = r_2 = 0.$$

  - For pure decisions of the form $(\sigma_1, i) \in \Sigma_1 \times \Sigma_2$ with $i > 0$, the distribution of $(o_\circ, r_1, r_2) \sim \widetilde{M}_{M,v}((\sigma_1, i))$ is given by:

    $$o_\circ = \widetilde{\bot}, \qquad r_2 = -r_1 = \begin{cases} -1 & : i \neq v \\ g^M(\sigma_1) & : i = v, \end{cases} \tag{30}$$

    where we recall that $g^M(\sigma_1) = \max_{\sigma'_1 \in \Sigma_1}\{f^M(\sigma'_1)\} - f^M(\sigma_1)$.

  - For general decisions $\pi \in \widetilde{\Pi}$, we can write $\pi = \pi_1 \times \pi_2$ for $\pi_k \in \widetilde{\Pi}_k$ for $k \in [2]$. Then the distribution $\widetilde{M}_{M,v}(\pi)$ is the distribution of $\widetilde{M}_{M,v}(\sigma)$ where $\sigma = (\sigma_1, \sigma_2)$ is distributed as: $\sigma_k \sim \pi_k$ for $k \in [2]$.

Next, let δ_1, δ_2 be defined given $T, N, \epsilon, \phi, \alpha$, as in the proof of Theorem D.3 (in particular, they are specified in (28)). We write $\mathscr{M}_1 = \mathscr{M}^{\delta_1,\beta_1}$ and $\mathscr{M}_2 = \mathscr{M}^{\delta_2,\beta_2}$, and correspondingly write $\mathcal{M}_1 = \widetilde{\mathcal{M}}^{\delta_1,\beta_1}$ and $\mathcal{M}_2 = \widetilde{\mathcal{M}}^{\delta_2,\beta_2}$. Moreover, we define the mapping $\mathscr{E} : \mathcal{M}_1 \to \mathcal{M}_2$ in an analogous manner to the definition in the proof of Theorem D.3. In particular, for each δ, β, we have $\mathcal{M}^{\delta,\beta} = \{M_1^{\delta,\beta}, \ldots, M_N^{\delta,\eta}\}$. First define $\mathscr{E}_0 : \mathcal{M}^{\delta_1,\beta_1} \to \mathcal{M}^{\delta_2,\beta_2}$ by $\mathscr{E}_0(M_i^{\delta_1,\beta_1}) = M_i^{\delta_2,\beta_2}$, for $i \in [N]$ (exactly as was done in the proof of Theorem D.3. Then for each model of the form $\widetilde{M}_{M,v} \in \widetilde{M}^{\delta_1,\beta_1} = \mathcal{M}_1$ (so that $M \in \mathcal{M}^{\delta_1,\beta_1}, v \in [V]$), define $\mathscr{E}(\widetilde{M}_{M,v}) := \widetilde{M}_{\mathscr{E}_0(M),v}$. We are now ready to verify the individual claims of the theorem:

Proof of Item 1. Consider any $\widetilde{M}_{M,v} \in \widetilde{\mathcal{M}}^{\delta_1,\beta_1}$ (so that $M \in \mathcal{M}^{\delta_1,\beta_1}, v \in [V]$). For any $(\sigma_1, \sigma_2) \in \Sigma_1 \times \Sigma_2$, we have, by definition of $\mathscr{E}$,

$$f_2^{\widetilde{M}_{M,v}}(\sigma_1, \sigma_2) = f_2^{\mathscr{E}(\widetilde{M}_{M,v})}(\sigma_1, \sigma_2) = \begin{cases} 0 & : \sigma_2 = 0 \\ -1 & : \sigma_2 \in [V]\backslash\{v\} \\ g^M(\sigma_1) & : \sigma_2 = v, \end{cases}$$

which establishes Item 1 since all instances are 2-player 0-sum instances.

Proof of Item 2. Consider any two models $\widetilde{M}_{M,v}, \widetilde{M}_{M',v'} \in \widetilde{\mathcal{M}}^{\delta_1,\beta_1}$ (so that $M, M' \in \mathcal{M}^{\delta_1,\beta_1}$, and $v, v' \in [V]$). For any $\pi_1 \in \Pi_1$, we have that

$$D_\phi(\widetilde{M}_{M,v}(\pi_1, 0) \| \widetilde{M}_{M',v'}(\pi_1, 0)) = D_\phi(M(\pi_1) \| M'(\pi_1))$$
$$= D_\phi(\mathscr{E}_0(M)(\pi_1) \| \mathscr{E}_0(M')(\pi_1)) = D_\phi(\widetilde{M}_{\mathscr{E}_0(M),v}(\pi_1, 0) \| \widetilde{M}_{\mathscr{E}_0(M'),v'}(\pi_1, 0))$$
$$= D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1, 0) \| \mathscr{E}(\widetilde{M}_{M',v'})(\pi_1, 0)), \tag{31}$$

where the first and third equalities follow by definition of $\widetilde{\mathcal{M}}^{\delta,\beta}$ above, the second equality follows by Item 2 of Theorem D.3 and the fact that our choice of $\delta_1, \beta_1, \delta_2, \beta_2$ is identical to that in the proof of Theorem D.3 (cf. Eq. (28) and Eq. (29)), and the fourth equality follows from definition of $\mathscr{E}$.

Next, for any $\sigma_1 \in \Sigma_1$ and $\sigma_2 \in \Sigma_2 \backslash \{0\}$, note that the distributions $\widetilde{M}_{M,v}(\sigma_1, \sigma_2)$ and $\mathscr{E}(\widetilde{M}_{M,v})(\sigma_1, \sigma_2) = \widetilde{M}_{\mathscr{E}_0(M),v}(\sigma_1, \sigma_2)$ are identical: the pure observation under both these distributions is $\widetilde{\perp}$ a.s., and the rewards are given by (30), where we have noted that $g^M(\sigma_1) = g^{\mathscr{E}_0(M)}(\sigma_1)$ for all $\sigma_1 \in \Sigma_1$. It follows that for any $\pi_1 \in \Pi_1$ and $\sigma_2 \in \Delta(\Sigma_2 \backslash \{0\})$, the distributions $\widetilde{M}_{M,v}(\pi_1, \pi_2)$ and $\mathscr{E}(\widetilde{M}_{M,v})(\pi_1, \pi_2)$ are identical. In a similar manner, we have that for any such π_1, π_2, the distributions $\widetilde{M}_{M',v'}(\pi_1, \pi_2)$ and $\mathscr{E}(\widetilde{M}_{M',v'})(\pi_1, \pi_2)$ are identical. Therefore,

$$D_\phi(\widetilde{M}_{M,v}(\pi_1, \pi_2) \| \widetilde{M}_{M',v'}(\pi_1, \pi_2)) = D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1, \pi_2) \| \mathscr{E}(\widetilde{M}_{M',v'})(\pi_1, \pi_2)). \tag{32}$$

Now consider any joint decision $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$. Let us write $\pi_2 = \pi_2(0) \cdot \mathbb{I}_0 + (1 - \pi_2(0)) \cdot \pi_2'$, where $\pi_2' \in \Delta(\Sigma_2 \backslash \{0\})$. Since, for any model $\widetilde{M} \in \widetilde{\mathcal{M}}^{\delta,\beta}$ (for any δ, β), the distributions $\widetilde{M}(\pi_1, 0)$ and $\widetilde{M}(\pi_1, \pi_2')$ have disjoint support (namely, under the second, the pure observation is always $\widetilde{\perp}$, and under the first, the pure observation is never $\widetilde{\perp}$), it follows from Lemma H.1 that for any two models $\widetilde{M}, \widetilde{M}' \in \mathcal{M}^{\delta,\beta}$,

$$D_\phi(\widetilde{M}(\pi_1, \pi_2) \| \widetilde{M}'(\pi_1, \pi_2)) = \pi_2(0) \cdot D_\phi(\widetilde{M}(\pi_1, 0) \| \widetilde{M}'(\pi_1, 0))$$
$$+ (1 - \pi_2(0)) \cdot D_\phi(\widetilde{M}(\pi_1, \pi_2') \| \widetilde{M}'(\pi_1, \pi_2')). \tag{33}$$

Then for the decision $(\pi_1, \pi_2) \in \Pi_1 \times \Pi_2$, with π_2' defined as above, we have

$$D_\phi(\widetilde{M}_{M,v}(\pi_1, \pi_2) \| \widetilde{M}_{M',v'}(\pi_1, \pi_2))$$
$$= \pi_2(0) \cdot D_\phi(\widetilde{M}_{M,v}(\pi_1, 0) \| \widetilde{M}_{M',v'}(\pi_1, 0)) + (1 - \pi_2(0)) \cdot D_\phi(\widetilde{M}_{M,v}(\pi_1, \pi_2') \| \widetilde{M}_{M',v'}(\pi_1, \pi_2'))$$
$$= \pi_2(0) \cdot D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1, 0) \| \mathscr{E}(\widetilde{M}_{M',v'})(\pi_1, 0)) + (1 - \pi_2(0)) \cdot D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1, \pi_2') \| \mathscr{E}(\widetilde{M}_{M',v'})(\pi_1, \pi_2'))$$
$$= D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1, \pi_2) \| \mathscr{E}(\widetilde{M}_{M',v'})(\pi_1, \pi_2)),$$

where the first and third equalities use (33), and the second equality uses (31) and (32). The above display verifies Item 2.

Proof of Item 3. For each $\delta, \beta \in (0, 1)$, the construction of $\mathscr{M}^{\delta,\beta}$ given $\mathscr{H}^{\delta,\beta}$ according to the construction in the proof of Theorem C.2, together with the conclusion of Theorem C.2, gives that, for all $T' \in \mathbb{N}$,

$$\mathfrak{M}(\mathscr{M}^{\delta,\beta}, T') \leq \mathfrak{M}(\mathscr{H}^{\delta,\beta}, T') \leq \mathfrak{M}(\mathscr{M}^{\delta,\beta}, T') + O((T' \log(T')/V)^{1/4}). \tag{34}$$

Then Item 3 of Theorem D.3, together with our choice of $\delta_1, \beta_1, \delta_2, \beta_2$ to mimic that in the proof of Theorem D.3, yields that for all T' with $T \le T' \le T^{3/2-2\epsilon} \cdot (C_\phi C_{\mathrm{prob}}^{1/2+\epsilon} \ln T)^{-1}$

$$
\begin{aligned}
\mathfrak{M}(\mathscr{M}_1, T') = \mathfrak{M}(\mathscr{M}^{\delta_1,\beta_1}, T') \le {}& \mathfrak{M}(\mathscr{H}^{\delta_1,\beta_1}, T') \le \frac{1}{T} + 2 \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2+\epsilon/(2\alpha)} \qquad (35)\\
\mathfrak{M}(\mathscr{M}_2, T') = \mathfrak{M}(\mathscr{M}^{\delta_2,\beta_2}, T') \ge {}& \mathfrak{M}(\mathscr{H}^{\delta_2,\beta_2}, T') - O((T'\log(T')/V)^{1/4}) \\
\ge {}& 2^{-2-2/\epsilon} \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2} - O((T'\log(T')/V)^{1/4}).
\end{aligned}
$$

Choosing $V = T^{100} \cdot 2^{8+8/\epsilon}$ ensures that

$$
\mathfrak{M}(\mathscr{M}_2, T') \ge 2^{-3-2/\epsilon} \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2}. \qquad (36)
$$

Together (35) and (36) verify Item 3.

∎

Appendix K. Proofs for Appendix E

Throughout this section, we consider an instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ of MA-DMSO which is an NE instance (Definition 1.1). It follows in particular that for any $M \in \mathcal{M}, \pi \in \Pi$, we have

$$
h^M(\pi) = \sum_{k=1}^K h_k^M(\pi) = \sum_{k=1}^K \sup_{\pi'_k \in \Pi_k} f_k^M(\pi'_k, \pi_{-k}) - f_k^M(\pi).
$$

K.1. Bounds for general games with convex decision spaces

Proof of Theorem E.1. For each $k \in [K]$ and $\pi_{-k} \in \Pi_{-k}$, define

$$
\widetilde{\mathcal{M}}_k(\pi_{-k}) := \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ M \in \mathcal{M}\}.
$$

It is straightforward from the definition of $\widetilde{\mathcal{M}}_k$ in (11) that for each $k \in [K]$, $\widetilde{\mathcal{M}}_k = \bigcup_{\pi_{-k} \in \Pi_{-k}} \widetilde{\mathcal{M}}_k(\pi_{-k})$, and therefore that $\bigcup_{\pi_{-k} \in \Pi_{-k}} \mathrm{co}(\widetilde{\mathcal{M}}_k(\pi_{-k})) \subseteq \mathrm{co}(\widetilde{\mathcal{M}}_k)$. For any $\overline{\pi}_{-k} \in \Pi_{-k}$ and $\overline{M} \in \mathrm{co}(\mathcal{M})$, we denote the corresponding element of $\mathrm{co}(\widetilde{\mathcal{M}}_k(\overline{\pi}_{-k}))$ by $(\overline{M}, \overline{\pi}_{-k})$. (In particular, $(\overline{M}, \overline{\pi}_{-k})$ is the model that sends $\pi_k \mapsto \overline{M}|_k(\pi_k, \overline{\pi}_{-k})$.) It then suffices to prove the following stronger result: for each $\overline{M} \in \mathrm{co}(\mathcal{M})$,

$$
\mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathscr{M}, \overline{M}) \le \sum_{k=1}^K \sup_{\overline{\pi}_{-k} \in \Pi_{-k}} \mathsf{r\text{-}dec}_{\gamma/K}^{\mathrm{o}}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})). \qquad (37)
$$

Next, note that for any $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$, the value function for the model $\pi_k \mapsto M|_k(\pi_k, \pi_{-k})$ is given by $f^{M|_k}(\pi_k) = f_k^M(\pi_k, \pi_{-k})$, for $\pi \in \Pi$ (this holds since the distribution of the reward

under $M|_k(\pi_k, \pi_{-k})$ is simply the distribution of agent k's reward under $M(\pi_k, \pi_{-k})$). Then for any $\overline{M} \in \mathrm{co}(\mathcal{M}), \overline{\pi}_{-k} \in \Pi_{-k}$, we have

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k}))$$

$$= \inf_{p_k \in \Delta(\Pi_k)} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{\pi_k \sim p_k} \left[\max_{\pi'_k \in \Pi_k} f_k^M(\pi'_k, \pi_{-k}) - f_k^M(\pi_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right]$$

$$\geq \inf_{p_k \in \Delta(\Pi_k)} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{\substack{\pi_k \sim p_k \\ a_k \sim \pi_k}} \left[\max_{\pi'_k \in \Pi_k} f_k^M(\pi'_k, \pi_{-k}) - f_k^M(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right]$$

$$= \inf_{\pi_k \in \Pi_k} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{a_k \sim \pi_k} \left[\max_{\pi'_k \in \Pi_k} f_k^M(\pi'_k, \pi_{-k}) - f_k^M(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right],$$

$$(38)$$

where the inequality uses joint convexity of the squared Hellinger distance, and the final inequality uses the fact that any distribution $p_k \in \Delta(\Pi_k)$ may be replaced by the singleton distribution for the decision $\widetilde{\pi}_k := \mathbb{E}_{\pi_k \sim p_k}[\pi_k]$, without changing the value of the expression.

Thus

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})) \geq \inf_{\pi_k \in \Pi_k} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{a_k \sim \pi_k} \left[h_k^M(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right].$$

Existence of fixed points. For each $k \in [K]$, define the set-valued function $\mathcal{C}_k : \Pi \to \mathcal{P}(\Pi_k)$ by

$$\mathcal{C}_k(\overline{\pi}) := \arg\min_{\pi_k \in \Pi_k} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{a_k \sim \pi_k} \left[h_k^M(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right].$$

Further, for $\pi_{-k} \in \Pi_{-k}, M \in \mathcal{M}$, define the function $G_{M, \pi_{-k}} : \Pi_k \times \Pi_{-k} \to \mathbb{R}$ by

$$G_{M, \pi_{-k}}(\pi_k, \overline{\pi}_{-k}) = \mathbb{E}_{a_k \sim \pi_k} \left[h_k^M(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right].$$

Assumption E.1 gives that for all a_k, the map $\overline{\pi}_{-k} \mapsto \overline{M}(a_k, \overline{\pi}_{-k})$ is linear. It follows by the dominated convergence theorem that for all M, π_{-k}, a_k, the function $\overline{\pi}_{-k} \mapsto D_\mathsf{H}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big)$ is continuous. Hence $G_{M, \pi_{-k}}(\pi_k, \overline{\pi}_{-k})$ is continuous in $(\pi_k, \overline{\pi}_{-k})$, and the function

$$\widetilde{G}_k(\pi_k, \overline{\pi}_{-k}) := \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} G_{M, \pi_{-k}}(\pi_k, \overline{\pi}_{-k})$$

is also continuous in $(\pi_k, \overline{\pi}_{-k})$. Furthermore, since, for each $\overline{\pi}_{-k}$, the function $G_{M, \pi_{-k}}(\pi_k, \overline{\pi}_{-k})$ is linear in π_k (Assumption E.1), $\widetilde{G}_k(\pi_k, \overline{\pi}_{-k})$ is convex in π_k. It follows that $\mathcal{C}_k(\overline{\pi}) = \arg\min_{\pi_k \in \Pi_k}\{\widetilde{G}_k(\pi_k, \overline{\pi}_{-k})\}$ is a closed, nonempty, and convex subset of Π_k for all $\overline{\pi}$. Furthermore, by continuity of $\widetilde{G}_k$ and Lemma H.8, we have that $\mathcal{C}_k(\overline{\pi})$ is upper hemicontinuous. By Lemma H.9, it follows that the mapping $\overline{\pi} \mapsto \mathcal{C}_1(\overline{\pi}) \times \cdots \times \mathcal{C}_K(\overline{\pi})$ has a fixed point, namely some $\overline{\pi} \in \Pi$ so that $\overline{\pi} \in \prod_{k \in [K]} \mathcal{C}_k(\overline{\pi})$.

Applying the fixed point strategy. Let $\overline{\pi} \in \prod_{k \in [K]} \mathcal{C}_k(\overline{\pi})$ be a fixed point of $\mathcal{C}_1 \times \cdots \times \mathcal{C}_K$. Then

$$
\begin{aligned}
\text{r-dec}_\gamma^\circ(\mathscr{M}, \overline{M}) &\leq \sup_{M \in \mathcal{M}} \left\{ h^M(\overline{\pi}) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big) \right\} \\
&= \sup_{M \in \mathcal{M}} \left\{ \sum_{k=1}^K h_k^M(\overline{\pi}) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big) \right\} \\
&\leq \sum_{k=1}^K \sup_{M \in \mathcal{M}} \left\{ h_k^M(\overline{\pi}) - \frac{\gamma}{K} \cdot D_{\mathsf{H}}^2\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big) \right\} \\
&\leq \sum_{k=1}^K \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \left\{ h_k^M(\overline{\pi}_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_{\mathsf{H}}^2\big(M(\overline{\pi}_k, \pi_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big) \right\} \\
&= \sum_{k=1}^K \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{a_k \sim \overline{\pi}_k} \left[h_k^M(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_{\mathsf{H}}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right] \\
&= \sum_{k=1}^K \inf_{\pi_k \in \Pi_k} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{a_k \sim \pi_k} \left[h_k^M(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_{\mathsf{H}}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right] \\
&\leq \sum_{k=1}^K \text{dec}_{\gamma/K}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})).
\end{aligned}
$$

Above, we have used the following facts:

1. The second equality uses Assumption E.1 to conclude that for all $a_k, \pi_k, \overline{\pi}_{-k}, M, \overline{M}$,

$$
\mathbb{P}_{o \sim M(a_k, \pi_{-k})}(\varphi(o) = a_k) = 1, \quad \mathbb{P}_{o \sim \overline{M}(a_k, \overline{\pi}_{-k})}(\varphi(o) = a_k) = 1,
$$

   thus allowing us to apply Lemma H.1 to give that

$$
D_{\mathsf{H}}^2\big(M(\overline{\pi}_k, \pi_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big) = \mathbb{E}_{a_k \sim \overline{\pi}_k} \left[D_{\mathsf{H}}^2\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right].
$$

2. The third equality follows from the fact that $\overline{\pi}_k \in \mathcal{C}_k(\overline{\pi})$ for all $k \in [K]$.

3. The final inequality follows from (38).

∎

K.2. Bounds for Markov games

Here, we prove Theorem E.2. The proof uses a number of technical lemmas which are stated and proven in the sequel.

Proof of Theorem E.2. As in the proof of Theorem E.1, for each $k \in [K]$ and $\pi_{-k} \in \Pi_{-k}$, we define

$$
\widetilde{\mathcal{M}}_k(\pi_{-k}) := \{ \pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ M \in \mathcal{M} \}.
$$

For any $\overline{\pi}_{-k} \in \Pi_{-k}$ and $\overline{M} \in \mathcal{M}$, we denote the corresponding element of $\widetilde{\mathcal{M}}_k(\overline{\pi}_{-k}) \subseteq \widetilde{\mathcal{M}}_k$ by $(\overline{M}, \overline{\pi}_{-k})$. We will prove the following stronger result: there is some constant $C' > 0$ so that for each $\overline{M} \in \mathcal{M}$,

$$\mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma}(\mathscr{M}, \overline{M}) \leq \frac{C'KH\log H}{\gamma} + \sum_{k=1}^{K} \sup_{\overline{\pi}_{-k} \in \Pi_{-k}} \mathsf{r\text{-}dec}^{\mathsf{o}}_{\gamma/(C'KH\log H)}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})). \quad (39)$$

Fix any $\epsilon > 0$. For each $k \in [K]$, let Π_k^{ϵ} be a finite ϵ-cover of Π_k in the sense that for all $\pi_k \in \Pi_k$, there is some element $\pi_k^{\epsilon} \in \Pi_k^{\epsilon}$ so that, for all $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$,

$$D_{\mathsf{H}}^2(M(\pi_k, \pi_{-k}), M(\pi_k^{\epsilon}, \pi_{-k})) \leq \epsilon^2.$$

Furthermore, we require that the mapping $\pi_k \mapsto \pi_k^{\epsilon}$ is measurable with respect to the Borel σ-algebra on Π_k. By finiteness of $\mathcal{S}, \mathcal{A}_k$, it is straightforward to see that such a finite cover Π_k^{ϵ} exists. The size of the cover Π_k^{ϵ} may depend on $|\mathcal{S}|, |\mathcal{A}|$, but this will not matter as $|\Pi_k^{\epsilon}|$ will not enter into our final bounds. (We introduce discretization here only to ensure that Π_k^{ϵ} is compact when applying Lemma H.8.)

We collect a few basic properties of Π_k^{ϵ} in the below lemma, proved at the end of the section:

Lemma K.1. *For any $\pi_k \in \Pi_k$, there is some $\pi_k^{\epsilon} \in \Pi_k^{\epsilon}$ so that the following holds. For any $M, \overline{M} \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$,*

$$D_{\mathsf{H}}^2\big(M(\pi_k^{\epsilon}, \pi_{-k}), \overline{M}(\pi_k^{\epsilon}, \overline{\pi}_{-k})\big) \geq \frac{1}{3} \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) - 2\epsilon^2$$

$$|h_k^M(\pi_k, \pi_{-k}) - h_k^M(\pi_k^{\epsilon}, \pi_{-k})| \leq \epsilon.$$

Existence of fixed points. Let $C > 0$ be the constant of Lemma K.4, and write $\gamma' = \gamma/(CKH\log H)$. For each $k \in [K]$, define the function $\mathcal{C}_k : \Pi \to \Delta(\Pi_k^{\epsilon})$ by

$$\mathcal{C}_k(\overline{\pi}) = \underset{p_k \in \Delta(\Pi_k^{\epsilon})}{\arg\min} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p_k}\big[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)\big] + \epsilon \cdot \|p_k\|_2^2,$$

where $\|p_k\|_2^2$ denotes the squared ℓ_2 norm of p_k, interpreted as a vector in the Euclidean space $\mathbb{R}^{|\Pi_k^{\epsilon}|}$.

Further, for $\pi_{-k} \in \Pi_{-k}, M \in \mathcal{M}$, define the function $G_{M, \pi_{-k}} : \Delta(\Pi_k) \times \Pi_{-k} \to \mathbb{R}$ by

$$G_{M, \pi_{-k}}(p_k, \overline{\pi}_{-k}) = \mathbb{E}_{\pi_k \sim p_k}\big[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)\big].$$

We may view $\overline{\pi}_{-k}$ as an element of $\Delta(\mathcal{A}_k)^{\mathcal{S} \times [H]}$, which is a subset of Euclidean space (since $\mathcal{A}_k, \mathcal{S}$ are assumed to be finite). Since there are finitely many states and actions, it follows from the dominated convergence theorem that for all M, π_k, π_{-k}, the function $\overline{\pi}_{-k} \mapsto D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)$ is continuous. Hence $G_{M, \pi_{-k}}(p_k, \overline{\pi}_{-k})$ is continuous in $(p_k, \overline{\pi}_{-k})$. Hence the function

$$\widetilde{G}_k(p_k, \overline{\pi}_{-k}) := \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} G_{M, \pi_{-k}}(p_k, \overline{\pi}_{-k}) + \epsilon \cdot \|p_k\|_2^2$$

is also continuous. Furthermore, $G_{M, \pi_{-k}}(p_k, \overline{\pi}_{-k})$ is linear in p_k (for fixed $\overline{\pi}_{-k}$), so $\widetilde{G}_k(p_k, \overline{\pi}_{-k})$ is strongly convex in p_k (for fixed $\overline{\pi}_{-k}$). Thus $\mathcal{C}_k(\overline{\pi}) = \arg\min_{p_k \in \Delta(\Pi_k^{\epsilon})}\{\widetilde{G}_k(p_k, \overline{\pi}_{-k})\}$ is a singleton for all $\overline{\pi}$. Furthermore, by continuity of $\widetilde{G}_k$, compactness of $\Delta(\Pi_k^{\epsilon})$ and Π_{-k}, and Lemma H.8,

we have that $\mathcal{C}_k(\overline{\pi})$ is upper hemicontinuous, which means, by single-valuedness, it is actually continuous.

Given $\overline{M} \in \mathcal{M}$, $\overline{\pi}_{-k} \in \Pi_{-k}$, note that the pure observation distribution of the model $\pi_k \mapsto \overline{M}(\pi_k, \overline{\pi}_{-k})$ is exactly that of an MDP, which we denote by $\overline{M}_{\overline{\pi}_{-k}}$: it has horizon H, state space $\mathcal{S}$, action space $\mathcal{A}_k$, and rewards and transitions given by those of $\overline{M}$ when each agent $k' \neq k$ acts according to $\overline{\pi}_{k', h}(\cdot | s)$ at each state s and step h (to be precise, the rewards of $\overline{M}_{\overline{\pi}_{-k}}$ are given by the rewards of agent k in $\overline{M}$). Note that the space of randomized nonstationary policies of $\overline{M}_{\overline{\pi}_{-k}}$ is Π_k (using Assumption E.2).

Since we do not assume convexity of Π_k, elements $p_k \in \Delta(\Pi_k^\epsilon)$ may not belong to Π_k. We next introduce a set of decisions in Π_k which are "equivalent" to p_k given a reference model $\overline{M}$ and a reference decision $\overline{\pi}_{-k}$. In particular, for $\overline{M} \in \mathcal{M}, \overline{\pi}_{-k} \in \Pi_k$, and $p_k \in \Delta(\Pi_k^\epsilon)$, let $\Pi^\star_{\overline{M}, \overline{\pi}_{-k}}(p_k) \subset \Pi_k$ be the set of all policies $\pi_k^\star \in \Pi_k$ which satisfy Eq. (44) of Lemma K.2 for p_k and $\pi_k \mapsto \overline{M}(\pi_k, \overline{\pi}_{-k})$. Note that $\Pi^\star_{\overline{M}, \overline{\pi}_{-k}}(p_k)$ is a nonempty convex set: as a subset of $\Delta(\mathcal{A}_k)^{\mathcal{S} \times [H]}$, it is a product of sets (one for each factor of $\Delta(\mathcal{A}_k)$), each of which is either a singleton or all of $\Delta(\mathcal{A}_k)$. It is straightforward from the definition that the map $(p_k, \overline{\pi}_{-k}) \mapsto \Pi^\star_{\overline{M}, \overline{\pi}_{-k}}(p_k)$ is upper hemicontinuous. Then Lemma K.4 gives that, for any $\overline{M}$ and $\overline{\pi}_{-k}$ and p_k, if $\pi_k^\star \in \Pi^\star_{\overline{M}, \overline{\pi}_{-k}}(p_k)$ is the corresponding policy in (44), then for $\gamma > 0$,

$$
\sup_{M \in \mathcal{M}, \, \pi_{-k} \in \Pi_{-k}} \left\{ h_k^M(\pi_k^\star, \pi_{-k}) - \frac{\gamma}{K} \cdot D_{\mathsf{H}}^2\big(M(\pi_k^\star, \pi_{-k}), \overline{M}(\pi_k^\star, \overline{\pi}_{-k})\big) \right\}
$$
$$
\leq \frac{1}{\gamma'} + \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p_k} \left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right]. \tag{40}
$$

Since the mapping $\overline{\pi} \mapsto \mathcal{C}_k(\overline{\pi}) \in \Delta(\Pi_k^\epsilon)$ is continuous, the composition $\mathcal{C}_k^\star(\overline{\pi}) := \Pi^\star_{\overline{M}, \overline{\pi}_{-k}}(\mathcal{C}_k(\overline{\pi}))$ is upper hemicontinuous. Thus, by Kakutani's fixed point theorem (Osborne and Rubinstein, 1994, Lemma 20.1), the set-valued mapping $C^\star(\overline{\pi}) := \mathcal{C}_1^\star(\overline{\pi}) \times \cdots \times \mathcal{C}_K^\star(\overline{\pi})$ has a fixed point.

Applying the fixed point strategy. Let $\overline{\pi} \in \Pi$ be a fixed point for $C^\star$, so that $\overline{\pi}_k \in \mathcal{C}_k^\star(\overline{\pi})$ for each $k \in [K]$. Then

$$
\mathsf{r\text{-}dec}_\gamma^\mathsf{o}(\mathscr{M}, \overline{M})
$$

$$
\leq \sup_{M \in \mathcal{M}} \left\{ \sum_{k=1}^K h_k^M(\overline{\pi}) - \gamma \cdot D_\mathsf{H}^2\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big) \right\}
$$

$$
\leq \sum_{k=1}^K \sup_{M \in \mathcal{M}} \left\{ h_k^M(\overline{\pi}_k, \overline{\pi}_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(\overline{\pi}_k, \overline{\pi}_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big) \right\}
$$

$$
\leq \sum_{k=1}^K \sup_{M \in \mathcal{M}, \, \pi_{-k} \in \Pi_{-k}} \left\{ h_k^M(\overline{\pi}_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D_\mathsf{H}^2\big(M(\overline{\pi}_k, \pi_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big) \right\}
$$

$$
\leq \frac{1}{\gamma'} + \sum_{k=1}^K \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim \mathcal{C}_k(\overline{\pi})} \left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_\mathsf{H}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right] \quad (41)
$$

$$
\leq \frac{1}{\gamma'} + \sum_{k=1}^K \epsilon + \inf_{p_k \in \Delta(\Pi_k^\epsilon)} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p_k} \left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_\mathsf{H}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right]
$$

$$
(42)
$$

$$
\leq \frac{1}{\gamma'} + \sum_{k=1}^K 2\epsilon + \gamma' \cdot 2\epsilon^2 + \inf_{p_k \in \Delta(\Pi_k)} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p_k} \left[h_k^M(\pi_k, \pi_{-k}) - \frac{\gamma'}{3} \cdot D_\mathsf{H}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right].
$$

$$
(43)
$$

where (41) uses Eq. (40) and the fact that $\overline{\pi}_k \in \mathcal{C}_k^\star(\overline{\pi}) = \Pi_{\overline{M}, \overline{\pi}_{-k}}^\star(\mathcal{C}_k(\overline{\pi}))$ for each k, and (42) uses the definition of $\mathcal{C}_k(\overline{\pi})$. Finally, (43) uses Lemma K.1, as follows: given any distribution $p_k \in \Delta(\Pi_k)$, we consider the distribution $p_k^\epsilon \in \Delta(\Pi_k^\epsilon)$ which is given by pushing forward p_k through the map $\pi_k \mapsto \pi_k^\epsilon$ (here we use that $\pi_k \mapsto \pi_k^\epsilon$ is measurable to ensure that $p_k^e p$ is well-defined). Then by Lemma K.1, for all $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}, \overline{\pi}_{-k} \in \Pi_{-k}$, we have

$$
\mathbb{E}_{\pi_k \sim p_k^\epsilon}[h_k^M(\pi_k, \pi_{-k})] \leq \mathbb{E}_{\pi_k \sim p_k}[h_k^M(\pi_k, \pi_{-k})] + \epsilon
$$

$$
-\gamma' \cdot \mathbb{E}_{\pi_k \sim p_k^\epsilon}[D_\mathsf{H}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)] \leq -\frac{\gamma'}{3} \cdot \mathbb{E}_{\pi_k \sim p_k}[D_\mathsf{H}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)] + \gamma' \cdot 2\epsilon^2.
$$

By taking $\epsilon \to 0$, we obtain that, for some constant $C > 0$,

$$
\mathsf{r\text{-}dec}_\gamma^\mathsf{o}(\mathscr{M}, \overline{M}) \leq \frac{CKH \log H}{\gamma} + \sum_{k=1}^K \mathsf{r\text{-}dec}_{\gamma/(CKH \log H)}^\mathsf{o}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k}))
$$

$$
\leq \frac{CKH \log H}{\gamma} + \sum_{k=1}^K \sup_{\widetilde{\pi}_{-k} \in \Pi_{-k}} \mathsf{r\text{-}dec}_{\gamma/(CKH \log H)}^\mathsf{o}(\widetilde{\mathcal{M}}_k, (\overline{M}, \widetilde{\pi}_{-k})),
$$

thus verifying (39). ∎

K.2.1. SUPPORTING LEMMAS

Proof of Lemma K.1. To establish the first property, we use the definition of Π_k^ϵ and the triangle inequality for Hellinger distance to conclude that

$$
\begin{aligned}
&D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \\
&\leq 3 \cdot \big(D_{\mathsf{H}}^2(M(\pi_k^\epsilon, \pi_{-k}), M(\pi_k, \pi_{-k})) + D_{\mathsf{H}}^2\big(\overline{M}(\pi_k^\epsilon, \overline{\pi}_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) + D_{\mathsf{H}}^2\big(M(\pi_k^\epsilon, \pi_{-k}), \overline{M}(\pi_k^\epsilon, \overline{\pi}_{-k})\big)\big) \\
&\leq 3 \cdot \big(D_{\mathsf{H}}^2(M(\pi_k^\epsilon, \pi_{-k}), M(\pi_k, \pi_{-k})) + 2\epsilon^2\big),
\end{aligned}
$$

and rearranging gives the first claimed inequality of the lemma.

To prove the second inequality, we note that for each $\pi_k \in \Pi_k$, the cover element $\pi_k^\epsilon \in \Pi_k^\epsilon$ satisfies the following: for all $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$

$$
|h_k^M(\pi_k, \pi_{-k}) - h_k^M(\pi_k^\epsilon, \pi_{-k})| = |f_k^M(\pi_k, \pi_{-k}) - f_k^M(\pi_k^\epsilon, \pi_{-k})| \leq D_{\mathsf{H}}(M(\pi_k, \pi_{-k}), M(\pi_k^\epsilon, \pi_{-k})) \leq \epsilon.
$$

∎

The following lemma shows that for any MDP $\overline{M}$ and distribution $p \in \Delta(\Pi_{\mathrm{RNS}})$, there exists a corresponding randomized policy in Π_{RNS} which induces identical occupancies in $\overline{M}$.

Lemma K.2. *Consider any finite-horizon MDP $\overline{M} = (\mathcal{S}, H, \mathcal{A}, P, R, \mu)$ with finite state and action spaces $\mathcal{S}, \mathcal{A}$. Let Π_{RNS} denote the set of randomized nonstationary policies of $\overline{M}$. Suppose $p \in \Delta(\Pi_{\mathrm{RNS}})$ is a distribution over Π_{RNS} with finite support. Consider any policy $\pi^\star \in \Pi_{\mathrm{RNS}}$ so that:*

$$
\forall a \in \mathcal{A}, \ s \in \mathcal{S} \text{ s.t. } \sum_{\pi' \in \Pi_{\mathrm{RNS}}} p(\pi') \cdot d_h^{\overline{M}, \pi'}(s) > 0: \quad \pi_h^\star(a|s) = \sum_{\pi \in \Pi_{\mathrm{RNS}}: \, p(\pi) > 0} \frac{p(\pi) \cdot d_h^{\overline{M}, \pi}(s)}{\sum_{\pi' \in \Pi_{\mathrm{RNS}}} p(\pi') \cdot d_h^{\overline{M}, \pi'}(s)} \cdot \pi_h(a|s).
$$

$$(44)$$

Then for all states $s \in \mathcal{S}$, $d_h^{\overline{M}, \pi^\star}(s) = \sum_{\pi \in \Pi_{\mathrm{RNS}}} p(\pi) \cdot d_h^{\overline{M}, \pi}(s)$, and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $d_h^{\overline{M}, \pi^\star}(s, a) = \sum_{\pi \in \Pi_{\mathrm{RNS}}} p(\pi) \cdot d_h^{\overline{M}, \pi}(s, a)$.

As a consequence, it follows that $V_1^{\overline{M}, \pi^\star} = \sum_{\pi \in \Pi_{\mathrm{RNS}}} p(\pi) \cdot V_1^{\overline{M}, \pi}$.

Proof of Lemma K.2. We drop the superscript $\overline{M}$ in all relevant quantities throughout the proof. We use induction on h, noting that the base case $h = 1$ is immediate since d_1^π is identical for all $\pi \in \Pi_{\mathrm{RNS}}$. Fix $p \in \Delta(\Pi_{\mathrm{RNS}})$, and let $\pi^\star$ be chosen as in Eq. (44). Assuming that the statement of

the lemma holds at step $h-1$, we compute

$$
\begin{aligned}
d_h^{\pi^\star}(s) &= \sum_{\substack{s',a': \\ d_{h-1}^{\pi^\star}(s')>0}} d_{h-1}^{\pi^\star}(s') \cdot \pi_{h-1}^\star(a'|s') \cdot P_{h-1}(s|s',a') \\
&= \sum_{\substack{s',a': \\ d_{h-1}^{\pi^\star}(s')>0}} \left(\sum_{\pi'} p(\pi') \cdot d_{h-1}^{\pi'}(s') \right) \cdot \sum_\pi \frac{p(\pi) \cdot d_{h-1}^\pi(s')}{\sum_{\pi'} p(\pi') \cdot d_{h-1}^{\pi'}(s)} \cdot \pi_{h-1}(a'|s') \cdot P_{h-1}(s|s',a') \\
&= \sum_\pi p(\pi) \cdot \sum_{\substack{s',a': \\ d_{h-1}^{\pi^\star}(s')>0}} d_{h-1}^\pi(s') \cdot \pi_{h-1}(a'|s') \cdot P_{h-1}(s|s',a') \\
&= \sum_\pi p(\pi) \cdot \sum_{s',a'} d_{h-1}^\pi(s') \cdot \pi_{h-1}(a'|s') \cdot P_{h-1}(s|s',a') \\
&= \sum_\pi p(\pi) \cdot d_h^\pi(s),
\end{aligned}
$$

where the second-to-last inequality follows since if $d_{h-1}^{\pi^\star}(s') = 0$, then (using the inductive hypothesis) for all π, $p(\pi) \cdot d_{h-1}^\pi(s') = 0$. The above chain of equalities then completes the inductive step. It then follows immediately from the definition of $\pi^\star$ that $d_h^{\pi^\star}(s,a) = \sum_{\pi \in \Pi_{\mathrm{RNS}}} p(\pi) \cdot d_h^\pi(s,a)$.

The final statement regarding the value functions follows since, for all policies π,

$$
V_1^\pi = \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^\pi(s,a) \cdot r_h(s,a).
$$

∎

The remaining lemmas establish certain technical properties for the policy $\pi^\star \in \Pi_{\mathrm{RNS}}$ constructed in Lemma K.2.

Lemma K.3. *There is a constant $C > 0$ so that the following holds. Consider any finite-horizon MDP $\overline{M} = (\mathcal{S}, H, \mathcal{A}, P^{\overline{M}}, R^{\overline{M}}, \mu^{\overline{M}})$ with finite state and action spaces $\mathcal{S}, \mathcal{A}$. Let Π_{RNS} denote the set of randomized nonstationary policies of $\overline{M}$, and let $p \in \Delta(\Pi_{\mathrm{RNS}})$ be a distribution of finite support. Consider any policy $\pi^\star \in \Pi_{\mathrm{RNS}}$ satisfying Eq. (44) for p. Then for any MDP $\mathcal{M} = (\mathcal{S}, H, \mathcal{A}, P^M, R^M, \mu^M)$,*

$$
\mathbb{E}_{\pi \sim p} \left[D_{\mathsf{H}}^2 \big(M(\pi), \overline{M}(\pi) \big) \right] \leq CH \log H \cdot D_{\mathsf{H}}^2 \big(M(\pi^\star), \overline{M}(\pi^\star) \big).
$$

Proof of Lemma K.3. For any $\pi \in \Pi_{\mathrm{RNS}}$, a full observation $(r, o_\circ) \sim M(\pi)$ consists of the trajectory $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H)$, where $s_1 \sim \mu^M$, $s_{h+1} \sim P_h^M(s_h, a_h)$ for $h \in [H-1]$, $r_h \sim R_h^M(s_h, a_h)$ for $h \in [H]$, and $a_h \sim \pi_h(s_h)$ for $h \in [H]$. We use the notation $\tau_{1:h}$ to denote the portion of a trajectory consisting of $(s_1, a_1, r_1, \ldots, s_h, a_h, r_h)$.

We use $\mathbb{P}^{M,\pi}$ to denote the distribution of the trajectory $\tau_H \sim M(\pi)$, and $\mathbb{P}^{\overline{M},\pi}$ to denote the distribution of the trajectory $\tau_H \sim \overline{M}(\pi)$. We use $\mathbb{E}^{M,\pi}[\cdot]$ and $\mathbb{E}^{\overline{M},\pi}[\cdot]$ to denote the corresponding

expectations. By Lemma A.13 of Foster et al. (2021), it holds that, for some constant $C > 0$,

$$\mathbb{E}_{\pi \sim p}\left[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right]$$

$$\leq C \log(H) \cdot \mathbb{E}_{\pi \sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D_{\mathsf{H}}^2\Big(\mathbb{P}^{M,\pi}(s_h|\tau_{1:h-1}), \mathbb{P}^{\overline{M},\pi}(s_h|\tau_{1:h-1})\Big)\right]$$

$$+ C \log(H) \cdot \mathbb{E}_{\pi \sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D_{\mathsf{H}}^2\Big(\mathbb{P}^{M,\pi}(r_h|\tau_{1:h-1}, s_h, a_h), \mathbb{P}^{\overline{M},\pi}(r_h|\tau_{1:h-1}, r_h, a_h)\Big)\right]$$

$$= C \log(H) \cdot D_{\mathsf{H}}^2\Big(\mu^M, \mu^{\overline{M}}\Big) + C \log(H) \cdot \mathbb{E}_{\pi \sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H-1} D_{\mathsf{H}}^2\Big(P_h^M(s_h, a_h), P_h^{\overline{M}}(s_h, a_h)\Big)\right]$$

$$+ C \log(H) \cdot \mathbb{E}_{\pi \sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D_{\mathsf{H}}^2\Big(R_h^M(s_h, a_h), R_h^{\overline{M}}(s_h, a_h)\Big)\right]. \tag{45}$$

By Lemma K.2 and the definition of $\pi^\star$, for each $h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$, it holds that

$$\mathbb{E}_{\pi \sim p}\left[d_h^{\overline{M},\pi}(s,a)\right] = d_h^{\overline{M},\pi^\star}(s,a).$$

Thus, we may replace the expectation over $\pi \sim p$, $(s_h, a_h) \sim \overline{M}(\pi)$ in (45) with $(s_h, a_h) \sim \overline{M}(\pi^\star)$, and obtain

$$\mathbb{E}_{\pi \sim p}\left[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right]$$

$$\leq C \log(H) \cdot \left(D_{\mathsf{H}}^2\Big(\mu^M, \mu^{\overline{M}}\Big) + \mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H-1} D_{\mathsf{H}}^2\Big(P_h^M(s_h, a_h), P_h^{\overline{M}}(s_h, a_h)\Big)\right]\right.$$

$$\left. + \mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H} D_{\mathsf{H}}^2\Big(R_h^M(s_h, a_h), R_h^{\overline{M}}(s_h, a_h)\Big)\right]\right).$$

By Foster et al. (2021, Lemma A.9) and the data processing inequality, we have that:

$$\mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H-1} D_{\mathsf{H}}^2\Big(P_h^M(s_h, a_h), P_h^{\overline{M}}(s_h, a_h)\Big)\right] \leq 4H \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big),$$

$$\mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H} D_{\mathsf{H}}^2\Big(R_h^M(s_h, a_h), R_h^{\overline{M}}(s_h, a_h)\Big)\right] \leq 4H \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big),$$

$$D_{\mathsf{H}}^2\Big(\mu^M, \mu^{\overline{M}}\Big) \leq D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big).$$

It then follows that, for some constant $C > 0$,

$$\mathbb{E}_{\pi \sim p}\left[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\right] \leq C \cdot H \log(H) \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big),$$

as desired. ∎

Lemma K.4. *There is a constant $C > 0$ so that the following holds. Consider any model class $\mathcal{M}$ consisting of MDPs of fixed horizon H, finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, and cumulative rewards bounded by $[0, 1]$. Let Π_{RNS} be the class of randomized nonstationary policies. Consider any $\overline{M} \in \mathcal{M}$ and finite-support distribution $p \in \Delta(\Pi_{\mathrm{RNS}})$, and let $\pi^\star \in \Pi_{\mathrm{RNS}}$ denote any policy satisfying Eq. (44) for $\overline{M}$ and p. Then for any $\gamma > 0$,*

$$\sup_{M \in \mathcal{M}} \left\{ f^M(\pi_M) - f^M(\pi^\star) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big) \right\}$$

$$\leq \frac{CH \log H}{\gamma} + \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \frac{\gamma}{CH \log H} \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big) \right]. \quad (46)$$

An immediate consequence of Lemma K.4 is that

$$\inf_{\pi \in \Pi_{\mathrm{RNS}}} \sup_{M \in \mathcal{M}} \left\{ f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big) \right\}$$

$$\leq \frac{CH \log H}{\gamma} + \inf_{p \in \Delta(\Pi_{\mathrm{RNS}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \frac{\gamma}{CH \log H} \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big) \right].$$

Proof of Lemma K.4. Consider any $\overline{M} \in \mathcal{M}$, finite-support $p \in \Delta(\Pi_{\mathrm{RNS}})$, and let $\pi^\star$ be defined as in the statement of the lemma. Lemma K.2 gives that $f^{\overline{M}}(\pi^\star) = \mathbb{E}_{\pi \sim p}\big[f^{\overline{M}}(\pi)\big]$. Let C be the constant from Lemma K.3, and let $C' = C + \frac{1}{2}$. Then for any $\gamma > 0$,

$$\sup_{M \in \mathcal{M}} f^M(\pi_M) - f^M(\pi^\star) - C'H \log H \cdot \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big)$$

$$\leq \sup_{M \in \mathcal{M}} f^M(\pi_M) - f^{\overline{M}}(\pi^\star) - C'H \log H \cdot \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big) + \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big)$$

$$= \sup_{M \in \mathcal{M}} f^M(\pi_M) - f^{\overline{M}}(\pi^\star) - CH \log H \cdot \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi^\star), \overline{M}(\pi^\star)\big) + \frac{1}{2\gamma}$$

$$\leq \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^{\overline{M}}(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big) \right] + \frac{1}{2\gamma} \quad (47)$$

$$\leq \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big) + \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big) \right] + \frac{1}{2\gamma}$$

$$= \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f^M(\pi_M) - f^M(\pi) - \frac{\gamma}{2} \cdot D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big) \right] + \frac{1}{\gamma}. \quad (48)$$

where (47) uses Lemma K.3. The statement of the proposition follows by replacing γ with $\gamma \cdot C'H \log H$. ∎

Appendix L. Proofs for upper bounds from Appendix F

In this section we prove Theorem F.1, which gives an upper bound for learning equilibria for CCE and CE instances in the MA-DMSO framework in a way that avoids the curse of multiple agents, i.e., avoids exponential scaling with the number of players K. In Appendix L.1, we describe the algorithm (Algorithm 1) used to prove Theorem F.1, which is based on the idea of *exploration-by-optimization*, used previously in Foster et al. (2022b); Lattimore and György (2021). In Appendices L.2 to L.4 we prove Theorem F.1; our proofs roughly follow those of Foster et al. (2022b), but require some subtle modifications to account for the multi-agent nature of our problem, as well as the more general notion of deviation sets Π'_k that we study.

L.1. The multi-agent exploration-by-optimization objective

We begin by describing the algorithm, *Multi-Agent Exploration-by-Optimization* (MAExO; Algorithm 1) used to prove Theorem F.1. The algorithm is a multi-agent counterpart to the exploration-by-optimization (ExO$^+$) algorithm given Foster et al. (2022b). At a high level, MAExO (as well as its precursor ExO$^+$) is a variant of EXP3, which applies the exponential weights algorithm to a sequence of reward estimators which act as importance-weighted estimates for the true reward function. However, unlike EXP3 and ExO$^+$, MAExO does not apply exponential weights to agents' pure policies themselves, but rather to their potential deviations Π'_k.

In particular, MAExO operates over T rounds of interaction with the environment. At each round $t \in [T]$, the algorithm first computes, for each player k, a *reference distribution $q_k^t \in \Delta(\Pi'_k)$* over their deviation space Π'_k, according to an exponential weights update given a sequence of vectors $\widehat{f}_k^1, \ldots, \widehat{f}_k^{t-1}$ constructed by the algorithm in previous rounds (Line 4). Roughly speaking, for $s \leq t - 1$, the entries $\widehat{f}_k^s(\pi'_k), \pi'_k \in \Pi'_k$, of these vectors can be interpreted as the potential gain in value that agent k could receive by deviating to π'_k, given adversarial choices of the other agents' decisions. Accordingly, the reference distribution q_k^t will put more mass on deviations which lead to larger gains in value.

Next, in Line 5, the players jointly solve an optimization problem. To define this optimization problem, we introduce some notation. For each $k \in [K]$, let $\mathcal{G}_k$ denote the set of all functions $g_k : \Pi'_k \times \Sigma \times \mathcal{O} \to \mathbb{R}$, and let $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_K$. Given $q \in \prod_{k=1}^K \Delta(\Pi'_k), \eta > 0, \pi \in \Pi, g \in \mathcal{G}$, $\pi^\star = (\pi_1^\star, \ldots, \pi_K^\star) \in \prod_{k=1}^K \Pi'_k$, and $M \in \mathcal{M}$, define

$$\Gamma_{q,\eta}(\pi, g; \pi^\star, M) := \mathbb{E}_{\sigma \sim \pi}\left[\sum_{k=1}^K f_k^M(U_k(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right] \tag{49}$$

$$+ \frac{1}{\eta} \cdot \sum_{k=1}^K \mathbb{E}_{\sigma \sim \pi, o \sim M(\sigma)} \mathbb{E}_{\pi'_k \sim q_k} \left[\exp\left(\frac{\eta}{\pi(\sigma)} \cdot \left(g_k(\pi'_k; \sigma, o) - g_k(\pi_k^\star; \sigma, o) \right) \right) - 1 \right].$$

With this definition, the optimization problem solved in Line 5 of MAExO is as follows:

$$(\pi^t, g^t) \leftarrow \underset{\pi \in \Pi, g \in \mathcal{G}}{\arg\min} \ \underset{M \in \mathcal{M}, \pi^\star \in \prod_{k=1}^K \Pi'_k}{\sup} \Gamma_{q^t, \eta}(\pi, g; \pi^\star, M). \tag{50}$$

The interpretation of the objective (49) and the optimization problem (50) is as follows. Roughly speaking, for each $k \in [K], \pi'_k \in \Pi'_k, \sigma \in \Sigma, o \in \mathcal{O}$, the value $g_k(\pi'_k; \sigma, o)$ for $g \in \mathcal{G}_k$ can be interpreted as an estimate of player k's gain in value by deviating to π'_k under joint decision profile σ, under an unknown model M which is "consistent with" the decision-observation pair (σ, o). Then, by solving (50), the algorithm wishes to find a joint decision $\pi^t \in \Pi$ and estimator $g^t = (g_1^t, \ldots, g_K^t) \in \mathcal{G}$, which, for each player $k \in [K]$, satisfies the following two properties:

- First, corresponding to the first term in (49), for a worst-case unknown model M and an unknown deviation $\pi_k^\star$, it should not be possible for player k to gain much value by deviating to $\pi_k^\star$ given the policy π^t. Here $\pi_k^\star$ should be interpreted as the best deviation in hindsight at the termination of the algorithm.

- Second, corresponding to the second term in (49): π^t and g_k^t should be chosen so that with high probability under $\sigma \sim \pi^t$, g_k^t does not underestimate the value gain in deviating to $\pi_k^\star$

as compared to a sample π'_k from the reference distribution q^t_k. The second term in (49) can be viewed as a term that regularizes the adversarial choice of $\pi^\star_k$, analogously to the term subtracting squared Hellinger distance in the offset DEC (see (6)): in particular, if $\pi^\star_k$ has significantly high value under the estimate g_k, then this term will be very negative, canceling out the (potentially large) first term.

Given (π^t, g^t) computed in Eq. (50), Algorithm 1 samples a decision $\sigma^t \sim \pi^t$ and receives an observation o^t from the true model. Finally, in Line 7, players construct their reward estimators $\widehat{f}^t_k$ (to be used in future iterations $t' > t$ to construct $q^{t'}_k$) using $g^t_k(\cdot; \sigma^t, o^t)$. Once all T rounds conclude, the algorithm outputs the joint decision $\widehat{\pi}$ which is the uniform average over the T pure decisions $\sigma^1, \ldots, \sigma^T$. We remark that Algorithm 1 is different from having each player run the exploration-by-optimization algorithm of Foster et al. (2022b): in the latter, agents each individually optimize their own objective, in contrast to the optimization problem in (50), which is solved for all agents simultaneously. This feature of MAExO allows us to obtain a guarantee scaling with $\mathsf{r\text{-}dec}^o_\gamma(\mathrm{co}(\mathscr{M}))$, which can be arbitrarily smaller than what one obtains by using the approach of Foster et al. (2022b) (see Proposition G.11).

In Definition L.1 below, we formalize the value of the minimax objective (50) computed in the course of Algorithm 1.

Algorithm 1 Multi-Agent Exploration by Optimization (MAExO)

1: **parameters**: Learning rate $\eta > 0$.

2: Initialize $\widehat{f}^0_k(\pi'_k) := 0$ for all $k \in [K]$, $\pi'_k \in \Pi'_k$.

3: **for** $t = 1, 2, \ldots T$ **do**

4:     For each agent $k \in [K]$, define $q^t_k \in \Delta(\Pi'_k)$ via exponential weights update: for $\pi'_k \in \Pi'_k$,

$$q^t_k(\pi'_k) := \frac{\exp\left(\eta \sum^{t-1}_{i=1} \widehat{f}^i_k(\pi'_k)\right)}{\sum_{\pi''_k \in \Pi'_k} \exp\left(\eta \sum^{t-1}_{i=1} \widehat{f}^i_k(\pi''_k)\right)}.$$

5:     Define $q^t = q^t_1 \times \cdots \times q^t_K$. The players jointly solve the following objective: `// Eq. (49)`

$$(\pi^t, g^t) \leftarrow \underset{\pi \in \Pi, g \in \mathcal{G}}{\arg\min} \ \underset{M \in \mathcal{M}, \pi^\star \in \Pi'}{\sup} \Gamma_{q^t, \eta}(\pi, g; \pi^\star, M).$$

6:     Sample $\sigma^t \sim \pi^t$, play σ^t, and observe $o^t \sim M^\star(\sigma^t)$.

7:     Each player $k \in [K]$ constructs their reward estimator $\widehat{f}^t_k$ as follows: for $\pi'_k \in \Pi'_k$,

$$\widehat{f}^t_k(\pi'_k) = \frac{g^t_k(\pi'_k; \sigma^t, o^t)}{\pi^t(\sigma^t)}.$$

8: **return** joint decision $\widehat{\pi} := \frac{1}{T} \sum^T_{t=1} \mathbb{I}_{\sigma^t}$.

Definition L.1 (Exploration-by-optimization objective). *Consider any instance $\mathcal{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ satisfying Assumption F.1. For any scale parameter $\eta > 0$ and distribution $q \in \prod_{k=1}^{K} \Delta(\Pi'_k)$, define*

$$\mathsf{exo}_\eta(\mathcal{M}, q) = \inf_{\pi \in \Pi, g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^\star \in \prod_{k=1}^{K} \Pi'_k} \Gamma_{q,\eta}(\pi, g; \pi^\star, M),$$

and let $\mathsf{exo}_\eta(\mathcal{M}) := \sup_{q \in \prod_{k=1}^{K} \Delta(\Pi'_k)} \mathsf{exo}_\eta(\mathcal{M}, q)$.

To prove Theorem F.1, we first (Appendix L.2) bound the performance of Algorithm 1 in terms of $\mathsf{exo}_\eta(\mathcal{M})$. Following this, in Appendix L.3 and Appendix L.4, we will upper bound $\mathsf{exo}_\eta(\mathcal{M})$ by $\mathsf{dec}^o_\gamma(\mathcal{M})$ for an appropriate choice of γ, using a quantity we call the *multi-agent (parametrized) information ratio* as an intermediary. Finally, in Appendix L.5, we put these pieces together and prove Theorem F.1.

L.2. Bounding the performance of Algorithm 1

The following result bounds the performance of Algorithm 1 (namely, the quantity $h^{M^\star}(\widehat{\pi})$) in terms of $\mathsf{exo}_\eta(\mathcal{M})$.

Lemma L.1. *For any $\eta > 0$, Algorithm 1 ensures that for all $\delta > 0$, with probability at least $1 - \delta$,*

$$h^{M^\star}(\widehat{\pi}) = \sum_{k=1}^{K} \max_{\pi'_k \in \Pi'_k} f_k^{M^\star}(U_k(\pi'_k, \widehat{\pi})) - f_k^{M^\star}(\widehat{\pi}) \leq \mathsf{exo}_\eta(\mathcal{M}) + \frac{2}{T\eta} \cdot \sum_{k=1}^{K} \log\left(\frac{K \cdot |\Pi'_k|}{\delta}\right).$$

Proof of Lemma L.1. For any $\pi^\star_k \in \Pi'_k$ and player $k \in [K]$, we define player k's *regret* with respect to the deviation $\pi^\star_k \in \Pi'_k$ as follows:

$$\mathrm{Reg}_k(\pi^\star_k) = \sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t}[f_k^{M^\star}(U_k(\pi^\star_k, \sigma^t)) - f_k^{M^\star}(\sigma^t)] = T \cdot \left(f_k^{M^\star}(U_k(\pi^\star_k, \widehat{\pi})) - f_k^{M^\star}(\widehat{\pi})\right),$$

where the second equality above uses the definition of $\widehat{\pi}$ in Line 8 of Algorithm 1 and the second property in Assumption F.1. Hence, it suffices to bound $\frac{1}{T} \cdot \sum_{k=1}^{K} \max_{\pi^\star_k \in \Pi'_k} \mathrm{Reg}_k(\pi^\star_k)$ to establish the statement of the lemma.

Throughout the proof we use the following convention: for functions $f_k : \Pi'_k \to \mathbb{R}$ (for instance, the reward estimators $\widehat{f}^t_k$ defined in Line 7 of Algorithm 1), we will view f_k as a vector in $\mathbb{R}^{|\Pi'_k|}$, whose coordinates are the values of $f_k(\pi'_k)$, for $\pi'_k \in \Pi'_k$. Furthermore, for each $\pi'_k \in \Pi'_k$, we write $e_{\pi'_k} \in \mathbb{R}^{|\Pi'_k|}$ to denote the corresponding unit vector whose π'_k-th entry is 1 and all other entries are 0.

By adding and subtracting $\sum_{t=1}^{T} \langle e_{\pi^\star_k}, \widehat{f}^t_k \rangle$, we obtain

$$\begin{aligned}
\mathrm{Reg}_k(\pi^\star_k) &= \sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t}[f_k^{M^\star}(U_k(\pi^\star_k, \sigma^t)) - f_k^{M^\star}(\sigma^t)] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t}[f_k^{M^\star}(U_k(\pi^\star_k, \sigma^t)) - f_k^{M^\star}(\sigma^t)] + \sum_{t=1}^{T} \langle e_{\pi^\star_k}, \widehat{f}^t_k \rangle - \sum_{t=1}^{T} \langle e_{\pi^\star_k}, \widehat{f}^t_k \rangle. \quad (51)
\end{aligned}$$

By Lemma L.2 and the definition of the multiplicative weights updates for q_k^t in Line 4 of Algorithm 1, it holds that

$$
\sum_{t=1}^{T} \langle e_{\pi_k^\star}, \widehat{f}_k^t \rangle
$$

$$
\leq \sum_{t=1}^{T} \langle q_k^{t+1}, \widehat{f}_k^t \rangle - \frac{1}{\eta} \sum_{t=1}^{T} D_{\mathsf{KL}} \left(q_k^{t+1} \,\|\, q_k^t \right) + \frac{1}{\eta} D_{\mathsf{KL}} \left(e_{\pi_k^\star} \,\|\, q_k^1 \right)
$$

$$
\leq \sum_{t=1}^{t} \langle q_k^{t+1}, \widehat{f}_k^t \rangle - \frac{1}{\eta} \sum_{t=1}^{T} D_{\mathsf{KL}} \left(q_k^{t+1} \,\|\, q_k^t \right) + \frac{\log |\Pi_k'|}{\eta}. \tag{52}
$$

By Lemma H.5, we have that for each $t \in [T]$,

$$
\langle q_k^{t+1}, \widehat{f}_k^t \rangle - \frac{1}{\eta} D_{\mathsf{KL}} \left(q_k^{t+1} \,\|\, q_k^t \right) \leq \frac{1}{\eta} \log \left(\sum_{\pi_k' \in \Pi_k'} q_k^t(\pi_k') \cdot \exp(\eta \cdot \widehat{f}_k^t(\pi_k')) \right).
$$

Using the above together with (52) and (51), we obtain

$$
\mathrm{Reg}_k(\pi_k^\star) \leq \sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t} [f_k^{M^\star}(U_k(\pi_k^\star, \sigma^t)) - f_k^{M^\star}(\sigma^t)] + \frac{1}{\eta} \log \left(\sum_{\pi_k' \in \Pi_k'} q_k^t(\pi_k') \cdot \exp(\eta \cdot \widehat{f}_k^t(\pi_k')) \right)
$$

$$
+ \frac{\log |\Pi_k'|}{\eta} - \sum_{t=1}^{T} \langle e_{\pi_k^\star}, \widehat{f}_k^t \rangle. \tag{53}
$$

Let $\mathscr{F}^t$ denote the σ-algebra generated by $(\sigma^1, o^1, \ldots, \sigma^t, o^t)$ (where the random variables σ^s, o^s are drawn as in Algorithm 1). Note that $\mathscr{F}^t$ is a filtration, and write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathscr{F}^t]$. For each $\pi_k^\star \in \Pi_k'$, we define a sequence of random variables, denoted $\{X_t(\pi_k^\star)\}_{t \in [T]}$, by

$$
X_t(\pi_k^\star) := \log \left(\sum_{\pi_k' \in \Pi_k'} q_k^t(\pi_k') \cdot \exp(\eta \cdot \widehat{f}_k^t(\pi_k')) \right) - \langle e_{\pi_k^\star}, \eta \cdot \widehat{f}_k^t \rangle.
$$

By Lemma H.7 and the union bound, with probability at least $1 - \delta/K$, it holds that for all $\pi_k^\star \in \Pi_k'$,

$$
\sum_{t=1}^{T} X_t(\pi_k^\star) \leq \sum_{t=1}^{T} \log \mathbb{E}_{t-1}[e^{X_t(\pi_k^\star)}] + \log \left(\frac{K \cdot |\Pi_k'|}{\delta} \right). \tag{54}
$$

Note that π^t, q^t are both measurable with respect to $\mathscr{F}^{t-1}$. Then, for any $\pi_k^\star \in \Pi_k'$ and any $t \in [T]$, we may compute

$$\log \mathbb{E}_{t-1}[e^{X_t(\pi_k^\star)}]$$

$$= \log \mathbb{E}_{t-1}\left[\exp\left(\log\left(\sum_{\pi_k' \in \Pi_k'} q_k^t(\pi_k') \cdot \exp(\eta \widehat{f}_k^t(\pi_k'))\right) - \eta \widehat{f}_k^t(\pi_k^\star)\right)\right]$$

$$= \log \mathbb{E}_{\sigma^t \sim \pi^t} \mathbb{E}_{o^t \sim M(\sigma^t)}\left[\mathbb{E}_{\pi_k' \sim q_k^t}\left[\exp(\eta \widehat{f}_k^t(\pi_k'))\right] \cdot \exp(-\eta \widehat{f}_k^t(\pi_k^\star))\right]$$

$$= \log \mathbb{E}_{\sigma^t \sim \pi^t} \mathbb{E}_{o^t \sim M(\sigma^t)}\left[\mathbb{E}_{\pi_k' \sim q_k^t}\left[\exp\left(\frac{\eta}{\pi^t(\sigma^t)} \cdot g_k^t(\pi_k'; \sigma^t, o^t)\right)\right] \cdot \exp\left(-\frac{\eta}{\pi^t(\sigma^t)} \cdot g_k^t(\pi_k^\star; \sigma^t, o^t)\right)\right]$$

$$\leq \mathbb{E}_{\sigma^t \sim \pi^t} \mathbb{E}_{o^t \sim M(\sigma^t)} \mathbb{E}_{\pi_k' \sim q_k^t}\left[\exp\left(\frac{\eta}{\pi^t(\sigma^t)} \cdot \left(g_k^t(\pi_k'; \sigma^t, o^t) - g_k^t(\pi_k^\star; \sigma^t, o^t)\right)\right)\right] - 1, \tag{55}$$

where the final inequality uses that $\log(x) \leq x - 1$ for all $x > 0$.

By (53), (54), and (55), and a union bound over $k \in [K]$, it follows that with probability at least $1 - \delta$, for all $\pi_1^\star \in \Pi_1', \ldots, \pi_K^\star \in \Pi_K'$, letting $\pi^\star = (\pi_1^\star, \ldots, \pi_K^\star)$,

$$\sum_{k=1}^K \mathrm{Reg}_k(\pi_k^\star)$$

$$\leq \sum_{k=1}^K \left(\frac{\log |\Pi_k'|}{\eta} + \frac{\log\left(\frac{K \cdot |\Pi_k'|}{\delta}\right)}{\eta}\right) + \sum_{t=1}^T \sum_{k=1}^K \mathbb{E}_{\sigma^t \sim \pi^t}\left[f_k^{M^\star}(U_k(\pi_k^\star, \sigma^t)) - f_k^{M^\star}(\sigma^t)\right]$$

$$+ \sum_{t=1}^T \sum_{k=1}^K \frac{1}{\eta}\left(\mathbb{E}_{\sigma^t \sim \pi^t} \mathbb{E}_{o^t \sim M(\sigma^t)} \mathbb{E}_{\pi_k' \sim q_k^t}\left[\exp\left(\frac{\eta}{\pi^t(\sigma^t)} \cdot \left(g_k^t(\pi_k'; \sigma^t, o^t) - g_k^t(\pi_k^\star; \sigma^t, o^t)\right)\right)\right] - 1\right)$$

$$\leq \frac{2}{\eta} \cdot \sum_{k=1}^K \log\left(\frac{K \cdot |\Pi_k'|}{\delta}\right) + \sum_{t=1}^T \Gamma_{q^t, \eta}(\pi^t, g^t; \pi^\star, M^\star)$$

$$\leq \frac{2}{\eta} \cdot \sum_{k=1}^K \log\left(\frac{K \cdot |\Pi_k'|}{\delta}\right) + \sum_{t=1}^T \sup_{\widetilde{\pi}^\star \in \Pi', \widetilde{M} \in \mathcal{M}} \Gamma_{q^t, \eta}(\pi^t, g^t; \widetilde{\pi}^\star, \widetilde{M})$$

$$\leq \frac{2}{\eta} \cdot \sum_{k=1}^K \log\left(\frac{K \cdot |\Pi_k'|}{\delta}\right) + T \cdot \mathsf{exo}_\eta(\mathscr{M}),$$

where the second inequality uses the definition of $\Gamma_{q^t, \eta}(\pi^t, g^t; \pi^\star, M^\star)$ in (49), and the final equality follows since π^t, g^t are chosen so as to minimize the multi-agent exploration-by-optimization objective (Line 5 of Algorithm 1). ∎

Lemma L.2. *Consider any $d \in \mathbb{N}$, and let $f^1, \ldots, f^T \in \mathbb{R}^d$ be an arbitrary sequence of vectors. For $\eta > 0$, let $q^1, \ldots, q^T \in \Delta^d$ denote the exponential weights update iterates with step size η when the reward vectors are given by $f^1, \ldots, f^T$; in particular, for $t \in [T]$:*

$$q^t(i) = \frac{\exp(\eta \sum_{s \leq t} f^s(i))}{\sum_{j=1}^d \exp(\eta \sum_{s \leq t} f^s(j))}. \tag{56}$$

Then for any $q \in \Delta^d$,

$$\sum_{t=1}^{T} \langle q, f^t \rangle \leq \sum_{t=1}^{T} \langle q^{t+1}, f^t \rangle - \frac{1}{\eta} \sum_{t=1}^{T} D_{\mathsf{KL}}(q^{t+1} \,\|\, q^t) + \frac{1}{\eta} D_{\mathsf{KL}}(q \,\|\, q^1).$$

Proof of Lemma L.2. By rearranging and telescoping, it suffices to show that, for each $t \in [T]$,

$$\langle q - q^{t+1}, f^t \rangle = \frac{1}{\eta} \cdot \left(D_{\mathsf{KL}}(q \,\|\, q^t) - D_{\mathsf{KL}}(q \,\|\, q^{t+1}) - D_{\mathsf{KL}}(q^{t+1} \,\|\, q^t) \right).$$

To establish this inequality, we note that the multiplicative weight updates (56) are equivalent to the following mirror descent updates with the negative entropy regularizer $\Phi(q) := \sum_{i=1}^{d} q_i \cdot \log q_i$:

$$\nabla \Phi(p^{t+1}) = \nabla \Phi(q^t) + \eta \cdot f^t, \quad q^{t+1} = \frac{p^{t+1}}{\langle \mathbf{1}, p^{t+1} \rangle},$$

where $\mathbf{1} \in \mathbb{R}^d$ denotes the all-ones vector. Using the fact that for all $x, y, z \in \Delta^d$ (Eq. (4.1) of Bubeck (2015))

$$\langle \nabla \Phi(y) - \nabla \Phi(x), x - z \rangle = D_{\mathsf{KL}}(z \,\|\, y) - D_{\mathsf{KL}}(z \,\|\, x) - D_{\mathsf{KL}}(x \,\|\, y)$$

with $z = q, y = q^t, x = q^{t+1}$, we obtain

$$
\begin{aligned}
\frac{1}{\eta} \cdot \left(D_{\mathsf{KL}}(q \,\|\, q^t) - D_{\mathsf{KL}}(q \,\|\, q^{t+1}) - D_{\mathsf{KL}}(q^{t+1} \,\|\, q^t) \right) &= \frac{1}{\eta} \cdot \langle \nabla \Phi(q^t) - \nabla \Phi(q^{t+1}), q^{t+1} - q \rangle \\
&= \frac{1}{\eta} \cdot \langle \nabla \Phi(q^t) - \nabla \Phi(p^{t+1}), q^{t+1} - q \rangle \\
&= \langle f^t, q - q^{t+1} \rangle,
\end{aligned}
$$

where in the second equality we have used that $\nabla \Phi(q^{t+1}) = \nabla \Phi(p^{t+1}) - \log \left(\langle \mathbf{1}, p^{t+1} \rangle \right) \cdot \mathbf{1}$. ∎

L.3. The multi-agent parametrized information ratio

In this section, we introduce a multi-agent version of the parametrized information ratio of (Foster et al., 2022b, Definition 3.1), and upper bound this information ratio by the DEC of the convex hull of $\mathscr{M}$. In the following section, we will upper bound $\mathsf{exo}_\eta(\mathscr{M})$ by this information ratio.

We first introduce some notation. We will wish to reason about the space of probability measures on $\mathcal{M} \times \Pi_1' \times \cdots \times \Pi_K'$. Since $|\mathcal{M}|$ may be infinite, to avoid measure-theoretic issues, we will slightly abuse notation by letting $\Delta(\mathcal{M} \times \Pi_1' \times \cdots \times \Pi_K')$ denote the set of *finitely supported* probability measures on $\mathcal{M} \times \Pi_1' \times \cdots \times \Pi_K'$. This convention ensures that for any function $h : \mathcal{M} \times \Pi_1' \times \cdots \times \Pi_K' \to \mathbb{R}$ and any $\mu \in \Delta(\mathcal{M} \times \Pi_1' \times \cdots \times \Pi_K')$, $\mathbb{E}_{(M, \pi_1', \dots, \pi_K') \sim \mu}[h(M, \pi_1', \dots, \pi_K')]$ is well-defined.

Consider any $k \in [K]$, a distribution $\mu \in \Delta(\mathcal{M} \times \Pi_1' \times \cdots \times \Pi_K')$, and a distribution $\pi \in \Delta(\Sigma)$. Let $\mathbb{P}$ denote the law of the process $(M, \pi_1^\star, \dots, \pi_K^\star) \sim \mu, \sigma \sim \pi$, and $o \sim M(\sigma)$. We introduce the following distributions, depending on μ and k:

- Define the distribution $\mu_{\mathrm{pr}}^k \in \Delta(\Pi_k')$ by $\mu_{\mathrm{pr}}^k(\pi_k') = \mathbb{P}(\pi_k^\star = \pi_k')$, for $\pi_k' \in \Pi_k'$.

- For each $\sigma \in \Sigma$ and $o \in \mathcal{O}$, define the distribution $\mu_{\mathrm{po}}^k \in \Delta(\Pi'_k)$ by $\mu_{\mathrm{po}}^k(\pi'_k; \sigma, o) = \mathbb{P}(\pi_k^\star = \pi'_k | (\sigma, o))$, for $\pi'_k \in \Pi'_k$.

The distribution μ_{pr}^k should be thought of as a prior distribution over the deviation $\pi_k^\star$, and the distribution $\mu_{\mathrm{po}}^k(\cdot; \sigma, o)$ should be thought of as a posterior distribution over $\pi_k^\star$ after observing the pure decision σ together with an observation $o \sim M(\sigma)$.

Definition L.2 (Multi-agent information ratio). *Given an instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ which is a generalized correlated equilibrium instance, the* parametrized multi-agent information ratio *of the instance $\mathscr{M}$ is defined as*

$$\mathsf{infr}_\gamma(\mathscr{M}) := \sup_{\mu \in \Delta(\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K)} \inf_{\pi \in \Pi} \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{(M, \pi_1^\star, \dots, \pi_K^\star) \sim \mu} \left[\sum_{k=1}^K f_k^M(U_k(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right]$$
$$- \gamma \cdot \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{o|\sigma} \left[\sum_{k=1}^K D_{\mathsf{H}}^2 \left(\mu_{\mathrm{po}}^k(\cdot; \sigma, o), \mu_{\mathrm{pr}}^k(\cdot) \right) \right].$$

In the above expression, when we write $U_k(\pi_k^\star, \sigma)$ and $f_k^M(\sigma)$, we view $\sigma \in \Sigma$ as an element of Π by associating it with the singleton distribution on σ, recalling that $\Pi = \Delta(\Sigma)$.

Lemma L.3 upper bounds the multi-agent information ratio in terms of the multi-agent offset DEC of the convex hull of a given instance.

Lemma L.3. *Consider any instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ which satisfies Assumption F.1, and for which $\mathrm{co}(\mathscr{M})$ satisfies Assumption 1.3. Then for all $\gamma > 0$,*

$$\mathsf{infr}_\gamma(\mathscr{M}) \le K \cdot \mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathrm{co}(\mathscr{M})).$$

Proof of Lemma L.3. We denote the pure decision sets of the instance $\mathscr{M}$ by $\Sigma_1, \dots, \Sigma_K$, and the joint decision set as $\Sigma = \Sigma_1 \times \cdots \times \Sigma_K$. Fix a prior $\mu \in \Delta(\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K)$ and a distribution $\pi \in \Delta(\Sigma)$. Recall our notation from above: let $\mathbb{P}$ denote the law of the process $\sigma \sim \pi, (M, \pi_1^\star, \dots, \pi_K^\star) \sim \mu, o \sim M(\sigma)$. For each $k \in [K]$, let $\mu_{\mathrm{pr}}^k(\pi'_k) = \mathbb{P}(\pi_k^\star = \pi'_k)$ and $\mu_{\mathrm{po}}^k(\pi'_k; \sigma, o) = \mathbb{P}(\pi_k^\star = \pi'_k | (\sigma, o))$.

Consider the value of the multi-agent information ratio given the choices for μ, π:

$$\mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{(M, \pi_1^\star, \dots, \pi_K^\star) \sim \mu} \left[\sum_{k=1}^K f_k^M(U_k(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right] - \gamma \cdot \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{o|\sigma} \left[\sum_{k=1}^K D_{\mathsf{H}}^2 \left(\mu_{\mathrm{po}}^k(\cdot; \sigma, o), \mu_{\mathrm{pr}}^k(\cdot) \right) \right].$$

For each $k \in [K]$, $\pi'_k \in \Pi'_k$, and $\pi \in \Pi$, define $\overline{M}_{\pi'_k}^k(\pi) := \mathbb{E}_\mu[M(\pi) | \pi_k^\star = \pi'_k]$. Further define $\overline{M}(\pi) = \mathbb{E}_\mu[M(\pi)]$. Note that $\overline{M}_{\pi'_k}^k(\sigma) = \mathbb{P}_{o|\sigma, \pi'_k}$ and $\overline{M}(\sigma) = \mathbb{P}_{o|\sigma}$.

To proceed, note that for each fixed $\sigma \in \Sigma$,

$$\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2 \left(\mu_{\mathrm{po}}^k(\cdot; \sigma, o), \mu_{\mathrm{pr}}(\cdot) \right) \right] = \mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2 \left(\mathbb{P}_{\pi_k^\star | \sigma, o}, \mathbb{P}_{\pi_k^\star} \right) \right]$$
$$= \mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2 \left(\mathbb{P}_{\pi_k^\star | \sigma, o}, \mathbb{P}_{\pi_k^\star | \sigma} \right) \right]$$
$$= \mathbb{E}_{\pi_k^\star \sim \mu} \left[D_{\mathsf{H}}^2 \left(\mathbb{P}_{o|\sigma, \pi_k^\star}, \mathbb{P}_{o|\sigma} \right) \right]$$
$$= \mathbb{E}_{\pi_k^\star \sim \mu} \left[D_{\mathsf{H}}^2 \left(\overline{M}_{\pi_k^\star}^k(\sigma), \overline{M}(\sigma) \right) \right], \tag{57}$$

98

where the second equality follows since $\sigma \sim \pi$ and $(\pi_1^\star, \ldots, \pi_K^\star) \sim \mu$ are (marginally) independent, and the third equality holds by Lemma H.2. Furthermore, we have that

$$
\mathbb{E}_{(M,\pi_1^\star,\ldots,\pi_K^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi_k^\star,\sigma)) - f_k^M(\sigma)\right] = \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{\pi_k^\star\sim\mu}\mathbb{E}_\mu\left[f_k^M(U_k(\pi_k^\star,\sigma)) - f_k^M(\sigma)|\pi_k^\star\right]
$$

$$
= \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{\pi_k^\star\sim\mu}\left[f_k^{\overline{M}_{\pi_k^\star}^k}(U_k(\pi_k^\star,\sigma)) - f_k^{\overline{M}_{\pi_k^\star}^k}(\pi)\right]
$$

$$
\leq \mathbb{E}_{\pi_k^\star\sim\mu}\max_{\pi_k'\in\Pi_k'}\mathbb{E}_{\sigma\sim\pi}\left[f_k^{\overline{M}_{\pi_k^\star}^k}(U_k(\pi_k',\sigma)) - f_k^{\overline{M}_{\pi_k^\star}^k}(\sigma)\right].
$$
$$(58)$$

Next, for any $\overline{M}_1 \in \mathrm{co}(\mathcal{M})$, we have, for $\gamma > 0$,

$$
\mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathrm{co}(\mathscr{M}), \overline{M}_1)
$$

$$
= \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathrm{co}(\mathcal{M})}\mathbb{E}_{\pi\sim p}\left[\sum_{k=1}^K\left(\max_{\pi_k'\in\Pi_k'}f_k^M(U_k(\pi_k',\pi)) - f_k^M(\pi)\right) - \gamma\cdot D_\mathsf{H}^2\big(M(\pi),\overline{M}_1(\pi)\big)\right]
$$

$$
= \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathrm{co}(\mathcal{M})}\mathbb{E}_{\pi\sim p}\left[\sum_{k=1}^K\max_{\pi_k'\in\Pi_k'}\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi_k',\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot D_\mathsf{H}^2\big(\mathbb{E}_{\sigma\sim\pi}[M(\sigma)],\mathbb{E}_{\sigma\sim\pi}[\overline{M}_1(\sigma)]\big)\right]
$$

$$
\geq \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathrm{co}(\mathcal{M})}\mathbb{E}_{\pi\sim p}\left[\sum_{k=1}^K\max_{\pi_k'\in\Pi_k'}\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi_k',\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\pi}[D_\mathsf{H}^2\big(M(\sigma),\overline{M}_1(\sigma)\big)]\right]
$$

$$
\geq \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathrm{co}(\mathcal{M})}\sum_{k=1}^K\max_{\pi_k'\in\Pi_k'}\mathbb{E}_{\pi\sim p}\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi_k',\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot \mathbb{E}_{\pi\sim p}\mathbb{E}_{\sigma\sim\pi}\left[D_\mathsf{H}^2\big(M(\sigma),\overline{M}_1(\sigma)\big)\right]
$$

$$
\geq \inf_{\pi\in\Delta(\Sigma)}\sup_{M\in\mathrm{co}(\mathcal{M})}\sum_{k=1}^K\max_{\pi_k'\in\Pi_k'}\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi_k',\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\pi}\left[D_\mathsf{H}^2\big(M(\sigma),\overline{M}_1(\sigma)\big)\right],
$$

where the second equality follows from Assumption F.1, the first inequality follows from convexity of squared Hellinger distance, the second inequality follows from Jensen's inequality, and the final inequality follows by replacing any $p \in \Delta(\Pi)$ with the decision $\overline{\pi} := \mathbb{E}_{\pi\sim p}[\pi] \in \Delta(\Sigma) = \Pi$.

By the above display, the following holds: for any $\overline{M}_1 \in \mathrm{co}(\mathcal{M})$, there is some $\pi \in \Pi$ so that, for each $\overline{M}_2 \in \mathrm{co}(\mathcal{M})$,

$$
\sum_{k=1}^K\max_{\pi_k'\in\Pi_k'}\mathbb{E}_{\sigma\sim\pi}\left[f_k^{\overline{M}_2}(U_k(\pi_k',\sigma)) - f_k^{\overline{M}_2}(\sigma)\right] - \gamma\cdot\mathbb{E}_{\sigma\sim\pi}\left[D_\mathsf{H}^2\big(\overline{M}_2(\sigma),\overline{M}_1(\sigma)\big)\right] \leq \mathsf{r\text{-}dec}_\gamma^\mathrm{o}(\mathrm{co}(\mathscr{M})).
$$
$$(59)$$

Since we have assumed that $\mathrm{co}(\mathscr{M})$ satisfies Assumption 1.3, the following holds: for each $k \in [K]$, $\pi \in \Pi$, and $\overline{M}_1 \in \mathrm{co}(\mathcal{M})$, we have (again using Assumption F.1)

$$
\max_{\pi_k'\in\Pi_k'}\mathbb{E}_{\sigma\sim\pi}\left[f_k^{\overline{M}_1}(U_k(\pi_k',\sigma)) - f_k^{\overline{M}_1}(\sigma)\right] = \max_{\pi_k'\in\Pi_k'}f_k^{\overline{M}_1}(U_k(\pi_k',\pi)) - f_k^{\overline{M}_1}(\pi) \geq 0. \qquad (60)
$$

Then, by (60) and (59), for each $k \in [K]$, we have that for any $\overline{M}_1 \in \mathrm{co}(\mathcal{M})$, there is $\pi \in \Pi$ so that for each $\overline{M}_2 \in \mathrm{co}(\mathcal{M})$,

$$\max_{\pi'_k \in \Pi'_k} \mathbb{E}_{\sigma \sim \pi} \left[f_k^{\overline{M}_2}(U_k(\pi'_k, \sigma)) - f_k^{\overline{M}_2}(\sigma) \right] - \gamma \cdot \mathbb{E}_{\sigma \sim \pi} \left[D_{\mathsf{H}}^2 \big(\overline{M}_2(\sigma), \overline{M}_1(\sigma) \big) \right] \leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathcal{M})).$$

(61)

Next, choose $\overline{\pi} \in \Pi$, given $\overline{M}_1 = \overline{M}$ to ensure that (59) holds for all $\overline{M}_2 \in \mathrm{co}(\mathcal{M})$. Then for each $k \in [K]$ and each $\pi_k^\star \in \Pi'_k$, choosing $\overline{M}_2 = \overline{M}_{\pi_k^\star}^k$ in (61),

$$\max_{\pi'_k \in \Pi'_k} \mathbb{E}_{\sigma \sim \overline{\pi}} \left[f_k^{\overline{M}_{\pi_k^\star}^k}(U_k(\pi'_k, \sigma)) - f_k^{\overline{M}_{\pi_k^\star}^k}(\sigma) \right] - \gamma \cdot \mathbb{E}_{\sigma \sim \overline{\pi}} \left[D_{\mathsf{H}}^2 \big(\overline{M}_{\pi_k^\star}^k(\sigma), \overline{M}(\sigma) \big) \right] \leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathcal{M})).$$

Taking expectation over $\pi_k^\star \sim \mu$ and using (57) and (58), we obtain

$$\mathbb{E}_{(M,\pi_k^\star) \sim \mu} \mathbb{E}_{\sigma \sim \overline{\pi}} \left[f_k^M(U_k(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right] - \gamma \cdot \mathbb{E}_{\sigma \sim \overline{\pi}} \mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2 \big(\mu_{\mathrm{po}}^k(\cdot; \sigma, o), \mu_{\mathrm{pr}}(\cdot) \big) \right]$$

$$\leq \mathbb{E}_{\pi_k^\star \sim \mu} \left[\max_{\pi'_k \in \Pi'_k} \mathbb{E}_{\sigma \sim \overline{\pi}} \left[f_k^{\overline{M}_{\pi_k^\star}^k}(U_k(\pi'_k, \sigma)) - f_k^{\overline{M}_{\pi_k^\star}^k}(\sigma) \right] - \gamma \cdot \mathbb{E}_{\sigma \sim \overline{\pi}} \left[D_{\mathsf{H}}^2 \big(\overline{M}_{\pi_k^\star}^k(\sigma), \overline{M}(\sigma) \big) \right] \right]$$

$$\leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathcal{M})).$$

Note that the choice of $\overline{\pi}$ depends only on $\overline{M}$, and in particular it does not depend on k. Therefore, we may sum the above display over $k \in [K]$, to obtain

$$\mathbb{E}_{(M,\pi_1^\star,\ldots,\pi_K^\star) \sim \mu} \mathbb{E}_{\sigma \sim \overline{\pi}} \left[\sum_{k=1}^K f_k^M(U_k(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right] - \gamma \cdot \mathbb{E}_{\sigma \sim \overline{\pi}} \mathbb{E}_{o|\sigma} \left[\sum_{k=1}^K D_{\mathsf{H}}^2 \big(\mu_{\mathrm{po}}^k(\cdot; \sigma, o), \mu_{\mathrm{pr}}(\cdot) \big) \right]$$

$$\leq K \cdot \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathcal{M})).$$

Using that the choice of $\mu \in \Delta(\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K)$ is arbitrary, we obtain that $\mathsf{infr}_\gamma(\mathcal{M}) \leq \mathsf{r\text{-}dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathcal{M}))$, as desired. ∎

L.4. Relating the multi-agent information ratio and exploration-by-optimization objective

In this section, we prove the following result, which upper bounds $\mathsf{exo}_\eta(\mathcal{M})$ by the multi-agent information ratio of $\mathcal{M}$, at scale $1/(8\eta)$.

Lemma L.4. *Consider any instance $\mathcal{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ satisfying Assumption F.1. Then for all $\eta > 0$,*

$$\mathsf{exo}_\eta(\mathcal{M}) \leq \mathsf{infr}_{1/(8\eta)}(\mathcal{M}).$$

Proof of Lemma L.4. Throughout the proof, we will denote the (finite) pure decision sets, as guaranteed by Assumption F.1, by $\Sigma_1, \ldots, \Sigma_K$, and the joint decision set by $\Sigma := \Sigma_1 \times \cdots \times \Sigma_K$. Additionally, we write $\Pi' := \prod_{k=1}^k \Pi'_k$ to denote the product of the deviation sets Π'_k. We can write

$$\mathsf{exo}_\eta(\mathcal{M}) = \sup_{q \in \prod_{k=1}^K \Delta(\Pi'_k)} \inf_{\pi \in \Delta(\Sigma), \, g \in \mathcal{G}} \sup_{\mu \in \Delta(\mathcal{M} \times \Pi')} \mathbb{E}_{(M,\pi^\star) \sim \mu} \left[\Gamma_{q,\eta}(\pi, g; \pi^\star, M) \right].$$

For $\alpha \geq \max\{1, 1/\eta\}$ and $\varepsilon \in (0, 1)$, define

$$\mathcal{G}_\alpha = \{(g_1, \ldots, g_k) \in \mathcal{G} \ : \ \|g_k\|_\infty \leq \alpha \ \forall k \in [K]\}, \qquad \mathcal{P}_\varepsilon = \{\pi \in \Pi \ : \ \pi(\sigma) \geq \epsilon |\Sigma|^{-1} \ \forall \sigma\}.$$

We will now use Sion's minimax theorem (Theorem H.1), with $\mathcal{X} = \mathcal{P}_\varepsilon \times \mathcal{G}_\alpha$ and $\mathcal{Y} = \Delta(\mathcal{M} \times \Pi')$, to interchange the $\inf_{\pi \in \Pi, g \in \mathcal{G}}$ and the $\sup_{\mu \in \Delta(\mathcal{M} \times \Pi')}$ in the definition ot $\mathsf{exo}_\eta(\mathscr{M})$ above. We first check that its preconditions hold:

- Let the set $\mathcal{P}_\varepsilon$ have the standard topology induced from Π, so that $\mathcal{P}_\varepsilon$ is compact, and let $\mathcal{G}_\alpha$ have the product topology. Tychanoff's theorem yields that $\mathcal{G}_\alpha$ is compact, and thus $\mathcal{X} = \mathcal{P}_\varepsilon \times \mathcal{G}_\alpha$ is compact. It is also clearly convex.

- Let us give $\mathcal{Y} = \Delta(\mathcal{M} \times \Pi')$ (which we recall is the space of finitely supported distributions on $\mathcal{M} \times \Pi'$) the weak topology, which is the coarsest topology so that the functional $\mu \mapsto \int \phi d\mu$ is continuous for all bounded functions $\phi : \mathcal{M} \times \Pi' \to \mathbb{R}$.

- To establish the remaining preconditions, we need that the mapping $(\pi, g, \mu) \mapsto \mathbb{E}_{(M,\pi^\star) \sim \mu}[\Gamma_{q,\eta}(\pi, g; \pi^\star, M)]$ is uniformly bounded for $(\pi, g) \in \mathcal{P}_\varepsilon \times \mathcal{G}_\alpha$ and $\mu \in \Delta(\mathcal{M} \times \Pi')$. This follows immediately from the definition of $\Gamma_{q,\eta}(\pi, g; \pi^\star, M)$ and the domains $\mathcal{P}_\varepsilon$ and $\mathcal{G}_\alpha$.

- Clearly, the map $\mu \mapsto \mathbb{E}_{(M,\pi^\star)}[\Gamma_{q,\eta}(\pi, g; \pi^\star, M)]$ is linear, and thus concave, for each π, g. Moreover, it is continuous by boundedness of $\Gamma_{q,\eta}(\pi, g; \pi^\star, M)$, and the fact that $\Delta(\mathcal{M} \times \Pi')$ has the weak topology.

- By Lemma L.5, the map $(\pi, g) \mapsto \mathbb{E}_{(M,\pi^\star) \sim \mu}[\Gamma_{q,\eta}(\pi, g; \pi^\star, M)]$ is convex in (π, g) for any fixed μ. Furthermore, it is continuous by definition of the product topology and since $\pi(\sigma)$ is uniformly bounded below for $\pi \in \mathcal{P}_\varepsilon$.

Having verified all of the conditions for Theorem H.1 to apply, we now have:

$$\begin{aligned}
\mathsf{exo}_\eta(\mathscr{M}) &\leq \sup_{q \in \prod_{k=1}^K \Delta(\Pi'_k)} \ \inf_{\pi \in \mathcal{P}_\varepsilon, g \in \mathcal{G}_\alpha} \ \sup_{\mu \in \Delta(\mathcal{M} \times \Pi')} \mathbb{E}_{(M,\pi^\star) \sim \mu}[\Gamma_{q,\eta}(\pi, g; \pi^\star, M)] \\
&= \sup_{q \in \prod_{k=1}^K \Delta(\Pi'_k)} \ \sup_{\mu \in \Delta(\mathcal{M} \times \Pi')} \ \inf_{\pi \in \mathcal{P}_\varepsilon, g \in \mathcal{G}_\alpha} \mathbb{E}_{(M,\pi^\star) \sim \mu} \left[\Gamma_{q,\eta}(\pi, g; \pi^\star, M)\right], \qquad (62)
\end{aligned}$$

where the inequality follows since we are restricting to smaller sets $\mathcal{G}_\alpha \subset \mathcal{G}$ and $\mathcal{P}_\varepsilon \subset \Pi$ in the infimum, and the equality uses Theorem H.1. Given $q \in \prod_{k=1}^K \Delta(\Pi'_k)$, $\mu \in \Delta(\mathcal{M} \times \Pi')$, $\pi \in \mathcal{P}_\varepsilon$,

consider the value of

$$\inf_{g \in \mathcal{G}_\alpha} \mathbb{E}_{(M,\pi^\star) \sim \mu}[\Gamma_{q,\eta}(\pi, g; \pi^\star, M)]$$

$$=\mathbb{E}_{(M,\pi^\star) \sim \mu} \mathbb{E}_{\sigma \sim \pi} \left[\sum_{k=1}^K f_k^M(U(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right]$$

$$+ \frac{1}{\eta} \inf_{g \in \mathcal{G}_\alpha} \sum_{k=1}^K \mathbb{E}_{(M,\pi^\star) \sim \mu} \mathbb{E}_{\sigma \sim \pi, o \sim M(\sigma)} \mathbb{E}_{\pi_k' \sim q_k} \left[\exp\left(\frac{\eta}{\pi(\sigma)} \cdot \left(g_k(\pi_k'; \sigma, o) - g_k(\pi_k^\star; \sigma, o) \right) \right) - 1 \right]$$

$$=\mathbb{E}_{(M,\pi^\star) \sim \mu} \mathbb{E}_{\sigma \sim \pi} \left[\sum_{k=1}^K f_k^M(U(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right]$$

$$+ \frac{1}{\eta} \sum_{k=1}^K \inf_{g_k \in \mathcal{G}_{k,\alpha}} \mathbb{E}_{(M,\pi^\star) \sim \mu} \mathbb{E}_{\sigma \sim \pi, o \sim M(\sigma)} \mathbb{E}_{\pi_k' \sim q_k} \left[\exp\left(\frac{\eta}{\pi(\sigma)} \cdot \left(g_k(\pi_k'; \sigma, o) - g_k(\pi_k^\star; \sigma, o) \right) \right) - 1 \right],$$

$$(63)$$

where we have used $\mathcal{G}_{k,\alpha}$ to denote $\{g_k \in \mathcal{G}_k : \|g_k\|_\infty \leq \alpha\}$, so that $\mathcal{G}_\alpha = \mathcal{G}_{1,\alpha} \times \cdots \times \mathcal{G}_{K,\alpha}$.

Let $\mathbb{P}$ be the law of the process $(M, \pi^\star) \sim \mu$, $\sigma \sim \pi$, $o \sim M(\sigma)$, and define, for $k \in [K]$, $\mu_{\mathrm{pr}}^k(\pi_k') = \mathbb{P}(\pi_k^\star = \pi_k')$, and $\mu_{\mathrm{po}}^k(\pi_k'; \sigma, o) = \mathbb{P}(\pi_k^\star = \pi_k'|(\sigma, o))$. For each $k \in [K]$, the term corresponding to agent k in the second term of (63) above can be rewritten as follows, using the definition of the posterior distribution $\mu_{\mathrm{po}}^k(\pi_k'; \sigma, o)$:

$$\inf_{g_k \in \mathcal{G}_{k,\alpha}} \mathbb{E}_{(M,\pi^\star) \sim \mu} \mathbb{E}_{\sigma \sim \pi, o \sim M(\sigma)} \mathbb{E}_{\pi_k' \sim q_k} \left[\exp\left(\frac{\eta}{\pi(\sigma)} \cdot \left(g_k(\pi_k'; \sigma, o) - g_k(\pi_k^\star; \sigma, o) \right) \right) - 1 \right]$$

$$= \inf_{g_k \in \mathcal{G}_{k,\alpha}} \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{o|\sigma} \left[\mathbb{E}_{\pi_k' \sim q_k} \left[\exp\left(\eta \cdot \frac{g_k(\pi_k'; \sigma, o)}{\pi(\sigma)} \right) \right] \cdot \mathbb{E}_{\pi_k^\star \sim \mu_{\mathrm{po}}^k(\cdot; \sigma, o)} \left[\exp\left(-\eta \cdot \frac{g_k(\pi_k^\star; \sigma, o)}{\pi(\sigma)} \right) \right] - 1 \right].$$

Given any $g_k \in \mathcal{G}_{k,\eta\alpha}$, we have that $(\pi_k', \sigma, o) \mapsto \frac{\pi(\sigma)}{\eta} \cdot g_k(\pi_k'; \sigma, o)$ and $(\pi_k^\star, \sigma, o) \mapsto \frac{\pi(\sigma)}{\eta} \cdot g_k(\pi_k^\star; \sigma, o)$ both belong to $\mathcal{G}_{k,\alpha}$, meaning that the above quantity is upper bounded by

$$\inf_{g_k \in \mathcal{G}_{k,\eta\alpha}} \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{o|\sigma} \left[\mathbb{E}_{\pi_k' \sim q_k} \left[\exp(g_k(\pi_k'; \sigma, o)) \right] \cdot \mathbb{E}_{\pi_k^\star \sim \mu_{\mathrm{po}}^k(\cdot; \sigma, o)} [\exp(-g_k(\pi_k^\star; \sigma, o))] - 1 \right].$$

This expression is equal to

$$V_k(\pi, q, \mu) := \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{o|\sigma} \inf_{g_k : \Pi_k' \to \mathbb{R}, \|g_k\|_\infty \leq \alpha\eta} \left\{ \mathbb{E}_{\pi_k' \sim q_k} [\exp(g_k(\pi_k'))] \cdot \mathbb{E}_{\pi_k^\star \sim \mu_{\mathrm{po}}^k(\cdot; \sigma, o)} [\exp(-g_k(\pi_k^\star))] - 1 \right\}.$$

By Lemma H.3, we have that for all π, q, μ,

$$V_k(\pi, q, \mu) = - \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{o|\sigma} \sup_{g_k : \Pi_k' \to \mathbb{R}, \|g_k\|_\infty \leq \alpha\eta} \left\{ -\mathbb{E}_{\pi_k' \sim q_k} [\exp(g_k(\pi_k'))] \cdot \mathbb{E}_{\pi_k^\star \sim \mu_{\mathrm{po}}^k(\cdot; \sigma, o)} [\exp(-g_k(\pi_k^\star))] + 1 \right\}$$

$$\leq -\frac{1}{2} \cdot \mathbb{E}_{\sigma \sim \pi} \mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\left(\mu_{\mathrm{po}}^k(\cdot; \sigma, o), q_k \right) \right] + 4e^{-\alpha\eta}. \tag{64}$$

Combining (62), (63), and (64), we obtain the following upper bound:

$$\mathsf{exo}_\eta(\mathscr{M}) \leq \sup_{q\in\prod_{k=1}^K \Delta(\Pi_k')} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi')} \inf_{\pi\in\mathcal{P}_\varepsilon} \left\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi} \left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma)) - f_k^M(\sigma) \right] \right.$$
$$\left. - \frac{1}{2\eta}\sum_{k=1}^K \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\Big(\mu_{\mathrm{po}}^k(\cdot;\sigma,o), q_k\Big) \right] + \frac{4K}{\eta}\cdot e^{-\alpha\eta} \right\}.$$

Since $f_k^M \in [0,1]$ for all k, M and $D_{\mathsf{H}}^2(\cdot,\cdot) \in [0,2]$, it follows that we may replace the $\inf_{\pi\in\mathcal{P}_\varepsilon}$ in the above expression with $\inf_{\pi\in\Pi}$ and pay an additive cost of $K\varepsilon\cdot(1+1/\eta)$, and so

$$\mathsf{exo}_\eta(\mathscr{M}) \leq \sup_{q\in\prod_{k=1}^K \Delta(\Pi_k')} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi')} \inf_{\pi\in\Pi} \left\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi} \left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma)) - f_k^M(\sigma) \right] \right.$$
$$\left. - \frac{1}{2\eta}\sum_{k=1}^K \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\Big(\mu_{\mathrm{po}}^k(\cdot;\sigma,o), q_k\Big) \right] + \frac{4}{\eta}\cdot e^{-\alpha\eta} + K\varepsilon\cdot(1+1/\eta) \right\}.$$

Since the above holds for any $\varepsilon\in(0,1)$ and $\alpha \geq \max\{1,1/\eta\}$, we may take the limits $\varepsilon\to 0, \alpha\to\infty$ to get

$$\mathsf{exo}_\eta(\mathscr{M}) \leq \sup_{q\in\prod_{k=1}^K \Delta(\Pi_k')} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi')} \inf_{\pi\in\Pi} \left\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi} \left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma)) - f_k^M(\sigma) \right] \right.$$
$$\left. - \frac{1}{2\eta}\sum_{k=1}^K \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\Big(\mu_{\mathrm{po}}^k(\cdot;\sigma,o), q_k\Big) \right] \right\}.$$

Next, for any choice of $q_k\in\Delta(\Pi_k')$, we have

$$\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\Big(\mu_{\mathrm{po}}^k(\cdot;\sigma,o), \mu_{\mathrm{pr}}^k\Big) \right] = \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\Big(\mu_{\mathrm{po}}^k(\cdot;\sigma,o), \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}[\mu_{\mathrm{po}}(\cdot;\sigma,o)]\Big) \right]$$
$$\leq 4\cdot\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\Big(\mu_{\mathrm{po}}^k(\cdot;\sigma,o), q_k\Big) \right],$$

where the equality uses that, for $\pi_k'\in\Pi_k'$, $\mu_{\mathrm{pr}}^k(\pi_k') = \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}[\mu_{\mathrm{po}}(\pi_k';\sigma,o)]$ (by Bayes' rule), and the inequality uses Lemma H.4.

Hence, we have

$$\mathsf{exo}_\eta(\mathscr{M}) \leq \sup_{\mu\in\Delta(\mathcal{M}\times\Pi')} \inf_{\pi\in\Pi} \left\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi} \left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma)) - f_k^M(\sigma) \right] - \frac{1}{8\eta}\sum_{k=1}^K \mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma} \left[D_{\mathsf{H}}^2\Big(\mu_{\mathrm{po}}^k(\cdot;\sigma,o), \right.\right.$$
$$= \mathsf{infr}_{1/(8\eta)}(\mathscr{M}),$$

as desired. ■

Lemma L.5. *For any fixed $\eta > 0$, $q\in\prod_{k=1}^K \Delta(\Pi_k')$, $M\in\mathcal{M}$ and $\pi^\star\in\Pi'$, the map $(\pi,g)\mapsto \Gamma_{q,\eta}(\pi,g;\pi^\star,M)$ is jointly convex with respect to $(\pi,g)\in\Pi\times\mathcal{G}$.*

Proof of Lemma L.5. Fix any $\eta, q, M, \pi^\star$ as in the statement of the lemma. Recall the definition of $\Gamma_{q,\eta}(\pi, g; \pi^\star, M)$ in (49). Since convexity is preserved under summation, it suffices to show that, for each k, the map from $\Pi \times \mathcal{G}_k \to \mathbb{R}$, given by

$$
(\pi, g_k) \mapsto \mathbb{E}_{\sigma \sim \pi} \left[f_k^M(U_k(\pi_k^\star, \sigma)) - f_k^M(\sigma) \right]
$$
$$
+ \frac{1}{\eta} \cdot \mathbb{E}_{\sigma \sim \pi, o \sim M(\sigma)} \mathbb{E}_{\pi_k' \sim q_k} \left[\exp \left(\frac{\eta}{\pi(\sigma)} \cdot (g_k(\pi_k'; \sigma, o) - g_k(\pi_k^\star; \sigma, o)) \right) - 1 \right]
$$

is convex. This follows directly from Lemma C.1 of Foster et al. (2022b). ∎

L.5. Putting everything together: Proof of Theorem F.1

The proof of Theorem F.1 is a straightforward consequence of the lemmas proven previously in this section.

Proof of Theorem F.1. Consider an instance $\mathscr{M}$ as in the statement of Theorem F.1. By Lemma L.3 and Lemma L.4, we have that, for any $\eta > 0$,

$$
\mathsf{exo}_\eta(\mathscr{M}) \leq \mathsf{infr}_{1/(8\eta)}(\mathscr{M}) \leq K \cdot \mathsf{r\text{-}dec}^{\mathrm{o}}_{1/(8\eta)}(\mathrm{co}(\mathscr{M})).
$$

On the other hand, Lemma L.1 gives that for any $\eta, \delta > 0$, Algorithm 1 run with the value η gives that with probability at least $1 - \delta$,

$$
\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \leq \mathsf{exo}_\eta(\mathscr{M}) + \frac{2K}{T\eta} \cdot \log \left(\frac{K \cdot \max_k |\Pi_k'|}{\delta} \right).
$$

Minimizing over $\eta > 0$ and substituting $\gamma = 1/(8\eta)$ yields that there is a value of η for which Algorithm 1 yields risk upper bounded as

$$
\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \leq K \cdot \inf_{\gamma > 0} \left\{ \mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathrm{co}(\mathscr{M})) + \frac{16\gamma}{T} \cdot \log \left(\frac{K \cdot \max_k |\Pi_k'|}{\delta} \right) \right\},
$$

which yields the claimed statement of Theorem F.1. ∎

Appendix M. Proofs for lower bounds from Appendix F

M.1. Proof of Proposition F.1

Proof of Proposition F.1. Fix $K \in \mathbb{N}$, and consider the K-player NE instance $\mathscr{M}$ of Example 1.1, where $\mathcal{A}_k = \{1, 2\}$ for each $k \in [K]$. Certainly we have $|\Pi_k'| = |\mathcal{A}_k| = 2$ for all k. By Proposition G.1, we have $\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathscr{M}^{\mathrm{NE}}) \leq O(K/\gamma)$ for all $\gamma > 0$. Finally, Rubinstein (2016) implies that there is no algorithm which draws $2^{o(K)}$ samples (each of which requires querying the true payoff function $a \mapsto (f_1^{M^\star}(a), \ldots, f_K^{M^\star}(a))$ once) and outputs a c_0-approximate Nash equilibrium with probability at least $2/3$, where $c_0 > 0$ is a sufficiently small universal constant; this yields the third claimed statement of Proposition F.1. ∎

M.2. Proof of Theorem F.2

In this section, we prove Theorem F.2. Before proving the result, we introduce some notation that will be useful in the remainder of the section.

- For integers $N \geq N' \geq 0$, we let $\binom{[N]}{N'}$ denote the set of all subsets of $[N] = \{1, 2, \ldots, N\}$ of size N'.

- For positive integers $n \leq n'$ let $[n, n'] = \{n, n+1, \ldots, n'\}$.

- For sets $\mathcal{X}, \mathcal{Y}$, $\mathcal{X} \sqcup \mathcal{Y}$ denotes the *disjoint union* of $\mathcal{X}$ and $\mathcal{Y}$; it is formally defined as $\{(x, 0) \; : \; x \in \mathcal{X}\} \cup \{(y, 1) \; : \; y \in \mathcal{Y}\}$.

- For finite sets $\mathcal{X}, \mathcal{Y}$, we let $\mathcal{X}^{\mathcal{Y}}$ denote the set of all functions $\phi : \mathcal{Y} \to \mathcal{X}$. Note that, in the case of $\mathcal{Y} = [n]$ for some $n \in \mathbb{N}$, the sets $\mathcal{X}^n$ (which is the n-fold product of $\mathcal{X}$) and $\mathcal{X}^{[n]}$ are in bijection. We will at times slightly abuse notation by identifying these two sets.

- For a finite set $\mathcal{X}$, let $\mathrm{Unif}(\mathcal{X})$ denote the uniform distribution over $\mathcal{X}$.

Proof of Theorem F.2. Fix $\epsilon > 0$ and $N \in \mathbb{N}$; by increasing the constant C_0 in the statement of the theorem, it is without loss of generality to assume that N is a multiple of 3. Set $N_1 = N/3$ and $N_2 = 2N/3 = N - N_1$. Define

$$k = N, \qquad q = n = \frac{2k}{\epsilon} = \frac{2N}{\epsilon}, \tag{65}$$

which ensures that $q^k \geq \binom{N}{N_1}$ for sufficiently large N. We write $\mathscr{T}_1 := \binom{[N]}{N_1}$ and $\mathscr{T}_2 := \binom{[N]}{N_2}$. We will now define a random function $\widetilde{\Phi} : \mathscr{T}_1 \cup \mathscr{T}_2 \to [q]^{[n]} \sqcup [q+1, 2q]^{[n+1, 2n]}$ so that $\widetilde{\Phi}$ maps $\mathscr{T}_1$ to $[q]^{[n]}$ and $\mathscr{T}_2$ to $[q+1, 2q]^{[n+1, 2n]}$. We will show that with positive probability, $\widetilde{\Phi}$ satisfies certain conditions.

1. First, let $\Gamma : \mathscr{T}_1 \cup \mathscr{T}_2 \to [q]^k \sqcup [q]^k$ denote a random function, defined as follows: Γ maps $\mathscr{T}_1$ to the first copy of $[q]^k$ (uniformly at random), and $\mathscr{T}_2$ to the second copy of $[q]^k$ (uniformly at random). In particular, for each $\mathcal{S} \in \mathscr{T}_1 \cup \mathscr{T}_2$, $\Gamma(\mathcal{S})$ are independent and chosen uniformly over their respective copies of $[q]^k$.

2. We next define a mapping $\Sigma : [q]^k \sqcup [q]^k \to [q]^{[n]} \sqcup [q+1, 2q]^{[n+1, 2n]}$ which maps the first copy of $[q]^k$ into $[q]^{[n]}$ and the second copy of $[q]^k$ into $[q+1, 2q]^{[n+1, 2n]}$ according to the Reed-Solomon code of Lemma M.6. (Here we have identified each of $[q]^{[n]}$ and $[q+1, 2q]^{[n+1, 2n]}$ with $[q]^n$ in the natural way.)

3. We then set $\widetilde{\Phi} = \Sigma \circ \Gamma$.

We next argue that there is some choice of Γ for which the resulting $\widetilde{\Phi}$ satisfies the following Conditions M.1 and M.2.

Condition M.1. *For each $i \in \{1, 2\}$, for all sets $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_i$ with $\mathcal{T} \neq \mathcal{T}'$, it holds that $d_{\mathrm{Ham}}(\widetilde{\Phi}(\mathcal{T}), \widetilde{\Phi}(\mathcal{T}')) \geq q - k + 1$.*

Condition M.2. *For any subset $\mathcal{Q} \subset [N]$ with $|\mathcal{Q}| \leq \sqrt{N}$,*

$$\forall a_1 \in [n], a_2 \in [q], \quad \mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_1)}\left(\widetilde{\Phi}(\mathcal{T})(a_1) = a_2 | \mathcal{Q} \subset \mathcal{T}\right) \leq 2/q \quad (66)$$

$$\forall a_1 \in [n+1, 2n], a_2 \in [q+1, 2q], \quad \mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_2)}\left(\widetilde{\Phi}(\mathcal{T})(a_1) = a_2 | \mathcal{Q} \subset \mathcal{T}\right) \leq 2/q. \quad (67)$$

To see that there exists such a choice for Γ, we make the following observations.

1. Since $q^k > 10 \cdot \binom{N}{N_1}^2$ whenever N is sufficiently large (by Eq. (65)), with probability at least $1 - \binom{N}{N_1}^2/q^k > 9/10$, the function Γ is injective. Conditioned on being injective, Lemma M.6 gives that Condition M.1 holds, since the action of Σ on each of the copies of $[q]^k$ is defined to be that of a Reed-Solomon code. Thus, Condition M.1 holds with probability at least $9/10$ over the choice of Γ.

2. Consider any fixed choice of $\mathcal{T} \in \mathscr{T}_1$. Note that, for each coordinate $a_1 \in [n]$, the mapping $\mathcal{T} \mapsto \widetilde{\Phi}(\mathcal{T})(a_1) = \Sigma(\Gamma(\mathcal{T}))(a_1)$, for $\mathcal{T} \in \mathscr{T}_1$, is distributed as a uniformly random function from $\mathscr{T}_1 \to [q]$ (with respect to the randomness in Γ). This fact follows from the final sentence of Lemma M.6 and the fact that Γ is a uniformly random function. Thus, by Lemma M.7 with $N_0 = N_1$ and a union bound over all n possible values of a_1, with probability $1 - n \cdot N^{\sqrt{N}+1} \cdot 2^{-\binom{5N/6}{N/6}/(Cq^2)}$ over the choice of Γ, for any subset $\mathcal{Q} \subset [N]$ of size $|\mathcal{Q}| \leq \sqrt{N}$, Eq. (66) holds. Similarly, an application of Lemma M.7 with $N_0 = N_2$ yields that with probability $1 - n \cdot N^{\sqrt{N}+1} \cdot 2^{-\binom{5N/6}{N/6}/(Cq^2)}$ over the choice of Γ, for any subset $\mathcal{Q} \subset [N]$ of size $|\mathcal{Q}| \leq \sqrt{N}$, Eq. (67) holds. Note that our choices of q, N, ϵ ensure that, for the constant C in Lemma M.7, as long as N is sufficiently large,

$$3 \log q \leq 6 \log(N/\epsilon) \leq 12 \log(N) \leq N/6 - C \leq \log\binom{5N/6}{N/6} - C,$$

meaning that it is valid to apply Lemma M.7. Finally, let us note that our choices for N, q ensure that as long as N is sufficiently large,

$$\binom{5N/6}{N/6} > Cq^2 \cdot \left(\log(2n) + \log(N^{\sqrt{N}+1}) + 5\right),$$

and therefore, (66) and (67) hold for all $\mathcal{Q} \subset [N]$ with $|\mathcal{Q}| \leq \sqrt{N}$, with probability at least $1 - 2^{-5}$. In particular, Condition M.2 holds with probability at least $1 - 2^{-5}$ over the random choice of Γ.

Summarizing the above points, with probability at least $1 - 1/10 - 2^{-5} > 0$ over the choice of Γ, Conditions M.1 and M.2 both hold. We pick any such Γ for which both conditions hold, and set $\Phi = \Sigma \circ \Gamma$.

We are now ready to define the 2-player instance $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$.

Policy space. Let $\Pi_1 = \{1, 2, \ldots, 2n\}$ and $\Pi_2 = \{0, 1, \ldots, 2q\}$, and write $\Pi = \Pi_1 \times \Pi_2$ to denote the joint policy space.

Deviation sets and switching functions. The deviation sets Π'_k and switching function U_k are set as in Definition 1.1 to make $\mathscr{M}$ a 2-player NE instance. To be concrete, we have $\Pi'_k = \Pi_k$ for each k, and $U_k(\pi'_k, \pi) = (\pi'_k, \pi_{-k})$.

Model class $\mathcal{M}$. The class $\mathcal{M}$ is indexed by $\mathscr{T}_1 \cup \mathscr{T}_2$. Given a set $\mathcal{T} \in \mathscr{T}_1 \cup \mathscr{T}_2$, we write the corresponding model as $M_\mathcal{T}$. We will often consider the decomposition $\mathcal{M} = \mathcal{M}_1 \sqcup \mathcal{M}_2$, where $\mathcal{M}_1 := \{M_\mathcal{T} : \mathcal{T} \in \mathscr{T}_1\}$ and $\mathcal{M}_2 := \{M_\mathcal{T} : \mathcal{T} \in \mathscr{T}_2\}$. For each $M_\mathcal{T} \in \mathcal{M}$, we need to specify the distributions $o = (r_1, r_2, o_\circ) \sim M_\mathcal{T}(\pi)$, for each $\pi \in \Pi$. To do so, we first define a mapping $\mathcal{B}^\star : \mathscr{T}_1 \cup \mathscr{T}_2 \to \mathcal{P}([2n] \times [2q]) \subset \mathcal{P}(\Pi)$, as follows: recall that Φ maps $\mathscr{T}_1$ to $[q]^{[n]}$ and $\mathscr{T}_2$ to $[q+1, 2q]^{[n+1,2n]}$. Then for $\mathcal{T} \in \mathscr{T}_1 \cup \mathscr{T}_2$, define $\mathcal{B}^\star(\mathcal{T}) \subset [2n] \times [2q] \subset \Pi$ by

$$\mathcal{B}^\star(\mathcal{T}) = \begin{cases} \{(i, \Phi(\mathcal{T})(i)) : i \in [n]\} & : M_\mathcal{T} \in \mathcal{M}_1 \\ \{(i, \Phi(\mathcal{T})(i)) : i \in [n+1, 2n]\} & : M_\mathcal{T} \in \mathcal{M}_2. \end{cases}$$

Note that here we view, for each set $\mathcal{T}$ in the domain of Φ, $\Phi(\mathcal{T})$ as a function mapping either $[n] \to [q]$ (for $M_\mathcal{T} \in \mathcal{M}_1$) or $[n+1, 2n] \to [q+1, 2q]$ (for $M_\mathcal{T} \in \mathcal{M}_2$).

We set the reward space to be $\mathcal{R} = [0, 1]$, and the pure observation space to be $\mathcal{O}_\circ = [N]$. Now, for each $M_\mathcal{T} \in \mathcal{M}$ and $\pi \in \Pi$, the full observation $o = (r_1, r_2, o_\circ) \sim M_\mathcal{T}(\pi)$ is drawn as follows:

- The pure observation $o_\circ \in \mathcal{O}_\circ$ is simply a uniformly random element of the set $\mathcal{T}$.

- The rewards are deterministic, i.e., we have $r_k = f_k^{M_\mathcal{T}}(\pi)$ for each $k \in [K]$, a.s. Moreover, we define

$$f_1^{M_\mathcal{T}}(\pi) = -f_2^{M_\mathcal{T}}(\pi) = \begin{cases} 0 & : \pi \in \Pi_1 \times \{0\} \\ 1 & : \pi \in (\Pi_1 \times \{1, 2, \ldots, 2q\}) \backslash \mathcal{B}^\star(\mathcal{T}) \\ -\delta & : \pi \in \mathcal{B}^\star(\mathcal{T}), \end{cases} \tag{68}$$

where we set $\delta := 10^{-3}$.

Establishing the claimed statements. It is immediate from definition of Π that $|\Pi| = 2n \cdot (2q + 1) = O(N^2/\epsilon^2)$, thus establishing the first claimed statement of the theorem. Next, Lemma M.1 below bounds $\mathsf{r\text{-}dec}_\gamma^\circ(\mathrm{co}(\mathscr{M}))$, establishing the second claimed statement.

Lemma M.1. *For any $\gamma > 0$, It holds that $\mathsf{r\text{-}dec}_\gamma^\circ(\mathrm{co}(\mathscr{M})) \le \epsilon$.*

The proof of Lemma M.1 uses that Φ satisfies Condition M.1. Finally, the third claimed statement is established by the following lemma.

Lemma M.2. *There is a constant $C > 0$ so that the following holds. For any algorithm that has at most $T \le \sqrt{N}/C$ rounds of interaction, there is some model $M^\star \in \mathcal{M}$ so that*

$$\mathbb{E}^{M^\star}[\mathbf{Risk}(T)] = \mathbb{E}^{M^\star}\left[h^{M^\star}(\widehat{\pi})\right] > \delta/100 = 10^{-5}.$$

Recall that above $\widehat{\pi}$ denotes the output policy of the algorithm.

The proof of Lemma M.2 uses that Φ satisfies Condition M.2. It remains to prove Lemmas M.1 and M.2; we do so in the remainder of this section. ∎

Proof of Lemma M.1. For $\overline{M} \in \mathrm{co}(\mathcal{M})$ and $i \in [N]$, let $\overline{M}[i] \in [0, 1]$ denote the probability $\mathbb{P}_{(r_1, r_2, o_\circ) \sim \overline{M}(\pi)}[o_\circ = i]$, for an arbitrary decision $\pi \in \Pi$ (note that the choice of decision does

not affect the distribution over the pure observation $o_\circ$). For any $\overline{M} \in \mathrm{co}(\mathcal{M})$, we define the set $\mathcal{T}(\overline{M}) \subset [N]$ as follows:

$$\mathcal{T}(\overline{M}) := \left\{ i \in [N] \ : \ \overline{M}[i] \geq 1/N \right\},$$

Now fix any $\overline{M} \in \mathrm{co}(\mathcal{M})$. Define $p^\star \in \Delta(\Pi)$ as follows, as a function of $\overline{M}$:

$$p^\star = \begin{cases} \mathrm{Unif}(\{(1,0),\ldots,(n,0)\}) & : |\mathcal{T}(\overline{M})| \geq N/2 \\ \mathrm{Unif}(\{(n+1,0),\ldots,(2n,0)\}) & : |\mathcal{T}(\overline{M})| < N/2. \end{cases}$$

We have that

$$\begin{aligned} \mathsf{r\text{-}dec}_\gamma^\circ(\mathrm{co}(\mathscr{M}),\overline{M}) &\leq \sup_{M \in \mathrm{co}(\mathcal{M})} \mathbb{E}_{\pi \sim p^\star} \left[h^M(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi),\overline{M}(\pi)\big) \right] \\ &= \sup_{M \in \mathrm{co}(\mathcal{M})} \mathbb{E}_{\pi \sim p^\star} \left[\sum_{k=1}^{2} \max_{\pi_k' \in \Pi_k} f_k^M(\pi_k',\pi_{-k}) - f_k^M(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi),\overline{M}(\pi)\big) \right] \\ &= \sup_{M \in \mathrm{co}(\mathcal{M})} \mathbb{E}_{\pi \sim p^\star} \left[\max_{\pi_2' \in \Pi_2} f_2^M(\pi_1,\pi_2') - f_2^M(\pi) - \gamma \cdot D_{\mathsf{H}}^2\big(M(\pi),\overline{M}(\pi)\big) \right], \end{aligned}$$

$$(69)$$

where the final equality follows because for all $M \in \mathrm{co}(\mathcal{M})$ and all π in the support of $p^\star$, $\max_{\pi_1' \in \Pi_1} f_1^M(\pi_1',\pi_2) = 0 = f_1^M(\pi)$.

Fix $\nu \in \Delta(\mathcal{M})$ so that $M = \overline{M}_\nu(\pi) := \mathbb{E}_{M \sim \nu}[M(\pi)]$ attains the supremum in Eq. (69). We consider the following possibilities:

Case 1. Suppose first that $|\mathcal{T}(\overline{M})| \geq N/2$. We consider the following sub-cases:

1. First suppose that $\nu(\mathcal{M}_1) \leq \frac{1}{1+\delta}$, where we recall that $\delta := 10^{-3}$ (Eq. (68)). Then for all $\pi_1 \in [n]$ and $\pi_2 \in [2q]$, it holds that

$$f_2^{\overline{M}_\nu}((\pi_1,\pi_2)) \leq \frac{1}{1+\delta} \cdot \delta - \left(1 - \frac{1}{1+\delta}\right) = 0, \tag{70}$$

since $f_2^{M_\mathcal{T}}((\pi_1,\pi_2))$ is only positive when $\pi \in \mathcal{B}^\star(\mathcal{T})$, which happens with probability at most $\frac{1}{1+\delta}$ under $M_\mathcal{T} \sim \nu$, as $\pi_1 \in [n]$; moreover, when it is positive, it is δ, and when it is not positive, it is -1. Using (70), since for all decisions π in the support of $p^\star$ (which have $\pi_1 \in [n]$ in this sub-case), the expression in (69) is bounded above by 0.

2. Next suppose that there is some model $M_\mathcal{T} \in \mathcal{M}_1$ so that $\nu(M_\mathcal{T}) \geq 14/15$. Thus, we must have $\sum_{j \in \mathcal{T}(\overline{M}) \setminus \mathcal{T}} \overline{M}_\nu[j] \leq \sum_{j \in [N] \setminus \mathcal{T}} \overline{M}_\nu[j] \leq \frac{1}{15}$. On the other hand, since $|\mathcal{T}(\overline{M})| \geq N/2$, we have $\sum_{j \in \mathcal{T}(\overline{M}) \setminus \mathcal{T}} \overline{M}[j] \geq \frac{1}{N} \cdot \frac{N}{6} = 1/6$. Thus, for any decision $\pi \in \Pi$,

$$D_{\mathsf{H}}^2\big(\overline{M}_\nu(\pi),\overline{M}(\pi)\big) \geq (D_{\mathsf{TV}}\big(\overline{M}_\nu(\pi),\overline{M}(\pi)\big))^2 \geq (1/6 - 1/15)^2 = 1/100,$$

Thus, as long as $\gamma \geq 100$, since $f_2^M(\pi_1,\pi_2') \leq 1$ for all π_1,π_2', if we recall that $M = \overline{M}_\nu$ is chosen to maximize the expression in (69), we have that this expression is bounded above by 0.

3. In the remaining case, we must have $\nu(\mathcal{M}_1) \geq \frac{1}{1+\delta}$, yet for each $M_\mathcal{T} \in \mathcal{M}_1, \nu(M_\mathcal{T}) < 14/15$. Suppose for the purpose of contradiction that

$$\mathbb{E}_{\pi \sim p^\star}\left[\max_{\pi_2' \in \Pi_2} f_2^{\overline{M}_\nu}(\pi_1, \pi_2') - f_2^{\overline{M}_\nu}(\pi)\right] = \mathbb{E}_{\pi \sim p^\star}\left[\max_{\pi_2' \in \Pi_2} f_2^{\overline{M}_\nu}(\pi_1, \pi_2')\right] > \epsilon. \quad (71)$$

Write $\mathcal{I} = \{\pi_1 \in [n] : \max_{\pi_2' \in \Pi_2} f_2^{\overline{M}_\nu}(\pi_1, \pi_2') \geq 0\}$; since $\max_{\pi_2' \in \Pi_2} f_2^{\overline{M}_\nu}(\pi_1, \pi_2') \leq 1$ for all π_1, (71) tells us that $|\mathcal{I}| \geq \epsilon n$. By construction, for each $\pi_1 \in \mathcal{I}$, there is at most one value of $\pi_2 \in [q]$ so that $f_2^{\overline{M}_\nu}(\pi_1, \pi_2) \geq 0$; let this value of π_2 be denoted by $\pi_2^\star(\pi_1)$, if such π_2 exists given π_1, and otherwise set $\pi_2^\star(\pi_1) = -1$.

Note that if $f_2^{\overline{M}_\nu}(\pi_1, \pi_2) \geq 0$ for any $\pi_1 \in \Pi_1$, then we must have that $\nu(\{M_\mathcal{T} \in \mathcal{M}_1 : \Phi(\mathcal{T})(\pi_1) = \pi_2\}) \geq \frac{1}{1+\delta} > 1 - \delta$. Therefore, for all π_1, if $\pi_2^\star(\pi_1) > 0$, then

$$\nu(\{M_\mathcal{T} \in \mathcal{M}_1 : \Phi(\mathcal{T})(\pi_1) = \pi_2^\star(\pi_1)\}) > 1 - \delta. \quad (72)$$

For each $M_\mathcal{T} \in \mathcal{M}_1$, define

$$\zeta(\mathcal{T}) := |\{\pi_1 \in \mathcal{I} : \Phi(\mathcal{T})(\pi_1) \neq \pi_2^\star(\pi_1)\}|.$$

We have that

$$|\mathcal{I}| - \sum_{M_\mathcal{T} \in \mathcal{M}_1} \nu(M_\mathcal{T})\zeta(\mathcal{T}) = \sum_{\pi_1 \in \mathcal{I}} \sum_{M_\mathcal{T} \in \mathcal{M}_1} \nu(M_\mathcal{T}) \cdot \mathbb{I}\{\Phi(\mathcal{T})(\pi_1) = \pi_2^\star(\pi_1)\} \geq |\mathcal{I}| \cdot (1 - \delta),$$

where the inequality uses (72). Thus, by Markov's inequality, for some subset $\mathcal{M}_1' \subset \mathcal{M}_1$, it holds that $\nu(\mathcal{M}_1 \backslash \mathcal{M}_1') \leq \sqrt{\delta}$ and for all $M_\mathcal{T} \in \mathcal{M}_1', \zeta(\mathcal{T}) \leq |\mathcal{I}| \cdot \sqrt{\delta}$. Since $\nu(\mathcal{M}_1) \geq 1 - \delta$, it follows that $\nu(\mathcal{M}_1') \geq 1 - \delta - \sqrt{\delta} \geq 1 - 2\sqrt{\delta}$. Since $1 - 2\sqrt{\delta} > 14/15$ by our choice of $\delta = 10^{-3}$, there must be at least two distinct elements of $\mathcal{M}_1'$, which we denote by $M_{\mathcal{T}_1}$ and $M_{\mathcal{T}_2}$.

To proceed, by definition of $\mathcal{M}_1'$, it holds that

$$|\{\pi_1 \in \mathcal{I} : \Phi(\mathcal{T}_1)(\pi_1) = \Phi(\mathcal{T}_2)(\pi_1) = \pi_2^\star(\pi_1)\}| \geq |\mathcal{I}| \cdot (1 - 2\sqrt{\delta}) \geq |\mathcal{I}|/2 \geq n\epsilon/2.$$

It follows that $d_{\mathrm{Ham}}(\Phi(\mathcal{T}_1), \Phi(\mathcal{T}_2)) \leq n - n\epsilon/2 = n(1 - \epsilon/2)$, which contradicts Condition M.1, since $n(1 - \epsilon/2) = \frac{2N}{\epsilon} \cdot (1 - \epsilon/2) < \frac{2N}{\epsilon} - N + 1 = q - k + 1$. Thus, (71) is false, and therefore the expression in (69) corresponding to choosing $M = \overline{M}_\nu$ is bounded above by ϵ.

Case 2. Now suppose that $|\mathcal{T}(\overline{M})| < N/2$. In this case an argument symmetric to that in the case that $|\mathcal{T}(\overline{M})| \geq N/2$ may be applied to establish the same upper bound on the multi-agent DEC. (In particular, the roles of $\mathcal{M}_1, \mathcal{M}_2$ are swapped; the symmetry arises from the fact that sets in $\mathscr{T}_1$ have size $N/3 = N/2 - N/6$ whereas sets in $\mathscr{T}_2$ have size $2N/3 = N/2 + N/6$.) Below we expand on the details for completeness.

1. If $\nu(\mathcal{M}_2) \leq \frac{1}{1+\delta}$, then for all $\pi_1 \in [n+1, 2n]$ and $\pi_2 \in [2q], f_2^{\overline{M}_\nu}((\pi_1, \pi_2)) \leq 0$, meaning that, since for all decisions π in the support of $p^\star, \pi_2 \in [n+1, 2n]$, the expression in (69) is non-positive.

2. Next suppose there is some model $M_{\mathcal{T}} \in \mathcal{M}_2$ so that $\nu(M_{\mathcal{T}}) \geq 14/15$. We must have that $\sum_{j \in \mathcal{T} \setminus \mathcal{T}(\overline{M})} \overline{M}[j] \leq |\mathcal{T} \setminus \mathcal{T}(\overline{M})| \cdot \frac{1}{N}$. On the other hand, since for each $i \in \mathcal{T}$ we have $M_{\mathcal{T}}[i] = 3/(2N)$, we have $\sum_{j \in \mathcal{T} \setminus \mathcal{T}(\overline{M})} \overline{M}_\nu[j] \geq \frac{14}{15} \cdot \frac{3}{2N} \cdot |\mathcal{T} \setminus \mathcal{T}(\overline{M})| \geq \frac{7}{5N} \cdot |\mathcal{T} \setminus \mathcal{T}(\overline{M})|$. Thus, for any $\pi \in \Pi$, since $|\mathcal{T}(\overline{M})| \leq N/2$ and $|\mathcal{T}| = 2N/3$ (as $M_{\mathcal{T}} \in \mathcal{M}_2$),

$$D_{\mathsf{H}}^2\big(\overline{M}_\nu(\pi), \overline{M}(\pi)\big) \geq \big(D_{\mathsf{TV}}\big(\overline{M}_\nu(\pi), \overline{M}(\pi)\big)\big)^2 \geq \left(|\mathcal{T} \setminus \mathcal{T}(\overline{M})| \cdot \left(\frac{7}{5N} - \frac{1}{N}\right)\right)^2$$

$$\geq \left(\frac{N}{6} \cdot \left(\frac{7}{5N} - \frac{1}{N}\right)\right)^2 = \frac{1}{225}.$$

Thus, as long as $\gamma \geq 225$, since $f_2^M(\pi_1, \pi_2') \leq 1$ for all π_1, π_2', the expression in (69) for $M = \overline{M}_\nu$ is bounded above by 0.

3. In the remaining case, we must have $\nu(\mathcal{M}_2) \geq \frac{1}{1+\delta}$, yet for each $M_{\mathcal{T}} \in \mathcal{M}_2$, $\nu(M_{\mathcal{T}}) < 14/15$. In this case, the expression in (69) for $M = \overline{M}_\nu$ is bounded above by ϵ, via an argument identical to the one in Item 3 above where one replaces all intances of $\mathcal{M}_1$ with $\mathcal{M}_2$.

Summarizing, we have shown that (69) is bounded above by ϵ for an arbitrary choice of $\overline{M}$, which completes the proof of the lemma. ∎

Proof of Lemma M.2.   Fix any $T \leq \sqrt{N}/C$ (for a constant C to be specified below), and consider any algorithm $(p,q) = \{(q^t(\cdot|\cdot), p(\cdot|\cdot)\}_{t=1}^T$. Recall that, for any model M, $\mathfrak{H}^T$ denotes the history of interaction between the algorithm (p,q) and the model M, and is defined by $\mathfrak{H}^T = (\pi^1, o^1), \ldots, (\pi^T, o^T)$. $\mathfrak{H}^T$ is associated with the measure space $(\Omega^T, \mathscr{F}^T)$. For each model $M \in \mathcal{M}$, we use the abbreviate $\mathbb{P}^M \equiv \mathbb{P}^{M,(p,q)}$ as the law of $\mathfrak{H}^T$, and write $\mathbb{E}^M$ for the corresponding expectation. We will show the stronger statement that the algorithm (p,q) has large risk for a *uniformly random* model $M^\star \in \mathcal{M}$; in particular,

$$\mathbb{E}_{M^\star \sim \mathrm{Unif}(\mathcal{M})} \mathbb{E}^{M^\star}\left[h^{M^\star}(\widehat{\pi})\right] > \delta/100. \tag{73}$$

Clearly (73) implies the statement of Lemma M.2.

In order to prove Lemma M.2, we first prove a few intermediate results. To start, we define an additional model M_0: the distribution of $(r_1, r_2, o_\circ) \sim M_0(\pi)$ are as follows:

- The rewards r_1, r_2 are given as in (68) with $\mathcal{B}^\star = \varnothing$; in particular, $r_k = f_k^{M_0}(\pi)$ are deterministic with

$$f_1^{M_0}(\pi) = -f_2^{M_0}(\pi) = \begin{cases} 0 & : \pi \in \Pi_1 \times \{0\} \\ 1 & : \pi \in \Pi_2 \times \{1, 2, \ldots, 2q\}. \end{cases}$$

- The pure observation $o_\circ \in [N]$ is a uniformly random element of $[N]$.

Next, recall that we write, for $\pi \in \Pi, k \in \{1, 2\}, M \in \mathcal{M}, h_k^M(\pi) = \max_{\pi_k' \in \Pi_k'} f_k^M(U_k(\pi_k', \pi)) - f_k^M(\pi)$.

Lemma M.3 below shows that for each $i \in \{1, 2\}$, under the model M_0, with constant probability either all models in $\mathcal{M}_1$ or all models in $\mathcal{M}_2$ have high risk with respect to the algorithm's output policy $\widehat{\pi}$.

Lemma M.3. *There is some $i \in \{1, 2\}$ (depending on the algorithm (p, q)) so that*

$$\mathbb{P}^{M_0} \left(\forall M \in \mathcal{M}_i : h^M(\widehat{\pi}) \geq \delta \right) \geq \frac{1}{2}.$$

The proof of Lemma M.3 is provided at the end of this section. Since M_0 is not in $\mathcal{M}$, Lemma M.3 is not enough to prove Lemma M.2; we will next use a series of change-of-measure arguments to reason about the history of interaction when the true model is a uniformly random model in M. In particular, for each model $M_\mathcal{T} \in \mathcal{M}$, we define an intermediate model $M_{\mathcal{T},0}$: the distribution of $(r_1, r_2, o_\circ) \sim M_{\mathcal{T},0}(\pi)$ is as follows:

- The rewards (r_1, r_2) are given identically to the rewards under $M_0(\pi)$ (in particular, they are deterministic).

- The pure observation $o_\circ$ is a uniformly random element of $\mathcal{T}$.

Lemma M.4 below shows that under a history drawn from $M_{\mathcal{T},0}$ for a uniformly random $\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)$, with high probability the algorithm will not query any decision belonging to $\mathcal{B}^\star(\mathcal{T}) \subset \Pi$; furthermore, the distribution of the history $\mathfrak{H}^T$ is close under M_0 and under $M_{\mathcal{T},0}$, again for a uniformly random $\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)$:

Lemma M.4. *For each $i \in \{1, 2\}$, the following holds:*

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_{\mathcal{T},0}} \left[\mathbb{I} \left\{ \{\pi^1, \dots, \pi^T\} \cap \mathcal{B}^\star(\mathcal{T}) \neq \varnothing \right\} \right] \leq \frac{2T}{q} + \frac{1}{100}. \tag{74}$$

Furthermore, for any measurable subset $\mathcal{F} \in \mathscr{F}^T$ of histories,

$$\left| \mathbb{E}^{M_0} \left[\mathbb{I} \left\{ \mathfrak{H}^T \in \mathcal{F} \right\} \right] - \mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_{\mathcal{T},0}} \left[\mathbb{I} \left\{ \mathfrak{H}^T \in \mathcal{F} \right\} \right] \right| \leq \frac{1}{100}. \tag{75}$$

The proof of Lemma M.4 is provided at the end of this section.

Next, Lemma M.5 shows that if, for some model $M_\mathcal{T}$, the algorithm does not query any decision in $\mathcal{B}^\star(\mathcal{T})$ with high probability, then the distribution of histories under $\mathbb{P}^{M_{\mathcal{T},0}}$ and $\mathbb{P}^{M_\mathcal{T}}$ are close.

Lemma M.5. *Fix some model $M_\mathcal{T} \in \mathcal{M}$ so that $\mathbb{P}^{M_{\mathcal{T},0}}(\{\pi^1, \dots, \pi^T\} \cap \mathcal{B}^\star(M_\mathcal{T}) \neq \varnothing) \leq \eta$ for some $\eta > 0$. Then $D_{\mathsf{TV}}(\mathbb{P}^{M_{\mathcal{T},0}}, \mathbb{P}^{M_\mathcal{T}}) \leq \eta$.*

The proof of Lemma M.5 is provided at the end of this section. Given the above lemmas, we now establish (73). Suppose for the purpose of contradiction that $\mathbb{E}_{M \sim \mathrm{Unif}(\mathcal{M})} \mathbb{E}^M \left[h^M(\widehat{\pi}) \right] \leq \delta/100$. Then by Markov's inequality, $\mathbb{E}_{M \sim \mathrm{Unif}(\mathcal{M})} \mathbb{E}^M \left[\mathbb{I} \left\{ h^M(\widehat{\pi}) \geq \delta \right\} \right] \leq 1/100$. Since $\mathrm{Unif}(\mathcal{M})$ is the uniform average of $\mathrm{Unif}(\mathcal{M}_1)$ and $\mathrm{Unif}(\mathcal{M}_2)$, it follows that for each $i \in \{1, 2\}$,

$$\mathbb{E}_{M \sim \mathrm{Unif}(\mathcal{M}_i)} \mathbb{E}^M \left[\mathbb{I} \left\{ h^M(\widehat{\pi}) \geq \delta \right\} \right] \leq 1/50. \tag{76}$$

We next note that Lemma M.3 gives that for some $i^\star \in \{1, 2\}$,

$$\mathbb{P}^{M_0} \left(\forall M \in \mathcal{M}_{i^\star} : h^M(\widehat{\pi}) \geq \delta \right) \geq \frac{1}{2}.$$

By the conclusion (75) of Lemma M.4, it follows that

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})} \mathbb{E}^{M_{\mathcal{T},0}} \left[\mathbb{I} \{ \forall M \in \mathcal{M}_{i^\star} : h^M(\widehat{\pi}) \geq \delta \} \right] \geq 1/2 - 1/100. \tag{77}$$

111

Next, by the statement (74) of Lemma M.4 and using that $2T \leq \sqrt{N}$ and $\sqrt{N}/q \leq 1/\sqrt{N} \leq 1/100$ for sufficiently large N,

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})} \mathbb{E}^{M_{\mathcal{T}},0} \left[\mathbb{I} \left\{ \{\pi^1, \ldots, \pi^T\} \cap \mathcal{B}^\star(M_{\mathcal{T}}) \neq \varnothing \right\} \right] \leq \frac{\sqrt{N}}{q} + \frac{1}{100} \leq \frac{1}{50}. \tag{78}$$

Now, for $\eta = 1/7$, let us write $\chi(\mathcal{T}) := \mathbb{I} \{\mathbb{E}^{M_{\mathcal{T}},0} [\mathbb{I}\{\{\pi^1, \ldots, \pi^T\} \cap \mathcal{B}^\star(M_{\mathcal{T}}) \neq \varnothing\}] > \eta\} \in \{0,1\}$; Eq. (78) together with Markov's inequality give that $\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})}[\chi(\mathcal{T})] \leq 1/7$.

Next, Lemma M.5 gives that, for all $\mathcal{T} \in \mathscr{T}_1 \cup \mathscr{T}_2$,

$$\mathbb{P}^{M_{\mathcal{T}}} \left(\forall M \in \mathcal{M}_{i^\star} : h^M(\widehat{\pi}) \geq \delta \right) \geq \mathbb{P}^{M_{\mathcal{T}},0} \left(\forall M \in \mathcal{M}_{i^\star} : h^M(\widehat{\pi}) \geq \delta \right) - \chi(\mathcal{T}) - \eta,$$

and taking expectation over $\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})$ and using that $\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})}[\chi(\mathcal{T})] \leq 1/7$ and the choice of $\eta = 1/7$ gives that

$$\mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})} \mathbb{E}^{M_{\mathcal{T}}} \left[\mathbb{I} \left\{ \forall M \in \mathcal{M}_{i^\star} : h^M(\widehat{\pi}) \geq \delta \right\} \right]$$
$$\geq \mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})} \mathbb{E}^{M_{\mathcal{T}},0} \left[\mathbb{I} \left\{ \forall M \in \mathcal{M}_{i^\star} : h^M(\widehat{\pi}) \geq \delta \right\} \right] - 2/7 \geq 1/2 - 1/100 - 2/7,$$

where the final inequality follows by Eq. (77). In particular, using that $M_{\mathcal{T}} \in \mathcal{M}_{i^\star}$ if $\mathcal{T} \in \mathscr{T}_{i^\star}$, we have

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_{i^\star})} \mathbb{E}^{M_{\mathcal{T}}} \left[\mathbb{I} \left\{ h^{M_{\mathcal{T}}}(\widehat{\pi}) \geq \delta \right\} \right] \geq 1/2 - 1/100 - 2/7 > 1/5,$$

which contradicts Eq. (76), thus completing the proof. ■

Proof of Lemma M.3. We write $\widetilde{\Pi} := \Pi_1 \times (\Pi_2 \setminus \{0\}) = \Pi_1 \times \{1, 2, \ldots, 2q\} \subset \Pi$. First, we claim that for all $\pi \in \widetilde{\Pi}$, and all $M \in \mathcal{M}$, it holds that $h^M(\pi) = h_1^M(\pi) + h_2^M(\pi) \geq \delta$. To see this, consider any $M = M_{\mathcal{T}} \in \mathcal{M}$, and we consider the following two cases:

- If $\pi \notin \mathcal{B}^\star(\mathcal{T})$, then $f_2^M((\pi_1, 0)) - f_2^M(\pi) = 0 - (-1) = 1$.

- If $\pi \in \mathcal{B}^\star(\mathcal{T})$, then there must be some $\pi_1' \in \Pi_1$ with $(\pi_1', \pi_2) \notin \mathcal{B}^\star(\mathcal{T})$, and so $f_1^M((\pi_1', \pi_2)) - f_1^M(\pi) = 1 - (-\delta) = 1 + \delta$.

Next, note that

$$\max \left\{ \mathbb{P}^{M_0} \left(\widehat{\pi} \in \widetilde{\Pi} \cup ([n] \times \{0\}) \right), \mathbb{P}^{M_0} \left(\widehat{\pi} \in \widetilde{\Pi} \cup ([n+1, 2n] \times \{0\}) \right) \right\} \geq 1/2.$$

Let us first suppose that $\mathbb{P}^{M_0} \left(\widehat{\pi} \in \widetilde{\Pi} \cup ([n] \times \{0\}) \right) \geq 1/2$. Note that if $M = M_{\mathcal{T}} \in \mathcal{M}_1$ and $\pi \in [n] \times \{0\}$, then $h^M(\pi) \geq h_2^M(\pi) = f_2^M((\pi_1, \Phi(\mathcal{T})(\pi_1))) - f_2^M(\pi) = \delta - 0 = \delta$. Moreover, the two bullet points above establish that if $\widehat{\pi} \in \widetilde{\Pi}$, then $h^M(\widehat{\pi}) \geq 1 > \delta$. Thus, in this case, we have established that $\mathbb{P}^{M_0} (\forall M \in \mathcal{M}_1, h^M(\widehat{\pi}) \geq \delta) \geq 1/2$.

In the other case, where $\mathbb{P}^{M_0} \left(\widehat{\pi} \in \widetilde{\Pi} \cup ([n+1, 2n] \times \{0\}) \right) \geq 1/2$, it follows inb a symmetric manner that, $\mathbb{P}^{M_0} (\forall M \in \mathcal{M}_2, h^M(\widehat{\pi}) \geq \delta) \geq 1/2$. ■

Proof of Lemma M.4. Fix any $i \in \{1, 2\}$. For a model $M \in \{M_0\} \cup \bigcup_{\mathcal{T} \in \mathscr{T}_i} \{M_{\mathcal{T},0}\}$, consider a draw of $\mathfrak{H}^T = (\pi^1, (r_1^1, r_2^1, o_\circ^1), \ldots, \pi^T, (r_1^T, r_2^T, o_\circ^T)) \sim \mathbb{P}^M$, where we have written out the full

observations $o^t = (r_1^t, r_2^t, o_\circ^t)$. Since the distribution of the pure observations $o_\circ^t \sim M(\pi)$ does not depend on the policy π, the distribution of $\mathfrak{H}^T$ is identical to the following one: first, $o_\circ^1, \ldots, o_\circ^T$ are drawn i.i.d. from $M(\pi_0)$ (for an arbitrary decision π_0), and then the decisions π^t are chosen adaptively, $\pi^t \sim q^t(\cdot | \mathfrak{H}^{t-1})$, with the rewards r_1^t, r_2^t being determined by π^t.

For any $\mathcal{T} \in \mathscr{T}_i$, and for any $t, t' \in [T]$ with $t \neq t'$, we have $\mathbb{P}^{M_{\mathcal{T},0}}[o_\circ^t = o_\circ^{t'}] = 1/|\mathcal{T}| \leq 3/N$. Thus

$$\mathbb{P}^{M_{\mathcal{T},0}} \left(\exists t \neq t' \ : \ o_\circ^t = o_\circ^{t'} \right) \leq T^2 \cdot 3/N \leq 1/100, \tag{79}$$

where the final inequality follows since $T \leq \sqrt{N/300}$ (as long as the constant C in the statement of Lemma M.2 is sufficiently large). Let $\mathcal{E} \in \mathscr{F}^T$ denote the event that for all $t \neq t'$, $o_\circ^t \neq o_\circ^{t'}$. The inequality (79) gives that

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_{\mathcal{T},0}} \left[\mathbb{I}\{\mathcal{E}\} \right] \geq 1 - 1/100. \tag{80}$$

In a similar manner, we also have that

$$\mathbb{E}^{M_0}[\mathbb{I}\{\mathcal{E}\}] \geq 1 - 1/100. \tag{81}$$

Now, we may compute

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_{\mathcal{T},0}} \left[\mathbb{I}\{\pi^t \in \mathcal{B}^\star(\mathcal{T})\} \mid \mathcal{E} \right]$$

$$= \sum_{\mathcal{T} \in \mathscr{T}_i} \frac{1}{|\mathscr{T}_i|} \sum_{\omega_1, \ldots, \omega_T \in [N]} \mathbb{P}^{M_{\mathcal{T},0}} \left(o_\circ^{1:T} = \omega_{1:T} \mid \mathcal{E} \right) \cdot \mathbb{E}^{M_{\mathcal{T},0}} \left[\mathbb{I}\{\pi^t \in \mathcal{B}^\star(\mathcal{T})\} \mid o_\circ^{1:T} = \omega_{1:T} \right]$$

$$= \sum_{\mathcal{T} \in \mathscr{T}_i} \frac{1}{|\mathscr{T}_i|} \sum_{\substack{\omega_1, \ldots, \omega_T \in \mathcal{T} \\ \text{distinct}}} \frac{1}{N_i(N_i-1)\cdots(N_i-T+1)} \cdot \mathbb{E}^{M_0} \left[\mathbb{I}\{\pi^t \in \mathcal{B}^\star(\mathcal{T})\} \mid o_\circ^{1:T} = \omega_{1:T} \right]$$

$$= \sum_{\substack{\omega_1, \ldots, \omega_T \in [N] \\ \text{distinct}}} \frac{1}{N(N-1)\cdots(N-T+1)} \mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_0} \left[\mathbb{I}\{\pi^t \in \mathcal{B}^\star(\mathcal{T})\} \mid o_\circ^{1:T} = \omega_{1:T}, \ \{\omega_1, \ldots, \omega_T\} \subset \mathcal{T} \right]$$

$$\tag{82}$$

$$\leq 2/q,$$

where:

- The second equality uses that the distribution of $\mathfrak{H}^T$ conditioned on $o^{1:T}$ is identical under $\mathbb{P}^{M_0}$ and $\mathbb{P}^{M_{\mathcal{T},0}}$.

- The third equality switches the order of summation and uses that $1/|\mathscr{T}_i| = \binom{N}{N_i}^{-1} = \frac{N_i!}{N(N-1)\cdots(N-N_i+1)}$, as well as the fact that the number of sets $\mathcal{T}$ containing any tuple $\omega_1, \ldots, \omega_T \in [N]$ of distinct integers is $\frac{(N-T)(N-T-1)\cdots(N-N_i+1)}{(N_i-T)!}$.

- The final inequality uses the fact that, for fixed $\omega_1, \ldots, \omega_T$, the distribution of $\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i) | \{\omega_1, \ldots, \omega_T\} \subset \mathcal{T}$ is independent of the distribution of $\mathfrak{H}^T \sim \mathbb{P}^{M_0} | o_\circ^{1:T} = \omega_{1:T}$. Moreover, the definition of $\mathcal{B}^\star(\mathcal{T})$ in terms of $\Phi(\mathcal{T})$ and the fact Φ satisfies Condition M.2 means that, for any fixed $\pi = (\pi_1, \pi_2) \in \Pi$ with $\pi_1 > 0$,

$$\mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)}(\pi \in \mathcal{B}^\star(\mathcal{T}) | \{\omega_1, \ldots, \omega_T\} \subset \mathcal{T}) = \mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)}(\Phi(\mathcal{T})(\pi_1) = \pi_2 | \{\omega_1, \ldots, \omega_T\} \subset \mathcal{T}) \leq 2/q,$$

where we take $\Phi(\mathcal{T})(\pi_1) = -1$ if π_1 is not in the domain of $\Phi(\mathcal{T})$. (Here we have also used that $T \leq \sqrt{N}$.) In particular, the above inequality holds with the random choice of $\pi^t \sim \mathbb{P}^{M_0}|o_{\circ}^{1:T} = \omega_{1:T}$ replacing π.

Taking a union bound over all T values of $t \in [T]$ and applying (80), the first claim (74) of the lemma follows.

To show the second claim (75) of the lemma, we note that for any fixed subset $\mathcal{F} \in \mathscr{F}^T$ (not depending on $\mathcal{T}$) the chain of equalities ending in (82) implies that

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_{\mathcal{T},0}}[\mathbb{I}\{\mathfrak{H}^T \in \mathcal{F}\} \mid \mathcal{E}] = \mathbb{E}^{M_0}[\mathbb{I}\{\mathfrak{H}^T \in \mathcal{F}\} \mid \mathcal{E}],$$

Eq. (75) follows from the above equality combined with (80) and (81). ∎

Proof of Lemma M.5. Let $\mathcal{E}$ denote the event that $\{\pi^1, \ldots, \pi^T\} \cap \mathcal{B}^\star(M_{\mathcal{T}}) = \varnothing$. Consider any subset $\mathcal{F} \subset \mathscr{F}^T$ of histories. Then

$$\mathbb{P}^{M_{\mathcal{T},0}}(\mathcal{E} \cap \mathcal{F}) = \mathbb{P}^{M_{\mathcal{T}}}(\mathcal{E} \cap \mathcal{F}),$$

which follows since for any decition $\pi \notin \mathcal{B}^\star(M_{\mathcal{T}})$, the distribution over the full observation $o \sim M_{\mathcal{T}}(\pi)$ and $o \sim M_{\mathcal{T},0}(\pi)$ is identical. The statement of the lemma then follows from Lemma H.6. ∎

M.3. Supplementary lemmas

The following lemma, which is an elementary fact from coding theory, states the dimension and distance properties of the *Reed-Solomon code*. To present it, we recall the definition of Hamming distance: for $q, n \in \mathbb{N}$, and $w, w' \in [q]^n$, we let $d_{\mathrm{Ham}}(w, w') = |\{i \in [n] : w_i \neq w'_i\}|$ to be the number of positions at which w, w' differ.

Lemma M.6 (Reed-Solomon code; Section 5.2 of Guruswami et al. (2022)). *Fix any integers n, q, k satisfying $q \geq k$. Then there is a mapping $\Phi : [q]^k \to [q]^n$ so that for any two vectors $v, v' \in [q]^k$ with $v \neq v'$, it holds that $d_{\mathrm{Ham}}(\Phi(v), \Phi(v')) \geq n - k + 1$.*

Furthermore, Φ may be chosen so that if $X \in [q]^k$ is uniformly random, then for each $i \in [n]$, the value $\Phi(X)_i \in [q]$ is uniformly random.

Lemma M.7 below shows that a certain type of *randomness extractor* exists.

Lemma M.7. *There is a sufficiently large constant $C \geq 1$ so that the following holds. Consider any positive integers N, N_0, R, q with $N_0 \leq 2N/3$, $R \leq N/6 \leq N_0 - N/6$, and $3 \log q \leq \log \binom{5N/6}{N/6} - C$. Let $\Psi : \binom{[N]}{N_0} \to [q]$ be a uniformly random function. Then with probability at least $1 - N^{R+1} \cdot 2^{-\binom{5N/6}{N/6}/(Cq^2)}$ over the choice of Ψ, for all subsets $\mathcal{Q} \subset [N]$ of size $|\mathcal{Q}| \leq R$, and all $j \in [q]$,*

$$\mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}\binom{[N]}{N_0}}(\Psi(\mathcal{T}) = j | \mathcal{Q} \subset \mathcal{T}) \leq \frac{2}{q}.$$

We clarify that the distribution of the uniformly random function $\Psi : \binom{[N]}{N_0} \to [q]$ in the above lemma statement is given as follows: for each $\mathcal{S} \in \binom{[N]}{N_0}$, $\Psi(\mathcal{S})$ is an independent random variable, distributed uniformly on $[q]$.

Proof of Lemma M.7. Since $R \leq N/6 \leq N_0 - N/6$ and $N_0 \leq 2N/3$, for any subset $\mathcal{Q} \subset [N]$ of size $|\mathcal{Q}| \leq R$, the distribution of $\mathcal{T} \sim \mathrm{Unif}\, \binom{[N]}{N_0}|\mathcal{Q} \subset \mathcal{T}$ puts mass at most $1/\binom{5N/6}{N/6}$ on any subset $\mathcal{T}$ (such a distribution is known as a *flat k-source* for some $k \geq \log \binom{5N/6}{N/6}$). By Vadhan (2012, Proposition 6.12) with $\varepsilon = 1/q$, for a sufficiently large constant C, as long as $3 \log q \leq \log \binom{5N/6}{N/6} - C$, with probability at least $1 - 2^{-\binom{5N/6}{N/6}/(Cq^2)}$ over the choice of Ψ, it holds that, for any fixed $\mathcal{Q}$ of size at most R, the distribution of $\Psi(\mathcal{T})$, with $\mathcal{T} \sim \mathrm{Unif}\, \binom{[N]}{N_0}|\mathcal{Q} \subset \mathcal{T}$, is $1/q$-close (in total variation distance) to uniform on $[q]$, which in particular implies that $\Psi(\mathcal{T}) = j$ with probability at most $2/q$ for any $j \in [q]$ (again under $\mathcal{T} \sim \mathrm{Unif}\, \binom{[N]}{N_0}|\mathcal{Q} \subset \mathcal{T}$).

Taking a union bound over all $\sum_{r=0}^{R} \binom{N}{R} \leq N^{R+1}$ possible sets $\mathcal{Q}$, we obtain that Ψ satisfies the desired property with probability at least $1 - N^{R+1} \cdot 2^{-\binom{5N/6}{N/6}/(Cq^2)}$. ∎