On the Role of Information Structure in Reinforcement Learning for Partially-Observable Sequential Teams and Games

Awni Altabaa, Zhuoran Yang

Department of Statistics & Data Science, Yale University

Last modified: December 25, 2023

Abstract

In a sequential decision-making problem, the *information structure* is the description of how events in the system occurring at different points in time affect each other. In particular, for each system variable, the information structure describes the subset of past events which affect it directly. Classical models of reinforcement learning (e.g., MDPs, POMDPs, Dec-POMDPs, and POMGs) assume a very simple and highly regular information structure, while more general models like predictive state representations don't explicitly model the information structure. Real-world sequential decision-making problems typically involve a complex and time-varying interdependence of system variables, requiring a rich and flexible representation of information structure. The control community has long recognized the importance of information structure, leading to the development of the celebrated Witsenhausen intrinsic model, and many impactful works since then. In this paper, we argue for the perspective that explicit representation information structures is an important component of analyzing and solving reinforcement learning problems. Taking inspiration from the control literature, we propose *partially-observable sequential teams* and *partially-observable sequential games* as reinforcement learning models with explicit representation of information structure, capturing classical models of reinforcement learning as special cases. We characterize the relationship between the rank of the observable dynamics of a sequential decision-making problem and its information structure through a graph-theoretic quantity of its DAG representation. This gives a clear and interpretable condition in terms of information structure for the tractability of a sequential decision-making problem. Finally, we propose provably sample-efficient algorithms for learning partially-observable sequential teams and games through a parameterization in a generalized predictive state representation.

1 Introduction

The *information structure* of a sequential decision-making problem is a description of how events in the system occurring at different points in time affect each other. In particular, in a causal sequential system, the information structure describes the subset of past events which have a direct effect on the present. This includes the information available to each agent at each time that they take an action as well as the information that affects the dynamics of the system. The control community has long recognized the importance of information structure, leading to the development of the celebrated Witsenhausen intrinsic model (H. S. Witsenhausen 1975), and extensive study since the 1970s (Hans S Witsenhausen 1971; Ho et al. 1972; Yoshikawa 1978; H. S. Witsenhausen 1988; Andersland and Teneketzis 1992; Teneketzis 1996; S. C. Tatikonda 2000; Mahajan and S. Tatikonda 2009; Mahajan and Teneketzis 2009; Mahajan, Martins, et al. 2012; Nayyar et al. 2014; Malikopoulos 2022; Nayyar et al. 2011).

In contrast to the control literature, reinforcement learning has so far primarily studied problems where the information structure is either fixed and highly-regular, or not explicitly considered. Classicaly, most research efforts have focused on fully observable models, such as MDPs in the single-agent setting or Markov teams/games in the multi-agent setting. Partially-observable models have received more attention in recent years, with significant progress being made. The models typically considered in the partially-observable setting are POMDPs in the single-agent setting or Dec-POMDPs/POMGs in the multi-agent settings. The highly-regular information structures of these models enable them to be studied more fruitfully and enable favorable learning results (e.g., Singh et al. 2000; Sutton et al. 2008; Munos and Szepesvári 2008; Abbasi-Yadkori and Szepesvári 2011; Lattimore and Hutter 2012).

In this paper, we argue for the perspective that information structure is an important component of analyzing and solving reinforcement learning problems. A rich and flexible representation of information structure is needed to faithfully represent real-world sequential decision-making problems. In real-world sequential decision-making problems, different agents will have different information available to them at different time points, and the system evolves according to a complicated and time-varying dependence on subsets of past variables. Moreover, the sequence in which events occur (i.e., the evolution of the system and the actions of different agents) may be irregular and unable to be naturally represented by classical models. Finally, the set of observable system variables which are available to the learning algorithm may also be irregular.

In this work, we present a framework which explicitly models the information structure of sequential decision-making problems and propose sample-efficient online reinforcement learning algorithms via a transformation to a generalized predictive state representation. Our contributions consist of the following:

1) Taking inspiration from the control literature, we introduce *partially-observable sequential teams* (POST) and *partially-observable sequential games* (POSG) as highly general models with an explicit representation of information structure. This forms a unifying framework which captures many commonly studied RL models as special cases, including MDPs, POMDPs, Dec-POMDPs, and POMGs.

2) We characterize the relationship between the rank of the observable dynamics and the information structure through a graph-theoretic quantity of its DAG representation. This gives a clear and interpretable condition in terms of the information structure for when a sequential decision-making problem is tractable or not. Moreover, this recovers known results on the tractability of various structured classes of sequential decision-making problems such as MDPs, POMDPs, Dec-POMDPs, etc.

3) We formalize a generalization of predictive state representations which can be used to learn and represent POSTs and POSGs. This generalized PSR formulation may be of independent interest as many of the analysis techniques (in particular through MLE) which have recently proven successful for standard PSRs (e.g., Liu, Chung, et al. 2022; Huang et al. 2023) carry over to our generalized PSR formulation. We identify a class of POSTs and POSGs for which we explicitly construct a (well-conditioned) generalized PSR representation.

4) Based on recent work (Huang et al. 2023), we propose a sample-efficient (and computationally oracle-efficient) UCB algorithm for generalized PSRs in both the team setting and the game settings. In particular, this allows us to learn to solve POSTs and POSGs.

2 Preliminaries

Notation. We use the convention that upper case letters denote random variables and lowercase letters denote realizations of those random variables (e.g., X_t is the random variable denoting the state at time t and $x_t \in \mathbb{X}_t$ is a particular realization). When clear from context, $\mathbb{P}\left[x_t\right]$ means $\mathbb{P}\left[X_t = x_t\right]$. We will tend to use blackboard symbols to denote the spaces that variables lie in (e.g., $\mathbb{X}_t$ for the space X_t lies in) and calligraphic symbols to denote sets (e.g., $\mathcal{S}$ for the indices of system variables). We use $\mathcal{P}\left(\mathbb{X}\right)$ to denote the space of probability measures on $\mathbb{A}$ and $\mathcal{P}\left(\mathbb{B} \mid \mathbb{A}\right)$ to denote the set of stochastic kernels from $\mathbb{A}$ to $\mathbb{B}$. $i : j$ denotes the set $\{i, i+1, \dots, j\}$. $\sigma_k(A)$ denotes the k-th largest eigenvalue of A.

Consider a sequential decision-making problem with observation index set $\mathcal{O}$ and action index set $\mathcal{A}$. We define the matrix norms $\left\|A\right\|_p = \max_{\left\|x\right\|_p = 1} \left\|Ax\right\|_p$, and $\left\|A\right\|_{\max} = \max_{ij} |A_{ij}|$. $A^\dagger$ is the Moore-Penrose pseudo-inverse. For a (joint) policy $\pi = (\pi_h : h \in \mathcal{A})$, we let $\pi(x_1, \dots, x_h) = \prod_{s \in \mathcal{A}_{1:h}} \pi_s(x_s | x_1, \dots, x_{s-1})$ and $\pi(x_{h+1}, \dots, x_{h'} | x_1, \dots, x_h) = \prod_{s \in \mathcal{A}_{h+1:h'}} \pi_s(x_s | x_1, \dots, x_{s-1})$. We define $\texttt{obs}(\cdot)$ (resp., $\texttt{act}(\cdot)$) as the operator which takes a trajectory as input and returns the observation (resp., action) component. That is, $\texttt{obs}(x_i, \dots, x_j) = (x_s : s \in \mathcal{O} \cap \{i, \dots, j\})$ and $\texttt{act}(x_i, \dots, x_j) = (x_s : s \in \mathcal{A} \cap \{i, \dots, j\})$. For a trajectory $\tau_h = (x_1, \dots, x_h)$, we define $\overline{\mathbb{P}}\left[\tau_h\right] \coloneqq \mathbb{P}\left[\texttt{obs}(\tau_h) \mid \mathrm{do}(\texttt{act}(\tau_h))\right]$ as the probability of the observations in τ_h

given that the actions in τ_h are executed. Note that the probability of a trajectory τ_h under a policy π is $\mathbb{P}^{\pi}(\tau_h) = \overline{\mathbb{P}}\left[\tau_h\right]\pi(\tau_h)$. Finally, we let $\nu_h(\pi, \pi')$ denote the policy which uses π until time h and uses π' from time h onwards.

2.1 (Low-Rank) Sequential Decision-Making Problems

Consider a controlled stochastic process $(X_1, \ldots, X_H)$, where X_h is a random variable corresponding to the variable at time h. At each time $h \in [H]$, the variable X_h may be either an 'observation' (i.e., observable system variable) or an 'action'. The dynamics of this stochastic process are described by a tuple $(H, \{\mathbb{X}_h\}_h, \mathcal{O}, \mathcal{A}, \mathbb{P})$, where H is the time horizon, $\mathbb{X}_h$ is the variable space at time h (i.e., $X_h \in \mathbb{X}_h$), $\mathcal{O} \subset [H]$ is the index set of observations (i.e., X_h is an observation if $h \in \mathcal{O}$), $\mathcal{A} \subset [H]$ is the index set of actions, and $\overline{\mathbb{P}}$ is a probability measure (kernel) which describes the the probability of any trajectory $X_1, \ldots, X_H$ given that the actions are executed. That is, $\overline{\mathbb{P}}\left[X_1, \ldots, X_H\right] = \mathbb{P}\left[\{X_h : h \in \mathcal{O}\} \mid \{\mathrm{do}(X_h) : h \in \mathcal{A}\}\right]$. We assume causality: the present observation is conditionally independent of the future given the past (i.e., $X_h \perp\!\!\!\perp X_{h+1:H} | X_{1:h}$).

We now define some notation. Let $\mathbb{H}_h = \prod_{s \in 1:h} \mathbb{X}_s$ denote the space of histories at time h and $\mathbb{F}_h = \prod_{s \in h+1:H} \mathbb{X}_s$ denote the space futures at time h. Similarly, let $\mathbb{H}_h^o = \mathtt{obs}(\mathbb{H}_h) = \prod_{s \in \mathcal{O}_{1:h}} \mathbb{X}_s$ denote the observation component of histories and let $\mathbb{H}_h^a = \mathtt{act}(\mathbb{H}_h) = \prod_{s \in \mathcal{A}_{1:h}} \mathbb{X}_s$ denote the action component. The observation and action components of the futures, $\mathbb{F}_h^p$ and $\mathbb{F}_h^a$ respectively, are defined similarly.

We define the *system dynamics matrix* $\boldsymbol{D}_h \in \mathbb{R}^{|\mathbb{H}_h| \times |\mathbb{F}_h|}$ as the matrix giving the probability of each possible pair of history and future at time h given the execution of the actions,

$$\left[\boldsymbol{D}_h\right]_{\tau_h, \omega_h} = \overline{\mathbb{P}}\left[\tau_h, \omega_h\right] = \mathbb{P}\left[\tau_h^o, \omega_h^o \mid \mathrm{do}(\tau_h^a, \omega_h^a)\right], \quad \tau_h \in \mathbb{H}_h, \omega_h \in \mathbb{F}_h, \tag{1}$$

where $\omega_h^o = \mathtt{obs}(\omega_h)$ are is the observation component of the future ω_h, $\omega_h^a = \mathtt{act}(\omega_h)$ is the action component, and similarly for τ_h^o, τ_h^a. Note that the actions are actively executed via the do-operation. Hence, the system dynamics matrices are independent of any action-selection criteria. Note that $\boldsymbol{D}_H \in \mathbb{R}^{|\mathbb{H}_H| \times 1}$ is defined as $\left[\boldsymbol{D}_H\right]_{\tau_H} = \overline{\mathbb{P}}\left[\tau_H\right]$, and $\boldsymbol{D}_0 = \boldsymbol{D}_H^\top$.

We introduce the notion of the *rank* of the dynamics. The rank of such a controlled stochastic process is the maximal rank of its dynamics matrices. This is a measure of the complexity of the dynamics.

Definition 1 (Rank of dynamics)**.** *The rank of the dynamics* $\{\boldsymbol{D}_h\}_{h \in [H]}$ *is* $r = \max_{h \in [H]} \mathrm{rank}(\boldsymbol{D}_h)$.

This defines the dynamics of the system. A sequential decision making problem is such a controlled stochastic process together with an *objective*. The objective is defined by a reward function $R : \mathbb{X}_1 \times \cdots \times \mathbb{X}_H \to [0, 1]$ mapping a trajectory to a reward in $[0, 1]$. The agent(s) can affect the dynamics of the system through their choice of actions or policies. Each action $X_h, h \in \mathcal{A}$ may be chosen by either a single agent or one of several agents (e.g., a team). The policy at time $h \in \mathcal{A}$ is a mapping $\pi_h : \mathbb{H}_{h-1} \to \mathcal{P}(\mathbb{X}_h)$ from previous observations to an action (or a distribution over actions, if randomized). The collection of policies at all time steps is denoted $\boldsymbol{\pi} = (\pi_h : h \in \mathcal{A})$, and induces a probability distribution over trajectories, denoted $\mathbb{P}^{\boldsymbol{\pi}}$. Then, the value of a policy $\boldsymbol{\pi}$ is the expected value of the reward under the measure $\mathbb{P}^{\boldsymbol{\pi}}$, $V(\boldsymbol{\pi}) \coloneqq \mathbb{E}^{\boldsymbol{\pi}}\left[R(X_1, \ldots, X_H)\right]$, where $\mathbb{E}^{\boldsymbol{\pi}}$ is the expectation associated with $\mathbb{P}^{\boldsymbol{\pi}}$.

Sequential decision-making problems as defined here capture many widely studied models, including MDPs, POMDPs, Dec-POMDPs, etc. The formalism of sequential decision-making problems introduced in this section is highly generic, but does not explicitly model the *information structure*. In the next section, we introduce the model of *sequential teams*, which explicitly models information structures. Later, in Section 4, we relate sequential teams back to the generic model of low-rank sequential decision-making problems, showing that the information structure of a sequential team implies a bound on the rank of its dynamics as per Definition 1.

2.2 (Generalized) Predictive State Representations

Predictive state representations (PSR) (Littman and Sutton 2001; Jaeger 2000) are a model of dynamical systems and sequential decision-making problems based on predicting future observations given the past, without explicitly modeling a latent state. One advantage of PSRs is that they are amenable to learning, as demonstrated by recent research efforts which propose provably-efficient algorithms (Liu, Chung, et al. 2022; Liu, Netrapalli, et al. 2022; Zhan et al. 2022; Huang et al. 2023).

In the standard formulation of sequential decision-making and predictive state representations, the sequence of observed variables is such that observations and actions always occur in an alternating manner (i.e., $o_t, a_t, o_{t+1}, a_{t+1}, \ldots$). In order to study our more general models with explicit representation of information structure, which will be introduced in the next section, we need a generalization of PSRs to allow for arbitrary order of observations and actions, as well as arbitrary variable spaces at each time point.

The "PSR rank" of a sequential decision-making problem coincides with the rank of its dynamics, as defined in Definition 1. Recall that the system dynamics matrix $\boldsymbol{D}_h \in \mathbb{R}^{|\mathbb{H}_h| \times |\mathbb{F}_h|}$ is indexed by all possible observable histories τ_h and futures ω_h. Denote the rank of the system dynamics at time h by $r_h := \operatorname{rank}(\boldsymbol{D}_h)$.

Definition 2 (Core test sets). *A core test set at time h is a subset of* $d_h \geq r_h$ *futures,* $\mathbb{Q}_h := \left\{q_h^1, \ldots, q_h^{d_h}\right\} \subset \mathbb{F}_h$, *such that the submatrix* $\boldsymbol{D}_h[\mathbb{Q}_h] \in \mathbb{R}^{|\mathbb{H}_h| \times d_h}$ *is full-rank,* $\operatorname{rank}(\boldsymbol{D}_h[\mathbb{Q}_h]) = \operatorname{rank}(\boldsymbol{D}_h) = r_h$.

A core test set implies the existence of a matrix $\boldsymbol{W}_h \in \mathbb{R}^{|\mathbb{F}_h| \times d_h}$ such that,

$$\boldsymbol{D}_h = \boldsymbol{D}_h[\mathbb{Q}_h] \cdot \boldsymbol{W}_h^\top. \tag{2}$$

Denote the τ_h-th row of $\boldsymbol{D}_h[\mathbb{Q}_h]$ by,

$$\psi_h(\tau_h) := \left(\overline{\mathbb{P}}\left[\tau_h, q_h^1\right], \ldots, \overline{\mathbb{P}}\left[\tau_h, q_h^{d_h}\right]\right) \in \mathbb{R}^{d_h}. \tag{3}$$

The vector $\psi_h(\tau_h)$ is a sufficient statistic for the history τ_h in predicting the probabilities of all futures conditioned on τ_h.

For any integer $d_h \geq r_h$, there exists a core test set of size d_h. In particular, for any low-rank sequential decision-making problem, there exists a minimal core test set of size r_h at each h. However, the minimal core test set depends on the system dynamics matrix $\boldsymbol{D}_h$, which is unknown in the learning setting. In the literature on reinforcement learning in general PSRs, it is assumed that a core test set is known. When $d_h > r_h$, the PSR is said to be overparameterized. We address the problem of obtaining a core test set of partially-observable sequential teams in Section 6. For a core test set $\mathbb{Q}_h$, let $\mathbb{Q}_h^A = \{\operatorname{act}(q) : q \in \mathbb{Q}_h\}$, where $\operatorname{act}(q)$ denotes the action components of the test $q \in \mathbb{Q}_h$. Let $Q_A = \max_h \left|\mathbb{Q}_h^A\right|$ and $d = \max_h d_h$.

Definition 3 (Generalized Predictive State Representations). *Consider a sequential decision-making problem* $(X_h \in \mathbb{X}_h)$ *where* $\mathcal{A}, \mathcal{O}$ *partition* $[H]$ *into actions and observations, respectively. Then, a predictive state representation of this sequential decision-making problem is a tuple* $\theta = \left(\{\mathbb{Q}_h\}_{0 \leq h \leq H-1}, \phi_H, \boldsymbol{M}, \psi_0\right)$ *given by*

1. $\{\mathbb{Q}_h\}_{0 \leq h \leq H-1}$ *are core test sets, including for* $h = 0$, *where* $\mathbb{Q}_0 = \left\{q_0^1, \ldots, q_0^{d_0}\right\} \subset \mathbb{F}_0$ *are core tests before the system begins.*
2. $\psi_0 \in \mathbb{R}^{\mathbb{Q}_0}$ *is the vector* $\psi(\emptyset) = \left(\overline{\mathbb{P}}\left[q_0^1\right], \ldots, \overline{\mathbb{P}}\left[q_0^{d_0}\right]\right)$.
3. $\boldsymbol{M} = \{M_h\}_{1 \leq h \leq H-1}$ *is a set of mappings* $M_h : \mathbb{X}_h \to \mathbb{R}^{d_{h+1} \times d_h}$, *from an observable/action to a matrix of size* $d_{h+1} \times d_h$
4. $\phi_H : \mathbb{X}_H \to \mathbb{R}^{d_{H-1}}$ *is a mapping from the final observation to a* d_{H-1}*-dimensional vector.*

This tuple satisfies,

$$\overline{\mathbb{P}}[x_1, \ldots, x_H] = \phi_H(x_H)^\top M_{H-1}(x_{H-1}) \cdots M_1(x_1)\psi_0 \tag{4}$$

$$\psi_h(x_1, \ldots, x_h) = M_h(x_h) \cdots M_1(x_1)\psi_0, \ \forall h \tag{5}$$

To obtain a probability for a trajectory $\tau_h = (x_1, \dots, x_h)$, with $h < H$, note that $\sum_{\omega_h \in \mathbb{F}_h} \overline{\mathbb{P}}\left[\tau_h, \omega_h\right] = |\mathbb{F}_h^a| \, \overline{\mathbb{P}}\left[\tau_h\right]$. Hence,

$$\begin{aligned} \overline{\mathbb{P}}\left[\tau_h\right] &= \frac{1}{|\mathbb{F}_h^a|} \sum_{\omega_h} \overline{\mathbb{P}}\left[\tau_h, \omega_h\right] \\ &= \frac{1}{\prod_{s \in h+1:H} (|\mathbb{X}_s| \, \mathbf{1}\{s \in \mathcal{A}\})} \sum_{x_H} \cdots \sum_{x_{h+1}} \phi_H^\top M_H(x_H) \cdots M_{h+1}(x_{h+1}) \psi_h(\tau_h). \end{aligned}$$

Thus, if we recursively define ϕ_h, $h < H$ via,

$$\frac{1}{|\mathbb{X}_h| \, \boldsymbol{\{} h \in \mathcal{A} \}} \sum_{x_h} \phi_h^\top M_h(x_h) = \phi_{h-1}^\top, \tag{6}$$

with ϕ_H as the terminating condition, then, we can obtain $\overline{\mathbb{P}}\left[\tau_h\right]$ for any $h < H$, via an inner product between ϕ_h and $\psi_h(\tau_h)$,

$$\overline{\mathbb{P}}\left[\tau_h\right] = \phi_h^\top \psi_h(\tau_h). \tag{7}$$

Finally, if we define $\overline{\psi}_h(\tau_h) = \psi_h(\tau_h) / \overline{\mathbb{P}}\left[\tau_h\right]$, then we obtain the *conditional* probability of the core tests given the history, $\overline{\psi}_h(\tau_h) = \Big(\overline{\mathbb{P}}(q_h^1 \,|\, \tau_h), \, \dots, \, \overline{\mathbb{P}}(q_h^{d_h} \,|\, \tau_h)\Big)$. $\overline{\psi}_h(\tau_h)$ is known as the prediction feature of the history τ_h (Littman and Sutton 2001).

An important condition for the learnability of PSR models, which was used in prior work (including Huang et al. 2023; Liu, Netrapalli, et al. 2022), is the so called "well-conditioning assumption". We state the corresponding assumption for our generalized PSR model.

Assumption 1 (γ-well-conditioned generalized PSR)**.** *A PSR model $\theta = \Big(\{\mathbb{Q}_h\}_{0 \le h \le H-1}, \phi_H, \boldsymbol{M}, \psi_0\Big)$, as defined in Definition 3, is said to be γ-well conditioned for $\gamma > 0$ if it satisfies*

1. *For any $h \in [H]$,*
$$\max_{\substack{z \in \mathbb{R}^{d_h} \\ \|z\|_1 \le 1}} \max_{\pi} \sum_{\omega_h \in \mathbb{F}_h} \pi(\omega_h | \tau_h) \left| m_h(\omega_h)^\top z \right| \le \frac{1}{\gamma}, \tag{8}$$
where $m_h(\omega_h)^\top = \phi_H(x_H)^\top M_{H-1}(x_{H-1}) \cdots M_{h+1}(x_{h+1})$ with $\omega_h = (x_{h+1}, \dots, x_H) \in \mathbb{F}_h$. The maximization is over policies π such that for any fixed future observations ω_h^o, $\sum_{\omega_h^a} \pi(\omega_h^o, \omega_h^a) = 1$.

2. *For any $h \in [H-1]$,*
$$\max_{\substack{z \in \mathbb{R}^{d_h} \\ \|z\|_1 \le 1}} \sum_{x_h \in \mathbb{X}_h} \|M_h(x_h) z\|_1 \, \pi(x_h) \le \frac{\left|\mathbb{Q}_{h+1}^A\right|}{\gamma},$$
where $\pi(x_h) = 1$ when $h \notin \mathcal{A}$ and $\sum_{x_h} \pi(x_h) = 1$ when $h \in \mathcal{A}$.

To understand this condition, recall that $m_h(\omega_h)^\top \psi_h(\tau_h) = \overline{\mathbb{P}}\left[\tau_h, \omega_h\right]$. We think of x in Assumption 1 as representing the error in estimating $\psi_h(\tau_h)$, the probabilities of core tests at time h given the history τ_h. Then the γ-well-conditioned assumption ensures that the error in estimating the overall PSR (i.e., the probability of a particular trajectory) does not blow up when the estimation error of the probability of core tests is small.

The following result states that any sequential decision-making problem of the form described in the previous section admits a PSR representation. The proof and explicit construction are given in Appendix A.

Proposition 1. *Let $(X_1, \dots, X_H)$ be any sequential decision-making problem with observation index set $\mathcal{O}$, action index set $\mathcal{A}$, and variable spaces $\{\mathbb{X}_h\}_{h \in [H]}$. Let $r_h = \operatorname{rank}(\boldsymbol{D}_h)$, where $D_h, h \in [H]$ are the system dynamics matrices. Then, there exists a PSR representation ψ_0, $\phi_H : \mathbb{X}_H \to \mathbb{R}^{r_{H-1}}$, $M_h : \mathbb{X}_h \to \mathbb{R}^{r_{h+1} \times r_h}, h \in [H-1]$, satisfying Definition 3.*

3 Information Structures, Sequential Teams, and POSTs

3.1 Sequential Teams

Sequential teams form a highly general model of sequential decision-making problems where information structures are considered explicitly. A sequential team is a controlled stochastic process consisting of a sequence of variables. Each variable is either a "system variable" or an "action variable." Unlike more specialized models of sequential decision-making, there is no restriction on the order of system variables and action variables. The information structure of a sequential team describes the dependence between these variables. The "information set" of a system variable describes the subset of past variables which directly affects it. The information set of an action variable describes the information available to the agent when choosing an action, hence defining the policy class they optimize over.

Definition 4 (Sequential Team Model)**.** *A sequential team is a controlled stochastic process that specifies the joint distribution of T variable $(X_t)_{t\in[T]}$, denoted by $\mathcal{M} \subset \mathcal{P}(X_1, \ldots, X_T)$, where T is a fixed integer. Here each X_t is either a system variable or an action variable. A sequential team is specified by the following components.*

1. ***Variable Structures.*** *The variables $\{X_t\}_{t\in[T]}$ are partitioned into two disjoint subsets — system variables and action variables. $\mathcal{S} \subset [T]$ indexes system variables and $\mathcal{A} \subset [T]$ indexes action variables, with $\mathcal{S} \cap \mathcal{A} = \emptyset$, $\mathcal{S} \cup \mathcal{A} = [T]$.*

2. ***Variable Spaces.*** *Let $\mathbb{X}_t$ be the space that the variable X_t takes values in, which is assumed to be finite for all $t \in [T]$.*

3. ***Information Structure.*** *For $t \in [T]$, the "information set" $\mathcal{I}_t \subset [t-1]$ of the variable X_t is the set of past variables which directly determines the distribution of X_t. Given the "information variable" $I_t := (X_s : s \in \mathcal{I}_t)$, X_t is conditionally independent of the past variables $\{X_s\}_{s\in[t-1]}$. We define $\mathbb{I}_t = \prod_{s\in\mathcal{I}_t} \mathbb{X}_s$ as the "information space" at time t. We denote realizations of I_t by $i_t = (x_s \in \mathbb{X}_s : s \in \mathcal{I}_t) \in \mathbb{I}_t$.*

4. ***System Kernels.*** *For any $t \in \mathcal{S}$, $\mathcal{T}_t$ is a mapping from $\mathbb{I}_t$ to $\mathcal{P}(\mathbb{X}_t)$ that specifies the conditional distribution of a system variable X_t given I_t. That is, $X_t \sim \mathcal{T}_t(\cdot | \{X_s, s \in \mathcal{I}_t\})$ for all $t \in \mathcal{S}$. If $\mathcal{I}_t = \emptyset$ then $\mathcal{T}_t$ is simply a (unconditional) distribution on $\mathbb{X}_t$.*

5. ***Decision Kernels.*** *Each agent chooses a decision kernel (policy) $\pi_t : \mathbb{I}_t \to \mathcal{P}(\mathbb{X}_t)$, specifying the distribution over actions at time $t \in \mathcal{A}$. That is, the action variable X_t at time $t \in \mathcal{A}$ satisfies $X_t \sim \pi_t(\cdot | \{X_s, s \in \mathcal{I}_t\})$. The joint policy is denoted by $\pi = (\pi_t)_{t\in\mathcal{A}}$.*

6. ***Reward Function.*** *At the end of an episode, the team receives the reward $R(x_1, \ldots, x_T)$, where $R : \prod_{t\in[T]} \mathbb{X}_t \to [0, 1]$ is the "reward function."*

Given the definition above, any set of decision kernels (joint policy) $\boldsymbol{\pi}$ induces a unique measure over $\mathbb{X}_1 \times \cdots \times \mathbb{X}_T$, which is given by

$$\mathbb{P}^{\boldsymbol{\pi}}\left[X_1 = x_1, \ldots X_T = x_t\right] = \prod_{t\in\mathcal{A}} \pi_t(x_t | \{x_s : s \in \mathcal{I}_t\}) \prod_{t\in\mathcal{S}} \mathcal{T}_t(x_t | \{x_s : s \in \mathcal{I}_t\}). \tag{9}$$

In the sequel, let $\mathbb{E}^{\boldsymbol{\pi}}$ denote the expectation with respect to $\mathbb{P}^{\boldsymbol{\pi}}$. The value of a policy $\boldsymbol{\pi}$ is defined as the expected value of the reward under $\mathbb{P}^{\boldsymbol{\pi}}$,

$$V(\boldsymbol{\pi}) := \mathbb{E}^{\boldsymbol{\pi}}\left[R(X_1, \ldots, X_T)\right]. \tag{10}$$

A sequential team can be viewed as a partial specification of the distribution on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_T$. That is, in Equation (24), the system kernels $\mathcal{T}_t \in \mathcal{P}\left(\mathbb{X}_t \mid \mathbb{I}_t\right), t \in \mathcal{S}$ are specified, while $\pi_t \in \mathcal{P}\left(\mathbb{X}_t \mid \mathbb{I}_t\right), t \in \mathcal{A}$ are not specified. The task of solving a sequential team is to maximize the expected reward with respect the possible choices of policies,

$$\sup_{\substack{\pi_t \in \mathcal{P}(\mathbb{X}_t \mid \mathbb{I}_t) \\ t\in\mathcal{A}}} \mathbb{E}^{\boldsymbol{\pi}}\left[R(X_1, \ldots, X_T)\right],$$

where $\mathcal{P}(\mathbb{X}_t|\mathbb{I}_t) =: \Gamma_t$ is the policy class for action $t \in \mathcal{A}$ and $\boldsymbol{\Gamma} = \times_{t\in\mathcal{A}}\Gamma_t$ is the joint policy class. When the variable spaces $\mathbb{X}_t$ are finite, this supremum is attained by a deterministic policy $\boldsymbol{\pi} = (\pi_t, t \in \mathcal{A})$, $\pi_t : \mathbb{I}_t \to \mathbb{X}_t$ (H. S. Witsenhausen 1975).

It is sometimes useful to distinguish between the "team form," consisting of the variable structure and information structure $(\mathcal{S}, \mathcal{A}, \{\mathcal{I}_t\}_t)$, and the "team type," which consists of the variable spaces, system kernels, and reward function $\left(\{\mathbb{X}_t\}_t, \{\mathcal{T}_t\}_{t\in\mathcal{S}}, R\right)$ (Mahajan and S. Tatikonda 2011).

In the control literature, there exists a taxonomy of decentralized systems. The model presented here falls within the class of dynamic sequential teams, and allows for non-classical information structures. The sequential team model is closely related Witsenhausen's model. It is a generalization of the model presented in (H. S. Witsenhausen 1988), which is itself equivalent to the intrinsic model of (H. S. Witsenhausen 1975). This is a very general model which captures many standard control problems as a special case, including MDPs, POMDPs, Dec-POMDPs, etc. This model can capture multiple agents acting in arbitrary environments, as long as the order in which agents act is predetermined and independent of the system dynamics—hence the name "sequential team".

In addition to allowing for arbitrary dependence on the past, the sequential team model can capture events occurring simultaneously. This is controlled by the specification of their information sets. For example, to represent m events occurring simultaneously, the corresponding variables can occupy any ordering of consecutive time points, $X_{t+1}, \dots, X_{t+m}$, as long as their information sets do not contain any of the other variables occurring at that time (i.e., $\mathcal{I}_s \subset [t]$ for all $s \in \{t+1, \dots, t+m\}$).

In the control setting, the sequential team model does not distinguish between agents and actions. That is, agents can be identified by the time in which they act, with each action viewed as being associated with a different agent. This is without loss of generality since the underlying 'identity' of an agent (i.e., the same agent acting multiple times) can be captured by the information structure. For example, the information sets can be specified in such a way so that for any $t \in \mathcal{A}$, $\mathcal{I}_t$ contains all variables which were observed by this agent in the past. In the game setting, the identity of the agent matters since it also determines the reward function associated with each action. We discuss this in Section 8.2.

Sequential teams were first formalized and studied in the control setting, where 'decentralization' refers to the information available to each agent when choosing its action, assuming full knowledge of the model. In the learning setting "decentralization" can refer to either decentralization in learning (i.e., the information each agent has available to it when deciding its policy) or decentralization in execution (i.e., the observations available to the agent when making it's decision; its policy class). Existing literature on multi-agent reinforcement learning has used the term "decentralized" in different ways (Zhang et al. 2021). The algorithm we will present here follows the "centralized-learning-decentralized-execution" paradigm, wherein the learning step is carried out in a centralized manner (with the observations of all agents aggregated) but within each episode, the information available to each agent is specified by the information structure.

3.2 Representation of the information structure as a Directed Acyclic Graph

The information structure of a sequential team can be naturally represented as a (labeled) directed acyclic graph (DAG). Given a sequential team form $(\mathcal{S}, \mathcal{A}, \{\mathcal{I}_t\}_t)$, its DAG representation is given by $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{L})$. The nodes of the graph are the set of variables, $\mathcal{V} = [T] = \mathcal{S} \cup \mathcal{A}$. The edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ of the DAG are given by

$$\mathcal{E} = \{(i,t) : t \in [T], i \in \mathcal{I}_t\}\,.$$

That is, there exists an edge from i to t if i is in the information set of t. Finally, $\mathcal{L}$ contains labels for each node as being a system variable (in $\mathcal{S}$) or an action variable (in $\mathcal{A}$). This DAG represents a graphical model for the sequential team. In particular, the probability distribution on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_T$ factors according to $\mathcal{G}$,

$$\mathbb{P}\left[X_1, \dots, X_T\right] = \prod_{t\in\mathcal{V}} \mathbb{P}\left[X_t \mid \mathrm{pa}(X_t)\right], \tag{11}$$

where $\mathrm{pa}(X_t)$ is the set of parents of X_t in $\mathcal{G}$ (which are $\mathcal{I}_t$). This representation of the information structure as a DAG will be crucial for our analysis of the dynamics of sequential teams in Section 4.

3.3 Partially-Observable Sequential Teams (POST)

The sequential team model defines the dynamics of a system and the associated control problem. We now introduce the notion of "observability" to this model in the context of reinforcement learning.

In general, the concept of "observability" plays two roles in reinforcement learning—one in control and another in model-estimation. In the former, observability refers to what information the agents can use to choose their actions, and hence defines their policy class. In the latter, observability refers to what information the learner has available to them when estimating a model of the environment. In the literature, these two aspects are not usually distinguished because they coincide in the models usually considered. For example, in a POMDP, the agents' information sets and the data used for learning both consist of the observation variables and action variables, but not the state variables.

The generality of the sequential team model allows us to generalize the notion of observability. In particular, in sequential teams, there exists T system variables, $X_1, \ldots, X_T$. We can define the "observable" system variables as a subset $\mathcal{O} \subset \mathcal{S}$ of the system variables which are available to the reinforcement learning algorithm. We assume that the system variables in the agents' information sets are always observable. That is, $\mathcal{O} \supset \bigcup_{t \in \mathcal{A}} (\mathcal{I}_t \cap \mathcal{S})$. This is a minimal assumption since we need to observe $i_t \in \mathbb{I}_t$ so that we can compute $\pi(x_t | i_t)$ and perform the planning step.

Hence, a partially-observable sequential team is a sequential team together with $\mathcal{O}$, the set of observable system variables. Then, the set of variables available to the reinforcement learning algorithm are the observable system variables together with the action variables, $\mathcal{U} := \mathcal{O} \cup \mathcal{A}$. We refer to $\mathcal{U}$ as the "observables", distinguished from the observable *system* variables $\mathcal{O}$. Finally, we assume that the reward function R is a function only of variables in $\mathcal{U}$. This is equivalent to assuming that the reward is observed. For example, we can introduce another variable equal to the reward, $X_{T+1} = R(X_1, \ldots, X_T)$, which is observable (see also the remark about the equivalent formulation with explicit reward variables $\mathcal{R}$, where the assumption becomes $\mathcal{R} \subset \mathcal{O}$).

The model-based reinforcement learning algorithm proposed in this paper will model the dynamics of the variables in $\mathcal{U}$. In particular, we will model the joint distribution over trajectories of these variables for each choice of actions.

We abbreviate "partially-observed sequential team" as POST. We now introduce some notation. Denote the unobservable variables by $\mathcal{O}^{\complement} = \mathcal{S} \setminus \mathcal{O}$. Let the time horizon over observable variables be $H := |\mathcal{O} \cup \mathcal{A}|$. When working with the observable variables, it will be more convenient to index variables by their order among observables $h \in [H]$ rather than their order among all variables. The observable variables can be indexed by $h \in [H]$ as follows,

$$\left(X_{t(h)}\right)_{h \in [H]} = \left(X_{t(1)}, \ldots, X_{t(H)}\right) = (X_t)_{t \in \mathcal{U}},$$

where $t : [H] \to \mathcal{U}$ maps the index over observables to the index over all variables. Finally, we define $\mathcal{U}_{i:j} := \mathcal{U} \cap \{t(i), \ldots, t(j)\}$ as the set of observables between the i-th and j-th observable. We similarly define $\mathcal{O}_{i:j}$ and $\mathcal{A}_{i:j}$ as the set of observable system variables (action variables, resp.) between the i-th and j-th observable.

A partially-observable sequential team is sequential decision-making problem as per the description in Section 2.1. We similarly define the set of histories at time h as $\mathbb{H}_h := \prod_{s \in \mathcal{U}_{1:h}} \mathbb{X}_s$, and the set of futures at time h as $\mathbb{F}_h := \prod_{s \in \mathcal{U}_{h+1:H}} \mathbb{X}_s$. A history $\tau_h \in \mathbb{H}_h$ takes the form $\tau_h = (x_s \in \mathbb{X}_s : s \in \mathcal{U}_{1:h})$, and a future $\omega_h \in \mathbb{F}_h$ takes the form $\omega_h = (x_s \in \mathbb{X}_s : s \in \mathcal{U}_{h+1:H})$. We separate the actions from other observations via $\tau_h^o = (x_s \in \mathbb{X}_s : s \in \mathcal{O}_{1:h})$, $\tau_h^a = (x_s \in \mathbb{X}_s : s \in \mathcal{A}_{1:h})$, $\omega_h^o = (x_s \in \mathbb{X}_s : s \in \mathcal{O}_{h+1:H})$, $\omega_h^a = (x_s \in \mathbb{X}_s : s \in \mathcal{A}_{h+1:H})$. We denote the observation and action components of the histories as $\mathbb{H}_h^o := \texttt{obs}(\mathbb{H}_h)$, $\mathbb{H}_h^a := \texttt{act}(\mathbb{H}_h)$, respectively, and define $\mathbb{F}_h^o, \mathbb{F}_h^a$ similarly. The (observable) dynamics matrix of a sequential team is defined by,

$$\begin{aligned} \left[\boldsymbol{D}_h\right]_{\tau_h, \omega_h} &:= \mathbb{P}\left[\tau_h^o, \omega_h^o \mid \mathrm{do}(\tau_h^a, \omega_h^a)\right] \\ &\equiv \sum_{\substack{x_s \in \mathbb{X}_s \\ s \in \mathcal{O}^{\complement}}} \prod_{t \in \mathcal{S}} \mathcal{T}_t\left(x_t \mid \{x_i, i \in \mathcal{I}_t\}\right). \end{aligned} \tag{12}$$

Note that the 'do' operator indicates that the actions in τ_h, ω_h are executed independently of the system dynamics (that is, the incoming edges into action variables are removed in the DAG $\mathcal{G}$ and the actions are predetermined rather than chosen by a policy). In the second line above, $(x_s, s \in \mathcal{U}) = (\tau_h, \omega_h)$ are the observable variables, and $(x_s, s \in \mathcal{O}^{\complement})$ are the unobservable system variables being marginalized over in the summation.

By Proposition 1, any POST admits a generalized PSR representation, as per Definition 3. In Section 6, we will explicitly construct a *well-conditioned* PSR representation. This enables sample-efficient learning, as demonstrated in Section 7.

4 Bounding the rank of system dynamics via information structure and d-separation

In this section, we will show that the information structure of a sequential team can be used to obtain a bound on the rank of the system dynamics matrices $\boldsymbol{D}_h$ (which coincides with the PSR rank). The main tool in doing so is the DAG representation $\mathcal{G}$ of the information structure $\{\mathcal{I}_t, t \in [T]\}$. To motivate this, we recall a classic result on PSRs (see e.g. Theorem 1 of Littman and Sutton 2001).

Example (POMDPs have PSR rank bounded by $|\mathbb{S}|$)**.** Consider a POMDP with states $s_t \in \mathbb{S}$, observations $o_t \in \mathbb{O}$, and actions $a_t \in \mathbb{A}$. The system dynamics are given by $\mathbb{P}\left[s_{t+1}, o_{t+1} \mid s_{1:t}, a_{1:t}, o_{1:t}\right] = \mathbb{P}\left[s_{t+1} \mid s_t, a_t\right] \mathbb{P}\left[o_{t+1} \mid s_{t+1}\right]$. We will derive a bound on the PSR rank of this partially observable system. For each history $\tau_t = (o_1, a_1, \dots, o_t, a_t)$ and a future $\omega_t = (o_{t+1}, a_{t+1}, \dots o_T, a_T)$, we have,

$$\boldsymbol{D}_{\tau_t, \omega_t} = \mathbb{P}\left[\tau_t^o, \omega_t^o \mid \mathrm{do}(\tau_t^a, \omega_t^a)\right] = \sum_{s_{t+1} \in \mathbb{S}} \mathbb{P}\left[\omega_t \mid s_{t+1}\right] \mathbb{P}\left[s_{t+1} \mid \tau_t\right] \mathbb{P}\left[\tau_t^o \mid \tau_t^a\right].$$

Hence, defining $\boldsymbol{D}_{t,1} := \left[\mathbb{P}\left[\omega_t \mid s_{t+1}\right]\right]_{\omega_t, s_{t+1}}$ and $\boldsymbol{D}_{t,2} := \left[\mathbb{P}\left[s_{t+1} \mid \tau_t\right] \mathbb{P}\left[\tau_t^o \mid \tau_t^a\right]\right]_{s_{t+1}, \tau_t}$, we have that $\boldsymbol{D}_t = \boldsymbol{D}_{t,1} \boldsymbol{D}_{t,2}$. Thus, $\mathrm{rank}(\boldsymbol{D}_t) \le |\mathbb{S}|$ for all t. Hence, the rank of the observable dynamics of a POMDP is bounded by the number of states.

□

In the above, the existence of a latent state implied a simplification of the dynamics and a bound on the rank. We will use the same intuitive idea to bound the rank of the observable dynamics of partially-observable sequential teams via their information structure.

Definition 5. *$\mathcal{G}^\dagger$ is the DAG obtained from $\mathcal{G}$ by removing all edges directed towards actions. That is, it consists of the edges $\mathcal{E}^\dagger := \mathcal{E} \setminus \{(x, a) : x \in \mathcal{N}, a \in \mathcal{A}\}$.*

Removing the incoming edges in the action variables in $\mathcal{G}^\dagger$ corresponds to the do operation in the definition of the system dynamics matrix in Equation (12).

Definition 6. *For each $h \in [H]$, let $\mathcal{I}_h^\dagger \subset [t(h)]$ be the minimal set of past variables (observed or unobserved) which d-separates the past observations $\left(X_{t(1)}, \dots, X_{t(h)}\right)$ from the future observations $\left(X_{t(h+1)}, \dots, X_{t(H)}\right)$ in the DAG $\mathcal{G}^\dagger$. Define $\mathbb{I}_h^\dagger := \prod_{s \in \mathcal{I}_h^\dagger} \mathbb{X}_s$ as the joint space of those variables.*

The notation $\mathcal{I}_h^\dagger$ is chosen to emphasize that this set depends on the information structure, $\mathcal{I} = \{\mathcal{I}_t, t \in \mathcal{N}\}$, and that it somehow simplifies or "inverts" the dynamics. Note also that $\mathcal{G}^\dagger$, and hence $\mathcal{I}_h^\dagger$, are independent of the information sets of action variables. That is, they only depend on the information structure of system variables.

Note that $\mathbb{I}_h^\dagger$ can contain system variables (observable or unobservable) as well as action variables. However, all action variables are assumed to be observable, and hence $\mathtt{act}(\mathbb{I}_h^\dagger) \subset \mathbb{H}_h$. In general, $\mathbb{I}_h^\dagger \not\subset \mathbb{H}_h$ since it may contain unobservable system variables.

Proposition 2 (Rank of Sequential Teams)**.** *A (partially-observable) sequential team has a rank bounded by,*

$$r \leq \max_{h \in [H]} \left| \mathbb{I}_h^\dagger \right| .$$

Proof. The proof is given in Appendix B. □

5 Examples

The procedure outlined in Proposition 2 gives a way to obtain a low-rank decomposition of any partially-observable sequential team. Since sequential teams are general models which capture commonly studied models of sequential decision making like MDPs, POMDPs, Dec-POMDPs, etc., this procedure recovers results about the PSR rank of those models. Crucially, it also allows us to study additional sequential decision making problems with more complex information structures. In this section, we discuss how classical sequential decision-making models are special cases of POSTs and how the rank of their system dynamics emerges from the graph-theoretic analysis of their information structure.

Decentralized POMDPs and POMGs. At each time t, the system variables of a decentralized POMDP (or POMG) consists of a latent state s_t, observations for each agent $o_t^1, \ldots, o_t^N$, and actions of each agent $a_t^1, \ldots, a_t^N$. The latent state transitions are Markovian and depend on the agents' joint action. The observations are sampled via a kernel conditional on the latent state. Each agent can use their own history of observations to choose an action. Thus, the information structure is given by,

$$\mathcal{I}(s_t) = \left\{s_{t-1}, a_{t-1}^1, \ldots, a_{t-1}^N\right\}, \mathcal{I}(o_t^i) = \left\{s_t\right\}, \mathcal{I}(a_t^i) = \left\{o_{1:t-1}^i, a_{1:t-1}^i\right\}.$$

Here, the observable variables are $\mathcal{U} = \left\{o_{1:T}^i, a_{1:T}^i, i \in [N]\right\}$[1]. By Proposition 2, we have $\mathcal{I}^\dagger(o_t^i) = \mathcal{I}^\dagger(a_t^i) = \left\{s_t\right\}, \forall t, i$, as shown in Figure 1a. Thus, the PSR rank of a Dec-POMDP is bounded by $|\mathbb{S}|$, where $\mathbb{S}$ is the state space.

Limited-memory information structures. Consider a sequential decision making problem with variables $o_t, a_t, t \in [T]$ and an information structure with m-length memory. That is, observations can only depend directly on at most m of the most recent observations and actions. That is, the information structure is

$$\mathcal{I}(o_t) = \left\{o_{t-m:t-1}, a_{t-m:t-1}\right\}, \mathcal{I}(a_t) = \left\{o_{1:t}, a_{1:t-1}\right\}.$$

The observables are all observations and actions, $\mathcal{U} = \{o_{1:T}, a_{1:T}\}$. By Proposition 2 we have that $\mathcal{I}^\dagger(o_t) = \{o_{t-m:t-1}, a_{t-m:t-1}\}$, as shown in Figure 1b. Hence, the PSR rank of this sequential decision-making process is bounded by $|\mathbb{O}|^m |\mathbb{A}|^m$.

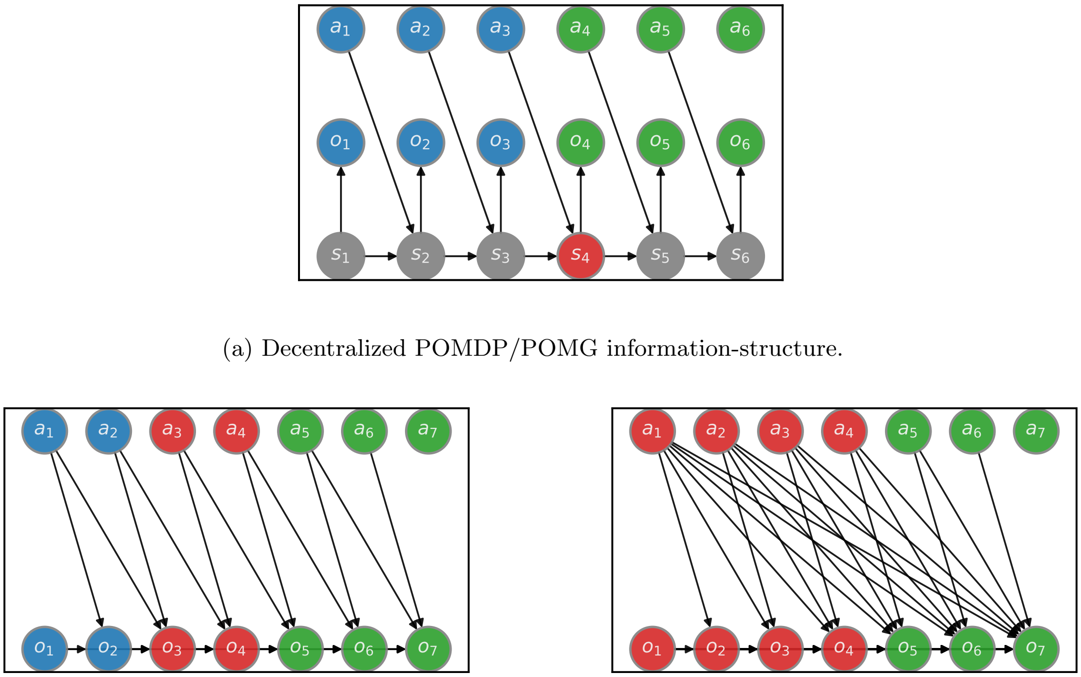
Fully-Connected Information Structures. Consider a sequential decision making problem with variables $o_t, a_t, t \in [T]$ and a fully-connected information structure. That is, each observation directly depends on the entire history of observations and actions. Thus, the information structure is

$$\mathcal{I}(o_t) = \left\{o_{1:t-1}, a_{1:t-1}\right\}, \mathcal{I}(a_t) = \left\{o_{1:t}, a_{1:t-1}\right\}.$$

The observables are all observations and actions, $\mathcal{U} = \{o_{1:T}, a_{1:T}\}$. By Proposition 2 we have that $\mathcal{I}^\dagger(o_t) = \{o_{1:t-1}, a_{1:t-1}\}$, as shown in Figure 1c. Hence, the PSR rank of this sequential decision-making process can be exponential in the time-horizon.

The examples above show that the tractability of a sequential decision-making problem (in terms of its rank) depends directly on its information structure. This gives an interpretation of why certain models, like POMDPs, are more tractable than those with arbitrary information structures. Previous work primarily

[1] Here, since we don't explicitly write the index sets $\mathcal{S}, \mathcal{O}, \mathcal{A}$ in the standard sequential teams formulation, we use the notation $\mathcal{I}(x)$ to mean the information set $\mathcal{I}_t$ where x occurs at time t. In the case of Dec-POMDPs, since some events occur simultaneously, there is not a unique ordering of variables. For example, (s_t, o_t^1, o_t^2) and (s_t, o_t^2, o_t^1) are both valid orderings. When mapping such models onto the sequential teams framework, we may choose any ordering arbitrarily. Similarly, we slightly abuse notation when defining the set of observables $\mathcal{U}$, where what we mean is the "time indices" of the variables in $\{\cdot\}$.


a_1 a_2 a_3 a_4 a_5 a_6
o_1 o_2 o_3 o_4 o_5 o_6
s_1 s_2 s_3 s_4 s_5 s_6
a_1 a_2 a_3 a_4 a_5 a_6 a_7
o_1 o_2 o_3 o_4 o_5 o_6 o_7
a_1 a_2 a_3 a_4 a_5 a_6 a_7
o_1 o_2 o_3 o_4 o_5 o_6 o_7

(a) Decentralized POMDP/POMG information-structure.

(b) Limited-memory ($m = 2$) information structures.

(c) Fully connected information structure.

Figure 1: DAG representation of various information structures. Grey nodes are represent unobservable variables, blue nodes represent past observable variables, green nodes represent future observable variables, and red nodes represent $\mathcal{I}_t^\dagger$. To find $\mathcal{I}_t^\dagger$, as per Proposition 2, we first remove the incoming edges into the action variables, then we find the minimal set among all past variables (both observable and unobservable) which d-separates the past observations from the future observations.

considers particular problem classes with fixed and highly regular information structures. In this work we argue for the importance of explicitly modeling the information structure of a sequential decision-making problem.

6 Constructing a PSR parameterization for POSTs

6.1 Core test sets for POSTs

In Section 4, we showed that the information structure $\mathcal{I} = \{\mathcal{I}_t,\, t \in \mathcal{N}\}$ of a sequential team can be used derive a decomposition of the system dynamics which bounds the rank of its observable dynamics, and hence its PSR rank. Another crucial ingredient for modeling partially-observable systems in the predictive state representation is the notion of a core test set, as defined in Definition 2. For systems with a simple and regular information structure such as a POMDP, a core test set may be simple to obtain. For example, undercomplete POMDPs with a full rank 1-step emission matrix admits the 1-step observation space as a core test set. The regularity of the information structure enables us to find a core test set without knowing much about the system dynamics.

For sequential teams with arbitrary information structures, obtaining a core test set is much more challenging without knowing the system dynamics. In this section, we identify a condition in terms of the information structure under which m-step futures are a core test set for partially-observable sequential teams.

We begin by introducing some notation. For each $h \in [H]$, we define candidate core test set of m-step future observations by

$$\mathbb{Q}_h^m := \prod_{s \in \mathcal{U}_{h+1:\min(h+m,H)}} \mathbb{X}_s. \tag{13}$$

Further, we define the matrix $\boldsymbol{G}_h \in \mathbb{R}^{|\mathbb{Q}_h^m| \times |\mathbb{I}_h^\dagger|}$ as encoding the probability of observing each m-step future conditioned on the separating information set $\mathbb{I}_h^\dagger$,

$$\begin{aligned} \boldsymbol{G}_h &= \left[\mathbb{P}\left[q^o \;\middle|\; i_h^\dagger;\, \mathrm{do}(q^a)\right]\right]_{q \in \mathbb{Q}_h^m,\, i_h^\dagger \in \mathbb{I}_h^\dagger} \\ &\equiv \left[\mathbb{P}\left[x_s, s \in \mathcal{O}_{h+1:h+m} \;\middle|\; \left\{x_s,\, s \in \mathcal{I}_h^\dagger\right\};\, \mathrm{do}\left(x_s,\, s \in \mathcal{A}_{h+1:h+m}\right)\right]\right]_{q \in \mathbb{Q}_h^m,\, i_h^\dagger \in \mathbb{I}_h^\dagger}, \end{aligned} \tag{14}$$

where in the second line $q \equiv (x_s,\, s \in \mathcal{U}_{h+1:h+m}) \in \mathbb{Q}_h^m$, and $i_h^\dagger \equiv \left(x_s,\, s \in \mathcal{I}_h^\dagger\right) \in \mathbb{I}_h^\dagger$, and q^o, q^a are the observation and action components of the test q, respectively.

We identify a condition on partially-observable sequential teams, called m-step $\mathcal{I}^\dagger$-weakly revealing, which we will show implies that the m-step futures are core test sets.

Definition 7 (m-step $\mathcal{I}^\dagger$-weakly revealing)**.** *We say that a sequential team is m-step $\mathcal{I}^\dagger$-weakly revealing if for all $h \in [H]$, $\mathrm{rank}(\boldsymbol{G}_h) = |\mathbb{I}_h^\dagger|$. Furthermore, we say that the sequential team is α-robustly m-step $\mathcal{I}^\dagger$-weakly revealing if for all $h \in [H-m+1]$, $\sigma_{|\mathbb{I}_h^\dagger|}(\boldsymbol{G}_h) \geq \alpha$.*

The $\mathcal{I}^\dagger$-weakly revealing condition is essentially an identifiability condition. If a POST is $\mathcal{I}^\dagger$-weakly revealing, then, at any time point, for any two mixtures of "separating-information" $\nu_1, \nu_2 \in \mathcal{P}(\mathbb{I}_h^\dagger)$ with disjoint support, the conditional distributions of the m-step futures is distinct (i.e., $\boldsymbol{G}_h \nu_1 \neq \boldsymbol{G}_h \nu_2$). That is, the future observations contain information that can distinguish between mixtures of "separating-information" (which can be thought of as a generalization of latent states). This description is equivalent to the condition that $\boldsymbol{G}_h$ is full-rank in Definition 7. The α-robust version of the $\mathcal{I}^\dagger$-weakly revealing condition requires that $\boldsymbol{G}_h$ is not only full rank, but that its $|\mathbb{I}_h^\dagger|$-th eigenvalue is bounded away from zero.

The condition holds whenever there exists a sequence of actions within the m-step futures such that executing these actions results in a sequence of observations which is informative about the "separating-information," $i_h^\dagger \in \mathbb{I}_h^\dagger$. In general, this condition will be harder to satisfy when $\mathbb{I}_h^\dagger$ is large since it would require the m-step future observations to encode more information. In particular, $\boldsymbol{G}_h$ cannot be full rank when $|\mathbb{Q}_h^m| < |\mathbb{I}_h^\dagger|$. Without prior knowledge about the dynamics, as a heuristic, we can choose m such that $|\mathbb{Q}_h^m| \geq |\mathbb{I}_h^\dagger|$. In general, it will be possible to find a smaller core test set when the d-separating set $\mathcal{I}_h^\dagger$ is small. This happens when the system dynamics contains state-like variables which are low-dimensional.

This condition is a generalization of the "weakly-revealing" condition for POMDPs introduced in (Liu, Chung, et al. 2022). That is, when the POST is a POMDP, the conditions are the same. Our formalization of the condition here extends this to much more general problems.

Recall that the vector of core test set probabilities for the history τ_h is given by the mapping $\overline{\psi}_h : \mathbb{H}_h \to \mathbb{R}^{|\mathbb{Q}_h^m|}$,

$$\overline{\psi}_h(\tau_h) \equiv \left[\mathbb{P}\left[q^o \mid \tau_h^o;\, \mathrm{do}(\tau_h^a),\, \mathrm{do}(q^a)\right]\right]_{q \in \mathbb{Q}_h^m} \in \mathbb{R}^{|\mathbb{Q}_h^m|}.$$

Define the mapping $m_h : \mathbb{F}_h \to \mathbb{R}^{|\mathbb{Q}_h^m|}$ as,

$$m_h(\omega_h) := (\boldsymbol{G}_h^\dagger)^\top \left[\overline{\mathbb{P}}[\omega_h \mid i_h^\dagger]\right]_{i_h^\dagger \in \mathbb{I}_h^\dagger} \tag{15}$$

The following lemma shows that the m-step futures $\mathbb{Q}_h^m$ are core test sets for any m-step $\mathcal{I}^\dagger$-weakly revealing POST. Moreover, given any future $\omega_h \in \mathbb{F}_h$ and history $\tau_h \in \mathbb{H}_h$, the conditional probability $\overline{\mathbb{P}}\left[\omega_h \mid \tau_h\right]$ can be written as a linear combination of the probabilities of the core tests given the history in $\overline{\psi}_h(\tau_h)$, with weights given by $m_h(\omega_h)$, depending only on ω_h.

Lemma 1 (Core test set for POSTs). *Suppose that the POST is m-step $\mathcal{I}^\dagger$-weakly revealing. Then, $\mathbb{Q}_h^m$ is a core test set for all $h \in [H]$. Furthermore, we have*

$$\overline{\mathbb{P}}\left[\tau_h, \omega_h\right] = \left\langle m_h(\omega_h), \psi_h(\tau_h) \right\rangle, \text{ and } \overline{\mathbb{P}}\left[\omega_h \mid \tau_h\right] = \left\langle m_h(\omega_h), \overline{\psi}_h(\tau_h) \right\rangle. \tag{16}$$

Proof. The proof is given in Appendix C. □

6.2 PSR parameterization of partially-observable sequential teams

Consider a POST which is m-step $\mathcal{I}^\dagger$-weakly revealing. Appendix C shows that the m-step futures $\mathbb{Q}_h^m$ are core test sets. In this section we will explicitly construct a PSR parameterization for this class sequential decision-making problems.

Let $d_h \coloneqq |\mathbb{Q}_h^m|$. The first observation is that the vector mappings $m_h : \mathbb{F}_h \to \mathbb{R}^{d_h}$ and $\psi_h : \mathbb{H}_h \to \mathbb{R}^{d_h}$ can be used to derive a recursive form of the dynamics of the POST. A direct corollary of Appendix C is the following.

Lemma 2. *For any $h \in [H]$, $\tau_h \in \mathbb{H}_h$, $x_{t(h+1)} \in \mathbb{X}_{t(h+1)}$, $\omega_{h+1} \in \mathbb{F}_{h+1}$, we have*

$$\overline{\mathbb{P}}\left[\tau_h, x_{t(h+1)}, \omega_{h+1}\right] = \left\langle m_h(x_{t(h+1)}, \omega_{h+1}), \psi_h(\tau_h) \right\rangle \tag{17}$$

Hence, given a history $\tau_h = (x_{t(1)}, \dots, x_{t(h)})$, having observed another variable $x_{t(h+1)}$, we can update our predictions of the future and obtain the probability of any future trajectory of the form $(\tau_h, x_{t(h+1)}, \omega_{h+1})$ for $\omega_{h+1} \in \mathbb{F}_{h+1}$. Note that $x_{t(h+1)}$ may be either an observation or an action. Hence, we can update our prediction of the future after deciding an action, and before receiving the next observation. This is in contrast to the standard PSR formulation where predictions of the future can only be updated with a *pair* of observation and action. Our formulation provides additional flexibility.

This means that, having observed $x_{t(h+1)}$, we can use the $m_h : \mathbb{F}_h \to \mathbb{R}^{d_h}$ mapping constructed in Appendix C to update the probability of any candidate future ω_{h+1}. We are particularly interested in updating the probabilities of the futures corresponding to the core test sets at the next time point. Thus, we define the matrix mapping $\boldsymbol{M}_h : \mathbb{X}_{t(h)} \to \mathbb{R}^{d_{h+1} \times d_h}$ by,

$$\left[M_h(x_{t(h)})\right]_{q,\cdot} = m_h(x_{t(h)}, q)^\top,\ q \in \mathbb{Q}_{h+1}. \tag{18}$$

That is, $M_h(x_{t(h)})$ is the matrix whose rows are indexed by the core tests at the $h+1$-th observable step, where the $q \in \mathbb{Q}_{h+1}$ row is the weights given by the m_h mapping for the future of $x_{t(h)}$ followed by q. This mapping enables us to update the probabilities of the core test sets.

Lemma 3. *For any $h \in [H-1]$, $\tau_h \in \mathbb{H}_h$, $x_{t(h+1)} \in \mathbb{X}_{t(h+1)}$, we have*

$$\psi_{h+1}(\tau_h, x_{t(h)}) = M_h(x_{t(h)})\psi_h(\tau_h). \tag{19}$$

Hence, for a history $\tau_h = \left(x_{t(1)}, \dots, x_{t(h)}\right) \in \mathbb{H}_h$, we have

$$\psi_h(\tau_h) = M_h(x_{t(h)}) \cdots M_1(x_{t(1)})\psi_0, \tag{20}$$

where $\psi_0 = \psi_0(\emptyset)$.

Finally, observe that $\mathbb{Q}_{H-1}^m = \mathbb{X}_{t(H)}$. Hence,

$$\psi_{H-1}(\tau_{H-1}) = \left(\overline{\mathbb{P}}\left[\tau_{h-1}, x_{t(H)}\right]\right)_{x_{t(H)} \in \mathbb{X}_{t(H)}} \in \mathbb{R}^{|\mathbb{X}_{t(H)}|}.$$

Thus, letting $\phi_H : \mathbb{X}_{t(H)} \to \mathbb{R}^{|\mathbb{X}_t(H)|}$ be $\phi_H(x_{t(H)}) = \boldsymbol{e}_{x_{t(H)}}$ (the canonical basis vector), yields

$$\overline{\mathbb{P}}\left[x_{t(h)} :\ h \in [H]\right] = \phi_H(x_{t(H)})^\top M_{H-1}(x_{t(H-1)}) \cdots M_1(x_{t(1)})\psi_0. \tag{21}$$

Hence, Equation (21) together with Equation (20) imply that $(\boldsymbol{M}, \phi_H, \psi_0)$ is a valid PSR representation for the partially-observable sequential team, under Definition 3. With $\mathbb{Q}_h^m$ as the core test sets, ϕ_H can be fixed to be the basis vectors, and the parameters are $\boldsymbol{M}$ and ψ_0.

Thus, we have shown that any $\mathcal{I}^\dagger$-weakly revealing POST has a generalized PSR representation with $\{\mathbb{Q}_h^m\}_h$ as core test sets. The next result shows that if the POST is α-robustly $\mathcal{I}^\dagger$-weakly revealing, then this PSR representation is *well-conditioned.*

Proposition 3 ($\mathcal{I}^\dagger$-weakly revealing POSTs are well-conditioned PSRs)**.** *Suppose a POST is α-robustly m-step $\mathcal{I}^\dagger$-weakly revealing. Then, the corresponding PSR as constructed above with core test sets consisting of m-step futures is γ-well-conditioned with $\gamma = \alpha / \max_h \sqrt{\left|\mathbb{I}_h^\dagger\right|}$*

Proof. The proof is given in Appendix C. □

7 Sample-efficient Reinforcement Learning for Dynamic Sequential Teams

We now introduce our model-based algorithm for learning generalized PSRs, including those representing partially-observable sequential teams. The algorithm is a slight generalization of Huang et al. 2023, extending the algorithm to our generalized notion of PSRs with arbitrary sequences of observations and actions. The algorithm involves the estimation of an upper confidence bound capturing the uncertainty in the estimated model and drives exploration so as to minimize this uncertainty. Our contribution in this section is to extend the algorithm and the theoretical guarantees to generalized PSRs.

We suppose that the core test sets $\{\mathbb{Q}_h\}_{0 \leq h \leq H-1}$ are known. For example, if the sequential decision-making problem is a partially-observable sequential team, Section 6 provides conditions under which m-step futures form core test sets. Let Θ be the set of generalized PSR representations with $\{\mathbb{Q}_h\}_{0 \leq h \leq H-1}$ as core test sets. That is,

$$\Theta = \{\theta = (\boldsymbol{M}, \psi_0, \phi_H) : \ \theta \text{ is a generalized PSR}\}$$

Recall that $d_h := |\mathbb{Q}_h|$ and $d = \max_h d_h$. Moreover, $\mathbb{Q}_h^A := \mathtt{act}(\mathbb{Q}_h)$ are the action components of the core test sets and $Q_A := \max_h \left|\mathbb{Q}_h^A\right|$ is the maximal size of those action components. We define the exploration action sequences at time h to be $\mathbb{Q}_{h-1}^{\exp} = \mathtt{act}(\mathbb{X}_h \times \mathbb{Q}_h \cup \mathbb{Q}_{h-1})$. Moreover, we define $\mathtt{u}_{h-1}^{\mathtt{exp}}$ as the policy, defined from time $h-1$ onwards, in which each selection of action sequences in $\mathbb{Q}_{h-1}^{\exp}$ are chosen uniformly at random. For a model θ and reward function R, we define the value of a policy under this model and reward as $V_\theta^R(\pi) = \sum_{\tau_H} R(\tau_H) \mathbb{P}_\theta^\pi(\tau_H)$.

The algorithmic description is given in Algorithm 1. At each iteration k, the learner collects a trajectory $\tau_H^{k,h}$ for each time index $h \in [H]$ by using a particular policy which drives exploration so as to better estimate the parameters associated with the h-th time step. To collect the trajectory $\tau_H^{k,h}$, the learner executes the policy at the previous iteration, π^{k-1}, until time $h-1$ collecting the trajectory $\tau_{h-1}^{k,h}$ then executes $\mathtt{u}_{h-1}^{\mathtt{exp}}$ which samples action sequences from $\mathbb{Q}_{h-1}^{\exp}$ uniformly. The particular choice of the exploratory action sequences $\mathbb{Q}_{h-1}^{\exp}$ comes out of the proof (see proof of Lemma 5 in the appendix). Intuitively, $\mathtt{act}(\mathbb{Q}_{h-1})$ allows us to estimate the prediction features $\overline{\psi}^*(\tau_{h-1}^{k,h}) = [\overline{\mathbb{P}}(q \,|\, \tau_{h-1}^{k,h})]_{q \in \mathbb{Q}_{h-1}}$, and $\mathtt{act}(\mathbb{X}_h \times \mathbb{Q}_h)$ allows us to estimate $M_h^*(x_h)\overline{\psi}^*(\tau_{h-1}^{k,h})$.

The collected trajectories are added to the dataset, together with the policies used to collect them. The next step is model estimation via (constrained) maximum likelihood estimation. The algorithm estimates a model $\widehat{\theta}^k$ by selecting any model in a constrained set $\mathcal{B}^k$ defined as,

$$\begin{aligned}
&\Theta_{\min}^k = \left\{\theta \in \Theta : \forall h, (\tau_h, \pi) \in \mathcal{D}_h^k, \mathbb{P}_\theta^\pi(\tau_h) \geq p_{\min}\right\}, \\
&\mathcal{B}^k = \left\{\theta \in \Theta_{\min}^k : \sum_{(\tau_H, \pi) \in \mathcal{D}^k} \log \mathbb{P}_\theta^\pi(\tau_H) \geq \max_{\theta' \in \Theta_{\min}^k} \sum_{(\tau_H, \pi) \in \mathcal{D}^k} \log \mathbb{P}_{\theta'}^\pi(\tau_H) - \beta\right\}.
\end{aligned} \tag{22}$$

The introduction of $\Theta^k_{\min}$ ensures that $\mathbb{P}^{\pi^{k-1}}_{\theta^*}(\tau^{k,h}_{h-1})$ is not too small so that the estimates of the prediction features $\overline{\psi}^*(\tau^{k,h}_{h-1}) = [\overline{\mathbb{P}}(q \mid \tau^{k,h}_{h-1})]_{q \in \mathbb{Q}_{h-1}}$ are accurate. This design differs from other MLE-based estimators (Liu, Chung, et al. 2022; Liu, Netrapalli, et al. 2022; Chen et al. 2022, e.g.,) due to the estimation of parameters capturing *conditional* probabilities.

Next, the algorithm chooses a policy which drives the algorithm to trajectories τ_h whose prediction features have so far been unexplored. To do this, Algorithm 1 constructs an upper confidence bound on the total variation distance between the estimated model and the true model. This is done via a bonus function $\widehat{b}^k(\tau_H)$,

$$\begin{aligned} \widehat{b}^k(\tau_H) &= \min\left\{ \alpha \sqrt{\sum_{h=0}^{H-1} \left\| \widehat{\psi}(\tau_h) \right\|^2_{(\widehat{U}^k_h)^{-1}}}, 1 \right\}, \quad \text{where,} \\ \widehat{U}^k_h &= \lambda I + \sum_{\tau_h \in \mathcal{D}_{h^k}} \widehat{\psi}^k(\tau_h) \widehat{\psi}^k(\tau_h)^\top, \end{aligned} \tag{23}$$

where λ and α are pre-specified parameters to the algorithm. Thus, the bonus function captures the degree of uncertainty in the estimated prediction features $\widehat{\overline{\psi}}^k(\tau_h)$. In particular, $\left\| \widehat{\psi}(\tau_h) \right\|^2_{(\widehat{U}^k_h)^{-1}}$ will be large when the prediction feature $\widehat{\psi}(\tau_h)$ lies far away from the empirical distribution of prediction features sampled in the dataset $\mathcal{D}^k_h$, whose covariance is captured by $\widehat{U}^k_h$.

The algorithm then chooses an exploration policy for the next iteration which maximizes this upper confidence bound, hence collecting trajectories which have high uncertainty in their prediction features. When the estimated model is sufficiently accurate on all trajectories, the algorithm terminates and returns the optimal policy with respect to the reward function R under the estimated model.

Algorithm 1: Learning Generalized PSRs (e.g., POSTs) via MLE and Exploration with UCB

```
for $k \leftarrow 1, \dots, K$ do
    for $h \leftarrow 1, \dots, H$ do
        Collect $\tau^{k,h}_H = (\omega^{k,h}_{h-1}, \tau^{k,h}_{h-1})$ using $\nu(\pi^{k-1}, \mathbf{u}_{\mathbb{Q}^{\texttt{exp}}_{h-1}})$.
        $\mathcal{D}^k_{h-1} \leftarrow \mathcal{D}^{k-1}_{h-1} \cup \left\{ \left( \tau^{k,h}_H, \nu(\pi^{k-1}, \mathbf{u}_{\mathbb{Q}^{\texttt{exp}}_{h-1}}) \right) \right\}$.
    end
    $\mathcal{D}^k = \{\mathcal{D}^k_h\}_{h=0}^{H-1}$
    Compute MLE $\widehat{\theta} \in \mathcal{B}^k$, where
        $\Theta^k_{\min} = \left\{ \theta : \forall h, (\tau_h, \pi) \in \mathcal{D}^k_h, \mathbb{P}^\pi_\theta(\tau_h) \geq p_{\min} \right\}$,
        $\mathcal{B}^k = \left\{ \theta \in \Theta^k_{\min} : \sum_{(\tau_H, \pi) \in \mathcal{D}^k} \log \mathbb{P}^\pi_\theta(\tau_H) \geq \max_{\theta' \in \Theta^k_{\min}} \sum_{(\tau_H, \pi) \in \mathcal{D}^k} \log \mathbb{P}^\pi_{\theta'}(\tau_H) - \beta \right\}$.
    Define the bonus function, $\widehat{b}^k(\tau_H) = \min\left\{ \alpha \sqrt{\sum_{h=0}^{H-1} \left\| \widehat{\psi}(\tau_h) \right\|^2_{(\widehat{U}^k_h)^{-1}}}, 1 \right\}$, where
     $\widehat{U}^k_h = \lambda I + \sum_{\tau_h \in \mathcal{D}_{h^k}} \widehat{\psi}^k(\tau_h) \widehat{\psi}^k(\tau_h)^\top$.
    Solve the planning problem $\pi^k = \arg\max_\pi V^{\widehat{b}^k}_{\widehat{\theta}^k}(\pi)$.
    if $V^{\widehat{b}^k}_{\widehat{\theta}^k}(\pi^k) \leq \epsilon/2$ then
        $\theta^\epsilon = \widehat{\theta}^k$. break.
    end
end
return $\overline{\pi} = \arg\max_\pi V^R_{\theta^\epsilon}(\pi)$
```

We extend Huang et al.'s theoretical guarantees to show that Algorithm 1 enjoys polynomial sample complexity for *generalized* PSRs (Definition 3).

Theorem 1. *Suppose Assumption 1 holds. Let* $p_{\min} = O\left(\frac{\delta}{KH\prod_{h=1}^{H}|\mathbb{X}_h|}\right)$, $\lambda = \frac{\gamma(\max_{s\in\mathcal{A}}|\mathbb{X}_s|)^2 Q_A \beta \max\left\{\sqrt{r}, Q_A\sqrt{H}/\gamma\right\}}{\sqrt{dH}}$, $\alpha = O\left(\frac{Q_A\sqrt{Hd}}{\gamma^2}\sqrt{\lambda} + \frac{\max_{s\in\mathcal{A}}|\mathbb{X}_s|Q_A\sqrt{\beta}}{\gamma}\right)$, *and let* $\beta = O(\log\left|\overline{\Theta}_\varepsilon\right|)$, *where* $\varepsilon = O(\frac{p_{\min}}{KH})$. *Then, with probability at least* $1-\delta$, *Algorithm 1 returns a model* θ^ϵ *and a policy* $\overline{\pi}$ *that satisfy*

$$V_{\theta^\epsilon}^R(\pi^*) - V_{\theta^\epsilon}^R(\overline{\pi}) \leq \varepsilon, \text{ and } \forall\pi,\ \mathtt{D_{TV}}\left(\mathbb{P}_{\theta^\epsilon}^{\pi}(\tau_H), \mathbb{P}_{\theta^*}^{\pi}(\tau_H)\right) \leq \varepsilon.$$

In addition, the algorithm terminates with a sample complexity of,

$$\tilde{O}\left(\left(r + \frac{Q_A^2 H}{\gamma^2}\right)\frac{rdH^3 \max_{s\in\mathcal{A}}|\mathbb{X}_s|^2 Q_A^4 \beta}{\gamma^4\epsilon^2}\right).$$

To apply this algorithm to a partially-observable sequential team, we can use the PSR parameterization constructed in Section 6. By Proposition 2 the PSR rank is bounded by $r \leq \max_h|\mathbb{I}_h^\dagger|$. If this POST is α-robustly $\mathcal{I}^\dagger$-weakly revealing, then by Appendix C the m-step futures $\{\mathbb{Q}_h^m\}$ form core test sets, and by Proposition 3 the corresponding PSR parameterization is γ-well-conditioned with $\gamma = \alpha/\max_h\sqrt{|\mathbb{I}_h^\dagger|}$. We have, $d = \max_h d_h = \max_h|\mathbb{Q}_h^m|$. The following corollary states that Algorithm 1 can learn a partially-observable sequential team with a sample complexity which is polynomial in the quantity $\max_h\left|\mathbb{I}_h^\dagger\right|$.

Corollary 1. *Suppose a partially-observable sequential team is* m*-step* α*-robustly* $\mathcal{I}^\dagger$*-weakly revealing as per Definition 7. Applying Algorithm 1 to this PSR representation, with parameters* $p_{\min}, \lambda, \alpha, \beta$ *chosen as in Theorem 1, returns a* ε*-optimal policy with a sample complexity of,*

$$\tilde{O}\left(\left(1 + \frac{Q_A^2 H}{\alpha^2}\right)\frac{\max_h\left|\mathbb{I}_h^\dagger\right|^7 \max_h|\mathbb{Q}_h^m| H^5 \max_{s\in\mathcal{A}}|\mathbb{X}_s|^2 \max_{s\in\mathcal{U}}|\mathbb{X}_s| Q_A^4}{\alpha^4\epsilon^2}\right).$$

We can interpret this result as saying that the information structure of a sequential decision-making problem, through the quantity $\max_h|\mathbb{I}_h^\dagger|$, is fundamentally a measure of the complexity of the dynamics which need to be modeled. As a result, learning is tractable when $\max_h|\mathbb{I}_h^\dagger|$ is of modest size, and intractable otherwise. Recall that $\max_h|\mathbb{I}_h^\dagger|$ is small when there exists "state-like" variables, whether they are observable or unobservable. In this sense, $\max_h|\mathbb{I}_h^\dagger|$ is a fundamental quantity which generalizes the notion of a "state."

8 Sample-efficient Reinforcement Learning for Dynamic Sequential Games

8.1 Sequential Decision-Making and PSRs in the Game Setting

The formulation of sequential decision-making presented in Section 2.1 applies to the team setting (considered in the previous section) as well as to the game setting, where different agents have different objectives. The description of the system dynamics is identical, described by the tuple $(H, \{\mathbb{X}_h\}_h, \mathcal{O}, \mathcal{A}, \mathbb{P})$. Similarly, since the formulation of generalized predictive state representations in Section 2.2 is independent of the objective of each agent and merely describes the dynamics of the system, it also applies to the game setting. The only difference in the game setting is in the formulation of the objectives of the agents. Let N be the number of agents and let $\mathcal{A}^1, \ldots, \mathcal{A}^N$ be a partition of the action index set $\mathcal{A}$ such that $\mathcal{A}^i$ are the actions taken by agent i. Each agent has their own objective specified by the reward function $R^i : \mathbb{X}_1 \times \cdots \mathbb{X}_H \to [0,1]$. In the next subsection, we introduce an explicit representation of information structure in the game setting through *partially-observable sequential teams.*

8.2 Sequential Games & Partially-Observable Sequential Games

In a sequential team, all agents share the same objective. In the game setting, different agents may have different objectives which compete with each other in interesting ways. Information structures play a crucial role in the study of games. The information available to each agent in making its decisions, compared to the information available to competing agents, determines how well it can achieve its objective. In particular, the information structure of a problem determines its equilibria. There has been a plethora of work in the game theory community studying such problems. In this paper, we tackle the problem of *learning* in dynamic sequential decision-making problems, studying the role of information structures. We begin by defining a generalization of sequential teams to the game-setting which we call sequential games.

A sequential game defines a controlled stochastic process defining the joint distribution of T variables, $(X_1, \dots, X_T)$. The dynamics of a sequential game are identical to a sequential team (Definition 4), with the same variable structure, variable spaces, information structure, system kernels, and decision kernels. In contrast to a sequential team, agents in a sequential game may have different objectives. In a sequential game, there exists N agents, with agent $i \in [N]$ deciding the actions at time $t \in \mathcal{A}^i$, where $\mathcal{A}^i \subset [N]$. Each agent has its own objective defined by a reward function $R^i : \mathbb{X}_1 \times \cdots \mathbb{X}_T \to [0, 1]$. This is defined formally below.

Definition 8 (Sequential Game)**.** *A sequential game is a controlled stochastic process that specifies the joint distribution of T variable $(X_t)_{t \in [T]}$, denoted by $\mathcal{M} \subset \mathcal{P}(X_1, \dots, X_T)$, where T is a fixed integer. Here each X_t is either a system variable or an action variable. A sequential team is specified by the following components.*

1. ***Variable Structures.*** *A partition of $[T]$ into system variables $\mathcal{S}$ and action variables $\mathcal{A}$.*
2. ***Information Structure.*** *Information sets $\mathcal{I}_t \subset [t-1]$ for each $t \in [T]$. The information spaces are defined in the same way, $\mathbb{I}_t = \prod_{s \in \mathcal{I}_t} \mathbb{X}_s$.*
3. ***System Kernels.*** *$\{\mathcal{T}_t : \mathbb{I}_t \to \mathcal{P}(\mathbb{X}_t) \colon t \in \mathcal{S}\}$ such that $X_t \sim \mathcal{T}_t(\cdot | \{X_s, s \in \mathcal{I}_t\})$.*
4. ***Decision Kernels.*** *$\{\pi_t \colon t \in \mathcal{A}\}$, where $\pi_t : \mathbb{I}_t \to \mathcal{P}(\mathbb{X}_t)$ determines how the action at time $t \in \mathcal{A}$ is chosen.*
5. ***Reward Structure.*** *Let N be the number of agents. Each agent may act several times. Denote by $\mathcal{A}^i \subset \mathcal{A}$ the index of action variables associated to agent $i \in [N]$. Each agent has a reward function $R^i : \mathbb{X}_1 \times \cdots \mathbb{X}_T \to [0, 1]$ which they aim to maximize.*

Denote by $\pi^i = (\pi_t : t \in \mathcal{A}^i)$ the collection of decision kernels belonging to agent i, one for each action they take. Denote by $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ the joint policies of all agents. Fixing $\boldsymbol{\pi}$ induces a probability distribution over $\mathbb{X}_1 \times \cdots \mathbb{X}_T$,

$$\mathbb{P}^{\boldsymbol{\pi}} \left[X_1 = x_1, \dots X_T = x_t \right] = \prod_{t \in \mathcal{A}} \pi_t(x_t | \{x_s : s \in \mathcal{I}_t\}) \prod_{t \in \mathcal{S}} \mathcal{T}_t(x_t | \{x_s : s \in \mathcal{I}_t\}). \tag{24}$$

The value of a policy $\boldsymbol{\pi}$ for agent $i \in [N]$ is defined as the expected value of their reward R_i under $\mathbb{P}^{\boldsymbol{\pi}}$,

$$V^i(\boldsymbol{\pi}) \equiv V^i(\pi^i, \boldsymbol{\pi}^{-i}) := \mathbb{E}^{\boldsymbol{\pi}} \left[R^i(X_1, \dots, X_T) \right], \tag{25}$$

where $\boldsymbol{\pi}^{-i} = (\pi^j : j \neq i)$.

To model randomized policies, which are potentially correlated, we introduce a random seed $\omega \in \Omega$ which is sampled at the beginning of an episode. Then, the policy at time $t \in \mathcal{A}$ can be modeled as a deterministic function mapping the seed ω and information variable $i_t \in \mathbb{I}_t$ to an action $\mathbb{X}_t$. That is, $\pi_t : \Omega \times \mathbb{I}_t \to \mathbb{X}_t$. To model independently randomized policies with each agent having private randomness, we consider the special case where the seed has the product structure $\omega = (\omega_1, \dots, \omega_N) \in \Omega_1 \times \cdots \times \Omega_N$, and ω_i is the seed belonging to agent $i \in [N]$. Then, for $t \in \mathcal{A}^i$, $\pi_t : \Omega_i \times \mathbb{I}_t \to \mathbb{X}_t$. For each agent $i \in [N]$, define the three policy spaces,

1. Deterministic policies, $\Gamma^i_{\text{det}} = \left\{ \pi^i : \pi^i = \left(\pi_t : \mathbb{I}_t \to \mathbb{X}_t, t \in \mathcal{A}^i \right) \right\}$,

2. Independently-randomized policies, $\Gamma^i_{\text{ind}} = \left\{\pi^i : \pi^i = \left(\pi_t : \Omega_i \times \mathbb{I}_t \to \mathbb{X}_t, t \in \mathcal{A}^i\right)\right\}$,

3. Correlated randomized policies, $\Gamma^i_{\text{cor}} = \left\{\pi^i : \pi^i = \left(\pi_t : \Omega \times \mathbb{I}_t \to \mathbb{X}_t, t \in \mathcal{A}^i\right)\right\}$.

Define the joint deterministic policy space, as $\mathbf{\Gamma}_{\text{det}} = \Gamma^1_{\text{det}} \times \cdots \times \Gamma^N_{\text{det}}$, and similarly for the independently-randomized policy spae $\mathbf{\Gamma}_{\text{ind}}$, and the correlated randomized policy space $\mathbf{\Gamma}_{\text{cor}}$.

Note that with the policies defined in this way, the probability of any trajectory τ_H under a joint policy $\boldsymbol{\pi}$ is $\mathbb{P}^{\boldsymbol{\pi}}(\tau_H) = \sum_\omega \overline{\mathbb{P}}\left[\tau_H\right] \boldsymbol{\pi}(\tau_H|\omega)\mathbb{P}\left[\omega\right]$, where $\overline{\mathbb{P}}\left[\tau_h\right] = \mathbb{P}\left[\tau^o_H \mid \tau^a_H\right]$ as before, and $\boldsymbol{\pi}(\tau_H \,|\, \omega) = \prod_{t \in \mathcal{A}} \mathbf{1}\{x_t = \pi_t(\{x_s, s \in \mathcal{I}_t\}, \omega)\}$. For each $t \in \mathcal{A}$, $\pi_t : \Omega \times \mathbb{I}_t \to \mathbb{X}_t$ is a deterministic function of the random seed and information variable. Recall that $\overline{\mathbb{P}}\left[\tau_H\right]$ is estimated by the model $\widehat{\theta}$, while the planner chooses $\boldsymbol{\pi}$ according to some objective. The form of $\mathbb{P}\left[\omega\right]$ is assumed to be known by the planner.

Recall that, for a sequential team with finite state and action spaces, there exists a deterministic policy achieving the optimal value. In the game setting, the existence of equilibria depends on the policy class and in general varies depending on the type of randomization. Hence, we will consider several different notions of equilibrium.

8.2.1 Notions of equilibrium

When studying games, a common question is to find an *equilibrium* within a particular policy space. At a high-level, an equilibrium is a joint policy where no agent can do better by deviating from their policy when the other agents keep their policies fixed. We begin by defining the notion of a *best-response*. Suppose that agent i's policy space is Π^i (e.g., Γ^i_{det}, Γ^i_{ind}, or Γ^i_{cor}). Then, we say that agent i's policy π^i is a best response to $\boldsymbol{\pi}^{-i}$ if there is no policy in Π^i which achieves a higher value. This is formalized in the definition below.

Definition 9 (Best response)**.** *For a joint policy $\boldsymbol{\pi}$, π^i is said to be a best-response to $\boldsymbol{\pi}^{-i}$ in the policy space Γ^i (e.g., Γ^i_{det}, Γ^i_{ind}, or Γ^i_{cor}), if $V^i(\pi^i, \boldsymbol{\pi}^{-i}) = \max_{\tilde{\pi}^i \in \Gamma^i} V^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}) =: V^{i,\dagger}(\boldsymbol{\pi}^{-i})$.*

This leads to the definition of two notions of equilibria. A *Nash Equilibrium* (NE) is a joint policy where all agents are best-responding in the space of independently-randomized policies. A *Coarse Correlated Equilibrium* (CCE) is a joint policy where all agents are best-responding in the space of correlated randomized policies. The difference between a NE and CCE is that the randomness in the joint policy must be independent in a NE but can be correlated in a CCE. Since $\Gamma_{\text{ind}} \subset \Gamma_{\text{cor}}$, coarse correlated equilibria are a generalization of Nash equilibria. We define them formally below.

Definition 10 (Nash Equilibrium)**.** *A joint policy $\boldsymbol{\pi} \in \Gamma_{\text{ind}}$ is said to ba a Nash equilibrium if for all agents $i \in [N]$, $V^i(\boldsymbol{\pi}) = \max_{\tilde{\pi}^i \in \Gamma^i_{\text{ind}}} V^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}) =: V^{i,\dagger}(\boldsymbol{\pi}^{-i})$. A joint policy $\boldsymbol{\pi} \in \Gamma_{\text{ind}}$ is said to an ε-approximate Nash equilibrium if $V^i(\boldsymbol{\pi}) \geq V^{i,\dagger}(\boldsymbol{\pi}^{-i}) - \varepsilon$ for all $i \in [N]$.*

Definition 11 (Coarse Correlated Equilibrium)**.** *A joint policy $\boldsymbol{\pi} \in \Gamma_{\text{cor}}$ is said to ba a coarse correlated equilibrium if for all agents $i \in [N]$, $V^i(\boldsymbol{\pi}) = \max_{\tilde{\pi}^i \in \Gamma^i_{\text{cor}}} V^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}) =: V^{i,\dagger}(\boldsymbol{\pi}^{-i})$. A joint policy $\boldsymbol{\pi} \in \Gamma_{\text{cor}}$ is said to an ε-approximate Nash equilibrium if $V^i(\boldsymbol{\pi}) \geq V^{i,\dagger}(\boldsymbol{\pi}^{-i}) - \varepsilon$ for all $i \in [N]$.*

Since we consider finite-space sequential games, an equilibrium is guaranteed to exist (Nash 1951).

We will propose an algorithm which can find a Nash equilibrium or coarse correlated equilibrium in a sample-efficient manner. We consider the self-play setting, in which the algorithm controls all agents, playing against itself. We define a notion of partial-observability for this kind of centralized algorithm analogously to POSTs.

8.2.2 Partially-observable sequential games (POSG)

As with sequential teams and partially-observable sequential teams, we can define a partially-observable variant of sequential games to model learning with partial observations. We define $\mathcal{O} \subset \mathcal{S}$ as the subset of system variables which are observable to the learning algorithm. Together with the action variables, the set of variables available to the algorithm are $\mathcal{U} := \mathcal{O} \cup \mathcal{A}$. Similar to POSTs, the observable trajectories are,

$$\left(X_{t(1)}, \ldots, X_{t(H)}\right) = \left(X_t\right)_{t \in \mathcal{U}},$$

where $H := |\mathcal{O} \cup \mathcal{A}|$. We assume that all agents' reward functions are only a function of variables in $\mathcal{U}$. This is equivalent to assuming that the rewards are observed since we can simply define a new variable equal to the reward of each agent.

8.2.3 Low-rank structure of Partially-observable sequential games and their PSR representation

The information structure of a partially-observable sequential game implies a bound on the rank of its system dynamics. When considering the rank of the dynamics of the observable system variables, the policies play no role since the dynamics matrix $\boldsymbol{D}_h$ is defined with the do-operation on the actions,

$$\begin{aligned} [\boldsymbol{D}_h]_{\tau_h,\omega_h} &:= \mathbb{P}\left[\tau_h^o, \omega_h^o \mid \mathrm{do}(\tau_h^a, \omega_h^a)\right] \\ &\equiv \sum_{\substack{x_s \in \mathbb{X}_s \\ s \in \mathcal{O}^{\complement}}} \prod_{t \in \mathcal{S}} \mathcal{T}_t\left(x_t \mid \{x_i, i \in \mathcal{I}_t\}\right). \end{aligned}$$

Since the rank depends only on the system dynamics and not the reward structure, the same results about the low-rank structure hold for POSGs.

Proposition (Rank of POSGs)**.** *A partially-observable sequential game has rank at most* $\max_{h \in [H]} \left|\mathbb{I}_h^{\dagger}\right|$, *where* $\left|\mathbb{I}_h^{\dagger}\right|$ *is the size of the minimal d-separating set at time h, as defined in Definitions 5 and 6.*

Moreover, for the same reason, an identical construction of a PSR representation applies for partially-observable sequential games.

Proposition ($\mathcal{I}^{\dagger}$-weakly revealing POSGs are well-conditioned PSRs)**.** *Suppose a POSG is α-robustly m-step $\mathcal{I}^{\dagger}$-weakly revealing. Then, the m-step futures constitute core test sets and the corresponding PSR is γ-well-conditioned with* $\gamma = \alpha / \max_h \sqrt{\left|\mathbb{I}_h^{\dagger}\right|}$

8.3 Reinforcement Learning Algorithm

We now introduce a sample-efficient reinforcement learning algorithm for learning well-conditioned generalized predictive state representations in the game setting with each agent having their own objective. In particular, since partially-observable observable sequential games with a $\mathcal{I}^{\dagger}$-weakly revealing information structure have well-conditioned generalized PSR representations, they can also be learned sample-efficiently by this algorithm.

The algorithm we propose is a *self-play* algorithm for learning an *equilibrium* of the dynamic game problem. That is, the algorithm specifies the policies of all agents during the learning phase, collecting the trajectory of observables $\mathcal{U}$ at each episode to improve its estimate of the system dynamics. This can be thought of as a centralized agent playing against itself.

The algorithmic description is presented in Algorithm 2. In the first stage of the algorithm, the centralized agent has a unified goal: to explore the environment through policies which maximize the bonus function $\widehat{b}^k(\tau_H)$ by visiting trajectories with imprecise estimates of their probability, as measured by the upper confidence bound on the total variation distance. This part is identical to Algorithm 1. Once the algorithm is sufficiently confident about the estimated probabilities of all trajectories, it computes the equilibrium using the estimated model directly. That is, `ComputeEquilibrium` computes either NE or CCE. The only difference in the exploration stage of the algorithm compared to Algorithm 1 is that the termination condition involves $\varepsilon/4$ rather than $\varepsilon/2$ in order to guarantee an ε-approximate equilibrium under the added complications of the game setting.

Theorem 2. *Suppose Assumption 1 holds. Let* $p_{\min} = O\left(\frac{\delta}{KH \prod_{h=1}^{H} |\mathbb{X}_h|}\right)$, $\lambda = \frac{\gamma(\max_{s \in \mathcal{A}} |\mathbb{X}_s|)^2 Q_A \beta \max\left\{\sqrt{r}, Q_A \sqrt{H}/\gamma\right\}}{\sqrt{dH}}$, $\alpha = O\left(\frac{Q_A \sqrt{Hd}}{\gamma^2}\sqrt{\lambda} + \frac{\max_{s \in \mathcal{A}} |\mathbb{X}_s| Q_A \sqrt{\beta}}{\gamma}\right)$, *and let* $\beta = O(\log\left|\overline{\Theta}_{\varepsilon}\right|)$, *where* $\varepsilon = O(\frac{p_{\min}}{KH})$. *Then, with probability*

Algorithm 2: Self-play UCB Algorithm for Sequential Games

```
for $k \leftarrow 1, \dots, K$ do
    for $h \leftarrow 1, \dots, H$ do
        Collect $\tau_H^{k,h} = (\omega_{h-1}^{k,h}, \tau_{h-1}^{k,h})$ using $\nu(\pi^{k-1}, \mathtt{u}_{\mathbb{Q}_{h-1}^{\mathtt{exp}}})$.
        $\mathcal{D}_{h-1}^k \leftarrow \mathcal{D}_{h-1}^{k-1} \cup \left\{ \left( \tau_H^{k,h}, \nu(\pi^{k-1}, \mathtt{u}_{\mathbb{Q}_{h-1}^{\mathtt{exp}}}) \right) \right\}$.
    end
    $\mathcal{D}^k = \left\{ \mathcal{D}_h^k \right\}_{h=0}^{H-1}$
    Compute MLE $\widehat{\theta} \in \mathcal{B}^k$, where
        $$\Theta_{\min}^k = \left\{ \theta : \forall h, (\tau_h, \pi) \in \mathcal{D}_h^k, \mathbb{P}_\theta^\pi(\tau_h) \geq p_{\min} \right\},$$
        $$\mathcal{B}^k = \left\{ \theta \in \Theta_{\min}^k : \sum_{(\tau_H, \pi) \in \mathcal{D}^k} \log \mathbb{P}_\theta^\pi(\tau_H) \geq \max_{\theta' \in \Theta_{\min}^k} \sum_{(\tau_H, \pi) \in \mathcal{D}^k} \log \mathbb{P}_{\theta'}^\pi(\tau_H) - \beta \right\}.$$
    Define the bonus function, $\widehat{b}^k(\tau_H) = \min \left\{ \alpha \sqrt{\sum_{h=0}^{H-1} \left\| \widehat{\psi}(\tau_h) \right\|_{(\widehat{U}_h^k)^{-1}}^2}, 1 \right\}$, where
     $\widehat{U}_h^k = \lambda I + \sum_{\tau_h \in \mathcal{D}_{h^k}} \widehat{\psi}^k(\tau_h) \widehat{\psi}^k(\tau_h)^\top$.
    Solve the planning problem $\pi^k = \arg\max_\pi V_{\widehat{\theta}^k}^{\widehat{b}^k}(\pi)$.
    if $V_{\widehat{\theta}^k}^{\widehat{b}^k}(\pi^k) \leq \epsilon/4$ then
        $\theta^\epsilon = \widehat{\theta}^k$. break.
    end
end
return $\overline{\pi} = \texttt{Equilibrium}(\theta^\epsilon, \left\{ R^1, \dots, R^N \right\})$
```

at least $1 - \delta$, Algorithm 2 returns a model θ^ϵ and a policy $\overline{\pi}$ which is an ε-approximate equilibrium (either NE or CCE). That is,

$$V_{\theta^*}^i(\overline{\pi}) \geq V_{\theta^*}^{i,\dagger}(\overline{\pi}^{-i}) - \varepsilon, \; \forall i \in [N].$$

In addition, the algorithm terminates with a sample complexity of,

$$\tilde{O}\left(\left(r + \frac{Q_A^2 H}{\gamma^2} \right) \frac{r d H^3 \max_{s \in \mathcal{A}} |\mathbb{X}_s|^2 Q_A^4 \beta}{\gamma^4 \epsilon^2} \right).$$

Corollary 2. *Suppose a partially-observable sequential game is m-step α-robustly $\mathcal{I}^\dagger$-weakly revealing as per Definition 7. Applying Algorithm 2 to this PSR representation, with parameters $p_{\min}, \lambda, \alpha, \beta$ chosen as in Theorem 2, returns a ε-approximate equilibrium with a sample complexity of,*

$$\tilde{O}\left(\left(1 + \frac{Q_A^2 H}{\alpha^2} \right) \frac{\max_h \left| \mathbb{I}_h^\dagger \right|^7 \max_h |\mathbb{Q}_h^m| H^5 \max_{s \in \mathcal{A}} |\mathbb{X}_s|^2 \max_{s \in \mathcal{U}} |\mathbb{X}_s| Q_A^4}{\alpha^4 \epsilon^2} \right).$$

9 Discussion

The importance of information structures has long been recognized in the control theory literature, and has been the subject of extensive study since the 1970s. In this paper, we argue that explicitly modeling the information structure is crucial in the learning setting as well, leading to deeper analysis.

In this paper, we studied the role of information structure in learning sequential decision-making problems. To facilitate this, we proposed the formalism of partially-observable sequential teams and games. These models capture classical models of reinforcement learning such as MDPs and POMDPs as special cases where the information structure is fixed and simple. Through a DAG representation of the information structure, we obtained an interpretable graph-theoretic quantity which describes the rank of the observable dynamics of these models. This analysis gives a condition in terms of the information structure for when learning is tractable. We also proposed a generalization of predictive state representation which allows us to employ recent learning results to POSTs and POSGs. We proposed provably sample-efficient algorithms for learning optimal policies in the team setting and equilibria in the game setting.

References

Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved algorithms for linear stochastic bandits". In: *Advances in neural information processing systems* 24.

Abbasi-Yadkori, Yasin and Csaba Szepesvári (2011). "Regret bounds for the adaptive control of linear quadratic systems". In: *Proceedings of the 24th Annual Conference on Learning Theory.* JMLR Workshop and Conference Proceedings, pp. 1–26.

Andersland, Mark S and Demosthenis Teneketzis (1992). "Information structures, causality, and nonsequential stochastic control I: Design-independent properties". In: *SIAM journal on control and optimization* 30.6, pp. 1447–1475.

Carpentier, Alexandra, Claire Vernade, and Yasin Abbasi-Yadkori (2020). "The elliptical potential lemma revisited". In: *arXiv preprint arXiv:2010.10182.*

Chen, Fan, Yu Bai, and Song Mei (Sept. 2022). *Partially Observable RL with B-Stability: Unified Structural Condition and Sharp Sample-Efficient Algorithms.* DOI: `10.48550/arXiv.2209.14990`. arXiv: `2209.14990 [cs, math, stat]`.

Dani, Varsha, Thomas P Hayes, and Sham M Kakade (2008). "Stochastic linear optimization under bandit feedback". In.

Geer, Sara van de (2006). "Rates of Convergence for Maximum Likelihood Estimators". In: *Applications of Empirical Process Theory.* Reprint. Cambridge Series on Statistical and Probabilistic Mathematics 6. Cambridge: Cambridge Univ. Pr.

Ho, Yu-Chi et al. (1972). "Team decision theory and information structures in optimal control problems–Part I". In: *IEEE Transactions on Automatic control* 17.1, pp. 15–22.

Huang, Ruiquan, Yingbin Liang, and Jing Yang (July 2023). *Provably Efficient UCB-type Algorithms For Learning Predictive State Representations.* arXiv: `2307.00405 [cs, stat]`.

Jaeger, Herbert (June 2000). "Observable Operator Models for Discrete Stochastic Time Series". In: *Neural Computation* 12.6, pp. 1371–1398. DOI: `10.1162/089976600300015411`.

Lattimore, Tor and Marcus Hutter (2012). "PAC bounds for discounted MDPs". In: *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23.* Springer, pp. 320–334.

Littman, Michael and Richard S Sutton (2001). "Predictive representations of state". In: *Advances in neural information processing systems* 14.

Liu, Qinghua, Alan Chung, et al. (May 2022). *When Is Partially Observable Reinforcement Learning Not Scary?* DOI: `10.48550/arXiv.2204.08967`. arXiv: `2204.08967 [cs, eess, stat]`.

Liu, Qinghua, Praneeth Netrapalli, et al. (Nov. 2022). *Optimistic MLE – A Generic Model-based Algorithm for Partially Observable Sequential Decision Making.* DOI: `10.48550/arXiv.2209.14997`. arXiv: `2209.14997 [cs, stat]`.

Mahajan, Aditya, Nuno C. Martins, et al. (Dec. 2012). "Information Structures in Optimal Decentralized Control". In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC).* Maui, HI, USA: IEEE, pp. 1291–1306. DOI: `10.1109/CDC.2012.6425819`.

Mahajan, Aditya and Sekhar Tatikonda (June 2009). "A Graphical Modeling Approach to Simplifying Sequential Teams". In: *2009 7th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, pp. 1–8. DOI: `10.1109/WIOPT.2009.5291560`.

– (2011). "An Axiomatic Approach for Simplification of Sequential Teams". In.

Mahajan, Aditya and Demosthenis Teneketzis (2009). "Optimal performance of networked control systems with nonclassical information structures". In: *SIAM Journal on Control and Optimization* 48.3, pp. 1377–1404.

Malikopoulos, Andreas A. (Nov. 2022). *On Team Decision Problems with Nonclassical Information Structures.* arXiv: 2101.10992 [math].

Munos, Rémi and Csaba Szepesvári (2008). "Finite-Time Bounds for Fitted Value Iteration." In: *Journal of Machine Learning Research* 9.5.

Nash, John (1951). "Non-cooperative games". In: *Annals of mathematics*, pp. 286–295.

Nayyar, Ashutosh, Aditya Mahajan, and Demosthenis Teneketzis (July 2011). "Optimal Control Strategies in Delayed Sharing Information Structures". In: *IEEE Transactions on Automatic Control* 56.7, pp. 1606–1620. DOI: 10.1109/TAC.2010.2089381.

– (2014). "The Common-Information Approach to Decentralized Stochastic Control". In: *Information and Control in Networks.* Ed. by Giacomo Como, Bo Bernhardsson, and Anders Rantzer. Vol. 450. Cham: Springer International Publishing, pp. 123–156. DOI: 10.1007/978-3-319-02150-8_4.

Singh, Satinder et al. (2000). "Convergence results for single-step on-policy reinforcement-learning algorithms". In: *Machine learning* 38, pp. 287–308.

Sutton, Richard S, Hamid Maei, and Csaba Szepesvári (2008). "A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation". In: *Advances in neural information processing systems* 21.

Tatikonda, Sekhar Chandra (2000). "Control under Communication Constraints". Thesis. Massachusetts Institute of Technology.

Teneketzis, Demosthenis (1996). "On information structures and nonsequential stochastic control". In: *CWI Quarterly* 9.3, pp. 241–260.

Witsenhausen, H. S. (1975). "The Intrinsic Model for Discrete Stochastic Control: Some Open Problems". In: *Control Theory, Numerical Methods and Computer Systems Modelling.* Ed. by M. Beckmann et al. Vol. 107. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 322–335. DOI: 10.1007/978-3-642-46317-4_24.

– (Feb. 1988). "Equivalent Stochastic Control Problems". In: *Mathematics of Control, Signals, and Systems* 1.1, pp. 3–11. DOI: 10.1007/BF02551232.

Witsenhausen, Hans S (1971). "On information structures, feedback and causality". In: *SIAM Journal on Control* 9.2, pp. 149–160.

Yoshikawa, Tsuneo (1978). "Decomposition of dynamic team decision problems". In: *IEEE Transactions on Automatic Control* 23.4, pp. 627–632.

Zhan, Wenhao et al. (Aug. 2022). *PAC Reinforcement Learning for Predictive State Representations.* arXiv: 2207.05738 [cs].

Zhang, Kaiqing, Zhuoran Yang, and Tamer Başar (2021). "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms". In: *Handbook of Reinforcement Learning and Control.* Ed. by Kyriakos G. Vamvoudakis et al. Vol. 325. Cham: Springer International Publishing, pp. 321–384. DOI: 10.1007/978-3-030-60990-0_12.

A Existence of Generalized PSR representations and their covering number

Notation. Let $(X_1, \dots, X_H)$ denote a controlled stochastic process associated with a sequential decision making problem. Let $\mathcal{O} \subset [H]$ denote the index set of observations and $\mathcal{A} \subset [H]$ denote the index set of actions, such that $\mathcal{O}, \mathcal{A}$ partition $[H]$. Let $\mathbb{H}_h = \prod_{s \in 1:h} \mathbb{X}_s$ denote the space of histories at time h and $\mathbb{F}_h = \prod_{s \in h+1:H} \mathbb{X}_s$ denote the space futures at time t. Similarly, let $\mathbb{H}_h^o = \texttt{obs}(\mathbb{H}_h) = \prod_{s \in \mathcal{O}_{1:h}} \mathbb{X}_s$ denote the observation component of histories and let $\mathbb{H}_h^a = \texttt{act}(\mathbb{H}_h) = \prod_{s \in \mathcal{A}_{1:h}} \mathbb{X}_s$ denote the action component. Define $\mathbb{F}_h^p$ and $\mathbb{F}_h^a$ similarly. Let $\boldsymbol{D}_h \in \mathbb{R}^{|\mathbb{H}_h| \times |\mathbb{F}_h|}$ denote the dynamics matrix, defined as $[\boldsymbol{D}_h]_{\tau_h, \omega_H} := \overline{\mathbb{P}}[\tau_h, \omega_h] = \mathbb{P}[\tau_h^o, \omega_h^o \mid (\tau_h^a, \omega_h^a)]$. $\boldsymbol{D}_H \in \mathbb{R}^{|\mathbb{H}_H| \times 1}$ is defined as $[\boldsymbol{D}_H]_{\tau_H} = \overline{\mathbb{P}}[\tau_H]$. Note that $\boldsymbol{D}_0 = \boldsymbol{D}_H^\top$.

Let Θ be the parameter space of a sequential decision making problem. An "optimistic net" of $\overline{\Theta}_\varepsilon$ is a finite set such that for any $\theta \in \Theta$, there exists $\overline{\theta} \in \overline{\Theta}_\varepsilon$ parameterizing a measure $\mathbb{P}_{\overline{\theta}}$ over $\mathbb{X}_1 \times \cdots \times \mathbb{X}_H$ such that,

$$\begin{aligned} &\text{(optimism)} \quad \forall \pi, h,\ \mathbb{P}_{\overline{\theta}}^\pi(\tau_h) \geq \mathbb{P}_\theta^\pi(\tau_h), \\ &(\varepsilon\text{-covering}) \quad \forall \pi, h\ \sum_{\tau_h} \left| \mathbb{P}_{\overline{\theta}}^\pi(\tau_h) - \mathbb{P}_\theta^\pi(\tau_h) \right| \leq \varepsilon. \end{aligned}$$

The cardinality of the optimistic net plays a role in the definition of β and the MLE analysis.

In this section we show that any rank-r sequential decision-making problem admits a (self-consistent) rank-r PSR representation (i.e., observable operator model). This implies a parameter space Θ. Next, we bound the cardinality of optimistic net $\overline{\Theta}_\varepsilon$ of this parameter space, bounding the log-covering number for the space of PSR representations. Such a result has been established in previous work for sequential decision-making problems with alternating observations and actions (Liu, Netrapalli, et al. 2022). Here, we follow a similar procedure to prove a slightly generalized result.

Proposition 4 (Existence of OOM representation)**.** *Consider a sequential decision-making problem with* $\operatorname{rank}(\boldsymbol{D}_h) = r_h$, $h \in 0 : H-1$. *There exists an OOM representation* $b_0, \{B_h(x_h)\}_{h \in [H], x_h \in \mathbb{X}_h}, \{v_h\}_{h \in 0:H}$ *such that,*

1. $B_h(x_h) \in \mathbb{R}^{r_h \times r_{h-1}}$ *and* $\|B_h(x_h)\|_2 \leq 1$ *for any* x_h.
2. $|b_0| \leq \sqrt{|\mathbb{H}_H^a|}$.
3. $\|v_h\|_2 \leq \sqrt{|\mathbb{F}_h^o| \,/\, |\mathbb{F}_h^a|}$.
4. *For any* h, $\frac{1}{|\mathbb{X}_h| \mathbf{1}\{h \in \mathcal{A}\}} v_h^\top \sum_{x_h \in \mathbb{X}_h} B_h(x_h) = v_{h-1}^\top$.
5. *For any* $\tau_h \in \mathbb{H}_h$, $\overline{\mathbb{P}}[\tau_h] = v_h^\top B_h(x_h) \cdots B_1(x_1) b_0$.

Proof. We construct the OOM representation via the singular value decomposition of the matrix $\boldsymbol{D}_h^\top$. Let $U_h \in \mathbb{R}^{|\mathbb{F}_h| \times r_h}, \Sigma_h \in \mathbb{R}^{r_h \times r_h}, V_h^\top \in \mathbb{R}^{r_h \times |\mathbb{H}_h|}$ be the SVD such that $\boldsymbol{D}_h^\top = U_h \Sigma_h V_h^\top$. Define $b_0, B_h, v_h^\top$ as follows,

$$b_0 = \|\boldsymbol{D}_0\|_2\,, \quad B_h(x_h) = U_h^\top \left[U_{h-1}\right]_{(x_h, \mathbb{F}_h), :}\,, \quad v_h^\top = \frac{1}{|\mathbb{F}_h^a|} \mathbf{1}^\top U_h.$$

Here, $[U_{h-1}]_{(x_h, \Omega_h), :}$ denotes an $|\mathbb{F}_h|$ by r_{h-1} submatrix of U_{h-1} consisting of the rows (x_h, ω_h), $\omega_h \in \mathbb{F}_h$ (i.e., the set of futures where the variable at time h is x_h). Note that $|\mathbb{F}_H^a| = 1$ by convention, a product over an empty set. We verify each property in turn.

First, $\|B_h(x_h)\|_2 = \left\|U_h^\top [U_{h-1}]_{(x_h,\mathbb{F}_h),:}\right\|_2 \leq 1$ since U_h, U_{h-1} are unitary matrices. Second,

$$\begin{aligned}|b_0| = \|\boldsymbol{D}_0\|_2 &= \sqrt{\sum_{\tau_H} \overline{\mathbb{P}}\left[\tau_H\right]^2}\\ &\leq \sqrt{\sum_{\tau_H} \overline{\mathbb{P}}\left[\tau_H\right]} = \sqrt{\sum_{\tau_H^a}\sum_{\tau_H^o} \mathbb{P}\left[\tau_H^o \mid \tau_H^a\right]} = \sqrt{\sum_{\tau_H^a} 1} = \sqrt{\prod_{s\in\mathcal{A}} |\mathbb{X}_s|},\end{aligned}$$

where the inequality is since $\overline{\mathbb{P}}\left[\tau_H\right] \in [0,1]$. For property 3, we have

$$\begin{aligned}\|v_h\|_2 &= \frac{1}{|\mathbb{F}_h^a|}\left\|\mathbf{1}^\top U_h\right\|_2\\ &\leq \frac{1}{|\mathbb{F}_h^a|}\|\mathbf{1}\|_2 = \frac{\sqrt{|\mathbb{F}_h|}}{|\mathbb{F}_h^a|} = \sqrt{|\mathbb{F}_h^o| \,/\, |\mathbb{F}_h^a|},\end{aligned}$$

where the inequality is since U_h is unitary, and the final equality is since $|\mathbb{F}_h| = |\mathbb{F}_h^o|\,|\mathbb{F}_h^a|$.

Next, to prove properties 4 and 5, we first show the following claim.

Claim. *For any history $\tau_h = (x_1,\ldots,x_h) \in \mathbb{H}_h$, $h \in 0:H$, we have $B_h(x_h)\cdots B_1(x_1)b_0 = U_h^\top \left[\boldsymbol{D}_h^\top\right]_{:,\tau_h}$.*

Proof of claim. We prove the claim by induction. In the base case, $h=0$, $\boldsymbol{D}_0^\top$ is a vector in $\mathbb{R}^{\mathbb{F}_0}$ (note that $\mathbb{F}_0 = \mathbb{H}_H$). Hence, U_0 is simply the normalized vector $U_0 = \boldsymbol{D}_0^\top / \left\|\boldsymbol{D}_0^\top\right\|_2$, and hence $U_0^\top \boldsymbol{D}_0^\top = \boldsymbol{D}_0\boldsymbol{D}_0^\top / \|\boldsymbol{D}_0\|_2 = \|\boldsymbol{D}_0\|_2 = b_0$. Proceeding by induction, suppose the claim holds for $h-1$. Then, we have,

$$\begin{aligned}B_h(x_h)\cdots B_1(x_1)b_0 &= B_h(x_h)U_{h-1}^\top\left[\boldsymbol{D}_{h-1}^\top\right]_{:,\tau_{h-1}}\\ &= U_h^\top [U_{h-1}]_{(x_h,\mathbb{F}_h),:} U_{h-1}^\top \left[\boldsymbol{D}_{h-1}^\top\right]_{:,\tau_{h-1}}\\ &= U_h^\top \left[U_{h-1}U_{h-1}^\top \boldsymbol{D}_{h-1}^\top\right]_{(x_h,\mathbb{F}_h),\tau_{h-1}}\\ &= U_h^\top \left[\boldsymbol{D}_{h-1}^\top\right]_{(x_h,\mathbb{F}_h),\tau_{h-1}}\\ &= U_h^\top \left[\boldsymbol{D}_h^\top\right]_{:,\tau_h},\end{aligned}$$

where the final equality is because $\left[\boldsymbol{D}_{h-1}^\top\right]_{(x_h,\omega_h),\tau_{h-1}} = \overline{\mathbb{P}}\left[\tau_{h-1}, x_h, \omega_h\right] = \overline{\mathbb{P}}\left[\tau_h, \omega_h\right] = \left[\boldsymbol{D}_h^\top\right]_{\omega_h,\tau_h}$. □

Using this fact, we can now show property 5 as follows,

$$\begin{aligned}v_h^\top B_h(x_h)\cdots B_1(x_1)b_0 &= \frac{1}{|\mathbb{F}_h^a|}\mathbf{1}^\top U_h U_h^\top \left[\boldsymbol{D}_h^\top\right]_{:,\tau_h}\\ &= \frac{1}{|\mathbb{F}_h^a|}\mathbf{1}^\top \left[\boldsymbol{D}_h^\top\right]_{:,\tau_h}\\ &= \frac{1}{|\mathbb{F}_h^a|}\sum_{\omega_h\in\mathbb{F}_h} \overline{\mathbb{P}}\left[\tau_h,\omega_h\right]\\ &= \frac{1}{|\mathbb{F}_h^a|}\sum_{\omega_h^a\in\mathbb{F}_h^a}\sum_{\omega_h^o\in\mathbb{F}_h^o} \mathbb{P}\left[\tau_h^o,\omega_h^o \mid \tau_h^a,\omega_h^a\right]\\ &= \frac{1}{|\mathbb{F}_h^a|}\mathbb{P}\left[\tau_h^o \mid \tau_h^a\right]\sum_{\omega_h^a\in\mathbb{F}_h^a}\sum_{\omega_h^o\in\mathbb{F}_h^o} \mathbb{P}\left[\omega_h^o \mid \omega_h^a,\tau_h^a,\tau_h^o\right]\\ &= \frac{1}{|\mathbb{F}_h^a|}\mathbb{P}\left[\tau_h^o \mid \tau_h^a\right]\sum_{\omega_h^a\in\mathbb{F}_h^a} 1\\ &= \mathbb{P}\left[\tau_h^o \mid \tau_h^a\right].\end{aligned}$$

Finally, it remains to show property 4. Consider the linear equation $x^\top U_h^\top \boldsymbol{D}_h^\top = \frac{1}{|\mathbb{F}_h^a|}\mathbf{1}^\top \boldsymbol{D}_h^\top$. Note that $U_h^\top \boldsymbol{D}_h^\top \in \mathbb{R}^{r_h \times |\mathbb{H}_h|}$ is rank r_h. Thus, this equation has a unique solution. Our strategy is to show that $v_h^\top$ and $v_{h+1}^\top \sum_{x_{h+1}} B_{h+1}(x_h)$ are both solutions to this linear equation, and hence $v_h^\top = v_{h+1}^\top \sum_{x_{h+1}} B_{h+1}(x_h)$. That $v_h^\top$ is a solution is clear by definition of v_h, $v_h^\top U_h^\top \boldsymbol{D}_h^\top = \frac{1}{|\mathbb{F}_h^a|}\mathbf{1}^\top U_h U_h^\top \boldsymbol{D}_h^\top = \frac{1}{|\mathbb{F}_h^a|}\mathbf{1}^\top \boldsymbol{D}_h^\top$. First, recall by the calculation above that $\frac{1}{|\mathbb{F}_h^a|}\mathbf{1}^\top \boldsymbol{D}_h^\top$ is a vector in $\mathbb{R}^{\mathbb{H}_h}$ where the τ_h-th entry is $\mathbb{P}\left[\tau_h^o \mid \tau_h^a\right]$. We will calculate the τ_h-th entry of the vector $x^\top U_h^\top \boldsymbol{D}_h$ when $x^\top = v_{h+1}^\top \sum_{x_{h+1}} B_{h+1}(x_{h+1})$,

$$\begin{aligned}
\left(v_{h+1}^\top \sum_{x_{h+1}} B_{h+1}(x_{h+1})\right)\left[U_h^\top D_h^\top\right]_{:,\tau_h} &= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\sum_{x_{h+1}} \mathbf{1}^\top U_{h+1} U_{h+1}^\top \left[U_h\right]_{(x_{h+1},\mathbb{F}_{h+1}),:}\left[U_h^\top D_h^\top\right]_{:,\tau_h} \\
&= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\sum_{x_{h+1}} \mathbf{1}^\top U_{h+1} U_{h+1}^\top \left[U_h U_h^\top D_h^\top\right]_{(x_{h+1},\mathbb{F}_{h+1}),\tau_h} \\
&= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\sum_{x_{h+1}} \left[\mathbf{1}^\top D_h^\top\right]_{(x_{h+1},\mathbb{F}_{h+1}),\tau_h} \\
&= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\sum_{x_{h+1}}\sum_{\omega_{h+1}} \overline{\mathbb{P}}\left[\tau_h, x_{h+1}, \omega_{h+1}\right] \\
&= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\sum_{\omega_h \in \mathbb{F}_h} \overline{\mathbb{P}}\left[\tau_h, \omega_h\right] \\
&= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\mathbb{P}\left[\tau_h^o \mid \tau_h^a\right]\sum_{\omega_h^a \in \mathbb{F}_h^a}\sum_{\tau_h^o \in \mathbb{F}_h^o} \mathbb{P}\left[\omega_h^o \mid \tau_h, \omega_h^a\right] \\
&= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\mathbb{P}\left[\tau_h^o \mid \tau_h^a\right]\sum_{\omega_h^a \in \mathbb{F}_h^a} 1 \\
&= \frac{1}{\left|\mathbb{F}_{h+1}^a\right|}\left|\mathbb{F}_h^a\right|\mathbb{P}\left[\tau_h^o \mid \tau_h^a\right] \\
&= \frac{1}{\left|\mathbb{X}_h\right|\mathbf{1}\{h \in \mathcal{A}\}}\mathbb{P}\left[\tau_h^o \mid \tau_h^a\right],
\end{aligned}$$

where the final inequality is since $|\mathbb{F}_h^a| = \prod_{s \in h+1:H}(|\mathbb{X}_s|\,\mathbf{1}\{s \in \mathcal{A}\})$. □

Corollary 3. *Consider a sequential decision-making problem with* $\operatorname{rank}(D_h) \leq r$. *Then, there exists an OOM representation* $b_0 \in \mathbb{R}^r$, $\{B_h(x_h)\}_{h \in [H], x_h \in \mathbb{X}_h} \subset \mathbb{R}^{r \times r}$, $v_H \in \mathbb{R}^r$ *such that,*

1. $\|B_h(x_h)\|_2 \leq 1$, $\forall h, x_h \in \mathbb{X}_h$, $\|b_0\|_2 \leq \sqrt{|\mathbb{H}_H^a|}$, *and* $\|v_H\|_2 \leq 1$.
2. *For any* $\tau_H \in \mathbb{H}_H$, $\overline{\mathbb{P}}\left[\tau_H\right] = v_H^\top B_H(x_H) \cdots B_1(x_1) b_0$.

Proof. In Proposition 4 we constructed such a representation with dimensions in terms of r_h instead of r. Since $r_h \leq r$, we can pad this representation with dummy columns and/or rows filled with zeros to obtain a representation with dimensions in terms of r. □

The MLE analysis in the next section depends on the size of the (optimistic) ε-covering number of the space of sequential decision-making problems (e.g., PSRs) Θ. $\overline{\Theta}_\varepsilon$ is said to be an optimistic ε-cover if for each $\theta \in \Theta$, there exists $\widehat{\theta} \in \overline{\Theta}_\varepsilon$ with an associated probability measure $\overline{\mathbb{P}}_{\widehat{\theta}}^{\varepsilon}$ such that,

$$\begin{gathered}
\forall h, \tau_h,\ \overline{\mathbb{P}}_{\widehat{\theta}}^{\varepsilon}(\tau_h) \geq \overline{\mathbb{P}}_\theta\left[\tau_h\right], \\
\forall h, \tau_h,\ \sum_{\tau_h}\left|\overline{\mathbb{P}}_{\widehat{\theta}}^{\varepsilon}(\tau_h) - \overline{\mathbb{P}}_\theta\left[\tau_h\right]\right| \leq \varepsilon.
\end{gathered}$$

The first condition ensures optimism and the second condition ensures that $\mathcal{C}_\delta$ ε-covers Θ. The next proposition bounds the size of $\left|\overline{\Theta}_\varepsilon\right|$.

Proposition 5 (Bracketing number of sequential decision making problems)**.** *Let $\mathfrak{M}$ be the set of all rank-r sequential decision-making problems with horizon of length H, observation index set $\mathcal{O} \subset [H]$, action index set $\mathcal{A} \subset [H]$, and variable spaces $\mathbb{X}_1, \dots, \mathbb{X}_H$. Then, the bracketing number of this set is bounded by,*

$$\log \mathcal{N}_\varepsilon(\mathfrak{M}) \le O\left(r^2 \max_s |\mathbb{X}_s| \, H^2 \log(r \max_s |\mathbb{X}_s| \, H/\varepsilon)\right).$$

Proof. Define the set of OOM representations constructed in Corollary 3,

$$\begin{aligned}\Theta \coloneqq \Bigg\{ b_0 \in \mathbb{R}^r, \{B_h(x_h)\}_{h,x_h}, v_H \in \mathbb{R}^r :\ & \|B_h(x_h)\|_2 \le 1, \forall h, x_h, \ \|b_0\|_2 \le \sqrt{|\mathbb{H}_H^a|}, \ \|v_H\|_2 \le 1, \\ & \text{and } \forall\, \tau_H \in \mathbb{H}_H, \ \overline{\mathbb{P}}_m[\tau_H] = v_H^\top B_H(x_H) \cdots B_1(x_1) b_0, \\ & \text{where } m \text{ is a sequential decision making problem in } \mathfrak{M} \Bigg\}.\end{aligned}$$

Let let $\mathcal{C}_\delta$ be a δ-cover of the above set with respect to the ℓ_∞-norm. For $\widehat{\theta} = (b_0, \{B_h(x_h)\}, v_H) \in \mathcal{C}_\delta$, define the ε-optimistic probabilities as,

$$\overline{\mathbb{P}}^{\varepsilon}_{\widehat{\theta}}(\tau_H) \coloneqq v_H^\top B_H(x_h) \cdots B_1(x_1) b_0 + \varepsilon/2$$

Let $\delta \coloneqq \varepsilon \max_s |\mathbb{X}_s|^{-cH}$, for c large enough such that for each $\theta \in \Theta$, there exists $\widehat{\theta} \in \mathcal{C}_\delta$ such that,

$$\begin{aligned}\forall h, \tau_h, \ & \overline{\mathbb{P}}^{\varepsilon}_{\widehat{\theta}}(\tau_h) \ge \overline{\mathbb{P}}_\theta[\tau_h], \\ \forall h, \tau_h, \ & \sum_{\tau_h} \left| \overline{\mathbb{P}}^{\varepsilon}_{\widehat{\theta}}(\tau_h) - \overline{\mathbb{P}}_\theta[\tau_h] \right| \le \varepsilon.\end{aligned}$$

We call $\mathcal{C}_\delta$ an optimistic ε-cover of Θ and denote it $\overline{\Theta}_\varepsilon$. Note that there exists c large enough such that the above holds since,

$$\begin{aligned}& \sum_{\tau_H} \left| \widehat{v}_H^\top \widehat{B}_H(x_H) \cdots B_1(x_1) \widehat{b}_0 - v_H^\top B_H(x_H) \cdots B_1(x_1) b_0 \right| \\ & \le \sum_{h=1}^{H} \sum_{\tau_H} \left| \widehat{v}_H^\top \widehat{B}_H(x_H) \cdots \widehat{B}_{h+1}(x_{h+1}) (\widehat{B}_h(x_h) - B_h(x_h)) B_{h-1}(x_{h-1}) \cdots B_1(x_1) b_0 \right| \\ & \quad + \sum_{\tau_H} \left| \widehat{v}_H^\top B_H(x_H) \cdots B_1(x_1)(\widehat{b}_0 - b_0) \right| \\ & \le \sum_h \sum_{\tau_H} r \left\| \widehat{B}_h(x_h) - B_h(x_h) \right\|_{\max} \sqrt{|\mathbb{H}_H^a|} + \sum_{\tau_H} \sqrt{r} \left\| \widehat{b}_0 - b_0 \right\|_\infty \\ & \le H \max_h |\mathbb{X}_h|^{H + |\mathcal{A}|/2} r\delta + \max_h |\mathbb{X}_h|^H \sqrt{r}\delta,\end{aligned}$$

where the second inequality uses $\|\widehat{v}_H\|_2 = \|v_H\|_2 = \|B_h(x_h)\|_2 = 1$, $\left\| \widehat{B}_h(x_h) - B_h(x_h) \right\|_2 \le r \left\| \widehat{B}_h(x_h) - B_h(x_h) \right\|_{\max} \le r\delta$, $\|b_0\|_2 \le \sqrt{|\mathbb{H}_H^a|}$, and $\left\| \widehat{b}_0 - b_0 \right\|_2 \le \sqrt{r} \left\| \widehat{b}_0 - b_0 \right\|_\infty \le \sqrt{r}\delta$. This shows that $\delta = \varepsilon \max_h |\mathbb{X}_h|^{-cH}$ achieves ε-optimistic covering for an absolute constant c large enough. It remains to bound the size of $\mathcal{C}_\delta = \overline{\Theta}_\varepsilon$.

Recall that $\|\cdot\|_\infty \le \|\cdot\|_2$ and that an interval $[-x, x]$ in $\mathbb{R}$ admits a δ-cover of size bounded by $2x/\delta$. Now, observe that $\max_{ij} |[B_h(x_h)]_{ij}| \le \|B_h(x_h)\|_2 \le 1$. Hence, for a fixed h, $\{B_h(x_h)\}_{x_h}$ admits a cover of size bounded by $(2/\delta)^{r^2 |\mathbb{X}_h|}$. Considering all h, the cover is bounded by $(2/\delta)^{r^2 \max_h |\mathbb{X}_h| H}$. For, b_0, we have $\|b_0\|_\infty \le \|b_0\|_2 \le \sqrt{|\mathbb{H}_H^a|}$, hence the covering number is bounded by $(2\sqrt{|\mathbb{H}_H^a|}/\delta)^r$. Finally for v_H, we have $\|v_H\|_\infty \le \|v_H\|_2 \le 1$, hence the covering number is bounded by $(2/\delta)^r$. Thus, we have,

$$\log \left|\overline{\Theta}_\varepsilon\right| \le O\left(r^2 \max_h |\mathbb{X}_h| \, H \log\left(\frac{1}{\delta}\right)\right).$$

Recalling that $\delta = \varepsilon \max_s |\mathbb{X}_s|^{-cH}$, we obtain that,

$$\log \left|\overline{\Theta}_\varepsilon\right| \le O\left(r^2 \max_h |\mathbb{X}_h| \, H^2 \log\left(\frac{\max_h |\mathbb{X}_h|}{\epsilon}\right)\right).$$

□

B Proofs of Section 4

Proposition (Restatement of Proposition 2)**.** *A (partially-observable) sequential team has a rank bounded by,*

$$r \le \max_{h\in[H]} \left|\mathbb{I}_h^\dagger\right|.$$

Proof. We have

$$\begin{aligned}
[\boldsymbol{D}_h]_{\tau_h,\omega_h} &= \mathbb{P}\left[\tau_h^o, \omega_h^o \mid \mathrm{do}(\tau_h^a, \tau_h^a)\right] \\
&= \mathbb{P}\left[\tau_h^o \mid \mathrm{do}\left(\tau_h^a\right)\right] \mathbb{P}\left[\omega_h^o \mid \tau_h^o; \mathrm{do}\left(\tau_h^a, \omega_h^a\right)\right] \\
&\overset{(a)}{=} \mathbb{P}\left[\tau_h^o \mid \mathrm{do}\left(\tau_h^a\right)\right] \sum_{\substack{x_k \in \mathbb{X}_k \\ k \in \mathcal{I}_h^\dagger}} \mathbb{P}\left[\left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\} \,\middle|\, \tau_h^o; \mathrm{do}\left(\tau_h^a, \omega_h^a\right)\right] \mathbb{P}\left[\omega_h^o \,\middle|\, \left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\}, \tau_h^o; \mathrm{do}\left(\tau_h^a, \omega_h^a\right)\right] \\
&\overset{(b)}{=} \sum_{\substack{x_k \in \mathbb{X}_k \\ k \in \mathcal{I}_h^\dagger}} \mathbb{P}\left[\tau_h^o \mid \mathrm{do}\left(\tau_h^a\right)\right] \mathbb{P}\left[\left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\} \,\middle|\, \tau_h^o; \mathrm{do}\left(\tau_h^a\right)\right] \mathbb{P}\left[\omega_h^o \,\middle|\, \left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\}, \tau_h^o; \mathrm{do}\left(\tau_h^a, \omega_h^a\right)\right] \\
&\overset{(c)}{=} \sum_{\substack{x_k \in \mathbb{X}_k \\ k \in \mathcal{I}_h^\dagger}} \mathbb{P}\left[\tau_h^o \mid \mathrm{do}\left(\tau_h^a\right)\right] \mathbb{P}\left[\left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\} \,\middle|\, \tau_h^o; \mathrm{do}\left(\tau_h^a\right)\right] \mathbb{P}\left[\omega_h^o \,\middle|\, \left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\}; \mathrm{do}\left(\omega_h^a\right)\right],
\end{aligned}$$

where step (a) is simply the law of total probability, step (b) is that $\{x_k, k \in \mathcal{I}_h^\dagger\}$ is conditionally independent of $\mathrm{do}(\omega_h^a)$ (future actions) given $(\tau_h^o; \mathrm{do}(\tau_h^a))$ (the past), and step (c) is that ω_h^o is conditionally independent of $(\tau_h^o; \mathrm{do}(\tau_h^a))$ given $\{x_k,\, k \in \mathcal{I}_h^\dagger\}$, by definition of $\mathcal{I}_h^\dagger$ as the minimal set which d-separates $(X_{t(1)}, \ldots, X_{t(h)})$ from $(X_{t(h+1)}, \ldots, X_{t(H)})$. Note that

$$\mathbb{P}\left[\left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\} \,\middle|\, \tau_h^o; \mathrm{do}\left(\tau_h^a\right)\right] = \mathbb{P}\left[\left\{x_k,\, k \in \mathcal{I}_h^\dagger \cap \mathcal{S}\right\} \,\middle|\, \tau_h^o; \mathrm{do}\left(\tau_h^a\right)\right] \mathbf{1}\left\{\left(x_k,\, k \in \mathcal{I}_h^\dagger \cap \mathcal{A}\right) \text{ matches } \tau_h^a\right\},$$

since the action components of $i_h^\dagger = (x_k, k \in \mathcal{I}_h^\dagger)$ are contained in the history τ_h.

Now define two matrices

$$\begin{aligned}
\boldsymbol{D}_{h,1} &\coloneqq \left[\mathbb{P}\left[\tau_h^o \mid \mathrm{do}\left(\tau_h^a\right)\right] \mathbb{P}\left[\left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\} \,\middle|\, \tau_h^o; \mathrm{do}\left(\tau_h^a\right)\right]\right]_{\tau_h, i_h^\dagger}, \quad \tau_h \in \mathbb{H}_h,\ i_h^\dagger \equiv \left(x_k,\, k \in \mathcal{I}_h^\dagger\right) \in \mathbb{I}_h^\dagger, \\
\boldsymbol{D}_{h,2} &\coloneqq \left[\mathbb{P}\left[\omega_h^o \,\middle|\, \left\{x_k,\, k \in \mathcal{I}_h^\dagger\right\}; \mathrm{do}\left(\omega_h^a\right)\right]\right]_{i_h^\dagger, \omega_h}, \quad i_h^\dagger \equiv \left(x_k,\, k \in \mathcal{I}_h^\dagger\right) \in \mathbb{I}_h^\dagger,\ \omega_h \in \mathbb{F}_h.
\end{aligned}$$

We have that $\boldsymbol{D}_h = \boldsymbol{D}_{h,1}\boldsymbol{D}_{h,2}$, where both $\boldsymbol{D}_{h,1}$ and $\boldsymbol{D}_{h,2}$ have rank upper bounded by $\left|\mathbb{I}_h^\dagger\right| = \prod_{s \in \mathcal{I}_h^\dagger} |\mathbb{X}_s|$. Hence, $\mathrm{rank}(\boldsymbol{D}_h) \le \left|\mathbb{I}_h^\dagger\right|$, and the result follows. □

C Proofs of Section 6

Lemma (Restatement of Appendix C)**.** *Suppose that the POST is m-step $\mathcal{I}^\dagger$-weakly revealing. Then, $\mathbb{Q}_h^m$ is a core test set for all $h \in [H]$. Furthermore, we have*

$$\overline{\mathbb{P}}\left[\tau_h, \omega_h\right] = \langle m_h(\omega_h), \psi_h(\tau_h)\rangle, \ \text{and} \ \ \overline{\mathbb{P}}\left[\omega_h \mid \tau_h\right] = \left\langle m_h(\omega_h), \overline{\psi}_h(\tau_h)\right\rangle. \tag{26}$$

Proof. Let $\tau_h \in \mathbb{H}_h$, $\omega_h \in \mathbb{F}_h$ be any history and future, respectively. By Proposition 2, recall that we have

$$\overline{\mathbb{P}}[\omega_h \,|\, \tau_h] = \sum_{i_h^\dagger \in \mathbb{I}_h^\dagger} \overline{\mathbb{P}}\left[\omega_h \,\middle|\, i_h^\dagger\right] \mathbb{P}\left[i_h^\dagger \,\middle|\, \tau_h\right]. \tag{27}$$

Recall that $i_h^\dagger$ may overlap with τ_h. In particular, the action component of $i_h^\dagger$ is contained in τ_h. Thus, $\mathbb{P}[i_h^\dagger \,|\, \tau_h] = \mathbb{P}[\{x_k,\, k \in \mathcal{I}_h^\dagger \setminus \mathcal{U}_{1:h}\} \,|\, \tau_h] \cdot \mathbf{1}\{(x_k,\, k \in \mathcal{I}_h^\dagger \cap \mathcal{U}_{1:h}) \text{ matches } \tau_h\}$. Note that $\mathcal{I}_h^\dagger \setminus \mathcal{U}_{1:h} \subset \mathcal{S}$ does not contain any actions. Hence, the summation over $\mathbb{I}_h^\dagger$ is equivalent to summing over its unobservable components with the restriction that its observable components match τ_h.

Define the mappings $\tilde{m}_h \colon \mathbb{F}_h \to \mathbb{R}^{|\mathbb{I}_h^\dagger|}$ and $p_h \colon \mathbb{H}_h \to \mathbb{R}^{|\mathbb{I}_h^\dagger|}$ by,

$$\tilde{m}_h(\omega_h) = \left[\overline{\mathbb{P}}\left[\omega_h \,\middle|\, i_h^\dagger\right]\right]_{i_h^\dagger \in \mathbb{I}_h^\dagger}, \quad p_h(\tau_h) = \left[\mathbb{P}\left[i_h^\dagger \,\middle|\, \tau_h\right]\right]_{i_h^\dagger \in \mathbb{I}_h^\dagger},$$

Then, we have

$$\overline{\mathbb{P}}\left[\omega_h \mid \tau_h\right] = \langle \tilde{m}_h(\omega_h), p_h(\tau_h) \rangle .$$

Recall that the vector of (conditional) core test set probabilities for the history τ_h is given by

$$\overline{\psi}_h(\tau_h) \equiv [\mathbb{P}\left[q^o \mid \tau_h^o;\, \mathrm{do}(\tau_h^a),\, \mathrm{do}(q^a)\right]]_{q \in \mathbb{Q}_h^m} \in \mathbb{R}^{|\mathbb{Q}_h^m|}.$$

By the definition of $\boldsymbol{G}_h$ and Equation (27), we have $\boldsymbol{G}_h\, p_h(\tau_h) = \psi_h(\tau_h)$, since, for $q \in \mathbb{Q}_h^m$,

$$\begin{aligned}
\left(\boldsymbol{G}_h\, p_h(\tau_h)\right)_q &= \sum_{i_h^\dagger} \left(\boldsymbol{G}_h\right)_{q, i_h^\dagger} \left(p_h(\tau_h)\right)_{i_h^\dagger} \\
&= \sum_{i_h^\dagger} \mathbb{P}\left[q^o \,\middle|\, i_h^\dagger;\, \mathrm{do}(q^a)\right] \mathbb{P}\left[i_h^\dagger \,\middle|\, \tau_h\right] \\
&= \mathbb{P}\left[q^o \mid \tau_h^o;\, \mathrm{do}(\tau_h^a), \mathrm{do}(q^a)\right] \\
&=: \left[\overline{\psi}_h(\tau_h)\right]_q
\end{aligned}$$

Since by assumption $\mathrm{rank}(\boldsymbol{G}_h) = \left|\mathbb{I}_h^\dagger\right|$, its pseudo-inverse $\boldsymbol{G}_h^\dagger$ is a left inverse of $\boldsymbol{G}_h$ (i.e., $\boldsymbol{G}_h^\dagger \boldsymbol{G}_h = I$). Hence, multiplying on the left by $\boldsymbol{G}_h^\dagger$, we obtain

$$p_h(\tau_h) = \boldsymbol{G}_h^\dagger \overline{\psi}_h(\tau_h).$$

Hence,

$$\begin{aligned}
\overline{\mathbb{P}}\left[\omega_h \mid \tau_h\right] &= \left\langle \tilde{m}_h(\omega_h), \boldsymbol{G}_h^\dagger \overline{\psi}_h(\tau_h) \right\rangle \\
&= \left\langle \underbrace{\left(\boldsymbol{G}_h^\dagger\right)^\top \tilde{m}_h(\omega_h)}_{m_h(\omega_h)}, \overline{\psi}_h(\tau_h) \right\rangle.
\end{aligned}$$

That $\overline{\mathbb{P}}\left[\tau_h, \omega_h\right] = \langle m_h(\omega_h), \psi_h(\tau_h) \rangle$ follows directly by noting the definition of $\overline{\psi}_h(\tau_h) := \psi_h(\tau_h) / \overline{\mathbb{P}}\left[\tau_h\right]$.

Hence, we have shown that for the test set $\mathbb{Q}_h^m$, the probability of each future ω_h given a history τ_h is a linear combination of the probabilities of each test in the core test set with weights $m_h(\omega_h) := (\boldsymbol{G}_h^\dagger)^\top \tilde{m}_h(\omega_h) \in \mathbb{R}^{|\mathbb{Q}_h^m|}$ depending only on the future and not the history. □

Proposition (Restatement of Proposition 3)**.** *Suppose a POST is α-robustly m-step $\mathcal{I}^\dagger$-weakly revealing. Then, the corresponding PSR as constructed above with core test sets consisting of m-step futures is γ-well-conditioned with $\gamma = \alpha / \max_h \sqrt{\left|\mathbb{I}_h^\dagger\right|}$*

Proof. We first show condition (1) in Assumption 1. Suppose $h > H - m$ and hence the core tests are the full futures, which have length smaller than m. Then for any $x \in \mathbb{R}^{d_h}$, $d_h = \prod_{s=h}^{H} |\mathbb{X}_s|$, we have

$$\max_{\pi} \sum_{\omega_h} \left| m_h(\omega_h)^\top x \right| \cdot \pi(\omega_h) = \max_{\pi} \sum_{\omega_h} |x[\omega_h]| \, \pi(\omega_h) \leq \|x\|_1 \,,$$

where $x[\omega_h]$ indexes the component of the vector x corresponding to the future ω_h.

Now suppose $h \leq H - m$ (and hence the core tests consist of m-step futures). Then, we have,

$$\begin{aligned}\max_{\pi} \sum_{\omega_h} \left| m_h(\omega_h)^\top x \right| \pi(\omega_h) &= \max_{\pi} \sum_{\omega_h} \left| m(\omega_h)^\top \boldsymbol{G}_h \boldsymbol{G}_h^\dagger x \right| \cdot \pi(\omega_h) \\ &\leq \max_{\pi} \sum_{\omega_h} \sum_{i^\dagger \in \mathbb{I}_h^\dagger} \left| m(\omega_h)^\top \boldsymbol{G}_h \boldsymbol{e}_{i^\dagger} \right| \left| \boldsymbol{e}_{i^\dagger}^\top \boldsymbol{G}_h^\dagger x \right| \cdot \pi(\omega_h).\end{aligned}$$

Now observe that for any policy π and any $i^\dagger \in \mathbb{I}_h^\dagger$, we have

$$\begin{aligned}\sum_{\omega_h} \left| m(\omega_h)^\top \boldsymbol{G}_h \boldsymbol{e}_{i^\dagger} \right| \cdot \pi(\omega_h) &= \sum_{\omega_h} \left| \tilde{m}(\omega_h)^\top \boldsymbol{G}_h^\dagger \boldsymbol{G}_h \boldsymbol{e}_{i^\dagger} \right| \cdot \pi(\omega_h) \\ &= \sum_{\omega_h} \overline{\mathbb{P}} \left[\omega_h \mid i^\dagger \right] \pi(\omega_h) \\ &= \sum_{\omega_h} \mathbb{P}^{\pi} \left[\omega_h \mid i^\dagger \right] = 1,\end{aligned}$$

where we used the definition of $m_h(\omega_h) := \tilde{m}_h(\omega_h)^\top \boldsymbol{G}_h^\dagger$, and $[\tilde{m}_h(\omega_h)]_{i^\dagger} := \overline{\mathbb{P}}[\omega_h \,|\, i^\dagger]$. Recall that $\pi(\omega_h)$ is such that for any fixed sequence of observations ω_h^o, $\sum_{\omega_h^a} \pi(\omega_h^o, \omega_h^a) = 1$.

Putting this observation together with the preceding inequality yields

$$\begin{aligned}&\max_{\pi} \sum_{\omega_h} \left| m_h(\omega_h)^\top x \right| \pi(\omega_h) \\ &\leq \sum_{i^\dagger \in \mathbb{I}_h^\dagger} \left| \boldsymbol{e}_{i^\dagger}^\top \boldsymbol{G}_h^\dagger x \right| \\ &= \left\| \boldsymbol{G}_h^\dagger x \right\|_1 \leq \left\| \boldsymbol{G}_h^\dagger \right\|_1 \cdot \|x\|_1 \\ &\leq \frac{\sqrt{\left| \mathbb{I}_h^\dagger \right|}}{\alpha} \|x\|_1 \,,\end{aligned}$$

where the final inequality is from the relation between the one-norm and two-norm $\left\| \boldsymbol{G}_h^\dagger \right\|_1 \leq \sqrt{\left| \mathbb{I}_h^\dagger \right|} \left\| \boldsymbol{G}_h^\dagger \right\|_2$, and $\left\| \boldsymbol{G}_h^\dagger \right\|_2 \leq \frac{1}{\alpha}$, by the assumption on its eigenvalues.

Now we show condition (2) in Assumption 1. For ease of notation, we denote $x_{t(h)}$ by x_h. When $h > H - m$, note that $[M_h(x_h)]_{q_{h+1}, q_h} = \mathbf{1}\{q_h = (x_h, q_{h+1})\}$, for all $q_h \in \mathbb{Q}_h, q_{h+1} \in \mathbb{Q}_{h+1}$. Hence,we have

$$\max_{\pi} \sum_{x_h} \|M_h(x_h) z\|_1 \, \pi(x_h) = \|z\|_1 \,.$$

Now, when $h \leq H - m$, by a similar line of reasoning to the proof for condition (1), we have,

$$\begin{aligned}
\max_{\pi} \sum_{x_h} \left\| M_h(x_h) z \right\|_1 \pi(x_h|\tau_{h-1}) &\leq \max_{\pi} \sum_{(x_h, q_{h+1}) \in \mathbb{X}_h \times \mathbb{Q}_{h+1}} \sum_{i^\dagger \in \mathbb{I}_h^\dagger} \left| e_{q_{h+1}}^\top M_h(x_h) \boldsymbol{G}_h e_{i^\dagger} \right| \cdot \left| e_{i^\dagger} \boldsymbol{G}_h^\dagger z \right| \pi(x_h|\tau_{h-1}) \\
&\overset{(a)}{=} \max_{\pi} \sum_{(x_h, q_{h+1}) \in \mathbb{X}_h \times \mathbb{Q}_{h+1}} \sum_{i^\dagger \in \mathbb{I}_h^\dagger} \left| m_h(x_{t(h)}, q_{h+1}) \boldsymbol{G}_h e_{i^\dagger} \right| \cdot \left| e_{i^\dagger} \boldsymbol{G}_h^\dagger z \right| \pi(x_h|\tau_{h-1}) \\
&\overset{(b)}{=} \max_{\pi} \sum_{i^\dagger} \left(\sum_{(x_h, q_{h+1})} \overline{\mathbb{P}} \left[x_h, q_{h+1} \mid i^\dagger \right] \pi(x_h|\tau_{h-1}) \right) \left| e_{i^\dagger}^\top \boldsymbol{G}_h^\dagger z \right|
\end{aligned}$$

where step (a) uses the definition of M_h and step (b) uses the definition of $m_h(\omega_h)^\top := \tilde{m}_h(\omega_h)^\top \boldsymbol{G}_h^\dagger$ and $[\tilde{m}_h(\omega_h)]_{i^\dagger} := \overline{\mathbb{P}}[\omega_h \,|\, i^\dagger]$. Now note that,

$$\begin{aligned}
\sum_{(x_h, q_{h+1})} \overline{\mathbb{P}} \left[x_h, q_{h+1} \mid i^\dagger \right] \pi(x_h|\tau_{h-1}) &= \sum_{x_h} \sum_{\mathtt{act}(q_{h+1})} \sum_{\mathtt{obs}(q_{h+1})} \overline{\mathbb{P}} \left[x_h, \mathtt{obs}(q_{h+1}) \mid i^\dagger, \mathtt{act}(q_{h+1}) \right] \pi(x_h) \\
&= \sum_{\mathtt{act}(q_{h+1})} 1 \\
&= \left| \mathbb{Q}_{h+1}^A \right|,
\end{aligned}$$

where the second line is since for any fixed action sequence, the sum over the probabilities of all observation sequences is 1.

Thus, putting this together, we obtain the following,

$$\begin{aligned}
\max_{\pi} \sum_{x_h} \left\| M_h(x_h) z \right\|_1 \pi(x_h|\tau_{h-1}) &\leq \left| \mathbb{Q}_{h+1}^A \right| \cdot \left\| \boldsymbol{G}_h^\dagger z \right\|_1 \\
&\leq \frac{\sqrt{\left| \mathbb{I}_h^\dagger \right|} \left| \mathbb{Q}_{h+1}^A \right|}{\alpha} \left\| z \right\|_1,
\end{aligned}$$

where the last line again follows by the assumption on the eigenvalues of $\boldsymbol{G}_h$. □

D Proof of Theorem 1: UCB Algorithm for Generalized PSRs (Team Setting)

D.1 Properties of Generalized PSRs

In the proof, we will consider PSRs of partially-observable sequential decision-making problem (with arbitrary order of observations and actions), rather than specializing PSRs induced by partially-observable sequential teams. Hence, in the notation, we will omit dependence on underlying unobserved system variables (e.g., we will use x_h rather than $x_{t(h)}$). The relevant PSR condition is Assumption 1. When a POST satisfies **??**, it satisfies this condition by Proposition 3.

Recall that a model $\theta = (\boldsymbol{M}, \psi_0, \phi_H)$ consists of operators $\boldsymbol{M} = \{M_h\}_{h=1}^{H-1}$, $M_h : \mathbb{X}_h \to \mathbb{R}^{d_{h+1} \times d_h}$, $\phi_H : \mathbb{X}_H \to \mathbb{R}^{d_{H-1}}$ (assumed to be the identity mapping), and ψ_0 (assumed to be known for the purposes of presentation).

Recall that, for any trajectory $\tau_{h-1} = (x_1, \ldots, x_{h-1})$, under model θ, we have

$$\begin{aligned}
M_h(x_h) \overline{\psi}_{h-1}(\tau_{h-1}) &= \frac{\psi_h(\tau_h)}{\overline{\mathbb{P}}_\theta \left(\tau_{h-1} \right)} \\
&= \frac{\psi_h(\tau_h)}{\overline{\mathbb{P}}_\theta \left(x_h \mid \tau_{h-1} \right) \overline{\mathbb{P}}_\theta \left(\tau_{h-1} \right)} \overline{\mathbb{P}}_\theta \left(x_h \mid \tau_{h-1} \right) \\
&= \overline{\psi}_h(\tau_h) \overline{\mathbb{P}}_\theta \left(x_h \mid \tau_{h-1} \right)
\end{aligned} \tag{28}$$

Notation. In what follows, we suppose θ is the true PSR model and $\widehat{\theta}$ is an estimated model. We indicate quantities associated with the estimated model by a 'hat'. E.g., $\widehat{M}_h, \widehat{m}_h, \widehat{\psi}_h$.

Proposition 6. *For any policy π and $\theta, \widehat{\theta} \in \Theta$, we have,*

$$\mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\pi}_{\widehat{\theta}}, \mathbb{P}^{\pi}_{\theta}\right) \leq \sum_{h=1}^{H} \sum_{\tau_H \in \mathbb{H}_H} \pi(\tau_h) \left| \widehat{m}_h(\omega_h)^\top \left(\widehat{M}_h(x_h) - M_h(x_h)\right) \psi_{h-1}(\tau_{h-1}) \right|,$$

$$\mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\pi}_{\widehat{\theta}}, \mathbb{P}^{\pi}_{\theta}\right) \leq \sum_{h=1}^{H} \sum_{\tau_H \in \mathbb{H}_H} \pi(\tau_h) \left| m_h(\omega_h)^\top \left(\widehat{M}_h(x_h) - M_h(x_h)\right) \widehat{\psi}_{h-1}(\tau_{h-1}) \right|,$$

Proof. The probability of any trajectory $\tau_H = (x_1, \dots, x_H)$ can be written in terms of products of the observable operators $M_h(x_h)$ of a PSR model (Equation (4)). Hence, we have,

$$\begin{aligned}
\mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\pi}_{\widehat{\theta}}, \mathbb{P}^{\pi}_{\theta}\right) &= \frac{1}{2} \sum_{\tau_H} \left| \mathbb{P}^{\pi}_{\widehat{\theta}}(\tau_H) - \mathbb{P}^{\pi}_{\theta}(\tau_H) \right| \\
&= \frac{1}{2} \sum_{\tau_H} \pi(\tau_H) \cdot \left| \left(\prod_{h=1}^{H} \widehat{M}_h(x_h) \right) \psi_0 - \left(\prod_{h=1}^{H} M_h(x_h) \right) \psi_0 \right| \\
&\leq \frac{1}{2} \sum_{\tau_H} \pi(\tau_H) \sum_{h=1}^{H} \left| \widehat{m}_h(x_{h+1:H})^\top \left(\widehat{M}_h(x_h) - M_h(x_h)\right) \psi_{h-1}(\tau_{h-1}) \right|,
\end{aligned}$$

where the second line follows by the triangle inequality after noting that for any trajectory $\tau_H = x_{1:H} \in \mathbb{H}_H$, the following holds for any $h = 1, \dots, H$,

$$\left(\prod_{h=1}^{H} \widehat{M}_h(x_h) \right) \psi_0 - \left(\prod_{h=1}^{H} M_h(x_h) \right) \psi_0 = \widehat{m}_h(x_{h+1:H})^\top \widehat{M}_h(x_h) \widehat{\psi}_{h-1}(x_{1:h-1}) - m_h(x_{h+1:H})^\top M_h(x_h) \psi_{h-1}(x_{1:h-1}).$$

By the same argument, we obtain the second inequality,

$$\begin{aligned}
\mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\pi}_{\widehat{\theta}}, \mathbb{P}^{\pi}_{\theta}\right) &= \frac{1}{2} \sum_{\tau_H} \left| \mathbb{P}^{\pi}_{\widehat{\theta}}(\tau_H) - \mathbb{P}^{\pi}_{\theta}(\tau_H) \right| \\
&\leq \frac{1}{2} \sum_{\tau_H} \pi(\tau_H) \sum_{h=1}^{H} \left| m_h(x_{h+1:H})^\top \left(\widehat{M}_h(x_h) - M_h(x_h)\right) \widehat{\psi}_{h-1}(\tau_{h-1}) \right|.
\end{aligned}$$

□

In this result, recall that we assume ψ_0 is known to the agent, to simplify the presentation. If ψ_0 was not known, there would be another term due to the estimation as $\widehat{\psi}_0$ (see (Liu, Netrapalli, et al. 2022, Lemma C.3)). Note that the sample complexity of estimating ψ_0 is small compared to learning the other parameters.

D.2 General Results on MLE

We will use some general results on maximum likelihood estimation in the remainder of the proof. We state them here for completeness. The proofs are given in (Huang et al. 2023) and use standard techniques on MLE analysis (Geer 2006).

The first proposition states that the log-likelihood of the true model θ^* is large compared to any other model $\overline{\theta} \in \overline{\Theta}$.

Proposition 7 (Proposition 4 of Huang et al. 2023)**.** *Fix $\varepsilon < \frac{1}{KH}$. With probability at least $1-\delta$, for any $\overline{\theta} \in \overline{\Theta}_\varepsilon$ and any $k \in [K]$, the following holds:*

$$\forall \overline{\theta} \in \overline{\Theta}_\varepsilon, \sum_h \sum_{(\tau_h,\pi)\in\mathcal{D}_h} \log \mathbb{P}^\pi_{\overline{\theta}}(\tau_h) - 3\log\frac{K\left|\overline{\Theta}_\varepsilon\right|}{\delta} \le \sum_h \sum_{(\tau_h,\pi)\in\mathcal{D}_h^k} \log \mathbb{P}^\pi_{\theta^*}(\tau_h)$$

$$\forall \overline{\theta} \in \overline{\Theta}_\varepsilon, \sum_{(\tau_H,\pi)\in\mathcal{D}^k} \log \mathbb{P}^\pi_{\overline{\theta}}(\tau_H) - 3\log\frac{K\left|\overline{\Theta}_\varepsilon\right|}{\delta} \le \sum_{(\tau_h,\pi)\in\mathcal{D}_h^k} \log \mathbb{P}^\pi_{\theta^*}(\tau_h)$$

The second proposition provides an upper bound on the total variation distance between the distributions of futures given histories on the empirical history of trajectories. This result ensures that the model estimated by Algorithm 1 is accurate on the sampled trajectories.

Proposition 8 (Proposition 5 in Huang et al. 2023)**.** *Fix $p_{\min}$ and $\varepsilon \le \frac{p_{\min}}{KH}$. Let $\Theta^k_{\min} = \left\{\theta : \forall h, (\tau_h,\pi) \in \mathcal{D}^k_h,\ \mathbb{P}^\pi_\theta(\tau_h) \ge p_{\min}\right\}$ Then, with probability at least $1-\delta$, for any $k \in [K], \theta \in \Theta^k_{\min}$, we have,*

$$\sum_h \sum_{(\tau_h,\pi)\in\mathcal{D}_h^k} \mathtt{D}^2_{\mathtt{TV}}\left(\mathbb{P}^\pi_\theta(\omega_h|\tau_h), \mathbb{P}^\pi_{\theta^*}(\omega_h|\tau_h)\right) \le 6 \sum_h \sum_{(\tau_h,\pi)\in\mathcal{D}_h^k} \log\frac{\mathbb{P}^\pi_{\theta^*}(\tau_H)}{\mathbb{P}^\pi_\theta(\tau_H)} + 31\log\frac{K\left|\overline{\Theta}_\varepsilon\right|}{\delta}.$$

The next proposition is standard in the analysis of maximum likelihood estimation. $\mathtt{D_H}$ denotes the Hellinger distance.

Proposition 9 (Proposition 6 of Huang et al. 2023)**.** *Let $\varepsilon < \frac{1}{K^2H^2}$. Then, with probability at least $1-\delta$, the following holds for all $\theta \in \Theta$ and $k \in [K]$,*

$$\sum_{\pi\in\mathcal{D}^k} \mathtt{D}^2_{\mathtt{H}}\left(\mathbb{P}^\pi_\theta(\tau_H), \mathbb{P}^\pi_{\theta^*}(\tau_H)\right) \le \frac{1}{2} \sum_{(\tau_H,\pi)\in\mathcal{D}^k} \log\frac{\mathbb{P}^\pi_{\theta^*}(\tau_H)}{\mathbb{P}^\pi_\theta(\tau_H)} + 2\log\frac{K\left|\overline{\Theta}_\varepsilon\right|}{\delta}.$$

The final proposition of this section states that when $p_{\min}$ is chosen as in Theorem 1, the true model θ^* lies in the constraint $\Theta^k_{\min}$ with high probability.

Proposition 10. *Fix $p_{\min} \le \frac{\delta}{KH\prod_{h=1}^H|\mathbb{X}_h|}$. Then, with probability at least $1-\delta$, we have $\theta^* \in \Theta^k_{\min}$ $\forall k$.*

Proof. For each $k \in [K]$, we have $\theta^* \in \Theta^k_{\min}$ if $\mathbb{P}^{\pi^k}_{\theta^*}(\tau^k_h) \ge p_{\min}$ for all $h \in [H]$, $(\tau^k_h, \pi^k) \in \mathcal{D}^k_h$. Note that the randomness in whether θ^* is in $\Theta^k_{\min}$ is over the sampling of (τ^k_h, π^k). For each $k, h, (\tau^k_h, \pi^k)$, we have

$$\begin{aligned}
\mathbb{P}\left[\mathbb{P}^{\pi^k}_{\theta^*}(\tau^k_h) < p_{\min}\right] &= \mathbb{E}_\pi\left[\mathbb{P}\left[\mathbb{P}^{\pi^k}_{\theta^*}(\tau^k_h) < p_{\min} \,\middle|\, \pi^k = \pi\right]\right] \\
&= \mathbb{E}_\pi\left[\sum_{\tau_h\in\mathbb{H}_h} \mathbb{P}^\pi_{\theta^*}(\tau^k_h = \tau_h)\mathbf{1}\{\mathbb{P}^\pi_{\theta^*}(\tau_h) < p_{\min}\}\right] \\
&\le \sum_{\tau_h\in\mathbb{H}_h} p_{\min} \\
&= |\mathbb{H}_h|\, p_{\min} \\
&\le \frac{\delta}{KH}.
\end{aligned}$$

In the above, the first line conditions on the randomness in π^k, the policy used while collecting the trajectory τ^k_h, and the second line considers the randomness over trajectories in that episode. Taking a union bound over $k \in [K]$, $h \in [H]$, and $(\tau_h, \pi) \in \mathcal{D}_h$ gives the result. □

Notation. In what follows, let $\mathcal{E}_\omega, \mathcal{E}_\pi, \mathcal{E}_{\min}$ be the events in Propositions 8 to 10, respectively. Let $\mathcal{E} = \mathcal{E}_\omega \cap \mathcal{E}_\pi \cap \mathcal{E}_{\min}$ be the intersection of all events. The results in Appendix D.2 guarantee the event $\mathcal{E}$ occurs with high probability, $\mathbb{P}[\mathcal{E}] \ge 1 - 3\delta$, by a union bound.

D.3 Estimation guarantee

The following result states that the estimated model is accurate on the past exploration policies and dataset of collected trajectories. This holds for both the conditional probabilities of futures given past trajectories in the dataset as well as over full trajectories. The result follows from Propositions 8 to 10 in the MLE analysis of the previous section.

Lemma 4. *Let $\beta = 31 \log \frac{K|\overline{\Theta}_\varepsilon|}{\delta}$, and suppose $\varepsilon \leq \frac{\delta}{K^2 H^2 \prod_h |\mathbb{X}_h|}$, where $\overline{\Theta}_\varepsilon$ is the optimistic ε-net in* **??**. *Then, under event $\mathcal{E}$, the following holds,*

$$\sum_h \sum_{(\tau_h, \pi) \in \mathcal{D}_h^k} \mathtt{D}^2_{\mathtt{TV}} \left(\mathbb{P}^\pi_{\widehat{\theta}^k}(\omega_h | \tau_h), \mathbb{P}^\pi_{\theta^*}(\omega_h | \tau_h) \right) \leq 7\beta, \textit{ and}$$

$$\sum_{\pi \in \mathcal{D}^k} \mathtt{D}^2_{\mathtt{H}} \left(\mathbb{P}^\pi_{\widehat{\theta}^k}(\tau_H), \mathbb{P}^\pi_{\theta^*}(\tau_H) \right) \leq 7\beta,$$

Proof. Lemma 1 of Huang et al. 2023. □

D.4 UCB for Total Variation Distance

Notation. Let $m^*, \{M_h^*\}_h$ be the observable operators of the true PSR θ^*, and let $\left\{ \widehat{M}_h^k \right\}_h$ be the estimates of the observable operators corresponding to $\widehat{\theta}^k$.

Recall that Proposition 6 shows that the total variation distance between the distribution over trajectories of two PSRs is bounded by the estimation error of the observable operators M_h. The following result constructs a bound on the estimation error of the observable operators $M_h(x_h)$. The proof is adapted from (Huang et al. 2023, Lemma 2) to our setting.

Lemma 5. *Under event $\mathcal{E}$, for any policy π and $k \in [K]$, we have,*

$$\sum_{\tau_H} \left| m^\star(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h) \right) \widehat{\psi}_{h-1}^k(\tau_{h-1}) \right| \pi(\tau_H) \leq \mathbb{E}^\pi_{\tau_{h-1} \sim \mathbb{P}_{\widehat{\theta}^k}} \left[\alpha_{h-1}^k \left\| \widehat{\overline{\psi}}_{h-1}^k(\tau_{h-1}) \right\|_{(\widehat{U}_{h-1}^k)^{-1}} \right]$$

where,

$$\widehat{U}_{h-1}^k = \lambda I + \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \left[\widehat{\overline{\psi}}_h^k(\tau_{h-1}) \widehat{\overline{\psi}}_h^k(\tau_{h-1})^\top \right]$$

$$\left(\alpha_{h-1}^k\right)^2 = \frac{4\lambda Q_A^2 d}{\gamma^4} + 4 \max_{s \in \mathcal{A}} |\mathbb{X}_s|^2 Q_A^2 \max\left(\frac{1}{\gamma^2}, \frac{Q_A^2}{\gamma^4} \right) \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \mathtt{D}^2_{\mathtt{TV}} \left(\mathbb{P}^{\mathtt{u}_{h-1}^{\mathtt{exp}}}_{\widehat{\theta}^k} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right), \mathbb{P}^{\mathtt{u}_{h-1}^{\mathtt{exp}}}_{\theta^*} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_h^a \right) \right)$$

Proof. To ease notation, we index the future trajectories $\omega_{h-1} = (x_h, \ldots, x_H) \in \mathbb{F}_{h-1}$ by i and history trajectories $\tau_{h-1} = (x_1, \ldots, x_{h-1}) \in \mathbb{H}_{h-1}$ by j. We denote $m^\star(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h) \right)$ as $w_i^\top$, $\widehat{\overline{\psi}}_h^k(\tau_{h-1})$ as x_j, and $\pi(\omega_{h-1} | \tau_{h-1})$ as $\pi_{i|j}$.

The following bound follows from the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sum_{\tau_H} \left| m^\star(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h) \right) \widehat{\psi}_{h-1}^k(\tau_{h-1}) \right| \pi(\tau_H) \\
&\overset{(a)}{=} \sum_{\omega_{h-1}} \sum_{\tau_{h-1}} \left| m^\star(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h) \right) \widehat{\overline{\psi}}_h^k(\tau_{h-1}) \right| \pi(\omega_{h-1}|\tau_{h-1}) \mathbb{P}_{\widehat{\theta}^k}^\pi(\tau_{h-1}) \\
&= \sum_i \sum_j \left| w_i^\top x_j \right| \pi_{i|j} \mathbb{P}_{\widehat{\theta}^k}^\pi(j) \\
&= \sum_i \sum_j \left(\pi_{i|j} \cdot \mathrm{sign}(w_i^\top x_j) w_i \right)^\top x_j \cdot \mathbb{P}_{\widehat{\theta}^k}^\pi(j) \\
&= \sum_j \left(\sum_i \pi_{i|j} \cdot \mathrm{sign}(w_i^\top x_j) w_i \right)^\top x_j \cdot \mathbb{P}_{\widehat{\theta}^k}^\pi(j) \\
&= \mathbb{E}_{j \sim \mathbb{P}_{\widehat{\theta}^k}^\pi} \left[\left(\sum_i \pi_{i|j} \cdot \mathrm{sign}(w_i^\top x_j) w_i \right)^\top x_j \right] \\
&\overset{(b)}{\le} \mathbb{E}_{j \sim \mathbb{P}_{\widehat{\theta}^k}^\pi} \left[\|x_j\|_{(\widehat{U}_{h-1}^k)^{-1}} \left\| \sum_i \pi_{i|j} \cdot \mathrm{sign}(w_i^\top x_j) \cdot w_i \right\|_{\widehat{U}_{h-1}^k} \right].
\end{aligned}$$

Step (a) follows from the fact that $\widehat{\psi}_{h-1}^k(\tau_{h-1}) = \widehat{\overline{\psi}}_h^k(\tau_{h-1}) \cdot (\widehat{\phi}_{h-1}^k)^\top \widehat{\psi}_{h-1}^k(\tau_{h-1}) = \widehat{\overline{\psi}}_h^k(\tau_{h-1}) \cdot \overline{\mathbb{P}}_{\widehat{\theta}^k}[\tau_{h-1}]$ and $\overline{\mathbb{P}}_{\widehat{\theta}^k}[\tau_{h-1}] \cdot \pi(\tau_H) = \pi(\omega_{h-1}|\tau_{h-1}) \cdot \mathbb{P}_{\widehat{\theta}^k}^\pi(\tau_{h-1})$. Step (b) is the Cauchy-Schwarz inequality.

Fix $\tau_{h-1} = j_0$. Let $I_1 := \left\|\sum_i \pi_{i|j_0} \cdot \mathrm{sign}(w_i^\top x_{j_0}) \cdot w_i \right\|_{\widehat{U}_{h-1}^k}^2$, which we bound next. By the definition of $\widehat{U}_{h-1}^k$, we partition this term into two parts,

$$I_1 = \underbrace{\lambda \left\| \sum_i \pi_{i|j_0} \cdot \mathrm{sign}(w_i^\top x_{j_0}) \cdot w_i \right\|_2^2}_{I_2} + \underbrace{\sum_{j \in D_{h-1}^\tau} \left[\left(\sum_i \pi_{i|j_0} \cdot \mathrm{sign}(w_i^\top x_{j_0}) \cdot w_i \right)^\top x_j \right]^2}_{I_3}.$$

We bound I_2 and I_3 separately. By the triangle inequality, $\sqrt{I_2}$ is bound by a sum of two terms,

$$\begin{aligned}
\sqrt{I_2} &= \sqrt{\lambda} \max_{z \in \mathbb{R}^{d_{h-1}} : \|z\|_2 = 1} \left| \sum_i \pi_{i|j_0} \cdot \mathrm{sign}(w_i^\top x_{j_0}) \cdot w_i^\top z \right| \\
&\overset{(a)}{\le} \sqrt{\lambda} \max_{\|z\|_2 = 1} \sum_{\omega_{h-1}} \left| m^\star(\omega_h^\top) \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h) \right) z \right| \pi(\omega_{h-1}|j_0) \\
&\overset{(b)}{\le} \sqrt{\lambda} \max_{\|z\|_2 = 1} \sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top \widehat{M}_h^k(x_h) z \right| \pi(\omega_{h-1}|j_0) \\
&\quad + \sqrt{\lambda} \max_{\|z\|_2 = 1} \sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top M_h^\star(x_h) z \right| \pi(\omega_{h-1}|j_0),
\end{aligned}$$

where step (a) is by the definition of $w_i^\top, \pi_{i|j_0}$ and the triangle inequality, and step (b) is by the triangle inequality.

Consider the first term. It can be bound via the definition of γ-well-conditioning as follows,

$$\begin{aligned}
&\max_{\|z\|_2=1} \sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top \widehat{M}_h^k(x_h) z \right| \pi(\omega_{h-1}|j_0) \\
&\quad = \max_{\|z\|_2=1} \sum_{x_h} \left(\sum_{\omega_h} \left| m^\star(\omega_h)^\top \widehat{M}_h^k(x_h) z \right| \pi(\omega_h|j_0, x_h) \right) \pi(x_h|j_0) \\
&\quad \overset{(a)}{\le} \frac{1}{\gamma} \max_{\|z\|_2=1} \sum_{x_h} \left\| \widehat{M}_h^k(x_h) z \right\|_1 \pi(x_h|j_0) \\
&\quad \overset{(b)}{\le} \frac{1}{\gamma} \max_{\|z\|_2=1} \frac{\left|\mathbb{Q}_{h+1}^A\right| \|z\|_1}{\gamma} \\
&\quad \overset{(c)}{\le} \frac{\sqrt{d} Q_A}{\gamma^2}
\end{aligned}$$

where step (a) is by the first condition in Assumption 1, step (b) is by the second condition of Assumption 1, and step (c) is by the fact that $\max_{z \in \mathbb{R}^{d_{h-1}} : \|z\|_2 = 1} \|z\|_1 = \sqrt{d_{h-1}} \le \sqrt{d}$ and $\left|\mathbb{Q}_{h+1}^A\right| \le Q_A$. In the above, note that we used the γ-well-conditioning of PSR $\widehat{\theta}^k$ in step (a) and the γ-well-conditioning of PSR θ^* in step (b). The second term in $\sqrt{I_2}$ admits an identical bound, simply by using the well-conditioning of the PSR θ^* in both steps. Hence, we have that

$$I_2 \le 4 \frac{\lambda d Q_A^2}{\gamma^4}. \tag{29}$$

Now we upper bound I_3,

$$\begin{aligned}
I_3 &\le \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \left(\sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h) \right) \widehat{\overline{\psi}}^k(\tau_{h-1}) \right| \pi(\omega_{h-1}|j_0) \right)^2 \\
&\le \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \Bigg(\underbrace{\sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top \left(\widehat{M}_h^k(x_h) \widehat{\overline{\psi}}^k(\tau_{h-1}) - M_h^\star(x_h) \overline{\psi}^\star(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0)}_{I_4} \\
&\qquad + \underbrace{\sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top M_h^\star(x_h) \left(\widehat{\overline{\psi}}^k(\tau_{h-1}) - \overline{\psi}^\star(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0)}_{I_5} \Bigg)^2 \\
&=: \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} (I_4 + I_5)^2
\end{aligned}$$

where the second equality follows from the triangle inequality by adding and subtracting $m^*(\omega_h)^\top M_h^*(x_h) \overline{\psi}^*(\tau_{h-1})$ inside the absolute value. We now bound each of I_4 and I_5.

$$
\begin{aligned}
I_4 &\coloneqq \sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top \left(\widehat{M}_h^k(x_h) \widehat{\overline{\psi}}^k(\tau_{h-1}) - M_h^\star(x_h) \overline{\psi}^\star(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0) \\
&\overset{(a)}{=} \sum_{\omega_{h-1}} \left| m^\star(\omega_h)^\top \left(\overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h \mid \tau_{h-1}\right] \widehat{\overline{\psi}}_h(\tau_h) - \overline{\mathbb{P}}_{\theta^*}\left[x_h \mid \tau_{h-1}\right] \overline{\psi}_h^*(\tau_h) \right) \right| \pi(\omega_{h-1}|j_0) \\
&\overset{(b)}{=} \sum_{x_h} \left(\sum_{\omega_h} \left| m^\star(\omega_h)^\top \left(\overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h \mid \tau_{h-1}\right] \widehat{\overline{\psi}}_h(\tau_h) - \overline{\mathbb{P}}_{\theta^*}\left[x_h \mid \tau_{h-1}\right] \overline{\psi}_h^*(\tau_h) \right) \right| \pi(\omega_h|j_0, x_h) \right) \pi(x_h|j_0) \\
&\overset{(c)}{\leq} \frac{1}{\gamma} \sum_{x_h} \left\| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h \mid \tau_{h-1}\right] \widehat{\overline{\psi}}_h(\tau_h) - \overline{\mathbb{P}}_{\theta^*}\left[x_h \mid \tau_{h-1}\right] \overline{\psi}_h^*(\tau_h) \right\|_1 \pi(x_h|j_0) \\
&\overset{(d)}{=} \frac{1}{\gamma} \sum_{x_h} \sum_{q_h \in \mathbb{Q}_h} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h, q_h \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[x_h, q_h \mid \tau_{h-1}\right] \right| \pi(x_h|j_0)
\end{aligned}
$$

where step (a) is by the fact that $M_h(x_h)\overline{\psi}_{h-1}(\tau_{h-1}) = \overline{\mathbb{P}}\left[x_h \mid \tau_{h-1}\right]\overline{\psi}(\tau_h)$, as shown in Equation (28), step (b) uses $\omega_{h-1} = (x_h, \omega_h)$ and $\pi(\omega_{h-1}|j_0) = \pi(x_h|j_0)\pi(\omega_h|j_0, x_h)$, step (c) is by Assumption 1, and step (d) follows by the definition $\overline{\psi}_h$, $\left[\overline{\psi}_h(\tau_h)\right]_l = \overline{\mathbb{P}}_\theta\left[q_h^l \mid \tau_h\right]$.

Now, we turn to bound the I_5 term. We have

$$
\begin{aligned}
I_5 &= \sum_{\omega_h} \sum_{x_h} \left| m_h^\star(\omega_h)^\top M_h^\star(x_h) \left(\widehat{\overline{\psi}}^k(\tau_{h-1}) - \overline{\psi}^\star(\tau_{h-1}) \right) \right| \pi(\omega_h|j_0, x_h)\pi(x_h|j_0) \\
&\overset{(a)}{=} \sum_{\omega_{h-1}} \left| m_{h-1}^\star(\omega_{h-1})^\top \left(\widehat{\overline{\psi}}^k(\tau_{h-1}) - \overline{\psi}^\star(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0) \\
&\overset{(a)}{\leq} \frac{1}{\gamma} \left\| \widehat{\overline{\psi}}^k(\tau_{h-1}) - \overline{\psi}^\star(\tau_{h-1}) \right\|_1 \\
&= \frac{1}{\gamma} \sum_{q_{h-1} \in \mathbb{Q}_{h-1}} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[q_{h-1} \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[q_{h-1} \mid \tau_{h-1}\right] \right|,
\end{aligned}
$$

where step (a) is since $m_h^*(\omega_h)^\top M_h^*(x_h) = m_{h-1}^*(\omega_{h-1})^\top$, step (b) is by the first condition of Assumption 1, and the final equality is again by the definition of $\overline{\psi}$.

Combining the above, we have that,

$$
\begin{aligned}
I_3 &\leq \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \left(I_4 + I_5\right)^2 \\
&\leq \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \left(\frac{1}{\gamma} \sum_{x_h \in \mathbb{X}_h} \sum_{q_h \in \mathbb{Q}_h} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h, q_h \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[x_h, q_h \mid \tau_{h-1}\right] \right| \pi(x_h|\tau_{h-1}) \right. \\
&\qquad\qquad \left. + \frac{1}{\gamma} \sum_{q_{h-1} \in \mathbb{Q}_{h-1}} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[q_{h-1} \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[q_{h-1} \mid \tau_{h-1}\right] \right| \right)^2 \\
&\leq \frac{1}{\gamma^2} \cdot \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \left(\sum_{(x_h, q_h) \in \mathbb{X}_h \times \mathbb{Q}_h} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h, q_h \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[x_h, q_h \mid \tau_{h-1}\right] \right| \pi(x_h|\tau_{h-1}) \right. \\
&\qquad\qquad \left. + \sum_{q_{h-1} \in \mathbb{Q}_{h-1}} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[q_{h-1} \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[q_{h-1} \mid \tau_{h-1}\right] \right| \right)^2
\end{aligned}
$$

Now, we decompose the summations above over $\mathbb{X}_h \times \mathbb{Q}_h$ and $\mathbb{Q}_{h-1}$ into separate summations over observation futures and action futures. That is, (x_h, q_h) is decomposed into $(\omega_{h-1}^a, \omega_{h-1}^o)$, where $\omega_{h-1}^a = \texttt{act}(x_h, q_h)$

and $\omega_{h-1}^o = \texttt{obs}(x_h, q_h)$, and the summations are over $\omega_{h-1}^a \in \texttt{act}(\mathbb{X}_h \times \mathbb{Q}_h)$ and $\omega_{h-1}^o \in \texttt{obs}(\mathbb{X}_h \times \mathbb{Q}_h)$. Similarly, q_{h-1} can be decomposed into $(q_{h-1}^o, q_{h-1}^a) \in \texttt{obs}(\mathbb{Q}_{h-1}) \times \texttt{act}(\mathbb{Q}_{h-1})$. Hence, the bound on I_3 can be written as,

$$\begin{aligned}
I_3 &\leq \frac{1}{\gamma^2} \cdot \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \left(\sum_{\omega_{h-1}^a} \sum_{\omega_{h-1}^o} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right] - \overline{\mathbb{P}}_{\theta^*} \left[\omega_{h-1}^o \mid \tau_{h-1}, \omega_h^a \right] \right| \pi(x_h | \tau_{h-1}) \right. \\
&\qquad \left. + \sum_{q_{h-1}^a} \sum_{q_{h-1}^o} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[q_{h-1}^o \mid \tau_{h-1}, q_{h-1}^a \right] - \overline{\mathbb{P}}_{\theta^*} \left[q_{h-1}^o \mid \tau_{h-1}, q_{h-1}^a \right] \right| \right)^2 \\
&\leq \frac{1}{\gamma^2} \cdot \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \left(\sum_{\omega_{h-1}^a \in \mathbb{Q}_{h-1}^{\text{exp}}} \sum_{\omega_{h-1}^o} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right] - \overline{\mathbb{P}}_{\theta^*} \left[\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right] \right| \right)^2 \\
&= \frac{1}{\gamma^2} \left| \mathbb{Q}_{h-1}^{\text{exp}} \right|^2 \cdot \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \texttt{D}_{\texttt{TV}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right), \mathbb{P}_{\theta^*}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right) \right).
\end{aligned}$$

Where the second inequality is by the definition of $\mathbb{Q}_{h-1}^{\texttt{exp}} = \texttt{act}\left(\mathbb{X}_h \times \mathbb{Q}_h \cup \mathbb{Q}_{h-1}\right)$. Here, the second summation is over $\omega_{h-1}^o \in \texttt{obs}\left(\mathbb{X}_h \times \mathbb{Q}_h \cup \mathbb{Q}_{h-1}\right)$. The final equality uses the fact the under the policy $\texttt{u}_{h-1}^{\texttt{exp}}$ the probability of each action sequence ω_{h-1}^o is $1/\left|\mathbb{Q}_{h-1}^{\texttt{exp}}\right|$. Note that $\left|\mathbb{Q}_{h-1}^{\texttt{exp}}\right| \leq \left|\texttt{act}\left(\mathbb{X}_h \times \mathbb{Q}_h\right)\right| + \left|\texttt{act}\left(\mathbb{Q}_{h-1}\right)\right|$, and hence we have $\left|\mathbb{Q}_{h-1}^{\texttt{exp}}\right| \leq 2 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right| Q_A$ for all h. Hence, we have,

$$I_3 \leq 4 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2 \frac{1}{\gamma^2} \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \texttt{D}_{\texttt{TV}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right), \mathbb{P}_{\theta^*}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_h^a \right) \right). \tag{30}$$

Putting this together with the bounds on I_2 and I_3, we get that,

$$\begin{aligned}
I_1 &\leq \frac{4\lambda Q_A^2 d}{\gamma^4} + 4 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2 \frac{1}{\gamma^2} \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \texttt{D}_{\texttt{TV}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right), \mathbb{P}_{\theta^*}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_h^a \right) \right) \\
&=: \left(\alpha_{h-1}^k\right)^2,
\end{aligned}$$

completing the proof. □

Lemma 6. *Under even $\mathcal{E}$, the total variation distance between the estimated model at iteration k, $\widehat{\theta}^k$, and the true model θ^*, is bounded by,*

$$\texttt{D}_{\texttt{TV}} \left(\mathbb{P}_{\widehat{\theta}^k}^{\pi}(\tau_H), \mathbb{P}_{\theta^*}^{\pi}(\tau_H) \right) \leq \alpha \cdot \mathbb{E}_{\tau_H \sim \mathbb{P}_{\widehat{\theta}^k}^{\pi}} \left[\sqrt{\sum_{h=0}^{H-1} \left\| \widehat{\psi}^k(\tau_h) \right\|_{(\widehat{U}_h^k)^{-1}}^2} \right], \tag{31}$$

for any policy π, where

$$\alpha^2 = \frac{4\lambda H Q_A^2 d}{\gamma^4} + 28 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2 \frac{1}{\gamma^2} \beta$$

Proof. Consider α_{h-1}^k in the previous lemma. We have that,

$$\begin{aligned}
&\sum_{h=1}^{H} \left(\alpha_{h-1}^k\right)^2 \\
&= \frac{4\lambda H Q_A^2 d}{\gamma^4} + 4 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2 \frac{1}{\gamma^2} \sum_{h=1}^{H} \sum_{\tau_{h-1} \in \mathcal{D}_{h-1}^k} \texttt{D}_{\texttt{TV}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a \right), \mathbb{P}_{\theta^*}^{\texttt{u}_{h-1}^{\texttt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_h^a \right) \right) \\
&\leq \frac{4\lambda H Q_A^2 d}{\gamma^4} + 4 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2 \frac{1}{\gamma^2} 7\beta =: \alpha^2,
\end{aligned}$$

where the inequality is by the bound on the total variation distance established in Lemma 4.

Now, by Proposition 6, the total variation distance is bounded by the estimation error:

$$\begin{aligned}
&\mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_H), \mathbb{P}^{\pi}_{\theta^*}(\tau_H)\right) \\
&\overset{(a)}{\leq} \sum_{h=1}^{H}\sum_{\tau_H}\left|m^{\star}(\omega_h)^{\top}\left(\widehat{M}^k_h(x_h) - M^{\star}_h(x_h)\right)\widehat{\psi}^k_{h-1}(\tau_{h-1})\right|\pi(\tau_H) \\
&\overset{(b)}{\leq} \sum_{h=1}^{H}\mathbb{E}^{\pi}_{\tau_{h-1}\sim\mathbb{P}_{\widehat{\theta}^k}}\left[\alpha^k_{h-1}\left\|\overset{\frown}{\overline{\psi}}{}^k_{h-1}(\tau_{h-1})\right\|_{(\widehat{U}^k_{h-1})^{-1}}\right] \\
&\overset{(c)}{\leq} \alpha\cdot\mathbb{E}_{\tau_H\sim\mathbb{P}^{\pi}_{\widehat{\theta}^k}}\left[\sqrt{\sum_{h=0}^{H-1}\left\|\overset{\frown}{\overline{\psi}}{}^k(\tau_h)\right\|^2_{(\widehat{U}^k_h)^{-1}}}\right],
\end{aligned}$$

where step (a) is by Proposition 6, step (b) is by Lemma 5, and step (c) is by the Cauchy-Schwarz inequality and the calculation above bounding $\sum_h\left(\alpha^k_{h-1}\right)^2$. □

A direct corollary is the following bound on the error in the estimated value function.

Corollary 4 (Upper confidence bound)**.** *Under the event $\mathcal{E}$, for any $k \in [K]$, any reward function $R : \prod_{h\in[H]}\mathbb{X}_h \to [0,1]$, and any policy π, we have,*

$$\left|V^R_{\widehat{\theta}^k}(\pi) - V^R_{\theta^*}(\pi)\right| \leq V^{\widehat{b}^k}_{\widehat{\theta}^k},$$

where $\widehat{b}^k(\tau_H) = \min\left\{\alpha\sqrt{\sum_h\left\|\overset{\frown}{\overline{\psi}}{}^k(\tau_h)\right\|^2_{(\widehat{U}^k_h)^{-1}}}, 1\right\}$.

Proof. By a direct calculation,

$$\begin{aligned}
\left|V^R_{\widehat{\theta}^k}(\pi) - V^R_{\theta^*}(\pi)\right| &= \left|\sum_{\tau_H}R(\tau_H)\mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_H) - \sum_{\tau_H}R(\tau_H)\mathbb{P}^{\pi}_{\theta^*}(\tau_H)\right| \\
&\overset{(a)}{\leq} \sum_{\tau_H}\left|\mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_H) - \mathbb{P}^{\pi}_{\theta^*}(\tau_H)\right| \\
&= \mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_H), \mathbb{P}^{\pi}_{\theta^*}(\tau_H)\right) \\
&\overset{(b)}{\leq} \alpha\cdot\mathbb{E}_{\tau_H\sim\mathbb{P}^{\pi}_{\widehat{\theta}^k}}\left[\sqrt{\sum_{h=0}^{H-1}\left\|\overset{\frown}{\overline{\psi}}{}^k(\tau_h)\right\|^2_{(\widehat{U}^k_h)^{-1}}}\right] \\
&\overset{(c)}{\leq} \alpha\sum_{\tau_H}\widehat{b}^k(\tau_H)\mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_H) \\
&=: V^{\widehat{b}^k}_{\widehat{\theta}^k}
\end{aligned}$$

where step (a) is by the triangle inequality and the fact that $R(\tau_H) \in [0,1]$, step (b) is by Lemma 6, and step (c) is by the definition of $\widehat{b}^k$. □

D.5 Sublinear Estimation

The next step is to prove that $\sum_{k=1}^{K}V^{\widehat{b}^k}_{\widehat{\theta}^k} = O(\sqrt{K})$. To do that, we first prove that the estimated features $\overset{\frown}{\overline{\psi}}{}^k$ and the ground-truth features can be related through the total-variation distance between the estimated model and the true model.

Lemma 7. *Under event $\mathcal{E}$, for any $k \in [K]$, we have:*

$$
\begin{aligned}
&\mathbb{E}_{\tau_H \sim \mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\sum_{h=0}^{H-1}\left\|\widehat{\overline{\psi^k}}(\tau_h)\right\|^2_{(\widehat{U}^k_h)^{-1}}}\right] \\
&\leq \frac{2HQ_A}{\sqrt{\lambda}}\mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\pi}_{\theta^*}(\tau_h), \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h)\right) + \left(1 + \frac{2\max_{s\in\mathcal{A}}|\mathbb{X}_s|\, Q_A\sqrt{7r\beta}}{\sqrt{\lambda}}\right)\sum_{h=0}^{H-1}\mathbb{E}_{\tau_h\sim\mathbb{P}^{\pi}_{\theta^*}}\left\|\overline{\psi^*}(\tau_h)\right\|_{(\widehat{U}^k_h)^{-1}}
\end{aligned}
$$

Proof. First, we recall the definition of $\widehat{U}^k_h$, and we define its ground-truth counterpart replacing estimated features with true features,

$$
\begin{aligned}
\widehat{U}^k_h &= \lambda I + \sum_{\tau\in\mathcal{D}^k_h}\widehat{\overline{\psi^k}}(\tau_h)\widehat{\overline{\psi^k}}(\tau_h)^\top, \\
U^k_h &= \lambda I + \sum_{\tau\in\mathcal{D}^k_h}\overline{\psi^*}(\tau_h)\overline{\psi^*}(\tau_h)^\top.
\end{aligned}
$$

For any trajectory $\tau_H \in \mathbb{H}_H$, we have,

$$
\begin{aligned}
\sqrt{\sum_{h=0}^{H-1}\left\|\widehat{\overline{\psi^k}}(\tau_h)\right\|^2_{(\widehat{U}^k_h)^{-1}}} &\overset{(a)}{\leq} \sum_{h=0}^{H-1}\left\|\widehat{\overline{\psi^k}}(\tau_h)\right\|_{(\widehat{U}^k_h)^{-1}} \\
&\leq \frac{1}{\sqrt{\lambda}}\sum_{h=0}^{H-1}\left\|\widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h)\right\|_2 + \sum_{h=0}^{H-1}\left(1 + \frac{\sqrt{r}\sqrt{\sum_{\tau_h\in\mathcal{D}^k_h}\left\|\widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h)\right\|^2_2}}{\sqrt{\lambda}}\right)\left\|\overline{\psi^*}(\tau_h)\right\|_{(\widehat{U}^k_h)^{-1}},
\end{aligned}
$$

where step (a) is simply that $\|x\|_2 \leq \|x\|_1$ and step (b) is by the identity (Huang et al. 2023, Lemma 13). Note that r is the rank of the PSR and $r \geq \operatorname{rank}(\{\widehat{\overline{\psi^k}}(\tau_h) : \tau_h \in \mathbb{H}_h\}), \operatorname{rank}(\{\overline{\psi^*}(\tau_h) : \tau_h \in \mathbb{H}_h\})$.

Moreover, we have,

$$
\begin{aligned}
\left\|\widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h)\right\|_2 &\leq \left\|\widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h)\right\|_1 \\
&\overset{(a)}{=} \sum_{q_h\in\mathbb{Q}_h}\left|\overline{\mathbb{P}}_{\widehat{\theta}^k}\left[q^o_h \mid \tau_h, q^a_h\right] - \overline{\mathbb{P}}_{\theta^*}\left[q^o_h \mid \tau_h, q^a_h\right]\right| \\
&\overset{(b)}{\leq} 2\max_{s\in\mathcal{A}}|\mathbb{X}_s|\, Q_A\mathtt{D}_{\mathtt{TV}}\left(\mathbb{P}^{\mathtt{u}^{\mathtt{exp}}_{h-1}}_{\widehat{\theta}^k}(\cdot|\tau_h), \mathbb{P}^{\mathtt{u}^{\mathtt{exp}}_{h-1}}_{\theta^*}(\cdot|\tau_h)\right),
\end{aligned}
$$

where we used the definition of $\overline{\psi}$ in (a) and the definition of the $\mathtt{u}^{\mathtt{exp}}_{h-1}$ in (b).

Now, by Lemma 4, we have,

$$
\begin{aligned}
\sqrt{\sum_{h=0}^{H-1}\left\|\widehat{\overline{\psi^k}}(\tau_h)\right\|^2_{(\widehat{U}^k_h)^{-1}}} &\leq \frac{1}{\sqrt{\lambda}}\sum_{h=0}^{H-1}\left\|\widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h)\right\|_2 + \sum_{h=0}^{H-1}\left(1 + \frac{\sqrt{r}\sqrt{\sum_{\tau_h\in\mathcal{D}^k_h}\left\|\widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h)\right\|^2_2}}{\sqrt{\lambda}}\right)\left\|\overline{\psi^*}(\tau_h)\right\|_{(\widehat{U}^k_h)^{-1}} \\
&\leq \frac{1}{\sqrt{\lambda}}\sum_{h=0}^{H-1}\left\|\widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h)\right\|_2 + \left(1 + \frac{2\max_{s\in\mathcal{A}}|\mathbb{X}_s|\, Q_A\sqrt{7r\beta}}{\sqrt{\lambda}}\right)\sum_{h=0}^{H-1}\left\|\overline{\psi^*}(\tau_h)\right\|_{(\widehat{U}^k_h)^{-1}},
\end{aligned}
$$

where the first line is combining the calculations above and the second line is by the estimation guarantee of Lemma 4.

The first term can be bounded in expectation under $\mathbb{P}^{\pi}_{\theta^*}$ for any π as follows,

$$\begin{aligned}
\sum_{h=0}^{H-1} \mathbb{E}_{\tau_h \sim \mathbb{P}^{\pi}_{\theta^*}} \left[\left\| \widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h) \right\|_2 \right] &\leq \sum_{h=0}^{H-1} \mathbb{E}_{\tau_h \sim \mathbb{P}^{\pi}_{\theta^*}} \left[\left\| \widehat{\overline{\psi^k}}(\tau_h) - \overline{\psi^*}(\tau_h) \right\|_1 \right] \\
&\leq \sum_{h=0}^{H-1} \sum_{\tau_h} \left\| \widehat{\overline{\psi^k}}(\tau_h) \left(\mathbb{P}^{\pi}_{\theta^*}(\tau_h) - \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right) + \widehat{\overline{\psi^k}}(\tau_h) \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) - \overline{\psi^*}(\tau_h) \mathbb{P}^{\pi}_{\theta^*}(\tau_h) \right\|_1 \\
&\overset{(a)}{\leq} \sum_{h=0}^{H-1} \sum_{\tau_h} \left\| \widehat{\overline{\psi^k}}(\tau_h) \right\|_1 \left| \mathbb{P}^{\pi}_{\theta^*}(\tau_h) - \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right| + \left\| \widehat{\overline{\psi^k}}(\tau_h) \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) - \overline{\psi^*}(\tau_h) \mathbb{P}^{\pi}_{\theta^*}(\tau_h) \right\|_1 \\
&\overset{(b)}{\leq} \sum_{h=0}^{H-1} \sum_{\tau_h} \left(\left\| \widehat{\overline{\psi^k}}(\tau_h) \right\|_1 \left| \mathbb{P}^{\pi}_{\theta^*}(\tau_h) - \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right| + \left\| \widehat{\psi^k}(\tau_h) - \psi^*(\tau_h) \right\|_1 \pi(\tau_h) \right) \\
&\overset{(c)}{\leq} 2Q_A \sum_{h=0}^{H-1} \mathtt{D_{TV}} \left(\mathbb{P}^{\pi}_{\theta^*}(\tau_h), \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right) \\
&\overset{(d)}{\leq} 2HQ_A \mathtt{D_{TV}} \left(\mathbb{P}^{\pi}_{\theta^*}(\tau_h), \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right),
\end{aligned}$$

where step (a) is the triangle inequality, step (b) is the definition of $\overline{\psi}(\tau_h)$, step (c) is since $\left\| \widehat{\overline{\psi^k}}(\tau_h) \right\|_1 \leq \left| \mathbb{Q}^A_h \right| \leq Q_A$ for any τ_h and the definition of $\psi(\tau_h)$, and step (d) is simply $\mathtt{D_{TV}} \left(\mathbb{P}^{\pi}_{\theta^*}(\tau_h), \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right) \geq \mathtt{D_{TV}} \left(\mathbb{P}^{\pi}_{\theta^*}(\tau_h), \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right)$. Putting this together concludes the proof,

$$\begin{aligned}
&\mathbb{E}_{\tau_H \sim \mathbb{P}^{\pi}_{\theta^*}} \left[\sqrt{ \sum_{h=0}^{H-1} \left\| \widehat{\overline{\psi^k}}(\tau_h) \right\|^2_{(\widehat{U}^k_h)^{-1}} } \right] \\
&\leq \frac{2HQ_A}{\sqrt{\lambda}} \mathtt{D_{TV}} \left(\mathbb{P}^{\pi}_{\theta^*}(\tau_h), \mathbb{P}^{\pi}_{\widehat{\theta}^k}(\tau_h) \right) + \left(1 + \frac{2 \max_{s \in \mathcal{A}} |\mathbb{X}_s| \, Q_A \sqrt{7r\beta}}{\sqrt{\lambda}} \right) \sum_{h=0}^{H-1} \mathbb{E}_{\tau_h \sim \mathbb{P}^{\pi}_{\theta^*}} \left\| \overline{\psi^*}(\tau_h) \right\|_{(\widehat{U}^k_h)^{-1}}.
\end{aligned}$$

□

The following lemma bounds the regret in the estimation error of the probability of trajectories. It can be proved via an ℓ_2 Eluder argumen. A significant portion of the proof is very similar to that of Proposition 6, involving an exchange of $\widehat{(\cdot)}$ and $(\cdot)^*$. We include the proof for completeness.

Lemma 8. *Under event $\mathcal{E}$, for any $h \in \{0, ..., H-1\}$, we have*

$$\sum_k \mathtt{D_{TV}} \left(\mathbb{P}^{\pi^k}_{\theta^\star}(\tau_H), \mathbb{P}^{\pi^k}_{\widehat{\theta}^k}(\tau_H) \right) \lesssim \frac{\max_{s \in \mathcal{A}} |\mathbb{X}_s| \, Q_A \sqrt{\beta}}{\gamma} \sqrt{rHK \log \left(1 + \frac{dQ_AK}{\gamma^4} \right)}.$$

Here, $a \lesssim b$ indicates that there is an absolute positive constant c s.t. $a \leq c \cdot b$.

Proof. Recall that, by the first inequality in Proposition 6, we have:

$$\mathtt{D_{TV}} \left(\mathbb{P}^{\pi^k}_{\theta^\star}(\tau_H), \mathbb{P}^{\pi^k}_{\widehat{\theta}^k}(\tau_H) \right) \leq \sum_{h=1}^{H} \sum_{\tau_H} \left| \widehat{m}^k(\omega_h)^\top \left(\widehat{M}^k_h(x_h) - M^\star_h(x_h) \right) \psi^\star(\tau_{h-1}) \right| \pi^k(\tau_H)$$

This is very similar to the inequality in Lemma 5, with the difference being that the quantities associated with estimated model and the true model are exchange. Since both correspond to a PSR, the analysis follows a similar series of steps. We will use analogous notation to Lemma 5. We index the future trajectory $\omega_{h-1} = (x_h, \ldots, x_H)$ by i and history trajectory $\tau_{h-1} = (x_1, \ldots, x_{h-1})$ by j. We denote $\widehat{m}^k(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h)\right)$ as w_i, $\overline{\psi}^\star(\tau_{h-1})$ as x_j, and $\pi(\omega_{h-1}|\tau_{h-1})$ as $\pi_{i|j}$.

Define the matrix,

$$\Lambda_h^k = \lambda_0 I + \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^t}} \left[x_j x_j^\top\right]$$

where λ_0 is a constant to be determined later.

For any policy π, using a similar calculation as in Lemma 5, we have,

$$\begin{aligned}
&\sum_{\tau_H} \left|\widehat{m}^k(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h)\right) \psi^\star(\tau_{h-1})\right| \pi^k(\tau_H) \\
&= \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left[\sum_i \pi_{i|j} \left|w_i^\top x_j\right|\right] \\
&= \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left[\left(\sum_i \pi_{i|j} \mathrm{sign}(w_i^\top x_j) w_i\right)^\top x_j\right] \\
&\leq \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left[\|x_j\|_{\Lambda_h^\dagger} \left\|\sum_i \pi_{i|j} \mathrm{sign}(w_i^\top x_j) w_i\right\|_{\Lambda_h}\right]
\end{aligned}$$

where the last line is the Cauchy-Schwarz inequality.

Fix $j = j_0$ and consider the term: $\left\|\sum_i \pi_{i|j_0} \mathrm{sign}(w_i^\top x_{j_0}) w_i\right\|_{\Lambda_h}$ in the above. This term can be partitioned in the same manner as in Lemma 5 by simply using the definition of Λ_h and expanding,

$$\begin{aligned}
&\left\|\sum_i \pi_{i|j_0} \mathrm{sign}(w_i^\top x_{j_0}) w_i\right\|_{\Lambda_h}^2 \\
&= \underbrace{\lambda_0 \left\|\sum_i \pi_{i|j_0} \cdot \mathrm{sign}(w_i^\top x_{j_0}) \cdot w_i\right\|_2^2}_{I_1} + \underbrace{\sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left[\left(\sum_i \pi_{i|j_0} \cdot \mathrm{sign}(w_i^\top x_{j_0}) \cdot w_i^\top x_j\right)^2\right]}_{I_2}
\end{aligned}$$

We bound each term separately. The process is nearly identical to the proof of Lemma 5, but we show it for completeness.

$\sqrt{I_1}$ is bounded by the sum of two terms,

$$\begin{aligned}
\sqrt{I_1} &= \sqrt{\lambda_0} \max_{z \in \mathbb{R}^{d_{h-1}} : \|z\|_2 = 1} \left|\sum_i \pi_{i|j_0} \cdot \mathrm{sign}(w_i^\top x_{j_0}) \cdot w_i^\top z\right| \\
&\overset{(a)}{\leq} \sqrt{\lambda_0} \max_{\|z\|_2=1} \sum_{\omega_{h-1}} \left|\widehat{m}^k(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h)\right) z\right| \pi(\omega_{h-1}|j_0) \\
&\overset{(b)}{\leq} \sqrt{\lambda_0} \max_{\|z\|_2=1} \sum_{\omega_{h-1}} \left|\widehat{m}^k(\omega_h)^\top \widehat{M}_h^k(x_h) z\right| \pi(\omega_{h-1}|j_0) \\
&\quad + \sqrt{\lambda_0} \max_{\|z\|_2=1} \sum_{\omega_{h-1}} \left|\widehat{m}^k(\omega_h)^\top M_h^\star(x_h) z\right| \pi(\omega_{h-1}|j_0),
\end{aligned}$$

where step (a) is the definition of $w_i, \pi_{i|j_0}$, and the triangle inequality, and step (b) is the triangle inequality.

Both terms can be bounded by the γ-well-conditioning assumption on $\widehat{\theta}^k$ and θ^*. Consider the first term,

$$
\begin{aligned}
&\max_{\|z\|_2=1} \sum_{\omega_{h-1}} \left| \widehat{m}^k(\omega_h)^\top \widehat{M}_h^k(x_h) z \right| \pi(\omega_{h-1}|j_0) \\
&= \max_{\|z\|_2=1} \sum_{x_h} \left(\sum_{\omega_h} \left| \widehat{m}^k(\omega_h)^\top \widehat{M}_h^k(x_h) z \right| \pi(\omega_h|j_0, x_h) \right) \pi(x_h|j_0) \\
&\overset{(a)}{\leq} \max_{\|z\|_2=1} \sum_{x_h} \frac{1}{\gamma} \left\| \widehat{M}_h^k(x_h) z \right\|_1 \pi(x_h|j_0) \\
&\overset{(b)}{\leq} \frac{1}{\gamma} \max_{\|z\|_2=1} \frac{\left|\mathbb{Q}_{h+1}^A\right| \|z\|_1}{\gamma} \\
&\overset{(c)}{\leq} \frac{\sqrt{d} Q_A}{\gamma^2}
\end{aligned}
$$

where step (a) is by the first condition in Assumption 1, step (b) is by the second condition of Assumption 1, and step (c) is by the fact that $\max_{z \in \mathbb{R}^{d_{h-1}}: \|z\|_2=1} \|z\|_1 = \sqrt{d_{h-1}} \leq \sqrt{d}$ and $\left|\mathbb{Q}_{h+1}^A\right| \leq Q_A$. In the above, note that we used the γ-well-conditioning of PSR $\widehat{\theta}^k$ in both step (a) and step (b). The second term in $\sqrt{I_1}$ admits an identical bound, simply by using the well-conditioning of the PSR $\widehat{\theta}^k$ in the first step and θ^* in the second step. Hence, we have that

$$
I_1 \leq 4 \frac{\lambda_0 d Q_A^2}{\gamma^4}. \tag{32}
$$

Now, we consider the term I_2

$$
\begin{aligned}
I_2 &\leq \sum_{t<k} \mathbb{E}_{\tau_{h-1} \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left[\left(\sum_{\omega_{h-1}} \left| \widehat{m}^k(\omega_h)^\top \left(\widehat{M}_h^k(x_h) - M_h^\star(x_h) \right) \overline{\psi}^*(\tau_{h-1}) \right| \pi(\omega_{h-1}|j_0) \right)^2 \right] \\
&\leq \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \Bigg[\Bigg(\underbrace{\sum_{\omega_{h-1}} \left| \widehat{m}^k(\omega_h)^\top \widehat{M}_h(x_h) \left(\overline{\psi}^\star(\tau_{h-1}) - \widehat{\overline{\psi}}^k(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0)}_{I_3} \\
&\quad + \underbrace{\sum_{\omega_{h-1}} \left| \widehat{m}^k(\omega_h)^\top \left(\widehat{M}_h^k(x_h) \widehat{\overline{\psi}}^k(\tau_{h-1}) - M_h^\star(x_h) \overline{\psi}^\star(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0) \Bigg)^2 \Bigg]}_{I_4} \\
&=: \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} (I_3 + I_4)^2
\end{aligned}
$$

where the first inequality is due to the absolute value inside the summation and second equality follows from the triangle inequality by adding and subtracting $\widehat{m}_h(\omega_h) \widehat{M}_h(x_h) \widehat{\overline{\psi}}^k(\tau_{h-1})$ inside the absolute value. We now bound each of I_3 and I_4.

First, we bound I_3 as follows,

$$\begin{aligned}
I_3 &= \sum_{\omega_{h-1}} \left| \widehat{m}_h^k(\omega_h)^\top \widehat{M}_h(x_h) \left(\overline{\psi}^\star(\tau_{h-1}) - \widehat{\overline{\psi}}^k(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0) \\
&\overset{(a)}{=} \sum_{\omega_{h-1}} \left| \widehat{m}_{h-1}^k(\omega_{h-1})^\top \left(\overline{\psi}^\star(\tau_{h-1}) - \widehat{\overline{\psi}}^k(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0) \\
&\overset{(b)}{\leq} \frac{1}{\gamma} \left\| \widehat{\overline{\psi}}^k(\tau_{h-1}) - \overline{\psi}^\star(\tau_{h-1}) \right\|_1 \\
&= \frac{1}{\gamma} \sum_{q_{h-1} \in \mathbb{Q}_{h-1}} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[q_{h-1} \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[q_{h-1} \mid \tau_{h-1}\right] \right|,
\end{aligned}$$

where step (a) is since $\widehat{m}(\omega_h)^\top \widehat{M}_h(x_h) = \widehat{m}(\omega_{h-1})^\top$, step (b) is by Assumption 1, and the final equality is by the definition of $\overline{\psi}$.

$$\begin{aligned}
I_4 &= \sum_{\omega_{h-1}} \left| \widehat{m}^k(\omega_h)^\top \left(\widehat{M}_h^k(x_h) \widehat{\overline{\psi}}^k(\tau_{h-1}) - M_h^\star(x_h) \overline{\psi}^\star(\tau_{h-1}) \right) \right| \pi(\omega_{h-1}|j_0) \\
&\overset{(a)}{=} \sum_{\omega_{h-1}} \left| \widehat{m}^k(\omega_h)^\top \left(\overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h \mid \tau_{h-1}\right] \widehat{\overline{\psi}}_h(\tau_h) - \overline{\mathbb{P}}_{\theta^*}\left[x_h \mid \tau_{h-1}\right] \overline{\psi}_h^*(\tau_h) \right) \right| \pi(\omega_{h-1}|j_0) \\
&= \sum_{x_h} \left(\sum_{\omega_h} \left| \widehat{m}^k(\omega_h)^\top \left(\overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h \mid \tau_{h-1}\right] \widehat{\overline{\psi}}_h(\tau_h) - \overline{\mathbb{P}}_{\theta^*}\left[x_h \mid \tau_{h-1}\right] \overline{\psi}_h^*(\tau_h) \right) \right| \pi(\omega_h|j_0, x_h) \right) \pi(x_h|j_0) \\
&\overset{(b)}{\leq} \frac{1}{\gamma} \sum_{x_h} \left\| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h \mid \tau_{h-1}\right] \widehat{\overline{\psi}}_h(\tau_h) - \overline{\mathbb{P}}_{\theta^*}\left[x_h \mid \tau_{h-1}\right] \overline{\psi}_h^*(\tau_h) \right\|_1 \pi(x_h|j_0) \\
&\overset{(c)}{=} \frac{1}{\gamma} \sum_{x_h} \sum_{q_h \in \mathbb{Q}_h} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k}\left[x_h, q_h \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*}\left[x_h, q_h \mid \tau_{h-1}\right] \right| \pi(x_h|j_0)
\end{aligned}$$

where step (a) is by the fact that $M_h(x_h)\overline{\psi}_{h-1}(\tau_{h-1}) = \overline{\mathbb{P}}\left[x_h \mid \tau_{h-1}\right] \overline{\psi}(\tau_h)$, as shown in Equation (28), step (b) is by Assumption 1, and step (c) is since $\left[\overline{\psi}_h(\tau_h)\right]_l = \overline{\mathbb{P}}_\theta\left[q_h^l \mid \tau_h\right]$.

Combining the above, we have that,

$$
\begin{aligned}
I_2 &\leq \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left(I_3 + I_4\right)^2 \\
&\leq \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \Bigg[\Bigg(\frac{1}{\gamma} \sum_{q_{h-1} \in \mathbb{Q}_{h-1}} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[q_{h-1} \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*} \left[q_{h-1} \mid \tau_{h-1}\right] \right| \\
&\qquad\qquad + \frac{1}{\gamma} \sum_{x_h} \sum_{q_h \in \mathbb{Q}_h} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[x_h, q_h \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*} \left[x_h, q_h \mid \tau_{h-1}\right] \right| \pi(x_h | j_0) \Bigg)^2 \Bigg] \\
&= \frac{1}{\gamma^2} \cdot \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \Bigg[\Bigg(\sum_{q_{h-1} \in \mathbb{Q}_{h-1}} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[q_{h-1} \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*} \left[q_{h-1} \mid \tau_{h-1}\right] \right| \\
&\qquad\qquad + \frac{1}{\gamma} \sum_{x_h} \sum_{q_h \in \mathbb{Q}_h} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[x_h, q_h \mid \tau_{h-1}\right] - \overline{\mathbb{P}}_{\theta^*} \left[x_h, q_h \mid \tau_{h-1}\right] \right| \pi(x_h | j_0) \Bigg)^2 \Bigg] \\
&\overset{(a)}{\leq} \frac{1}{\gamma^2} \cdot \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left(\sum_{\omega_{h-1}^a \in \mathbb{Q}_{h-1}^{\exp}} \sum_{\omega_{h-1}^o} \left| \overline{\mathbb{P}}_{\widehat{\theta}^k} \left[\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a\right] - \overline{\mathbb{P}}_{\theta^*} \left[\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h}^a\right] \right| \right)^2 \\
&= \frac{\left|\mathbb{Q}_{h-1}^{\exp}\right|^2}{\gamma^2} \cdot \sum_{t<k} \mathbb{E}_{j \sim \mathbb{P}_{\theta^\star}^{\pi^k}} \left[\mathtt{D}_{\mathtt{TV}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\mathtt{u}_{h-1}^{\mathtt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h-1}^a\right), \mathbb{P}_{\theta^*}^{\mathtt{u}_{h-1}^{\mathtt{exp}}} \left(\omega_{h-1}^o \mid \tau_{h-1}, \omega_{h}^a\right) \right) \right] \\
&\overset{(b)}{\leq} \frac{4 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2}{\gamma^2} \cdot \sum_{t<k} \mathtt{D}_{\mathtt{H}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\nu_h(\pi^t, \mathtt{u}_{h-1}^{\mathtt{exp}})} \left(\tau_H\right), \mathbb{P}_{\theta^*}^{\nu_h(\pi^t, \mathtt{u}_{h-1}^{\mathtt{exp}})} \left(\tau_H\right) \right),
\end{aligned}
$$

where step (a) follows from the definition of $\mathbb{Q}_{h-1}^{\exp}$ (same as Lemma 5), and step (b) is because the Hellinger distance bounds the total variation distance and since $\left|\mathbb{Q}_{h-1}^{\mathtt{exp}}\right| \leq 2 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right| Q_A$. Hence, we have,

$$
I_2 \leq 4 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2 \frac{1}{\gamma^2} \sum_{t<k} \mathtt{D}_{\mathtt{H}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\nu_h(\pi^t, u_{\mathbb{Q}_{h-1}^{\mathtt{exp}}})} \left(\tau_H\right), \mathbb{P}_{\theta^*}^{\nu_h(\pi^t, u_{\mathbb{Q}_{h-1}^{\mathtt{exp}}})} \left(\tau_H\right) \right)
$$

Now, combining the bound on I_1 and I_2 allows us to finally bound $\left\|\sum_i \pi_{i|j} \operatorname{sign}(w_i^\top x_j) w_i\right\|_{\Lambda_h}^2$ as follows,

$$
\begin{aligned}
&\left\| \sum_i \pi_{i|j} \operatorname{sign}(w_i^\top x_j) w_i \right\|_{\Lambda_h}^2 \\
&\leq \frac{4 \lambda_0 Q_A^2 d}{\gamma^4} + \frac{4 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2}{\gamma^2} \cdot \sum_{t<k} \mathtt{D}_{\mathtt{H}}^2 \left(\mathbb{P}_{\widehat{\theta}^k}^{\nu_h(\pi^t, \mathtt{u}_{h-1}^{\mathtt{exp}})} \left(\tau_H\right), \mathbb{P}_{\theta^*}^{\nu_h(\pi^t, \mathtt{u}_{h-1}^{\mathtt{exp}})} \left(\tau_H\right) \right) \\
&=: \left(\tilde{\alpha}_{h-1}^k\right)^2,
\end{aligned}
$$

We choose $\lambda_0 = \frac{\gamma^4}{4 Q_A^2 d}$, and recalling Lemma 4, we have:

$$
\sum_h \left(\tilde{\alpha}_{h-1}^k\right)^2 \leq \frac{28 \max_{s \in \mathcal{A}} \left|\mathbb{X}_s\right|^2 Q_A^2 \beta}{\gamma^2} =: \tilde{\alpha}^2
$$

Thus, we have,

$$\begin{aligned}
\mathtt{D_{TV}}\left(\mathbb{P}^{\pi^k}_{\theta^\star}(\tau_H), \mathbb{P}^{\pi^k}_{\widehat{\theta}^k}(\tau_H)\right) &\leq \sum_{h=1}^{H}\sum_{\tau_H}\left|\widehat{m}^k(\omega_h)^\top\left(\widehat{M}^k_h(x_h) - M^\star_h(x_h)\right)\psi^\star(\tau_{h-1})\right|\pi^k(\tau_H) \\
&\leq \sum_{h=1}^{H}\mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\left\|\overline{\psi}^*(\tau_{h-1})\right\|_{\Lambda_h^\dagger}\left\|\sum_i \pi_{i|j}\mathrm{sign}(w_i^\top x_j)w_i\right\|_{\Lambda_h}\right] \\
&\leq \mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\sum_{h=1}^{H}\left\|\overline{\psi}^*(\tau_{h-1})\right\|_{\Lambda_h^\dagger}\left\|\sum_i \pi_{i|j}\mathrm{sign}(w_i^\top x_j)w_i\right\|_{\Lambda_h}\right] \\
&\overset{(a)}{\leq} \mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\sqrt{\sum_{h=1}^{H}\left\|\overline{\psi}^*(\tau_{h-1})\right\|^2_{\Lambda_h^\dagger}}\sqrt{\sum_{h=1}^{H}\left\|\sum_i \pi_{i|j}\mathrm{sign}(w_i^\top x_j)w_i\right\|^2_{\Lambda_h}}\right] \\
&\overset{(b)}{\leq} \tilde{\alpha}\cdot\mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\sqrt{\sum_{h=1}^{H}\left\|\overline{\psi}^*(\tau_{h-1})\right\|^2_{\Lambda_h^\dagger}}\right] \\
&\leq \tilde{\alpha}\cdot\sqrt{\sum_{h=1}^{H}\mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\left\|\overline{\psi}^*(\tau_{h-1})\right\|^2_{\Lambda_h^\dagger}\right]},
\end{aligned}$$

where step (a) is by the Cauchy-Schwarz inequality and step (b) is by the bound established above. Since the total variation distance is bounded above by 2, we have

$$\mathtt{D_{TV}}\left(\mathbb{P}^{\pi^k}_{\theta^\star}(\tau_H), \mathbb{P}^{\pi^k}_{\widehat{\theta}^k}(\tau_H)\right) \leq \min\left\{\tilde{\alpha}\cdot\sqrt{\sum_{h=1}^{H}\mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\left\|\overline{\psi}^*(\tau_{h-1})\right\|^2_{\Lambda_h^\dagger}\right]}, 2\right\}.$$

Finally, the proof is completed by summing over k using the elliptical potential lemma as follows,

$$\begin{aligned}
\sum_{k=1}^{K}\mathtt{D_{TV}}\left(\mathbb{P}^{\pi^k}_{\theta^\star}(\tau_H), \mathbb{P}^{\pi^k}_{\widehat{\theta}^k}(\tau_H)\right) &\leq \sum_{k=1}^{K}\min\left\{\tilde{\alpha}\cdot\sqrt{\sum_{h=1}^{H}\mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\left\|\overline{\psi}^*(\tau_{h-1})\right\|^2_{\Lambda_h^\dagger}\right]}, 2\right\} \\
&\overset{(a)}{\leq} \sqrt{K}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\tilde{\alpha}^2\cdot\mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\left\|\overline{\psi}^*(\tau_{h-1})\right\|^2_{\Lambda_h^\dagger}\right], 4\right\}} \\
&\leq \sqrt{K}\tilde{\alpha}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\mathbb{E}_{\tau_{h-1}\sim\mathbb{P}^{\pi^k}_{\theta^\star}}\left[\left\|\overline{\psi}^*(\tau_{h-1})\right\|^2_{\Lambda_h^\dagger}\right], 4/\tilde{\alpha}^2\right\}} \\
&\overset{(b)}{\leq} \sqrt{KH}\tilde{\alpha}\sqrt{(1+4/\tilde{\alpha}^2)r\log(1+K/\lambda_0)} \\
&\overset{(c)}{\leq} \left(\sqrt{28(1+4/\tilde{\alpha}^2)}\right)\frac{\max_{s\in\mathcal{A}}|\mathbb{X}_s|\,Q_A}{\gamma}\sqrt{rKH\beta\log(1+K/\lambda_0)} \\
&\overset{(d)}{\lesssim} \frac{\max_{s\in\mathcal{A}}|\mathbb{X}_s|\,Q_A}{\gamma}\sqrt{rKH\beta\log(1+dQ_AK/\gamma)}.
\end{aligned}$$

Here, step (a) is uses the relationship between the ℓ_1 and ℓ_2 norms $\|\cdot\|_1 \leq \sqrt{d}\,\|\cdot\|_2$. Step (b) is by the elliptical potential lemma (Huang et al. 2023, Lemma 14; see also Dani et al. 2008; Abbasi-Yadkori, Pál, et al. 2011; Carpentier et al. 2020). Step (c) uses the bound on $\tilde{\alpha}$ established above. Step (d) uses the definition of λ_0 and the fact that $\sqrt{28(1+4/\tilde{\alpha}^2)}$ is bounded by an absolute constant. □

The following lemma shows that $\sum_{k=1}^{V^{\widehat{b}^k}_{\widehat{\theta}^k}(\pi^k)} = O(\sqrt{K})$.

Lemma 9. *Under the event $\mathcal{E}$, with probability at least $1-\delta$, we have:*

$$\sum_{k=1}^{K} V^{\pi^k}_{\widehat{\theta}^k,\bar{b}^k} \lesssim \left(\sqrt{r}+\frac{Q_A\sqrt{H}}{\gamma}\right)\frac{|\mathcal{A}|\,Q_A^2 H\sqrt{drH\beta K\beta_0}}{\gamma^2}$$

where $\beta_0 = \max\{\log(1+K/\lambda), \log(1+dQ_AK/\gamma)\}$, *and* $\lambda = \frac{\gamma|\mathcal{A}|Q_A\beta\max\{\sqrt{r},Q_A\sqrt{H}/\gamma\}}{\sqrt{dH}}$

Proof. Lemma 6 of Huang et al. 2023. □

D.6 Proof of Theorem 1

Theorem (Restatement of Theorem 1)**.** *Suppose Assumption 1 holds. Let* $p_{\min} = O\left(\frac{\delta}{KH\prod_{h=1}^{H}|\mathbb{X}_h|}\right)$, $\lambda = \frac{\gamma(\max_{s\in\mathcal{A}}|\mathbb{X}_s|)^2 Q_A\beta\max\left\{\sqrt{r},Q_A\sqrt{H}/\gamma\right\}}{\sqrt{dH}}$, $\alpha = O\left(\frac{Q_A\sqrt{Hd}}{\gamma^2}\sqrt{\lambda}+\frac{\max_{s\in\mathcal{A}}|\mathbb{X}_s|Q_A\sqrt{\beta}}{\gamma}\right)$, *and let* $\beta = O(\log\left|\overline{\Theta}_\varepsilon\right|)$, *where* $\varepsilon = O(\frac{p_{\min}}{KH})$. *Then, with probability at least $1-\delta$, Algorithm 1 returns a model θ^ϵ and a policy $\overline{\pi}$ that satisfy*

$$V^R_{\theta^\epsilon}(\pi^*) - V^R_{\theta^\epsilon}(\overline{\pi}) \le \varepsilon, \ \textit{and}\ \forall\pi,\ \mathtt{D_{TV}}\left(\mathbb{P}^\pi_{\theta^\epsilon}(\tau_H), \mathbb{P}^\pi_{\theta^*}(\tau_H)\right) \le \varepsilon.$$

In addition, the algorithm terminates with a sample complexity of,

$$\tilde{O}\left(\left(r+\frac{Q_A^2H}{\gamma^2}\right)\frac{rdH^3\max_{s\in\mathcal{A}}|\mathbb{X}_s|^2 Q_A^4\beta}{\gamma^4\epsilon^2}\right).$$

Proof. By Propositions 8 to 10, the event $\mathcal{E}$ occurs with high probability, $\mathbb{P}[\mathcal{E}] \ge 1-3\delta$. Suppose $\mathcal{E}$ holds. Then, by the upper confidence bound established in Corollary 4, if Algorithm 1 terminates, then the following must hold,

$$\forall\pi,\ \mathtt{D_{TV}}\left(\mathbb{P}^\pi_{\theta^\epsilon}(\tau_H), \mathbb{P}^\pi_{\theta^*}(\tau_H)\right) = 2\max_R\left|V^R_{\theta^\epsilon}(\pi) - V^R_{\theta^*}(\pi)\right| \le V^{\widehat{b}^\epsilon}_{\theta^\epsilon}(\pi) \le \epsilon,$$

where the maximization is over reward functions $R : \mathbb{H}_H \to [0,1]$. The last inequality is simply the termination condition of Algorithm 1.

Now, the difference between the optimal value and the value of $\overline{\pi}$ (the policy returned by the algorithm) can be bounded as follows,

$$\begin{aligned} V^R_{\theta^*}(\pi^*) - V^R_{\theta^*}(\overline{\pi}) &= V^R_{\theta^*}(\pi^*) - V^R_{\theta^\epsilon}(\pi^*) + V^R_{\theta^\epsilon}(\pi^*) - V^R_{\theta^\epsilon}(\overline{\pi}) + V^R_{\theta^\epsilon}(\overline{\pi}) - V^R_{\theta^*}(\overline{\pi}) \\ &\le 2\max_\pi V^{\widehat{b}^\epsilon}_{\theta^\epsilon}(\pi) \le \epsilon, \end{aligned}$$

where the inequality follows from the fact that $\overline{\pi} = \arg\max_\pi V^R_{\theta^\epsilon}(\pi)$ and by Corollary 4.

Recall that by Lemma 9, we have,

$$\sum_{k=1}^{K} V^{\pi^k}_{\widehat{\theta}^k,\bar{b}^k} \lesssim \left(\sqrt{r}+\frac{Q_A\sqrt{H}}{\gamma}\right)\frac{\max_{s\in\mathcal{A}}|\mathbb{X}_s|\,Q_A^2 H\sqrt{rdHK\beta\beta_0}}{\gamma^2}.$$

By the pigeon-hole principle and the termination condition of Algorithm 1, the algorithm must terminate within

$$K = \tilde{O}\left(\left(r+\frac{Q_A^2H}{\gamma^2}\right)\frac{rdH^2\max_{s\in\mathcal{A}}|\mathbb{X}_s|^2 Q_A^4\beta}{\gamma^4\epsilon^2}\right)$$

episodes. Since each episode contains H iterations, this implies a sample complexity of

$$K = \tilde{O}\left(\left(r+\frac{Q_A^2H}{\gamma^2}\right)\frac{rdH^3\max_{s\in\mathcal{A}}|\mathbb{X}_s|^2 Q_A^4\beta}{\gamma^4\epsilon^2}\right).$$

□

E Proof of Theorem 2: UCB Algorithm for Generalized PSRs (Game Setting)

Theorem (Restatement of Theorem 2)**.** *Suppose Assumption 1 holds. Let* $p_{\min} = O\left(\frac{\delta}{KH\prod_{h=1}^{H}|\mathbb{X}_h|}\right)$, $\lambda = \frac{\gamma(\max_{s\in\mathcal{A}}|\mathbb{X}_s|)^2 Q_A\beta\max\left\{\sqrt{r}, Q_A\sqrt{H}/\gamma\right\}}{\sqrt{dH}}$, $\alpha = O\left(\frac{Q_A\sqrt{Hd}}{\gamma^2}\sqrt{\lambda} + \frac{\max_{s\in\mathcal{A}}|\mathbb{X}_s|Q_A\sqrt{\beta}}{\gamma}\right)$, *and let* $\beta = O(\log\left|\overline{\Theta}_\varepsilon\right|)$, *where* $\varepsilon = O(\frac{p_{\min}}{KH})$. *Then, with probability at least* $1-\delta$, *Algorithm 2 returns a model* θ^ϵ *and a policy* $\overline{\pi}$ *which is an* ε*-approximate equilibrium (either NE or CCE). That is,*

$$V_{\theta^*}^i(\overline{\pi}) \geq V_{\theta^*}^{i,\dagger}(\overline{\pi}^{-i}) - \varepsilon, \,\forall i \in [N].$$

In addition, the algorithm terminates with a sample complexity of,

$$\tilde{O}\left(\left(r + \frac{Q_A^2 H}{\gamma^2}\right)\frac{rdH^3\max_{s\in\mathcal{A}}|\mathbb{X}_s|^2 Q_A^4\beta}{\gamma^4\epsilon^2}\right).$$

Proof. Recall that the model-estimation portion of Algorithm 2 is identical to Algorithm 1. Hence, by Theorem 1, the returned estimated model θ^ε satisfies,

$$\mathtt{D_{TV}}\left(\mathbb{P}_{\theta^\varepsilon}^{\boldsymbol{\pi}}(\tau_H), \mathbb{P}_{\theta^*}^{\boldsymbol{\pi}}(\tau_H)\right) \leq \varepsilon/2,$$

for any collection of policies $\boldsymbol{\pi} = (\pi^i : i \in [N])$. This implies that $V_{\theta^*}^i(\overline{\pi}) \geq V_{\theta^\varepsilon}^i(\overline{\pi}) - \varepsilon/2$ for all $i \in [N]$.

Consider first the case of Nash equilibrium. We have that $\overline{\pi}$ is a Nash equilibrium under θ^ε. That is, for all $i \in [N]$,

$$V_{\theta^\varepsilon}^i(\overline{\pi}) = \max_{\tilde{\pi}^i \in \Gamma_{\text{ind}^i}} V_{\theta^\varepsilon}^i(\tilde{\pi}^i, \overline{\pi}^{-i}) =: V_{\theta^\varepsilon}^{i,\dagger}(\overline{\pi}^{-i}).$$

Moreover, note that,

$$\begin{aligned}
\left|V_{\theta^\varepsilon}^{i,\dagger}(\overline{\pi}^{-i}) - V_{\theta^*}^{i,\dagger}(\overline{\pi}^{-i})\right| &= \left|\max_{\tilde{\pi}^i} V_{\theta^\varepsilon}^i(\tilde{\pi}^i, \overline{\pi}^{-i}) - \max_{\tilde{\pi}^i} V_{\theta^*}^i(\tilde{\pi}^i, \overline{\pi}^{-i})\right| \\
&\leq \max_{\tilde{\pi}^i}\left|V_{\theta^\varepsilon}^i(\tilde{\pi}^i, \overline{\pi}^{-i}) - V_{\theta^*}^i(\tilde{\pi}^i, \overline{\pi}^{-i})\right| \\
&\leq \varepsilon/2,
\end{aligned}$$

where the final inequality is since $\mathtt{D_{TV}}\left(\mathbb{P}_{\theta^\varepsilon}^{\boldsymbol{\pi}}(\tau_H), \mathbb{P}_{\theta^*}^{\boldsymbol{\pi}}(\tau_H)\right) \leq \varepsilon/2$ for any $\boldsymbol{\pi}$. Thus, $V_{\theta^\varepsilon}^{i,\dagger}(\overline{\pi}^{-i}) \geq V_{\theta^*}^{i,\dagger}(\overline{\pi}^{-i}) - \varepsilon/2$.

Putting this together, we have,

$$\begin{aligned}
V_{\theta^*}^i(\overline{\pi}) &\geq V_{\theta^\varepsilon}^i(\overline{\pi}) - \varepsilon/2 \\
&= V_{\theta^\varepsilon}^{i,\dagger}(\overline{\pi}^{-i}) - \varepsilon/2 \\
&\geq V_{\theta^*}^{i,\dagger}(\overline{\pi}^{-i}) - \varepsilon.
\end{aligned}$$

Hence, $\overline{\pi}$ is an ε-approximate Nash equilibrium. The proof for coarse correlated equilibria is identical, with the maximization over Γ_{ind}^i replaced by Γ_{cor}^i. □