Lower Bounds for Learning in Revealing POMDPs

Fan Chen [1] **Huan Wang** [2] **Caiming Xiong** [2] **Song Mei** [3] **Yu Bai** [2]

Abstract

This paper studies the fundamental limits of reinforcement learning (RL) in the challenging *partially observable* setting. While it is well-established that learning in Partially Observable Markov Decision Processes (POMDPs) requires exponentially many samples in the worst case, a surge of recent work shows that polynomial sample complexities are achievable under the *revealing condition*—A natural condition that requires the observables to reveal some information about the unobserved latent states. However, the fundamental limits for learning in revealing POMDPs are much less understood, with existing lower bounds being rather preliminary and having substantial gaps from the current best upper bounds.

We establish strong PAC and regret lower bounds for learning in revealing POMDPs. Our lower bounds scale polynomially in all relevant problem parameters in a multiplicative fashion, and achieve significantly smaller gaps against the current best upper bounds, providing a solid starting point for future studies. In particular, for *multi-step* revealing POMDPs, we show that (1) the latent state-space dependence is at least $\Omega(S^{1.5})$ in the PAC sample complexity, which is notably harder than the $\widetilde{\Theta}(S)$ scaling for fully-observable MDPs; (2) Any polynomial sublinear regret is at least $\Omega(T^{2/3})$, suggesting its fundamental difference from the *single-step* case where $\widetilde{\mathcal{O}}(\sqrt{T})$ regret is achievable. Technically, our hard instance construction adapts techniques in *distribution testing*, which is new to the RL literature and may be of independent interest. We also complement our results with new sharp regret upper bounds for *strongly B-stable PSRs*, which include single-step revealing POMDPs as a special case.

[1]Peking University [2]Salesforce AI Research [3]UC Berkeley. Correspondence to: Fan Chen <chern@pku.edu.cn>, Song Mei <songmei@berkeley.edu>, Yu Bai <yu.bai@salesforce.com>.

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

1. Introduction

Partial observability—where the agent can only observe partial information about the true underlying state of the system—is ubiquitous in real-world applications of Reinforcement Learning (RL) and constitutes a central challenge to RL (Kaelbling et al., 1998; Sutton & Barto, 2018). It is known that learning in the standard model of Partially Observable Markov Decision Processes (POMDPs) is much more challenging than its fully observable counterpart—Finding a near-optimal policy in long-horizon POMDPs requires a number of samples at least exponential in the horizon length in the worst-case (Krishnamurthy et al., 2016). Such an exponential hardness originates from the fact that the agent may not observe any useful information about the true underlying state of the system, without further restrictions on the structure of the POMDP. This is in stark contrast to learning fully observable (tabular) MDPs where polynomially many samples are necessary and sufficient without further assumptions (Kearns & Singh, 2002; Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018; Zhang et al., 2020; Domingues et al., 2021).

Towards circumventing this hardness result, recent work seeks additional structural conditions that permit sample-efficient learning. One natural proposal is the *revealing condition* (Jin et al., 2020a; Liu et al., 2022a), which at a high level requires the observables (observations and actions) to reveal some information about the underlying latent state, thus ruling out the aforementioned worst-case situation where the observables are completely uninformative. Concretely, the *single-step* revealing condition (Jin et al., 2020a) requires the (immediate) emission probabilities of the latent states to be well-conditioned, in the sense that different states are probabilistically distinguishable from their emissions. The *multi-step* revealing condition (Liu et al., 2022a) generalizes the single-step case by requiring the well conditioning of the multi-step *emission-action* probabilities—the probabilities of observing a *sequence* of observations in the next $m \geqslant 2$ steps, conditioned on taking a specific *sequence* of actions at the current latent state.

Sample-efficient algorithms for learning single-step and multi-step revealing POMDPs are initially designed by Jin et al. (2020a) and Liu et al. (2022a), and subsequently developed in a surge of recent work (Cai et al., 2022; Wang et al.,

Table 1. A summary of lower bounds and current best upper bounds for learning revealing POMDPs, with our contributions highlighted in gray cells. The rates presented here only focus on the dependence in S, O, A, α^{-1}, and T (or ε^{-1}), and omit $\mathrm{poly}(H)$ and all polylog factors. We also assume $O \geqslant \Omega(SA)$ (in our upper bounds) and $A^H \gg \mathrm{poly}(H, S, O, A^m, \alpha^{-1}, T)$ to simplify the presentation. For regret lower bounds, we additional ignore the min with T (due to the trivial $O(T)$ regret upper bound). *Obtained by an explore-then-exploit conversion.

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step α-revealing	$\widetilde{\mathcal{O}}\Big(\frac{S^2OA}{\alpha^2\varepsilon^2}\Big)$ (Chen et al., 2022a)	$\Omega\Big(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\Big)$ (Theorem 4)	$\widetilde{\mathcal{O}}\Big(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\Big)$ (Theorem 8)	$\Omega\Big(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\Big)$ (Corollary 7)
m-step ($m \geqslant 2$) α-revealing	$\widetilde{\mathcal{O}}\Big(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\Big)$ (Chen et al., 2022a)	$\Omega\Big(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\Big)$ (Theorem 5)	$\widetilde{\mathcal{O}}\Big(\Big(\frac{S^2OA^m}{\alpha^2}\Big)^{1/3}T^{2/3}\Big)$ (Chen et al., 2022a)*	$\Omega\Big(\Big(\frac{SO^{1/2}A^m}{\alpha^2}\Big)^{1/3}T^{2/3}\Big)$ (Theorem 6)

2022; Uehara et al., 2022b; Zhan et al., 2022; Chen et al., 2022a; Liu et al., 2022b; Zhong et al., 2022). For finding an ε near-optimal policy in m-step revealing POMDPs, these results obtain PAC sample complexities (required episodes of play) that scale polynomially with the number of states, observations, action sequences (of length m), the horizon, $(1/\alpha)$ where $\alpha > 0$ is the *revealing constant*, and $(1/\varepsilon)$, with the current best rate given by Chen et al. (2022a).

Despite this progress, the fundamental limit for learning in revealing POMDPs remains rather poorly understood. First, lower bounds for revealing POMDPs are currently scarce, with existing lower bounds either being rather preliminary in its rates (Liu et al., 2022a), or following by direct reduction from fully observable settings, which does not exhibit the challenge of partial observability (cf. Section 2.2 for detailed discussions). Such lower bounds leave open many fundamental questions, such as the dependence on α in the optimal PAC sample complexity: the current best lower bound scales in α^{-1} while the current best upper bound requires α^{-2}. Second, the current best upper bounds for learning revealing POMDPs are mostly obtained by general-purpose algorithms not specially tailored to POMDPs (Chen et al., 2022a; Liu et al., 2022b; Zhong et al., 2022). These algorithms admit unified analysis frameworks for a large number of RL problems including revealing POMDPs, and it is unclear whether these analyses (and the resulting upper bounds) unveil fundamental limits of revealing POMDPs.

This paper establishes strong sample complexity lower bounds for learning revealing POMDPs. Our contributions can be summarized as follows.

- We establish PAC lower bounds for learning both single-step (Section 3.1) and multi-step (Section 3.2) revealing POMDPs. Our lower bounds are the first to scale with all relevant problem parameters in a multiplicative fashion, and settles several open questions about the fundamental limits for learning revealing POMDPs. Notably, our PAC lower bound for the multi-step case scales as $\Omega(S^{1.5})$, where S is the size of the latent state-space, which is notably harder than fully observable MDPs where $\widetilde{\Theta}(S)$ is the minimax optimal scaling. Further, our lower bounds exhibit rather mild gaps from the current best upper bounds, which could serve as a starting point for further fine-grained studies.
- We establish regret lower bounds for the same settings. Perhaps surprisingly, we show an $\Omega(T^{2/3})$ regret lower bound for multi-step revealing POMDPs (Section 4). Our construction unveils some new insights about the multi-step case, and suggests its fundamental difference from the single-step case in which $\widetilde{\mathcal{O}}(\sqrt{T})$ regret is achievable.
- Technically, our lower bounds are obtained by embedding *uniformity testing* problems into revealing POMDPs, in particular into an *m-step revealing combination lock* which is the core of our hard instance constructions (Section 5). The proof further uses information-theoretic techniques such as Ingster's method for bounding certain divergences, which are new to the RL literature.
- We discuss some additional interesting implications to RL theory in general, in particular to the Decision-Estimation Coefficients (DEC) framework (Section 6.2).

We illustrate our main results against the current best upper bounds in Table 1.

1.1. Related work

Hardness of learning general POMDPs It is well-established that learning a near-optimal policy in POMDPs is computationally hard in the worst case (Papadimitriou & Tsitsiklis, 1987; Mossel & Roch, 2005). With regard to learning, Krishnamurthy et al. (2016); Jin et al. (2020a) used the combination lock hard instance to show that learning episodic POMDPs requires a sample size at least exponential in the horizon H. Kearns et al. (1999); Even-Dar et al. (2005) developed algorithms for learning episodic POMDPs that admit sample complexity scaling with A^H. A similar sample complexity can also be obtained by bounding the

Bellman rank (Jiang et al., 2017; Du et al., 2021; Jin et al., 2021) or coverability (Xie et al., 2022).

Revealing POMDPs Jin et al. (2020a) proposed the single-step revealing condition in under-complete POMDPs and showed that it is a sufficient condition for sample-efficient learning of POMDPs by designing a spectral type learning algorithm. Liu et al. (2022a;c) proposed the multi-step revealing condition to the over-complete POMDPs and developed the optimistic maximum likelihood estimation (OMLE) algorithm for efficient learning. Cai et al. (2022); Wang et al. (2022) extended these results to efficient learning of linear POMDPs under variants of the revealing condition. Golowich et al. (2022b;a) showed that approximate planning under the observable condition, a variant of the revealing condition, admits quasi-polynomial time algorithms.

The only existing lower bound for learning revealing POMDPs is provided by Liu et al. (2022a), which modified the combination lock hard instance (Krishnamurthy et al., 2016) to construct an m-step 1-revealing POMDP and show an $\Omega(A^{m-1})$ sample complexity lower bound for learning a $1/2$-optimal policy. Our lower bound improves substantially over theirs using a much more sophisticated hard instance construction that integrates the combination lock with the tree hard instance for learning MDPs (Domingues et al., 2021) and the hard instance for uniformity testing (Paninski, 2008; Canonne, 2020). Similar to the lower bound for uniformity testing, the proof of our lower bound builds on Ingster's method (Ingster & Suslina, 2012).

Other structural conditions Other conditions that enable sample-efficient learning of POMDPs include reactiveness (Jiang et al., 2017), decodablity (Efroni et al., 2022), structured latent MDPs (Kwon et al., 2021), learning short-memory policies (Uehara et al., 2022b), deterministic transitions (Uehara et al., 2022a), and regular predictive state representations (PSRs) (Zhan et al., 2022). Chen et al. (2022a); Liu et al. (2022b); Zhong et al. (2022) propose unified structural conditions for PSRs, which encompasses most existing tractable classes including revealing POMDPs, decodable POMDPs, and regular PSRs.

2. Preliminaries

POMDPs An episodic Partially Observable Markov Decision Process (POMDP) is specified by a tuple $M = \{H, \mathcal{S}, \mathcal{O}, \mathcal{A}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, \mu_1\}$, where $H \in \mathbb{Z}_{\geqslant 1}$ is the horizon length; $(\mathcal{S}, \mathcal{O}, \mathcal{A})$ are the spaces of (latent) states, observations, and actions with cardinality (S, O, A) respectively; $\mathbb{O}_h(\cdot|\cdot) : \mathcal{S} \to \Delta(\mathcal{O})$ is the emission dynamics at step h (which we identify as an emission matrix $\mathbb{O}_h \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}}$); $\mathbb{T}_h(\cdot|\cdot,\cdot) : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamics over the latent states (which we identify as a transition matrix $\mathbb{T}_h \in \mathbb{R}^{\mathcal{S}\times(\mathcal{S}\times\mathcal{A})}$); $r_h(\cdot,\cdot) : \mathcal{O}\times\mathcal{A} \to [0,1]$ is the (possibly random) reward function; $\mu_1 = \mathbb{T}_0(\cdot) \in \Delta(\mathcal{S})$ specifies the distribution of initial state. At each step $h \in [H]$, given latent state s_h (which the agent does not observe), the system emits observation $o_h \sim \mathbb{O}_h(\cdot|s_h)$, receives action $a_h \in \mathcal{A}$ from the agent, emits reward $r_h(o_h, a_h)$, and then transits to the next latent state $s_{h+1} \sim \mathbb{T}_h(\cdot|s_h, a_h)$ in a Markovian fashion.

We use $\tau = (o_1, a_1, \ldots, o_H, a_H) = (o_{1:H}, a_{1:H})$ to denote a full history of observations and actions observed by the agent, and $\tau_h = (o_{1:h}, a_{1:h})$ to denote a partial history up to step $h \in [H]$. A policy is given by a collection of distributions over actions $\pi = \{\pi_h(\cdot|\tau_{h-1}, o_h) \in \Delta(\mathcal{A})\}_{h,\tau_{h-1},o_h}$, where $\pi_h(\cdot|\tau_{h-1}, o_h)$ specifies the distribution of a_h given the history (τ_{h-1}, o_h). We denote Π as the set of all policies. The value function of any policy π is denoted as $V_M(\pi) = \mathbb{E}_M^\pi[\sum_{h=1}^H r_h(o_h, a_h)]$, where $\mathbb{E}_M^\pi$ specifies the law of $(o_{1:H}, a_{1:H})$ under model M and policy π. The optimal value function of model M is denoted as $V_M^\star = \max_{\pi\in\Pi} V_M(\pi)$. Without loss of generality, we assume that the total rewards are bounded by one, i.e. $\sum_{h\in[H]} r_h(o_h, a_h) \leqslant 1$ for any $(o_{1:H}, a_{1:H}) \in (\mathcal{O}\times\mathcal{A})^H$.

Learning goals We consider learning POMDPs from bandit feedback (exploration setting) where the agent plays with a fixed (unknown) POMDP model M for $T \in \mathbb{N}_+$ episodes. In each episode, the agent plays some policy $\pi^{(t)}$, and observes the trajectory $\tau^{(t)}$ and the rewards $r_{1:H}^{(t)}$.

We consider the two standard learning goals of PAC learning and no-regret learning. In PAC learning, the goal is to output a near-optimal policy $\widehat{\pi}$ so that $V_M^\star - V_M(\widehat{\pi}) \leqslant \varepsilon$ within as few episodes of play as possible. In no-regret learning, the goal is to minimize the regret

$$\mathbf{Regret}(T) := \sum_{t=1}^T \left(V_M^\star - V_M\left(\pi^{(t)}\right)\right),$$

and an algorithm is called no-regret if $\mathbf{Regret}(T) = o(T)$ is sublinear in T. It is known that no-regret learning is no easier than PAC learning, as any no-regret algorithm can be turned to a PAC learning algorithm by the standard online-to-batch conversion (e.g. Jin et al. (2018)) that outputs the average policy $\widehat{\pi} := \frac{1}{T}\sum_{t=1}^T \pi^{(t)}$ after T episodes of play.

2.1. Revealing POMDPs

We consider revealing POMDPs (Jin et al., 2020a; Liu et al., 2022a), a structured subclass of POMDPs that is known to be sample-efficiently learnable. For any $m \geqslant 1$, define the *m-step emission-action matrix* $\mathbb{M}_{h,m} \in \mathbb{R}^{\mathcal{O}^m\mathcal{A}^{m-1}\times\mathcal{S}}$ of a POMDP M at step $h \in [H-m+1]$ as

$$\begin{aligned}&[\mathbb{M}_{h,m}]_{(\mathbf{o},\mathbf{a}),s}\\ &:= \mathbb{P}_M(o_{h:h+m-1} = \mathbf{o}|s_h = s, a_{h:h+m-2} = \mathbf{a}).\end{aligned} \tag{1}$$

In the special case where $m = 1$ (the *single-step* case), we have $\mathbb{M}_{h,1} = \mathbb{O}_h \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}}$, i.e. the emission-action matrix reduces to the emission matrix. For $m \geqslant 2$, the m-step emission-action matrix $\mathbb{M}_{h,m}$ generalizes the emission matrix by encoding the *emission-action probabilities*, i.e. probabilities of observing any observation sequence $\mathbf{o} \in \mathcal{O}^m$, starting from any latent state $s \in \mathcal{S}$ and taking any action sequence $\mathbf{a} \in \mathcal{A}^{m-1}$ in the next $m-1$ steps.

A POMDP is called m-step revealing if its emission-action matrices $\{\mathbb{M}_{h,m}\}_{h\in[H-m+1]}$ admit *generalized left inverses* with bounded operator norm.

Definition 1 (m-step α-revealing POMDPs)**.** *For $m \geqslant 1$ and $\alpha > 0$, a POMDP model M is called* m-step revealing*, if there exists matrices $\mathbb{M}^+_{h,m} \in \mathbb{R}^{\mathcal{S}\times\mathcal{O}^m\mathcal{A}^{m-1}}$ satisfying $\mathbb{M}^+_{h,m}\mathbb{M}_{h,m}\mathbb{T}_{h-1} = \mathbb{T}_{h-1}$ (generalized left inverse of $\mathbb{M}_{h,m}$) for any $h \in [H-m+1]$. Furthermore, the POMDP model M is called m-step α-revealing if each $\mathbb{M}^+_{h,m}$ further admits $(* \to 1)$-operator norm bounded by α^{-1}:*

$$\|\mathbb{M}^+_{h,m}\|_{*\to 1} := \max_{\|\mathbf{x}\|_* \leqslant 1} \|\mathbb{M}^+_{h,m}\mathbf{x}\|_1 \leqslant \alpha^{-1}, \tag{2}$$

where for any vector $\mathbf{x} = (\mathbf{x}(\mathbf{o},\mathbf{a}))_{\mathbf{o}\in\mathcal{O}^m,\mathbf{a}\in\mathcal{A}^{m-1}}$, we denote its star-norm by

$$\|\mathbf{x}\|_* := \left[\sum_{\mathbf{a}\in\mathcal{A}^{m-1}}\left(\sum_{\mathbf{o}\in\mathcal{O}^m}|\mathbf{x}(\mathbf{o},\mathbf{a})|\right)^2\right]^{1/2}.$$

Let $\alpha_m(M)$—the m-step revealing constant of model M—denote the maximum possible $\alpha > 0$ such that (2) holds, so that M is m-step α-revealing iff $\alpha_m(M) \geqslant \alpha$.

In Definition 1, the existence of a generalized left inverse requires the matrix $\mathbb{M}_{h,m}$ to have full rank in the column space of $\mathbb{T}_{h-1}$, which ensures that different states reachable from the previous step are information-theoretically distinguishable from the next m observations and $m-1$ actions. The revealing condition—as a quantitative version of this full rank condition—ensures that states can be probabilistically "revealed" from the observables, and enables sample-efficient learning (Liu et al., 2022a).

Our choice of the $(* \to 1)$-norm in (2) is different from existing work (Liu et al., 2022a;b; Chen et al., 2022a); however, it enables a tighter gap between our lower bounds and existing upper bounds. In addition, $(* \to 1)$-norm has natural probabilistic interpretations: The m-step emission matrix $\mathbb{M}_{h,m}$ maps a distribution over $\mathcal{S}$ to a collection of A^{m-1} distributions over $\mathcal{O}^m$. Then, the 1-norm over $\mathbb{R}^{\mathcal{S}}$ and the $*$-norm over $\mathbb{R}^{\mathcal{O}^m\times\mathcal{A}^{m-1}}$ directly correspond to the TV distance (and its aggregated version over $\mathcal{A}^{m-1}$), which is arguably a more natural choice than the ℓ_2 norm in (Liu et al., 2022a). Finally, we remark that the choice of the norms is not important when only polynomial learnability (not the exact rate of the polynomial) is of consideration, due to the equivalence between norms up to dimension factors.

Single-step vs. multi-step We highlight that when $m = 1$, the emission-action matrix $\mathbb{M}_{h,1} = \mathbb{O}_h$ does not involve the effect of actions. This turns out to make it qualitatively different from the *multi-step* cases where $m \geqslant 2$, which will be reflected in our results.

Additionally, we show that any m-step α-revealing POMDP is also $(m+1)$-step α-revealing, but not vice versa (proof in Appendix C.1; this result is intuitive yet we were unable to find it in the literature). Therefore, as m increases, the class of m-step revealing POMDPs becomes strictly larger and thus no easier to learn.

Proposition 2 (m-step revealing $\subsetneq$ $(m+1)$-step revealing)**.** *For any $m \geqslant 1$ and any POMDP M with horizon $H \geqslant m+1$, we have $\alpha_{m+1}(M) \geqslant \alpha_m(M)$. Consequently, any m-step α-revealing POMDP is also an $(m+1)$-step α-revealing POMDP. Conversely, there exists an $(m+1)$-step revealing POMDP that is not an m-step revealing POMDP.*

2.2. Known upper and lower bounds

Upper bounds Learning revealing POMDPs is known to admit polynomial sample complexity upper bounds (Liu et al., 2022a;b; Chen et al., 2022a). The current best PAC sample complexity for learning revealing POMDPs is given in the following result, which follows directly by adapting the results of Chen et al. (2022a;b) to our definition of the revealing condition (cf. Appendix C.2).

Theorem 3 (PAC upper bound for revealing POMDPs (Chen et al., 2022a))**.** *There exists algorithms (*OMLE*,* EXPLORATIVE E2D *&* MOPS*) that can find an ε-optimal policy of any m-step α-revealing POMDP w.h.p. within*

$$T \leqslant \widetilde{\mathcal{O}}\left(\frac{S^2OA^m(1+SA/O)H^3}{\alpha^2\varepsilon^2}\right) \tag{3}$$

episodes of play.

Lower bounds Existing lower bounds for learning revealing POMDPs are scarce and preliminary. The only existing PAC lower bound for m-step α-revealing POMDPs is

$$\Omega(\min\{\tfrac{1}{\alpha H}, A^{H-1}\} + A^{m-1})$$

given by Liu et al. (2022a, Theorem 6 & 9) for learning an $\varepsilon = \Theta(1)$-optimal policy, which does not scale with either the model parameters S, O or $(1/\varepsilon)$ for small ε.

In addition, revealing POMDPs subsume two fully observable models as special cases: (fully observable) MDPs with H steps, $\min\{S, O\}$ states, and A actions (with $\alpha = 1$); and contextual bandits with O contexts and A actions. By standard PAC lower bounds (Dann & Brunskill, 2015; Lattimore & Szepesvári, 2020; Domingues et al., 2021) in both

settings[1], this implies an

$$\Omega\big((H \min\{S, O\}A + OA)/\varepsilon^2\big)$$

PAC lower bound for m-step α-revealing POMDPs for any $m \geqslant 1$ and $\alpha \leqslant 1$.

Both lower bounds above exhibit substantial gaps from the upper bound (3). Indeed, the upper bound scales *multiplicatively* in S, A^m, O, α^{-1} and $1/\varepsilon^2$, whereas the lower bounds combined are far smaller than this multiplicative scaling.

3. PAC lower bounds

We establish PAC lower bounds for both single-step (Section 3.1) and multi-step (Section 3.2) revealing POMDPs. We first state and discuss our results, and then provide a proof overview for the multi-step case in Section 5.

3.1. Single-step revealing POMDPs

We begin by establishing the PAC lower bound for the single-step case. The proof can be found in Appendix E.

Theorem 4 (PAC lower bound for single-step revealing POMDPs)**.** *For any $O \geqslant S \geqslant 5$, $A \geqslant 3$, $H \geqslant 4\log_2 S$, $\alpha \in (0, \frac{1}{5H}]$, $\varepsilon \in (0, 0.01]$, there exists a family $\mathcal{M}$ of single-step revealing POMDPs with $|\mathcal{S}| \leqslant S$, $|\mathcal{O}| \leqslant O$, $|\mathcal{A}| = A$, and $\alpha_1(M) \geqslant \alpha$ for all $M \in \mathcal{M}$, such that for any algorithm $\mathfrak{A}$ that interacts with the environment for T episodes and returns a π^{out} such that $V_M^\star - V_M(\pi^{\mathrm{out}}) < \varepsilon$ with probability at least $3/4$ for all $M \in \mathcal{M}$, we must have*

$$T \geqslant c \cdot \min\left\{\frac{SO^{1/2}AH}{\alpha^2\varepsilon^2}, \frac{SA^{H/2}H}{\varepsilon^2}\right\}, \tag{4}$$

where $c > 0$ is an absolute constant.

The lower bound in Theorem 4 (and subsequent lower bounds) involves the minimum over two terms, where the second term "caps" the lower bound by an exponential scaling[2] in H and is less important. The main term $\Omega(S\sqrt{O}AH/(\alpha^2\varepsilon^2))$ scales polynomially in $1/\alpha^2$, $1/\varepsilon^2$, and (S, O, A) in a *multiplicative* fashion. This is the first such result for revealing POMDPs and improves substantially over existing lower bounds (cf. Section 2.2).

Implications Theorem 4 shows that, the multiplicative dependence on $(S, A, O, 1/\alpha, 1/\varepsilon)$ in the the current best PAC upper bound $\widetilde{\mathcal{O}}(S^2OA(1 + SA/O)/(\alpha^2\varepsilon^2))$ (Theorem 3; ignoring H) is indeed necessary, and settles several open questions about learning revealing POMDPs:

- It settles the optimal dependence on α to be $\Theta(\alpha^{-2})$ (combining our lower bound with the $\mathcal{O}(\alpha^{-2})$ upper bound), whereas the previous best lower bound on α is $\Omega(\alpha^{-1})$ (Liu et al., 2022a).
- For joint dependence on (α, ε), it shows that $1/(\alpha^2\varepsilon^2)$ samples are necessary. This rules out possibilities for better rates—such as the $\widetilde{\mathcal{O}}(\max\{1/\alpha^2, 1/\varepsilon^2\})$ upper bound for single-step revealing POMDPs with *deterministic transitions* (Jin et al., 2020a)—in the general case.
- It necessitates a $\mathrm{poly}(O)$ factor as multiplicative upon the other parameters (most importantly $1/(\alpha^2\varepsilon^2)$) in the sample complexity, which confirms that large observation spaces do impact learning in a strong sense.

Finally, compared with the current best PAC upper bound, the lower bound $\Omega(SO^{1/2}A/(\alpha^2\varepsilon^2))$ captures all the parameters and is a $S\sqrt{O}$-factor away in the rich-observation regime where $O \geqslant \Omega(SA)$. This provides a solid starting point for future studies.

Remark on requiring $O \geqslant S$ All of our results require $O \geqslant S$ due to the tree structure in our construction. In the general case (where we may have $O < S$), all our lower bounds still hold with S replaced by $\min\{S, O\}$. In addition, it is potentially possible to strengthen the lower bound when $O < S$, which however may significantly complicate the constructions, and hence are left for future work.

3.2. Multi-step revealing POMDPs

Using similar hard instance constructions (more details in Section 5), we establish the PAC lower bound for the multi-step case with $m \geqslant 2$ (proof in Appendix G).

Theorem 5 (PAC lower bound for multi-step revealing POMDPs)**.** *For any $m \geqslant 2$, $O \geqslant S \geqslant 10$, $A \geqslant 3$, $H \geqslant 8\log_2 S + 2m$, $\alpha \in (0, 0.1]$, $\varepsilon \in (0, 0.01]$, there exists a family $\mathcal{M}$ of m-step revealing POMDPs with $|\mathcal{S}| \leqslant S$, $|\mathcal{O}| \leqslant O$, $|\mathcal{A}| = A$, and $\alpha_m(M) \geqslant \alpha$ for all $M \in \mathcal{M}$, such that any algorithm $\mathfrak{A}$ that interacts with the environment and returns a π^{out} such that $V_M^\star - V_M(\pi^{\mathrm{out}}) < \varepsilon$ with probability at least $3/4$ for all $M \in \mathcal{M}$, we must have*

$$T \geqslant c_m \cdot \min\left\{\frac{(S^{1.5} \vee SA)O^{1/2}A^{m-1}H}{\alpha^2\varepsilon^2}, \frac{SA^{H/2}H}{\varepsilon^2}\right\},$$

where $c_m = c_0/m$ for some absolute constant $c_0 > 0$.

The main difference in the multi-step case (Theorem 5) is in its higher A dependence $\Omega(A^{m-1})$, which suggests that the A^m dependence in the upper bound (Theorem 3) is morally unimprovable. Also, the $S^{1.5}$ scaling in Theorem 5 is higher than Theorem 4, which makes the result qualitatively stronger than the single-step case even aside from the A-dependence. This happens since the hard instance here is actually a strengthening—instead of a direct adaptation—of

[1] With total reward scaled to $[0, 1]$.

[2] A $\widetilde{\mathcal{O}}(\mathrm{poly}(S, O, H)A^H/\varepsilon^2)$ PAC upper bound is indeed achievable for *any* POMDP (not necessarily revealing) (Even-Dar et al., 2005); see also the discussions in Uehara et al. (2022b).

the single-step case, by leveraging the nature of multi-step revealing; see Section 5.3 for a discussion.

Again, compared with the current best PAC upper bound $S^2OA^m(1+SA/O)/(\alpha^2\varepsilon^2)$ (Theorem 3), the lower bound in Theorem 5 has an $\sqrt{SOA} \wedge S\sqrt{O}$ gap from the current best upper bound. We believe that the $\sqrt{SO}$ factor in this gap is unimprovable from the lower bound side under the current hard instance; see Section 6.3 for a discussion.

$\sqrt{O}$ dependence Our lower bounds for both the single-step and the multi-step cases scale as $\sqrt{O}$ in its O-dependence. Such a scaling comes from the complexity of the *uniformity testing* task of size $\mathcal{O}(O)$, embedded in the revealing POMDP hard instances, whose sample complexity is $\Theta(\sqrt{O}/\varepsilon^2)$ (Paninski, 2008; Diakonikolas et al., 2014; Canonne, 2020). The construction of the hard instances will be described in detail in Section 5.

4. Regret lower bound for multi-step case

We now turn to establishing regret lower bounds. We show that surprisingly, for m-step revealing POMDPs with any $m \geqslant 2$, a non-trivial polynomial regret (neither linear in T nor exponential in H) has to be at least $\Omega(T^{2/3})$. The proof can be found in Appendix F.

Theorem 6 ($\Omega(T^{2/3})$ regret lower bound for multi-step revealing POMDPs)**.** *For any $m \geqslant 2, O \geqslant S \geqslant 8$, $A \geqslant 3$, $H \geqslant 8\log_2 S + 2m$, $\alpha \in (0, 0.1]$, $T \geqslant 1$, there exists a family $\mathcal{M}$ of m-step revealing POMDPs with $|\mathcal{S}| \leqslant S$, $|\mathcal{O}| \leqslant O$, $|\mathcal{A}| = A$, and $\alpha_m(M) \geqslant \alpha$ for all $M \in \mathcal{M}$, such that for any algorithm $\mathfrak{A}$, it holds that*

$$\max_{M\in\mathcal{M}} \mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \geqslant c_m \cdot \min\left\{\left(\frac{SO^{1/2}A^mH}{\alpha^2}\right)^{1/3} T^{2/3}, \sqrt{SA^{H/2}HT}, T\right\},$$

where $c_m = c_0/m$ for some absolute constant $c_0 > 0$.

Currently, the best sublinear regret (polynomial in other problem parameters) is indeed $T^{2/3}$ by a standard explore-then-exploit style conversion from the PAC result (Chen et al., 2022a). Theorem 6 rules out possibilities for obtaining an improvement (e.g. to $\sqrt{T}$) by showing that $T^{2/3}$ is rather a fundamental limit.

Proof intuition The hard instance used in Theorem 6 is the same as one of the PAC hard instances (see Section 5). However, Theorem 6 relies on a key new observation that leads to the $\Omega(T^{2/3})$ regret lower bound. Specifically, for multi-step revealing POMDPs, we can design a hard instance such that the following two kinds of action sequences (of length $m-1$) are *disjoint*:

- *Revealing* action sequences, which yield observations that reveal information about the true latent state;
- *High-reward* action sequences.

The multi-step revealing condition (Definition 1) permits such constructions. Intuitively, this is since its requirement that $\mathbb{M}_{h,m} \in \mathbb{R}^{\mathcal{O}^m\mathcal{A}^{m-1}\times\mathcal{S}}$ admits a generalized left inverse is fairly liberal, and can be achieved by carefully designing the emission-action probabilities over a *subset* of action sequences. In other words, the multi-step revealing condition allows only *some* action sequences to be revealing, such as the ones that receive rather suboptimal rewards.

Such a hard instance forbids an efficient exploration-exploitation tradeoff, as exploration (taking revealing actions) and exploitation (taking high-reward actions) cannot be simultaneously done. Consequently, the best thing to do is simply an explore-then-exploit type algorithm[3] whose regret is typically $\Theta(T^{2/3})$ (Lattimore & Szepesvári, 2020).

Difference from the single-step case Theorem 6 demonstrates a fundamental difference between the multi-step and single-step settings, as single-step revealing POMDPs are known to admit $\widetilde{\mathcal{O}}(\sqrt{T})$ regret upper bounds (Liu et al., 2022a). Intuitively, the difference is that in single-step revealing POMDPs, the agent does not need to take specific actions to acquire information about the latent state, so that information acquisition (exploration) and taking high-reward actions (exploitation) *can* always be achieved simultaneously.

Towards $\sqrt{T}$ regret under stronger assumptions It is natural to ask whether the $\Omega(T^{2/3})$ lower bound can be circumvented by suitably strengthening the multi-step revealing condition (yet still weaker than single-step revealing). Based on our intuitions above, a possible direction is to additionally require that *all* action sequences (of length $m-1$) must reveal information about the latent state. We leave this as a question for future work.

5. Proof overview

We now provide a technical overview of the hard instance constructions and the lower bound proofs. We present a simplified version of the multi-step revealing hard instance in Appendix F that is used for proving both the PAC and the regret lower bounds (Theorem 5 & 6). For simplicity, we describe our construction in the 2-step case ($m = 2$); a schematic plot of the resulting POMDP is given in Figure 1.

[3] Alternatively, a bandit-style algorithm that does not take revealing actions but instead attempts to identify the optimal policy directly by brute-force trying, which corresponds to the $\sqrt{A^HT}$ term in Theorem 6.

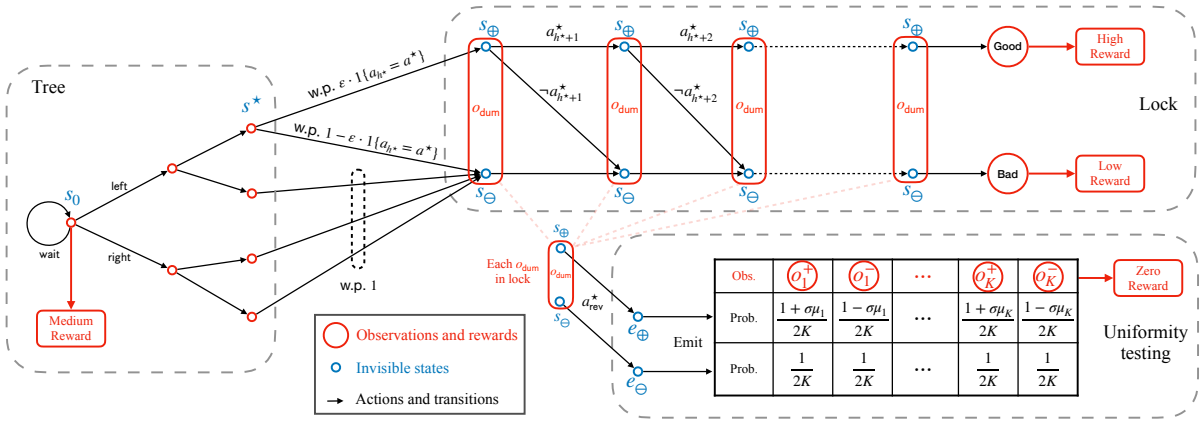

Figure 1. Schematic plot of a simplified version of our hard instance for 2-step revealing POMDPs. The instance consists of three components: tree, lock, and uniformity testing. In the tree, all transitions are deterministic and fully observable, and the agent fully controls how to transit from s_0 to a leaf node. The tree transits stochastically to the lock if any action is taken at any leaf of the tree, but there is a unique (unknown) state $s^\star$, step $h^\star$, and action $a^\star$ at which the agent to transit to $s_\oplus$ with positive probability. In the lock, the agent cannot observe the latent states $\{s_\oplus, s_\ominus\}$, and they need to enter the correct password $\mathbf{a}^\star$ to stay at $s_\oplus$ to eventually receive a high reward. The agent may also take the revealing action $a^\star_{\mathsf{rev}}$ at any o_{dum} to transit to the uniformity testing component, in which they will receive an observation that slightly reveals whether the previous latent state is $s_\oplus$ or $s_\ominus$. See Section 5.1 for a more detailed description.

5.1. Construction of hard instance

A main challenge for obtaining our lower bounds—compared with existing lower bounds in fully observable settings—is to characterize the difficulty of partial observability, i.e. the dependence on O and α^{-1}.

2-step revealing combination lock To reflect this difficulty, the basic component we design is a "2-step revealing combination lock" (cf. the "Lock" part in Figure 1), which is a modification of the non-revealing combination lock of Liu et al. (2022a); Jin et al. (2020a). This lock consists of two hidden states $s_\oplus, s_\ominus$ and an (unknown) sequence of "correct" actions (i.e. the "password") $\mathbf{a}^\star_{h^\star+1:H}$. The only way to stay at $s_\oplus$ is to take the correct action $\mathbf{a}^\star_h$ at each step h, and only state $s_\oplus$ at step H gives a high reward. Therefore, the task of learning the optimal policy is equivalent to identifying the correct action $\mathbf{a}^\star_h$ at each step. We make the hidden states $s_\oplus, s_\ominus$ non-observable (emit dummy observations o_{dum}), so that a naive strategy for the agent is to guess the sequence $\mathbf{a}^\star$ from scratch, which incurs an $\exp(\Omega(H))$ sample complexity.

A central ingredient of our design is a unique (known) *revealing action* $a^\star_{\mathsf{rev}}$ at each step that is always distinct from the correct action. Taking $a^\star_{\mathsf{rev}}$ will transit from latent state $s_\oplus$ to $e_\oplus$ which then emits an observation from distribution $\mu_\oplus \in \Delta(\mathcal{O})$, and similarly from $s_\ominus$ to $e_\ominus$ which then emits an observation from distribution $\mu_\ominus \in \Delta(\mathcal{O})$. After this (single) emission, the system deterministically transits to an absorbing terminal state with reward 0.

Uniformity testing We adapt techniques from the uniformity testing (Canonne, 2020; 2022) literature to pick $\{\mu_\oplus, \mu_\ominus\}$ that are as hard to distinguish as possible, yet ensuring that the POMDP still satisfies the α-revealing condition. Concretely, picking $\mu_\ominus = \mathrm{Unif}(\mathcal{O})$ to be the uniform distribution over $\mathcal{O}$[4], it is known that testing $\mu_\ominus$ from a nearby $\mu_\oplus$ with $D_{\mathrm{TV}}(\mu_\oplus, \mu_\ominus) \asymp \sigma$ requires $\Theta(\sqrt{O}/\sigma^2)$ samples (Paninski, 2008). Further, the worst-case prior for $\mu_\oplus$ takes form $\mu_\oplus = \mathrm{Unif}(\mathcal{O}) + \sigma\mu/O$, where $\mu \sim \mathrm{Unif}(\{(+1,-1),(-1,+1)\}^{O/2})$. We adopt such choices of $\mu_\ominus$ and $\mu_\oplus$ in our hard instance (cf. the "Uniformity testing" part in Figure 1), which can also ensure that the POMDP is $\Theta(\sigma^{-1})$-revealing.

Tree MDP; rewards To additionally exhibit an HSA factor in the lower bound, we further embed a fully observable *tree MDP* (Domingues et al., 2021) before the combination lock. The tree is a balanced binary tree with S leaf nodes, with deterministic transitions (so that which leaf node to arrive at is fully determined by the action sequence) and full observability. All leaf nodes of the tree will transit to the combination lock (i.e. one of $\{s_\oplus, s_\ominus\}$). However, there exists a unique $(h^\star, s^\star, \mathbf{a}^\star)$ such that only taking $a_{h^\star} = a^\star$ at $s_{h^\star} = s^\star$ and step $h^\star$ has a probability ε of transiting to $s_\oplus$; all other choices at leaf nodes transit to $s_\ominus$ with probability one (cf. the "Tree" part in Figure 1).

We further design the reward function so that the agent must identify the underlying parameters $(h^\star, s^\star, a^\star)$ correctly to

[4]Technically, we pick $\mu_\oplus, \mu_\ominus$ to be uniformity testing hard instances on *subset* of $\mathcal{O}$ with size $2K = \Theta(O)$. Here we use the full set $\mathcal{O}$ for simplicity of presentation.

learn a $\Theta(\varepsilon)$ near-optimal policy.

5.2. Calculation of lower bound

Base on our construction, to learn an ε near-optimal policy in this hard instance, the agent has to identify $(h^\star, s^\star, a^\star)$, which can only be achieved by trying all "entrances" (s, a, h) and testing between

$$H_0 : \mathbb{P}(s_{h+1} = s_\oplus | s_h = s, a_h = a) = 0,$$
$$H_1 : \mathbb{P}(s_{h+1} = s_\oplus | s_h = s, a_h = a) = \varepsilon.$$

for each entrance. As we have illustrated, to achieve this, the agent has to either (1) guess the password $\mathbf{a}^\star$ from scratch (using $\Omega\big(A^{H-h}/\varepsilon^2\big)$ samples), or (2) take $a^\star_{\mathsf{rev}}$ and perform uniformity testing using the observations. The latter task turns out to be equivalent to testing between

$$H_0' = \mu_\ominus, \quad H_1' = \varepsilon\mu_\oplus + (1-\varepsilon)\mu_\ominus,$$

where $\mu_\ominus$ is the uniform distribution over $2K = \Theta(O)$ elements, and $\mu_\oplus$ is drawn from the worst-case prior for uniformity testing. Distinguishing between H_0' and H_1' is a uniformity testing task with parameter $\sigma\varepsilon$, which requires $n \geqslant \Omega(\sqrt{O}/(\varepsilon\sigma)^2)$ samples (Paninski, 2008).

With careful information-theoretic arguments, the arguments above will result in a PAC lower bound

$$\Theta(SAH) \times \Omega\Big(\min\Big\{\frac{\sqrt{O}}{\sigma^2\varepsilon^2}, \frac{A^{\Theta(H)}}{\varepsilon^2}\Big\}\Big),$$

for learning 2-step $\Theta(\sigma^{-1})$-revealing POMDPs. This rate is similar as (though slightly worse than) our actual PAC lower bound (Theorem 5). The same hard instance further yields a $\Omega(T^{2/3})$ regret lower bound (though slightly worse rate than Theorem 6); see a calculation in Appendix F.8.

We remark that the above calculations are heuristic; rigorizing these arguments relies on information-theoretic arguments—in our case Ingster's method (Ingster & Suslina, 2012) (cf. Appendix D & Lemma E.5 as an example)—for bounding the divergences between distributions induced by an arbitrary algorithm on different hard instances.

5.3. Remark on actual constructions

The above 2-step hard instance is a simplification of the actual ones used in the proofs of Theorem 5 & 6 in several aspects. The actual constructions are slightly more sophisticated, with the following additional ingredients:

- For the m-step case, to obtain a lower bound that scales with A^m, we modify the construction above so that the agent can take $a^\star_{\mathsf{rev}}$ only once per $(m-1)$-steps, and replace $a^\star_{\mathsf{rev}}$ by a set $|\mathcal{A}_{\mathsf{rev}}| = \Theta(A)$ of revealing actions, which collectively lead to an $A^{m-1} \times A = A^m$ factor.
- We further obtain an extra $\sqrt{S}$ factor in Theorem 5 by replacing the single combination lock with $\Theta(S)$ parallel locks that *share the same password* but *differ in their emission probabilities*. We show that learning in this setting is least as hard as uniformity testing over $\Theta(SO)$ elements, which leads to the extra $\sqrt{S}$ factor.

6. Discussions

6.1. Regret for single-step case

As we have discussed, single-step revealing POMDPs cannot possibly admit a $\Omega(T^{2/3})$ regret lower bound like the multi-step case, as a $\widetilde{\mathcal{O}}(\sqrt{T})$ upper bound is achievable. Nevertheless, we obtain a matching $\Omega(\sqrt{T})$ regret lower bound by a direct reduction from the PAC lower bound (Theorem 4) using Markov's inequality and standard online-to-batch conversion, which we state as follows.

Corollary 7 (Regret lower bound for single-step revealing POMDPs)**.** *Under the same setting as Theorem 4, the same family $\mathcal{M}$ of single-step α-revealing POMDPs there satisfy that for any algorithm $\mathfrak{A}$,*

$$\max_{M\in\mathcal{M}} \mathbb{E}^{\mathfrak{A}}_M[\mathbf{Regret}] \geqslant c_0 \cdot \min\Big\{\sqrt{\frac{SO^{1/2}AH}{\alpha^2}T}, \sqrt{SA^{H/2}HT}, T\Big\}, \tag{5}$$

where $c_0 > 0$ is an absolute constant.

To contrast this lower bound, the current best regret upper bound for single-step revealing POMDPs is $\widetilde{\mathcal{O}}(\sqrt{S^3O^3A^2(1+SA/O)\alpha^{-4}\cdot T} \times \mathrm{poly}(H))$ (Liu et al., 2022b)[5], which is at least a $\sqrt{S^2O^{2.5}A\alpha^{-2}}$-factor larger than the main term in (5). Here we present a much sharper regret upper bound, reducing this gap to $\sqrt{SO^{1.5}}$ and importantly settling the dependence on α.

Theorem 8 (Regret upper bound for single-step revealing POMDPs)**.** *There exists algorithms (*OMLE*,* E2D-TA*, and* MOPS*) that can interact with any single-step α-revealing POMDP M and achieve regret*

$$\mathbf{Regret} \leqslant \widetilde{\mathcal{O}}\Big(\sqrt{\tfrac{S^2O^2A(1+SA/O)H^3}{\alpha^2}\cdot T}\Big) \tag{6}$$

with high probability.

We establish Theorem 8 on a broader class of sequential decision problems termed as *strongly B-stable PSRs* (cf. Appendix H.1), which include single-step revealing POMDPs as a special case. The proof is largely parallel to the analysis of PAC learning for B-stable PSRs (Chen et al., 2022a), and can be found in Appendix H.

[5] Converted from their result whose revealing constant is defined in $(2 \to 2)$-norm.

6.2. Implications on the DEC approach

The Decision-Estimation Coefficient (DEC) (Foster et al., 2021) offers another potential approach for establishing sample complexity lower bounds for any general RL problem. However, here we demonstrate that for revealing POMDPs, any lower bound given by the DEC will necessarily be strictly weaker than our lower bounds.

For example, for PAC learning, the Explorative DEC (EDEC) of m-step revealing POMDPs is known to admit an *upper bound* $\mathrm{edec}_\gamma \leqslant \widetilde{\mathcal{O}}(SA^mH^2\alpha^{-2}/\gamma)$ (Chen et al. (2022a); see also Proposition C.2), and consequently any PAC lower bound obtained by *lower bounding* the EDEC is at most $\Omega(SA^mH^2\alpha^{-2}/\varepsilon^2)$ (Chen et al., 2022b). Such a lower bound would be necessarily smaller than our Theorem 5 by at least a factor of $\sqrt{O}(1 \vee \sqrt{S}/A)$, and importantly does not scale polynomially in O.

Our lower bounds have additional interesting implications on the DEC theory in that, while algorithms such as the E2D achieve sample complexity upper bounds in terms of the DEC and log covering number for the *model class* (Foster et al., 2021; Chen et al., 2022b), without further assumptions, this log covering number cannot be replaced by that of either the *value class* or the *policy class*, giving negative answers to the corresponding questions left open in Foster et al. (2021) (cf. Appendix I.1 for a detailed discussion).

6.3. Towards closing the gaps

Finally, as an important open question, our lower bounds still have mild gaps from the current best upper bounds, importantly in the (S, O) dependence. For example, for multi-step revealing POMDPs, the (first term in the) PAC lower bound $\Omega(S^{1.5}\sqrt{O}A^{m-1}/(\alpha^2\varepsilon^2))$ (Theorem 5) still has a $\sqrt{SO}A$ gap from the upper bound (Theorem 3). While we believe that the A factor is an analysis artifact that may be removed, the remaining $\sqrt{SO}$ factor *cannot* be obtained in the lower bound if we stick to the current family of hard instances—There exists an algorithm *specially tailored to this family* that achieves an $\widetilde{\mathcal{O}}(S^{1.5}\sqrt{O}A^m/(\alpha^2\varepsilon^2))$ upper bound, by brute-force enumeration in the tree and uniformity testing in the combination lock (Appendix I.2).

Closing this $\sqrt{SO}$ gap may require either stronger lower bounds with alternative hard instances—e.g. by embedding other problems in distribution testing (Canonne, 2020)—or sharper upper bounds, which we leave as future work.

7. Conclusion

This paper establishes sample complexity lower bounds for partially observable reinforcement learning in the important tractable class of revealing POMDPs. Our lower bounds are the first to scale polynomially in the number of states, actions, observations, and the revealing constant in a multiplicative fashion, and suggest rather mild gaps between the lower bounds and current best upper bounds. Our work provides a strong foundation for future fine-grained studies and opens up many interesting questions, such as closing the gaps (from either side), or strengthening the multi-step revealing assumption meaningfully to allow a $\sqrt{T}$ regret.

Acknowledgement

S.M. is supported in part by NSF DMS 2210827 and NSF CCF 2315725.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Cai, Q., Yang, Z., and Wang, Z. Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency. In *International Conference on Machine Learning*, pp. 2485–2522. PMLR, 2022.

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

Canonne, C. L. Topics and techniques in distribution testing. 2022.

Chen, F., Bai, Y., and Mei, S. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*, 2022a.

Chen, F., Mei, S., and Bai, Y. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022b.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Diakonikolas, I., Kane, D. M., and Nikishkin, V. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1841–1854. SIAM, 2014.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S. Provable reinforcement learning with a short-term memory. *arXiv preprint arXiv:2202.03983*, 2022.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Reinforcement learning in pomdps without resets. 2005.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

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

Golowich, N., Moitra, A., and Rohatgi, D. Learning in observable pomdps, without computationally intractable oracles. *arXiv preprint arXiv:2206.03446*, 2022a.

Golowich, N., Moitra, A., and Rohatgi, D. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022b.

Ingster, Y. and Suslina, I. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169. Springer Science & Business Media, 2012.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL http://jmlr.org/papers/v11/jaksch10a.html.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Kakade, S., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Kearns, M., Mansour, Y., and Ng, A. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34: 24523–24534, 2021.

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

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022a.

Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*, 2022b.

Liu, Q., Szepesvári, C., and Jin, C. Sample-efficient reinforcement learning of partially observable markov games. *arXiv preprint arXiv:2206.01315*, 2022c.

Mossel, E. and Roch, S. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 366–375, 2005.

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

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Uehara, M., Sekhari, A., Lee, J. D., Kallus, N., and Sun, W. Computationally efficient pac rl in pomdps with latent determinism and conditional embeddings. *arXiv preprint arXiv:2206.12081*, 2022a.

Uehara, M., Sekhari, A., Lee, J. D., Kallus, N., and Sun, W. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*, 2022b.

Van de Geer, S. A. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Embed to control partially observed systems: Representation learning with provable sample efficiency. *arXiv preprint arXiv:2205.13476*, 2022.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2022.

Contents

A. Technical tools

Lemma A.1. *For positive real numbers $A, B, T, \varepsilon_0 > 0$, it holds that*

$$\sup_{\varepsilon\in(0,\varepsilon_0]}\left(\varepsilon T \wedge \frac{A}{\varepsilon^2} \wedge \frac{B}{\varepsilon}\right) \geqslant A^{1/3}T^{2/3} \wedge \sqrt{BT} \wedge \varepsilon_0 T.$$

Proof of Lemma A.1. Suppose that $R > 0$ is such that $R \geqslant \varepsilon T \wedge \frac{A}{\varepsilon^2} \wedge \frac{B}{\varepsilon}$ for all $\varepsilon \in (0, \varepsilon_0]$. Then for each $\varepsilon \in (0, \varepsilon_0]$,

either $\varepsilon \leqslant \frac{R}{T}$, or $\varepsilon \geqslant \sqrt{\frac{A}{R}}$, or $\varepsilon \geqslant \frac{B}{R}$. Thus,

$$(0, \varepsilon_0] \subseteq (0, \frac{R}{T}] \cup [\sqrt{\frac{A}{R}}, +\infty) \cup [\frac{B}{R}, +\infty).$$

Therefore, either $\frac{R}{T} \geqslant \varepsilon_0$, or $\sqrt{\frac{A}{R}} \leqslant \frac{R}{T}$, or $\frac{B}{R} \leqslant \frac{R}{T}$. Combining these three cases together, we obtain

$$R \geqslant \varepsilon_0 T \wedge A^{1/3}T^{2/3} \wedge \sqrt{BT}.$$

□

Lemma A.2. *Suppose that $(R_t)_{t\geqslant 1}$ is a sequence of positive random variables adapted to filtration $(\mathcal{F}_t)_{t\geqslant 1}$ and T is a stopping time (i.e. for $t \geqslant 1$, R_t is $\mathcal{F}_t$-measurable and the event $\{\mathsf{T} \leqslant t\} \in \mathcal{F}_t$). Then it holds that*

$$\mathbb{E}\left[\prod_{t=1}^{\mathsf{T}} R_t \times \prod_{t=1}^{\mathsf{T}} \mathbb{E}[R_t|\mathcal{F}_{t-1}]^{-1}\right] = 1.$$

Equivalently,

$$\mathbb{E}\left[\prod_{t=1}^{\mathsf{T}} R_t \times \exp\left(-\sum_{t=1}^{\mathsf{T}} \log \mathbb{E}[R_t|\mathcal{F}_{t-1}]\right)\right] = 1.$$

Lemma A.2 follows immediately from iteratively applications of the tower properties.

Lemma A.3. *Suppose that random variable X is σ-sub-Gaussian, i.e. $\mathbb{E}[\exp(tX)] \leqslant \exp\left(\frac{\sigma^2 t^2}{2}\right)$ for any $t \in \mathbb{R}$. Then for all $t \geqslant 0$, we have*

$$\mathbb{E}[\exp(t\,|X|)] \leqslant \exp\left(\max\left\{\sigma^2 t^2, \frac{4}{3}\sigma t\right\}\right).$$

Proof of Lemma A.3. For any $x \geqslant 1$, we have

$$\mathbb{E}[\exp(t\,|X|)] \leqslant \mathbb{E}[\exp(xt\,|X|)]^{\frac{1}{x}} \leqslant (\mathbb{E}[\exp(xtX)] + \mathbb{E}[\exp(-xtX)])^{\frac{1}{x}} \leqslant 2^{\frac{1}{x}} \exp\left(\frac{\sigma^2 t^2 x}{2}\right) = \exp\left(\frac{\sigma^2 t^2 x}{2} + \frac{\log 2}{x}\right).$$

We consider two cases: 1. If $\sigma t \geqslant \sqrt{2\log 2}$, then by taking $x = 1$ in the above inequality, we have $\mathbb{E}[\exp(t\,|X|)] \leqslant \exp(\sigma^2 t^2)$. 2. If $\sigma t < \sqrt{2\log 2}$, then by taking $x = \frac{\sqrt{2\log 2}}{\sigma t} > 1$ in the above inequality, we have $\mathbb{E}[\exp(t\,|X|)] \leqslant \exp(\sqrt{2\log 2}\sigma t) \leqslant \exp(\frac{4}{3}\sigma t)$. Combining these two cases completes the proof. □

For probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $(\mathcal{X}, \mathcal{F})$ with a base measure μ, we define the TV distance and the Hellinger distance between $\mathbb{P}, \mathbb{Q}$ as

$$D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A\in\mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2}\int_{\mathcal{X}} \left|\frac{d\mathbb{P}}{d\mu}(x) - \frac{d\mathbb{Q}}{d\mu}(x)\right| d\mu(x),$$
$$D_{\mathrm{H}}^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \left(\sqrt{\frac{d\mathbb{P}}{d\mu}} - \sqrt{\frac{d\mathbb{Q}}{d\mu}}\right)^2 d\mu.$$

When $\mathbb{P} \ll \mathbb{Q}$, we can also define the KL-divergence and the χ^2-divergence between $\mathbb{P}, \mathbb{Q}$ as

$$\mathrm{KL}(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right], \qquad \chi^2(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^2\right] - 1.$$

Lemma A.4. *Suppose $\mathbb{P}, \mathbb{Q}, \mathbb{P}', \mathbb{Q}'$ are four probability measures on $(\mathcal{X}, \mathcal{F})$, and Ω is an event such that $\mathbb{P}|_\Omega = \mathbb{P}'|_\Omega$, $\mathbb{Q}|_\Omega = \mathbb{Q}'|_\Omega$. Then it holds that*

$$D_{\mathrm{TV}}\left(\mathbb{P}', \mathbb{Q}'\right) \geqslant D_{\mathrm{TV}}\left(\mathbb{P}, \mathbb{Q}\right) - \mathbb{P}(\Omega^c).$$

Proof of Lemma A.4. Let μ be a base measure on $(\mathcal{X}, \mathcal{F})$ such that $\mathbb{P}, \mathbb{P}', \mathbb{Q}, \mathbb{Q}'$ have densities with respect to μ (for example, $\mu = (\mathbb{P} + \mathbb{P}' + \mathbb{Q} + \mathbb{Q}')/4$). For notation simplicity, we use $\mathbb{P}(x)$ to stand for $d\mathbb{P}(x)/d\mu(x)$ and use dx to stand for $\mu(dx)$. Then we have

$$\begin{aligned}
2D_{\mathrm{TV}}\left(\mathbb{P}', \mathbb{Q}'\right) =& \int_{\mathcal{X}} \left|\mathbb{P}'(x) - \mathbb{Q}'(x)\right| dx = \int_{\Omega} \left|\mathbb{P}'(x) - \mathbb{Q}'(x)\right| dx + \int_{\Omega^c} \left|\mathbb{P}'(x) - \mathbb{Q}'(x)\right| dx \\
\geqslant& \int_{\Omega} \left|\mathbb{P}'(x) - \mathbb{Q}'(x)\right| dx + \left|\mathbb{P}'(\Omega^c) - \mathbb{Q}'(\Omega^c)\right| \\
=& \int_{\Omega} \left|\mathbb{P}(x) - \mathbb{Q}(x)\right| dx + \left|\mathbb{P}(\Omega^c) - \mathbb{Q}(\Omega^c)\right| \\
\geqslant& \int_{\Omega} \left|\mathbb{P}(x) - \mathbb{Q}(x)\right| dx + \mathbb{P}(\Omega^c) + \mathbb{Q}(\Omega^c) - 2\mathbb{P}(\Omega^c) \\
\geqslant& \int_{\Omega} \left|\mathbb{P}(x) - \mathbb{Q}(x)\right| dx + \int_{\Omega^c} \left|\mathbb{P}(x) - \mathbb{Q}(x)\right| dx - 2\mathbb{P}(\Omega^c) \\
=& 2D_{\mathrm{TV}}\left(\mathbb{P}, \mathbb{Q}\right) - 2\mathbb{P}(\Omega^c).
\end{aligned}$$

This completes the proof. □

Lemma A.5 (Divergence inequalities, see e.g. Sason & Verdú (2016))**.** *For two probability measures $\mathbb{P}, \mathbb{Q}$ on $(\mathcal{X}, \mathcal{F})$, it holds that*

$$2D_{\mathrm{TV}}\left(\mathbb{P}, \mathbb{Q}\right)^2 \leqslant \mathrm{KL}(\mathbb{P} \,\|\, \mathbb{Q}) \leqslant \log\left(1 + \chi^2(\mathbb{P} \,\|\, \mathbb{Q})\right).$$

Lemma A.6 (Hellinger conditioning lemma, see e.g. Chen et al. (2022a, Lemma A.1))**.** *For any pair of random variables (X, Y), it holds that*

$$\mathbb{E}_{X \sim \mathbb{P}_X}\left[D_{\mathrm{H}}^2\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right] \leqslant 2D_{\mathrm{H}}^2\left(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}\right).$$

B. Basics of predictive state representations and B-stability

The following notations for predictive state representations (PSRs) and the B-stability condition are extracted from (Chen et al., 2022a).

Sequential decision processes with observations An episodic sequential decision process is specified by a tuple $\left\{H, \mathcal{O}, \mathcal{A}, \mathbb{P}, \{r_h\}_{h\in[H]}\right\}$, where $H \in \mathbb{Z}_{\geqslant 1}$ is the horizon length; $\mathcal{O}$ is the observation space; $\mathcal{A}$ is the action space; $\mathbb{P}$ specifies the transition dynamics, such that the initial observation follows $o_1 \sim \mathbb{P}_0(\cdot) \in \Delta(\mathcal{O})$, and given the *history* $\tau_h := (o_1, a_1, \cdots, o_h, a_h)$ up to step h, the observation follows $o_{h+1} \sim \mathbb{P}(\cdot|\tau_h)$; $r_h : \mathcal{O} \times \mathcal{A} \to [0, 1]$ is the reward function at h-th step, which we assume is a known deterministic function of (o_h, a_h).

In an episodic sequential decision process, a policy $\pi = \{\pi_h : (\mathcal{O}\times\mathcal{A})^{h-1}\times\mathcal{O} \to \Delta(\mathcal{A})\}_{h\in[H]}$ is a collection of H functions. At step $h \in [H]$, an agent running policy π observes the observation o_h and takes action $a_h \sim \pi_h(\cdot|\tau_{h-1}, o_h) \in \Delta(\mathcal{A})$ based on the history $(\tau_{h-1}, o_h) = (o_1, a_1, \ldots, o_{h-1}, a_{h-1}, o_h)$. The agent then receives their reward $r_h(o_h, a_h)$, and the environment generates the next observation $o_{h+1} \sim \mathbb{P}(\cdot|\tau_h)$ based on $\tau_h = (o_1, a_1, \cdots, o_h, a_h)$ (if $h < H$). The episode terminates immediately after a_H is taken.

For any $\tau_h = (o_1, a_1, \cdots, o_h, a_h)$, we write

$$\begin{aligned}
\mathbb{P}(\tau_h) :=& \mathbb{P}(o_{1:h}|a_{1:h}) = \prod_{h' \leqslant h} \mathbb{P}(o_{h'}|\tau_{h'-1}), \\
\pi(\tau_h) :=& \prod_{h' \leqslant h} \pi_{h'}(a_{h'}|\tau_{h'-1}, o_{h'}), \\
\mathbb{P}^\pi(\tau_h) :=& \mathbb{P}(\tau_h) \times \pi(\tau_h).
\end{aligned}$$

Then $\mathbb{P}^\pi(\tau_h)$ is the probability of observing τ_h (for the first h steps) when executing π.

PSR, core test sets, and predictive states A *test* t is a sequence of future observations and actions (i.e. $t \in \mathfrak{T} := \bigcup_{W\in\mathbb{Z}_{\geqslant 1}} \mathcal{O}^W \times \mathcal{A}^{W-1}$). For some test $t_h = (o_{h:h+W-1}, a_{h:h+W-2})$ with length $W \geqslant 1$, we define the probability of test t_h being successful conditioned on (reachable) history τ_{h-1} as $\mathbb{P}(t_h|\tau_{h-1}) := \mathbb{P}(o_{h:h+W-1}|\tau_{h-1}; \mathrm{do}(a_{h:h+W-2}))$, i.e., the probability of observing $o_{h:h+W-1}$ if the agent deterministically executes actions $a_{h:h+W-2}$, conditioned on history τ_{h-1}. We follow the convention that, if $\mathbb{P}^\pi(\tau_{h-1}) = 0$ for any π, then $\mathbb{P}(t|\tau_{h-1}) = 0$.

Definition B.1 (PSR, core test sets, and predictive states)**.** *For any $h \in [H]$, we say a set $\mathcal{U}_h \subset \mathfrak{T}$ is a* core test set *at step h if the following holds: For any $W \in \mathbb{Z}_{\geqslant 1}$, any possible future (i.e., test) $t_h = (o_{h:h+W-1}, a_{h:h+W-2}) \in \mathcal{O}^W \times \mathcal{A}^{W-1}$, there exists a vector $b_{t_h,h} \in \mathbb{R}^{\mathcal{U}_h}$ such that*

$$\mathbb{P}(t_h|\tau_{h-1}) = \langle b_{t_h,h}, [\mathbb{P}(t|\tau_{h-1})]_{t\in\mathcal{U}_h}\rangle, \qquad \forall \tau_{h-1} \in \mathcal{T}^{h-1} := (\mathcal{O}\times\mathcal{A})^{h-1}. \tag{7}$$

We refer to the vector $\mathbf{q}(\tau_{h-1}) := [\mathbb{P}(t|\tau_{h-1})]_{t\in\mathcal{U}_h}$ as the predictive state *at step h (with convention $\mathbf{q}(\tau_{h-1}) = 0$ if τ_{h-1} is not reachable), and $\mathbf{q}_0 := [\mathbb{P}(t)]_{t\in\mathcal{U}_1}$ as the initial predictive state. A (linear) PSR is a sequential decision process equipped with a core test set $\{\mathcal{U}_h\}_{h\in[H]}$.*

Define $\mathcal{U}_{A,h} := \{\mathbf{a} : (\mathbf{o},\mathbf{a}) \in \mathcal{U}_h \text{ for some } \mathbf{o} \in \bigcup_{W\in\mathbb{N}^+} \mathcal{O}^W\}$ as the set of "core actions" (possibly including an empty sequence) in $\mathcal{U}_h$, with $U_A := \max_{h\in[H]} |\mathcal{U}_{A,h}|$. Further define $\mathcal{U}_{H+1} := \{o_{\mathrm{dum}}\}$ for notational simplicity. The core test sets $(\mathcal{U}_h)_{h\in[H]}$ are assumed to be known and the same within a PSR model class.

Definition B.2 (PSR rank)**.** *Given a PSR, its* PSR rank *is defined as $d_{\mathsf{PSR}} := \max_{h\in[H]} \mathrm{rank}(D_h)$, where $D_h := [\mathbf{q}(\tau_h)]_{\tau_h\in\mathcal{T}^h} \in \mathbb{R}^{\mathcal{U}_{h+1}\times\mathcal{T}^h}$ is the matrix formed by predictive states at step $h \in [H]$.*

For POMDP, it is clear that $d_{\mathsf{PSR}} \leqslant S$, regardless of the core test sets.

B-representation (Chen et al., 2022a) introduced the notion of *B-representation* of PSR, which plays a fundamental role in their general structural condition and their analysis.

Definition B.3 (B-representation)**.** *A B-representation of a PSR with core test set $(\mathcal{U}_h)_{h\in[H]}$ is a set of matrices $\{(\mathbf{B}_h(o_h,a_h) \in \mathbb{R}^{\mathcal{U}_{h+1}\times\mathcal{U}_h})_{h,o_h,a_h}, \mathbf{q}_0 \in \mathbb{R}^{\mathcal{U}_1}\}$ such that for any $0 \leqslant h \leqslant H$, policy π, history $\tau_h = (o_{1:h}, a_{1:h}) \in \mathcal{T}^h$, and core test $t_{h+1} = (o_{h+1:h+W}, a_{h+1:h+W-1}) \in \mathcal{U}_{h+1}$, the quantity $\mathbb{P}(\tau_h, t_{h+1})$, i.e. the probability of observing $o_{1:h+W}$ upon taking actions $a_{1:h+W-1}$, admits the decomposition*

$$\mathbb{P}(\tau_h, t_{h+1}) = \mathbb{P}(o_{1:h+W}|\mathrm{do}(a_{1:h+W-1})) = \mathbf{e}_{t_{h+1}}^\top \cdot \mathbf{B}_{h:1}(\tau_h)\cdot\mathbf{q}_0, \tag{8}$$

where $\mathbf{e}_{t_{h+1}} \in \mathbb{R}^{\mathcal{U}_{h+1}}$ is the indicator vector of $t_{h+1} \in \mathcal{U}_{h+1}$, and

$$\mathbf{B}_{h:1}(\tau_h) := \mathbf{B}_h(o_h,a_h)\mathbf{B}_{h-1}(o_{h-1},a_{h-1})\cdots\mathbf{B}_1(o_1,a_1).$$

Based on the B-representations of PSRs, (Chen et al., 2022a) proposed the following structural condition for sample-efficient learning in PSRs.

Definition B.4 (B-stability (Chen et al., 2022a))**.** *A PSR is* B-stable with parameter $\Lambda_{\mathsf{B}} \geqslant 1$ *(henceforth also Λ_{B}-stable) if it admits a B-representation such that for all step $h \in [H]$, policy π, and $x \in \mathbb{R}^{\mathcal{U}_h}$, we have*

$$\sum_{\tau_{h:H}=(o_h,a_h,\cdots,o_H,a_H)} \pi(\tau_{h:H}) \times |\mathbf{B}_H(o_H,a_H)\cdots\mathbf{B}_h(o_h,a_h)x| \leqslant \Lambda_{\mathsf{B}} \max\left\{\|x\|_*, \|x\|_{\Pi'}\right\}, \tag{9}$$

where for any vector $x = (x(t))_{t\in\mathcal{U}_h}$, we denote its $(1,2)$-norm by

$$\|x\|_* := \left(\sum_{\mathbf{a}\in\mathcal{U}_{A,h}} \left(\sum_{\mathbf{o}:(\mathbf{o},\mathbf{a})\in\mathcal{U}_h} |x(\mathbf{o},\mathbf{a})|\right)^2\right)^{1/2},$$

and its Π'-norm by

$$\|x\|_{\Pi'} := \max_{\bar\pi} \sum_{t\in\overline{\mathcal{U}}_h} \bar\pi(t)\,|x(t)|\,,$$

where $\overline{\mathcal{U}}_h := \{t \in \mathcal{U}_h : \nexists t' \in \mathcal{U}_h \text{ such that } t \text{ is a prefix of } t'\}$.

Equivalently, (9) can be written as $\|\mathcal{B}x\|_\Pi \leqslant \Lambda_{\mathsf{B}} \max\left\{\|x\|_*, \|x\|_{\Pi'}\right\}$, where for each step h, vector $x \in \mathbb{R}^{\mathcal{U}_h}$, we write

$$\|\mathcal{B}_{H:h}x\|_\Pi := \max_\pi \sum_{\tau_{h:H}} \pi(\tau_{h:H}) \times |\mathbf{B}_{H:h}(\tau_{h:H})x|\,. \tag{10}$$

Chen et al. (2022a) showed that B-stability enables sample efficiency of PAC-learning, and we summarize the results in the following theorem.

Theorem B.5 (PAC upper bound for learning PSRs)**.** *Suppose Θ is a PSR class with the same core test sets $\{\mathcal{U}_h\}_{h\in[H]}$, and each $\theta \in \Theta$ admits a B-representation that is Λ_{B}-stable and has PSR rank at most d. Then there exists algorithms (*OMLE/EXPLORATIVE E2D/MOPS*) that can find an ε-optimal policy with probability at least $1-\delta$, within*

$$T \leqslant \widetilde{\mathcal{O}}\left(\frac{\Lambda_{\mathsf{B}}^2 dAU_A H^2 \log(\mathcal{N}_\Theta(1/T)/\delta)}{\varepsilon^2}\right) \tag{11}$$

episodes of play, where $\mathcal{N}_\Theta$ is the covering number of Θ (cf. Chen et al. (2022a, Definition A.4)).

When Θ is a subclass of POMDPs, we have $\log \mathcal{N}_\Theta(1/T) = \widetilde{\mathcal{O}}\big(H(S^2A+SO)\big)$ (Chen et al., 2022a). Therefore, to deduce Theorem 3 from the above general theorem, it remains to upper bound Λ_{B} for m-step α-revealing POMDPs, which is done in Appendix C.2.

C. Proofs for Section 2

C.1. Proof of Proposition 2

Fix any POMDP M, and we first show that $\alpha_{m+1}(M) \geqslant \alpha_m(M)$. By the definition of $\alpha_{m+1}(M)$ (Definition 1), it suffices to show the following result.

Lemma C.1. *For any $h \in [H-m]$, and any choice of generalized left inverse $\mathbb{M}_{h,m}^+$ (of $\mathbb{M}_{h,m}$), the matrix $\mathbb{M}_{h,m+1}$ admits a generalized left inverse $\mathbb{M}_{h,m+1}^+$ such that*

$$\left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to 1} \leqslant \left\|\mathbb{M}_{h,m}^+\right\|_{*\to 1}.$$

The converse part of Proposition 2 can be shown directly by examples. In particular, our construction in Appendix F readily provides such an example (see Remark F.10).

Proof of Lemma C.1. Fix an arbitrary action $\widetilde{a} \in \mathcal{A}$. Consider the matrix $F_{\widetilde{a}} \in \mathbb{R}^{\mathcal{O}^m\mathcal{A}^{m-1}\times\mathcal{O}^{m+1}\mathcal{A}^m}$ defined as (the unique matrix associated with) the following linear operator:

$$[F_{\widetilde{a}}\mathbf{x}](\mathbf{o}_{h:h+m-1}, \mathbf{a}_{h:h+m-2}) := \sum_{o\in\mathcal{O}} \mathbf{x}(\mathbf{o}_{h:h+m-1}o, \mathbf{a}_{h:h+m-2}\widetilde{a}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{O}^{m+1}\mathcal{A}^m}.$$

We first show that $F_{\widetilde{a}}\mathbb{M}_{h,m+1} = \mathbb{M}_{h,m}$. Indeed,

$$\begin{aligned}
&[F_{\widetilde{a}}\mathbb{M}_{h,m+1}]_{\mathbf{o}_{h:h+m-1}\mathbf{a}_{h:h+m-2},s} = \sum_{o\in\mathcal{O}} [\mathbb{M}_{h,m+1}]_{(\mathbf{o}_{h:h+m-1}o)(\mathbf{a}_{h:h+m-2}\widetilde{a}),s} \\
&= \sum_{o\in\mathcal{O}} \mathbb{P}(o_{h:h+m} = \mathbf{o}_{h:h+m-1}o | a_{h:h+m-1} = \mathbf{a}_{h:h+m-2}\widetilde{a}, s_h = s) \\
&= \mathbb{P}(o_{h:h+m-1} = \mathbf{o}_{h:h+m-1} | a_{h:h+m-2} = \mathbf{a}_{h:h+m-2}, s_h = s) = [\mathbb{M}_{h,m}]_{\mathbf{o}_{h:h+m-1}\mathbf{a}_{h:h+m-2},s}
\end{aligned}$$

for any $(\mathbf{o}_{h:h+m-1}\mathbf{a}_{h:h+m-2}, s)$, which verifies the claim. Therefore, for any generalized left inverse $\mathbb{M}_{h,m}^+$, we can take

$$\mathbb{M}_{h,m+1}^+ := \mathbb{M}_{h,m}^+ F_{\widetilde{a}}.$$

This matrix satisfies $\mathbb{M}_{h,m+1}^+\mathbb{M}_{h,m+1}\mathbb{T}_{h-1} = \mathbb{M}_{h,m}^+F_{\widetilde{a}}\mathbb{M}_{h,m+1}\mathbb{T}_{h-1} = \mathbb{M}_{h,m}^+\mathbb{M}_{h,m}\mathbb{T}_{h-1} = \mathbb{T}_{h-1}$ and is thus indeed a generalized left inverse of $\mathbb{M}_{h,m+1}$. Further,

$$\left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to 1} = \left\|\mathbb{M}_{h,m}^+F_{\widetilde{a}}\right\|_{*\to 1} \leqslant \left\|\mathbb{M}_{h,m}^+\right\|_{*\to 1} \|F_{\widetilde{a}}\|_{*\to *},$$

so it remains to show that $\|F_{\widetilde{a}}\|_{*\to *} \leqslant 1$. To see this, note that for any $\mathbf{x} \in \mathbb{R}^{\mathcal{O}^{m+1}\mathcal{A}^m}$ with $\|\mathbf{x}\|_*^2 \leqslant 1$, we have

$$\begin{aligned}
\|F_{\widetilde{a}}\mathbf{x}\|_*^2 &= \sum_{\mathbf{a}_{h:h+m-2}\in\mathcal{A}^{m-1}} \left(\sum_{\mathbf{o}_{h:h+m-1}\in\mathcal{O}^m} \left| \sum_{o\in\mathcal{O}} \mathbf{x}(\mathbf{o}_{h:h+m-1}o, \mathbf{a}_{h:h+m-2}\widetilde{a}) \right| \right)^2 \\
&\leqslant \sum_{\mathbf{a}_{h:h+m-2}\in\mathcal{A}^{m-1}} \left(\sum_{\mathbf{o}_{h:h+m}\in\mathcal{O}^{m+1}} |\mathbf{x}(\mathbf{o}_{h:h+m}, \mathbf{a}_{h:h+m-2}\widetilde{a})| \right)^2 \\
&\leqslant \sum_{\mathbf{a}_{h:h+m-1}\in\mathcal{A}^{m}} \left(\sum_{\mathbf{o}_{h:h+m}\in\mathcal{O}^{m+1}} |\mathbf{x}(\mathbf{o}_{h:h+m}, \mathbf{a}_{h:h+m-1})| \right)^2 = \|\mathbf{x}\|_*^2 .
\end{aligned}$$

This proves $\|F_{\widetilde{a}}\|_{*\to *} \leqslant 1$ and thus the desired result. □

C.2. Proof of Theorem 3

We will deduce Theorem 3 from the general result (Theorem B.5) of learning PSRs (Chen et al., 2022a). To apply Theorem B.5, we first invoke the following proposition, which basically states that any m-step α-revealing POMDP is B-stable with $\Lambda_{\mathsf{B}} \leqslant \alpha^{-1}$.

Proposition C.2. *Any m-step α-revealing POMDP is a α^{-1}-stable PSR with core test set $\mathcal{U}_h = (\mathcal{O}\times\mathcal{A})^{\min\{m-1, H-h\}}\times\mathcal{O}$, i.e. it admits a $\Lambda_{\mathsf{B}} \leqslant \alpha^{-1}$-stable B-representation.*

Therefore, for $\mathcal{M}$ a class of m-step α-revealing POMDPs, $\mathcal{M}$ is also a class of PSRs with common core test sets, such that each $M \in \mathcal{M}$ is α^{-1}-stable, has PSR rank at most S and $U_A = A^{m-1}$. Then, Theorem B.5 implies that an ε-optimal policy of $\mathcal{M}$ can be learned using OMLE, EXPLORATIVE E2D, or MOPS, with sample complexity

$$\widetilde{\mathcal{O}}\left(\frac{SA^m H^2 \log(\mathcal{N}_{\mathcal{M}}(1/T)/\delta)}{\alpha^2\varepsilon^2}\right),$$

and we also have $\log\mathcal{N}_{\mathcal{M}}(1/T) = \widetilde{\mathcal{O}}\big(H(S^2A + SO)\big)$ (Chen et al., 2022a). Combining these facts completes the proof of Theorem 3. □

Proof of Proposition C.2. Chen et al. (2022a, Appendix B.3.3) showed that any m-step α-revealing POMDP M is a α^{-1}-stable PSR with core test set $\mathcal{U}_h = (\mathcal{O}\times\mathcal{A})^{\min\{m-1,H-h\}}\times\mathcal{O}$, and explicitly constructed the following B-representation for it: when $h \leqslant H-m$, set

$$\mathbf{B}_h(o,a) = \mathbb{M}_{h+1}\mathbb{T}_{h,a}\,\mathrm{diag}\,(\mathbb{O}_h(o|\cdot))\mathbb{M}_h^+, \qquad h\in[H-m], \tag{12}$$

and when $h > H-m$, take

$$\mathbf{B}_h(o_h,a_h) = [\mathbb{1}(t_h = (o_h,a_h,t_{h+1}))]_{(t_{h+1},t_h)\in\mathcal{U}_{h+1}\times\mathcal{U}_h} \in \mathbb{R}^{\mathcal{U}_{h+1}\times\mathcal{U}_h}, \tag{13}$$

where $\mathbb{1}(t_h = (o_h,a_h,t_{h+1}))$ is 1 if t_h equals to (o_h,a_h,t_{h+1}), and 0 otherwise.

Then, by Chen et al. (2022a, Lemma B.13), for any $1\leqslant h\leqslant H$, $x\in\mathbb{R}^{|\mathcal{U}_h|}$, it holds that

$$\|\mathcal{B}_{H:h}x\|_\Pi = \max_\pi \sum_{\tau_{h:H}} \|\mathbf{B}_H(o_H,a_H)\cdots\mathbf{B}_h(o_h,a_h)x\|_1 \times \pi(\tau_{h:H}) \leqslant \max\left\{\left\|\mathbb{M}_h^+x\right\|_1, \|x\|_{\Pi'}\right\} \leqslant \alpha^{-1}\max\left\{\|x\|_*, \|x\|_{\Pi'}\right\}.$$

Therefore, B-representation provided in (12) and (13) is indeed α^{-1}-stable, and hence completes the proof. □

D. Basics of Ingster's method

In this section, we first introduce the basic notations frequently used in our analysis of hard instances, and then state Ingster's method for proving information-theoretic lower bounds (Ingster & Suslina, 2012). Recall that we have introduced the formulation of sequential decision process in Appendix B.

Algorithms for sequential decision processes An algorithm $\mathfrak{A}$ for sequential decision processes (with a fixed number of episodes T) is specified by a collection of HT functions $\mathfrak{A} = \{\pi^{\mathfrak{A}}_{t,h}\}_{h\in[H],t\in[T]}$, where $\pi^{\mathfrak{A}}_{t,h}$ maps the tuple of all past histories and the current observation $(\tau^{(1)}, \cdots, \tau^{(t-1)}, \tau^{(t)}_{h-1}, o^{(t)}_h)$ to a distribution over actions $\Delta(\mathcal{A})$ from which we sample the next action $a^{(t)}_h \sim \pi^{\mathfrak{A}}_{t,h}(\cdot|\tau^{(1:t-1)}, \tau^{(t)}_{h-1}, o^{(t)}_h)$. At the end of interaction, the algorithm output a $\pi^{\mathrm{out}} \in \Pi$ by taking $\pi^{\mathrm{out}} = \pi^{\mathfrak{A}}_{\mathsf{output}}(\tau^{1:T})$.

For any algorithm $\mathfrak{A}$ (with a fixed number of episodes T), we write $\mathbb{P}^{\mathfrak{A}}_M$ to be the law of $(\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(T)})$ under the model M and the algorithm $\mathfrak{A}$. We remark that although our formulation seems only to allow deterministic algorithms where each $\pi^{\mathfrak{A}}_{t,h}$ is a deterministic mapping to $\Delta(\mathcal{A})$, our formulation indeed allows randomized algorithms: any randomized algorithm can be written as a mixture of deterministic algorithm $\mathfrak{B}(\omega)$ parameterized by ω which satisfies a distribution $\omega \sim \zeta$; furthermore, for any $\mathfrak{B}(\omega)$ and ζ, there exists a deterministic algorithm $\mathfrak{A}$ such that the marginal laws of $\tau^{1:T}$ induced by $\mathfrak{B}$ and $\mathfrak{A}$ are the same, i.e., $\mathbb{E}_{\omega\sim\zeta}[\mathbb{P}^{\mathfrak{B}(\omega)}_M(\,\cdot\,)] = \mathbb{P}^{\mathfrak{A}}_M(\,\cdot\,)$.

Algorithms with a random stopping time Our analysis requires us to consider algorithms with a random stopping time. An algorithm $\mathfrak{A}$ with a random stopping time (with at most T interaction) is specified by a collection of HT functions $\{\pi^{\mathfrak{A}}_{t,h}\}_{h\in[H],t\in[T]}$ along with an exit criterion exit, where $\pi^{\mathfrak{A}}_{t,h}$ is the strategy at t-th episode and h-th step, and exit is a deterministic function such that

$$\mathsf{exit}(\tau^{(1)}, \cdots, \tau^{(t)}) \in \{\mathsf{TRUE}, \mathsf{FALSE}\}.$$

Once $\mathsf{exit}(\tau^{(1)}, \cdots, \tau^{(\mathsf{T})}) = \mathsf{TRUE}$ or $\mathsf{T} = T$, the algorithm $\mathfrak{A}$ terminates at the end of the T-th episode. The random variable T (induced by the exit criterion exit) is clearly a stopping time. We write $\mathbb{P}^{\mathfrak{A}}_M$ to be the law of $(\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(\mathsf{T})})$ under the model M and the algorithm $\mathfrak{A}$.

The following lemma and discussions hold for algorithms with or without a random stopping time.

Lemma D.1 (Ingster's method)**.** *For a family of sequential decision processes $(\mathbb{P}_M)_{M\in\mathcal{M}}$, a distribution ζ over $\mathcal{M}$, a reference model $0 \in \mathcal{M}$, and an algorithm $\mathfrak{A}$ that interacts with the environment for T episodes (where T is stopping time), it holds that*

$$1 + \chi^2(\mathbb{E}_{M\sim\zeta}\big[\mathbb{P}^{\mathfrak{A}}_M\big] \,\|\, \mathbb{P}^{\mathfrak{A}}_0) = \mathbb{E}_{M,M'\sim_{\mathrm{iid}}\zeta}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}^{\mathfrak{A}}_0}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right].$$

Proof. We only need to consider the case $\mathfrak{A}$ has a random stopping time T. By our definition, $\mathbb{P}^{\mathfrak{A}}_M$ is supported on the following set:

$$\Omega_0 := \Big\{\omega = \tau^{(1:\mathsf{T})} : \forall t < \mathsf{T}, \mathsf{exit}(\tau^{(1:t)}) = \mathsf{FALSE}, \text{ and either } \mathsf{T} = T \text{ or } \mathsf{exit}(\tau^{(1:\mathsf{T})}) = \mathsf{TRUE}\Big\}.$$

For any $(\tau^{(1)}, \cdots, \tau^{(\mathsf{T})}) \in \Omega_0$, we have

$$\begin{aligned}
\mathbb{P}^{\mathfrak{A}}_M(\tau^{(1)}, \cdots, \tau^{(\mathsf{T})}) &= \prod_{t=1}^{\mathsf{T}}\mathbb{P}^{\mathfrak{A}}_M(\tau^{(t)}|\tau^{(1:t-1)}) \\
&= \prod_{t=1}^{\mathsf{T}}\prod_{h=1}^{H}\mathbb{P}_M(o^{(t)}_h|\tau^{(t)}_{1:h}) \times \pi^{\mathfrak{A}}_{t,h}(a^{(t)}_h|\tau^{(1:t-1)}, \tau^{(t)}_{1:h}, o^{(t)}_h) \qquad (14)\\
&= \prod_{t=1}^{\mathsf{T}}\mathbb{P}_M(\tau^{(t)}) \times \prod_{t=1}^{\mathsf{T}}\prod_{h=1}^{H}\pi^{\mathfrak{A}}_{t,h}(a^{(t)}_h|\tau^{(1:t-1)}, \tau^{(t)}_{1:h}, o^{(t)}_h).
\end{aligned}$$

Therefore, by definition of χ^2 divergence, we have

$$1 + \chi^2(\mathbb{E}_{M\sim\zeta}\big[\mathbb{P}^{\mathfrak{A}}_M\big] \,\|\, \mathbb{P}^{\mathfrak{A}}_0) = \mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}^{\mathfrak{A}}_0}\left[\left(\frac{\mathbb{E}_{M\sim\zeta}\big[\mathbb{P}^{\mathfrak{A}}_M(\tau^{(1)}, \cdots, \tau^{(\mathsf{T})})\big]}{\mathbb{P}^{\mathfrak{A}}_0(\tau^{(1)}, \cdots, \tau^{(\mathsf{T})})}\right)^2\right]$$

$$= \mathbb{E}_{M,M'\sim\zeta}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\frac{\mathbb{P}_M^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})\mathbb{P}_{M'}^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})}{\mathbb{P}_0^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})^2}\right]$$
$$= \mathbb{E}_{M,M'\sim\zeta}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right],$$

where the last equality is due to (14). This proves the lemma. □

Therefore, in order to upper bound $\chi^2(\mathbb{E}_{M\sim\zeta}[\mathbb{P}_M^{\mathfrak{A}}] \,\|\, \mathbb{P}_0^{\mathfrak{A}})$, we just need to upper bound the quantity

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right] = \mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\prod_{h=1}^{H}\frac{\mathbb{P}_M(o_h^{(t)}|\tau_{h-1}^{(t)})\mathbb{P}_{M'}(o_h^{(t)}|\tau_{h-1}^{(t)})}{\mathbb{P}_0(o_h^{(t)}|\tau_{h-1}^{(t)})^2}\right]. \tag{15}$$

At this aim, we will leverage the following fact (which is due to Lemma A.2 and (15)):

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\cdot\exp\left(-\sum_{t=1}^{\mathsf{T}}\sum_{h=1}^{H}\log I_{M,M'}(\tau_{h-1}^{(t)})\right)\right] = 1, \tag{16}$$

where $I_{M,M'}(\tau_{h-1})$ is defined as

$$I_{M,M'}(\tau_{h-1}) := \mathbb{E}_0\left[\frac{\mathbb{P}_M(o_h|\tau_{h-1})\mathbb{P}_{M'}(o_h|\tau_{h-1})}{\mathbb{P}_0(o_h|\tau_{h-1})^2}\middle|\tau_{h-1}\right]. \tag{17}$$

Early stopped algorithm Consider an algorithm $\mathfrak{A}$ that interacts with the environment for a fixed number of episodes T and consider an exit criterion exit. We define the early stopped algorithm $\mathfrak{A}(\mathsf{exit})$, which executes the algorithm $\mathfrak{A}$ until $\mathsf{exit} = \mathsf{TRUE}$ is satisfied (or T is reached). Clearly, $\mathfrak{A}(\mathsf{exit})$ is an algorithm with a random stopping time. We have the following lemma regarding how much the TV distance $D_{\mathrm{TV}}\left(\mathbb{E}_{M\sim\zeta}[\mathbb{P}_M^{\mathfrak{A}}], \mathbb{P}_0^{\mathfrak{A}}\right)$ is perturbed after changing the algorithm $\mathfrak{A}$ to its stopped version $\mathfrak{A}(\mathsf{exit})$.

Lemma D.2. *It holds that*

$$D_{\mathrm{TV}}\left(\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}\right], \mathbb{P}_0^{\mathfrak{A}(\mathsf{exit})}\right) \geqslant D_{\mathrm{TV}}\left(\mathbb{E}_{M\sim\zeta}[\mathbb{P}_M^{\mathfrak{A}}], \mathbb{P}_0^{\mathfrak{A}}\right) - \mathbb{P}_0^{\mathfrak{A}}(\exists t < T, \mathsf{exit}(\tau^{(1:t)}) = \mathsf{TRUE}).$$

Proof. We consider the event $\Omega = \{\omega = \tau^{(1:T)} : \forall t < T, \mathsf{exit}(\tau^{(1:t)}) = \mathsf{FALSE}\}$. To prove this lemma, we only need to verify that $\mathbb{P}_M^{\mathfrak{A}}|_\Omega = \mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}|_\Omega$ and then apply Lemma A.4.

Indeed, for $\omega = \tau^{(1:T)} \in \Omega$, we have that for all $t < T$, $\mathsf{exit}(\tau^{(1:t)}) = \mathsf{FALSE}$. Then, by (14) we have

$$\mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}(\tau^{(1:T)}) = \prod_{t=1}^{T}\mathbb{P}_M(\tau^{(t)}) \times \prod_{t=1}^{T}\prod_{h=1}^{H}\pi_{t,h}^{\mathfrak{A}}(a_h^{(t)}|\tau^{(1:t-1)}, \tau_{1:h}^{(t)}, o_h^{(t)}) = \mathbb{P}_M^{\mathfrak{A}}(\tau^{(1:T)}),$$

and thus $\mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}(\omega) = \mathbb{P}_M^{\mathfrak{A}}(\omega)$ for any $\omega \in \Omega$. Applying Lemma A.4 proves the lemma. □

E. Proof of Theorem 4

We first construct a family of hard instances in Appendix E.1. We state the PAC lower bound of this family of hard instances in Proposition E.1. Theorem 4 then follows from Proposition E.1 as a direct corollary.

E.1. Construction of hard instances and proof of Theorem 4

We consider the following family of single-step revealing POMDPs $\mathcal{M}$ that admits a tuple of hyperparameters $(\varepsilon, \sigma, n, K, H)$. All POMDPs in $\mathcal{M}$ have the same horizon length H, the state space $\mathcal{S}$, the action space $\mathcal{A}$, and the observation space $\mathcal{O}$, defined as follows.

- The state space $\mathcal{S} = \mathcal{S}_{\mathsf{tree}} \bigsqcup \{s_\oplus, s_\ominus\}$, where $\mathcal{S}_{\mathsf{tree}}$ is a binary tree with level n (so that $|\mathcal{S}_{\mathsf{tree}}| = 2^n - 1$). Let s_0 be the root of $\mathcal{S}_{\mathsf{tree}}$, and $\mathcal{S}_{\mathsf{leaf}}$ be the set of leaves of $\mathcal{S}_{\mathsf{tree}}$, with $|\mathcal{S}_{\mathsf{leaf}}| = 2^{n-1}$.

- The observation space $\mathcal{O} = \mathcal{S}_{\mathsf{tree}} \bigsqcup \{o_1^+, o_1^-, \cdots, o_K^+, o_K^-\} \bigsqcup \{\mathsf{good}, \mathsf{bad}\}$. Note that here we slightly abuse notations, reusing $\mathcal{S}_{\mathsf{tree}}$ to denote both a set of states and the corresponding set of observations, in the sense that each state $s \in \mathcal{S}_{\mathsf{tree}} \subset \mathcal{S}$ corresponds to a unique observation $o_s \in \mathcal{S}_{\mathsf{tree}} \subset \mathcal{O}$, which we also denote as s when it is clear from the context.
- The action space $\mathcal{A} = \{0, 1, \cdots, A-1\}$.

Model parameters Each non-null POMDP model $M = M_{\theta,\mu} \in \mathcal{M} \backslash \{M_0\}$ is specified by two parameters (θ, μ). Here $\mu \in \{-1, +1\}^K$, and $\theta = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$, where

- $s^\star \in \mathcal{S}_{\mathsf{leaf}}$, $a^\star \in \mathcal{A}_c := \{1, \cdots, A-1\}$.
- $h^\star \in \{n+1, \cdots, H-1\}$.
- $\mathbf{a}^\star = (\mathbf{a}^\star_{h^\star+1}, \ldots, \mathbf{a}^\star_{H-1}) \in \mathcal{A}^{H-h^\star-1}$ is an action sequence indexed by $h^\star+1, \cdots, H-1$.

For any POMDP $M_{\theta,\mu}$, its emmision and transition dynamics $\mathbb{P}_{\theta,\mu} := \mathbb{P}_{M_{\theta,\mu}}$ are defined as follows.

Emission dynamics

- At states $s \in \mathcal{S}_{\mathsf{tree}}$, the agent always receives (the unique observation corresponding to) s itself as the observation.
- At state $s_\oplus$ and steps $h < H$, the emission dynamics is given by

$$\mathbb{O}_{h;\mu}(o_i^+|s_\oplus) = \frac{1+\sigma\mu_i}{2K}, \qquad \mathbb{O}_{h;\mu}(o_i^-|s_\oplus) = \frac{1-\sigma\mu_i}{2K}, \qquad \forall i \in [K].$$

- At state $s_\ominus$ and steps $h < H$, the observation is uniformly drawn from $\mathcal{O}_o := \{o_1^+, o_1^-, \cdots, o_K^+, o_K^-\}$:

$$\mathbb{O}_h(o_i^+|s_\ominus) = \mathbb{O}_h(o_i^-|s_\ominus) = \frac{1}{2K}, \qquad \forall i \in [K].$$

  Here we omit the subscript μ to emphasize that the dynamic does not depend on μ.
- At step H, the emission dynamics at $\{s_\oplus, s_\ominus\}$ is given by

$$\mathbb{O}_H(\mathsf{good}|s_\oplus) = \frac{3}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_\oplus) = \frac{1}{4},$$
$$\mathbb{O}_H(\mathsf{good}|s_\ominus) = \frac{1}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_\ominus) = \frac{3}{4}.$$

Transition dynamics In each episode, the agent always begins at s_0.

- At any node $s \in \mathcal{S}_{\mathsf{tree}} \backslash \mathcal{S}_{\mathsf{leaf}}$, there are three types of available actions: $\mathsf{wait} = 0$, $\mathsf{left} = 1$ and $\mathsf{right} = 2$, such that the agent can take wait to stay at s, left to transit to the left child of s, and right to transit to the right child of s.[6]
- At any $s \in \mathcal{S}_{\mathsf{leaf}}$, the agent can take action $\mathsf{wait} = 0$ to stay at s (i.e. $\mathbb{P}(s|s, \mathsf{wait}) = 1$); otherwise, for $s \in \mathcal{S}_{\mathsf{leaf}}$, $h \in [H-1]$, $a \neq \mathsf{wait}$ (i.e. $a \in \mathcal{A}_c$),

$$\mathbb{P}_{h;\theta}(s_\oplus|s,a) = \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star),$$
$$\mathbb{P}_{h;\theta}(s_\ominus|s,a) = 1 - \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star),$$

  where we use subscript θ to emphasize the dependence of the transition probability $\mathbb{P}_{h;\theta}$ on θ. In words, at step h, state $s \in \mathcal{S}_{\mathsf{leaf}}$, and after $a \in \mathcal{A}_c$ is taken, any leaf node will transit to one of $\{s_\oplus, s_\ominus\}$, and only taking $a^\star$ at state $s^\star$ and step $h^\star$ can transit to the state $s_\oplus$ with a small probability ε; in any other case, the system will transit to the state $s_\ominus$ with probability one.
- At state $s_\oplus$, we set

$$\mathbb{P}_{h;\theta}(s_\oplus|s_\oplus,a) = \begin{cases} 1, & a = \mathbf{a}^\star_h, \\ 0, & a \neq \mathbf{a}^\star_h, \end{cases}, \qquad \mathbb{P}_{h;\theta}(s_\ominus|s_\oplus,a) = \begin{cases} 0, & a = \mathbf{a}^\star_h, \\ 1, & a \neq \mathbf{a}^\star_h. \end{cases}$$

- The state $s_\ominus$ is an absorbing state, i.e. $\mathbb{P}_h(s_\ominus|s_\ominus,a) = 1$ for all $a \in \mathcal{A}$.

[6] For action $a \in \{3, \cdots, A-1\}$, a has the same effect as wait.

Reward The reward function is known (and only depends on the observation): at the first $H-1$ steps, no reward is given; at step H, we set $r_H(\mathsf{good})=1$, $r_H(\mathsf{bad})=0$, $r_H(s_0)=(1+\varepsilon)/4$, and $r_H(o)=0$ for any other $o\in\mathcal{O}$.

Reference model We use M_0 (or simply 0) to refer to the null model (reference model). The null model M_0 has transition and emission the same as any non-null model, except that the agent always arrives at $s_\ominus$ by taking any action $a\neq\mathsf{wait}$ at $s\in\mathcal{S}_{\mathsf{leaf}}$ and $h\in[H-1]$ (i.e., $\mathbb{P}_{h;M_0}(s_\ominus|s,a)=1$ for any $s\in\mathcal{S}_{\mathsf{leaf}}$, $a\in\mathcal{A}_c$, $h\in[H-1]$). In this model, $s_\oplus$ is not reachable, and hence we do not need to specify the emission dynamics at $s_\oplus$.

We present the PAC-learning sample complexity lower bound of the above POMDP model class $\mathcal{M}$ in the following proposition, which we prove in Appendix E.2.

Proposition E.1. *For given $\varepsilon\in(0,0.1]$, $\sigma\in(0,\frac{1}{2H}]$, $n\geqslant 1$, $K\geqslant 1$, $H\geqslant 4n$, the model class $\mathcal{M}$ we construct above satisfies the following properties:*

1. $|\mathcal{S}|=2^n+1$, $|\mathcal{O}|=2^n+2K+1$, $|\mathcal{A}|=A$.
2. *For each $M\in\mathcal{M}$ (including the null model M_0), M is single-step revealing with $\alpha_1(M)^{-1}\leqslant 1+\frac{2}{\sigma}$.*
3. $\log|\mathcal{M}|\leqslant K\log 2+H\log A+\log(SAH)$.
4. *Suppose algorithm $\mathfrak{A}$ interacts with the environment for T episodes and returns π^{out} such that*

$$\mathbb{P}_M^{\mathfrak{A}}\left(V_M^\star-V_M(\pi^{\mathrm{out}})<\frac{\varepsilon}{8}\right)\geqslant\frac{3}{4}$$

for any $M\in\mathcal{M}$. Then it must hold that

$$T\geqslant\frac{1}{20000}\min\left\{\frac{|\mathcal{S}_{\mathsf{leaf}}|\,K^{1/2}AH}{\sigma^2\varepsilon^2},\frac{|\mathcal{S}_{\mathsf{leaf}}|\,A^{H/2}H}{\varepsilon^2}\right\},$$

where we recall that $|\mathcal{S}_{\mathsf{leaf}}|=2^{n-1}$.

Proof of Theorem 4 In Proposition E.1, suitably choosing σ,n,K, and choosing a rescaled ε, we obtain Theorem 4. More specifically, we can take $n\geqslant 1$ to be the largest integer such that $2^n\leqslant\min\{S-1,(O-1)/2\}$, and take $K=\left\lfloor\frac{O-2^n-1}{2}\right\rfloor\geqslant\frac{O-1}{4}$, $\varepsilon'=\varepsilon/8$, and $\sigma=\frac{2}{\alpha^{-1}-1}\leqslant\frac{1}{2H}$. Applying Proposition E.1 to the parameters $(\varepsilon',\sigma,n,K,H)$ completes the proof of Theorem 4. □

E.2. Proof of Proposition E.1

All propositions and lemmas stated in this section are proved in Appendix E.3-E.6.

Claim 1 follows directly by the counting the number of states, observations, and actions in construction of $\mathcal{M}$. Claim 3 follows as we have $|\mathcal{M}|=|\{(h^\star,s^\star,a^\star,\mathbf{a}^\star)\}|\times|\{\pm1\}^K|+1\leqslant HSA\times A^H\times 2^K$. Taking logarithm yields the claim.

Claim 2 follows directly by the following proposition with proof in Appendix E.3.

Proposition E.2. *For any $M\in\mathcal{M}$, M is single-step revealing with $\alpha_1(M)^{-1}\leqslant\frac{2}{\sigma}+1$.*

We now prove Claim 4 (the sample complexity lower bound). We begin by using the following lemma to relate the PAC learning problem to a testing problem, using the structure of $\mathcal{M}$. Intuitively, the lemma states that a near-optimal policy of any $M\neq 0$ cannot "stay" at s_0, whereas a near-optimal policy of model $M=0$ has to "stay" at s_0. The proof of the lemma is contained in Appendix E.4.

Lemma E.3 (Relating policy suboptimality to the probability of staying)**.** *For any $M\in\mathcal{M}$ such that $M\neq 0$ and any policy π, it holds that*

$$V_M^\star-V_M(\pi)\geqslant\frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H=s_0).\tag{18}$$

On the other hand, for the reference model 0 and any policy π, we have

$$V_0^\star-V_0(\pi)\geqslant\frac{\varepsilon}{4}\mathbb{P}_0^\pi(o_H\neq s_0).\tag{19}$$

Notice that the probability $\mathbb{P}^{\pi}_{M}(o_H = s_0)$ actually does not depend on the model $M \in \mathcal{M}$, i.e.

$$\mathbb{P}^{\pi}_{M}(o_H = s_0) = \mathbb{P}^{\pi}_{0}(o_H = s_0).$$

This is because once the agent leaves s_0, it will never come back (for any model $M \in \mathcal{M}$). In the following, we define $w(\pi) := \mathbb{P}^{\pi}_{0}(o_H = s_0)$. Note that π^{out} is the output policy that depends on the observation histories $\tau^{1:T}$, and thus $w(\pi^{\mathrm{out}})$ is a deterministic function of the observation histories $\tau^{1:T}$.

By Lemma E.3 and our assumption that $\mathbb{P}^{\mathfrak{A}}_{M}\big(V^{\star}_{M} - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\big) \geqslant \frac{3}{4}$ for any $M \in \mathcal{M}$, we have

$$\mathbb{P}^{\mathfrak{A}}_{0}\left(1 - w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) \geqslant \frac{3}{4}, \qquad \text{while} \qquad \mathbb{P}^{\mathfrak{A}}_{M}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) \geqslant \frac{3}{4}, \qquad \forall M \neq 0.$$

Now we consider $\mu \sim \mathrm{Unif}(\{\pm 1\}^K)$ to be the uniform prior over the parameter μ. For any fixed θ, we consider averaging the above quantity over the non-null models $M = (\theta, \mu)$ when $\mu \sim \mathrm{Unif}(\{\pm 1\}^K)$,

$$\mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\big]\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) = \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right)\right] \geqslant \frac{3}{4}.$$

However, we also have

$$\mathbb{P}^{\mathfrak{A}}_{0}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) = 1 - \mathbb{P}^{\mathfrak{A}}_{0}\left(w(\pi^{\mathrm{out}}) \geqslant \frac{1}{2}\right) \leqslant 1 - \mathbb{P}^{\mathfrak{A}}_{0}\left(w(\pi^{\mathrm{out}}) > \frac{1}{2}\right) \leqslant \frac{1}{4}.$$

Thus by the definition of TV distance we must have

$$D_{\mathrm{TV}}\left(\mathbb{P}^{\mathfrak{A}}_{0}, \mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\big]\right) \geqslant \left|\mathbb{P}^{\mathfrak{A}}_{0}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) - \mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\big]\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right)\right| \geqslant \frac{1}{2}. \tag{20}$$

As the core of the proof, we now use (20) to derive our lower bound on T. Recall that $\mathbb{P}^{\mathfrak{A}}_{M}$ is the law of $(\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(\mathsf{T})})$ induced by letting $\mathfrak{A}$ interact with the model M. For any event $E \subseteq (\mathcal{O} \times \mathcal{A})^H$, we denote the visitation count of E as

$$N(E) := \sum_{t=1}^{T} \mathbb{1}(\tau^{(t)} \in E).$$

Since $N(E)$ is a function of $\tau^{(1:T)}$, we can talk about its expectation under the distribution $\mathbb{P}^{\mathfrak{A}}_{M}$ for any $M \in \mathcal{M}$. We present the following lemma on the lower bound of the expected visitation count of some good events, whose proofs are contained in Appendix E.5.

Lemma E.4. *Fix a $\theta = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$. We consider events*

$$E^{\theta}_{\mathsf{rev},h} := \big\{o_{h^\star} = s^\star, a_{h^\star:h} = (a^\star, \mathbf{a}^\star_{h^\star+1:h})\big\},\ \forall h \in \{h^\star + 1, \ldots, H-2\},$$
$$E^{\theta}_{\mathsf{correct}} := \{o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star)\}.$$

Then for any algorithm $\mathfrak{A}$ with $\delta := D_{\mathrm{TV}}\left(\mathbb{P}^{\mathfrak{A}}_{0}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\right]\right) > 0$, we have

$$\textit{either} \sum_{h=h^\star}^{H-2} \mathbb{E}^{\mathfrak{A}}_{0}\big[N(E^{\theta}_{\mathsf{rev},h})\big] \geqslant \frac{\delta^3\sqrt{K}}{54\varepsilon^2\sigma^2} - \frac{H\delta}{6}, \quad \textit{or} \quad \mathbb{E}^{\mathfrak{A}}_{0}\big[N(E^{\theta}_{\mathsf{correct}})\big] \geqslant \frac{\delta^3}{54\varepsilon^2} - \frac{\delta}{6}.$$

Applying Lemma E.4 for any parameter tuple $\theta = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$ with $\delta = \frac{1}{2}$, we obtain

$$\text{either} \sum_{h=h^\star}^{H-2} \mathbb{E}^{\mathfrak{A}}_{0}\left[N\left(E^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}_{\mathsf{rev},h}\right)\right] \geqslant \frac{\sqrt{K}}{1000\varepsilon^2\sigma^2}, \quad \text{or} \quad \mathbb{E}^{\mathfrak{A}}_{0}\left[N\left(E^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}_{\mathsf{correct}}\right)\right] \geqslant \frac{1}{1000\varepsilon^2}, \tag{21}$$

by our choice that $\varepsilon \in (0, 0.1]$ and $\sigma \in (0, \frac{1}{2H}]$.

Fix a tuple $(h^\star, s^\star, a^\star)$ with $s^\star \in \mathcal{S}_{\mathsf{leaf}}, a^\star \in \mathcal{A}_c, h^\star \in [n+1, \frac{H}{2}]$. By (21), we know that for all $\mathbf{a} \in \mathcal{A}^{H-h^\star-1}$, it holds that

$$\sum_{h=h^\star}^{H-2} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev},h}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] + A^{H-h^\star-1} \cdot \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] \geqslant \frac{1}{1000}\min\left\{\frac{\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{A^{H/2-1}}{\varepsilon^2}\right\} =: \omega. \tag{22}$$

Notice that by definition,

$$\begin{aligned}\sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] =& \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}))]\\ =&\mathbb{E}_0^{\mathfrak{A}}\left[\sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}))\right]\\ =&\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)],\end{aligned}$$

and similarly for each $h \in [h^\star, H-2]$, it holds

$$\begin{aligned}\sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev},h}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] =& \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h}=(a^\star,\mathbf{a}_{h^\star+1:h}))]\\ =& \sum_{\mathbf{a}_{h^\star+1:h}\in\mathcal{A}^{h-h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h}=(a^\star,\mathbf{a}_{h^\star+1:h}))]\cdot \sum_{\mathbf{a}_{h+1:H-1}\in\mathcal{A}^{H-h-1}} 1\\ =&\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)]\cdot A^{H-h-1}.\end{aligned}$$

Therefore, summing the bound (22) over all $\mathbf{a} \in \mathcal{A}^{H-h^\star-1}$, we get

$$\begin{aligned}A^{H-h^\star-1}\omega = \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \omega \leqslant& \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \left[\sum_{h=h^\star}^{H-2} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev},h}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] + A^{H-h^\star-1}\cdot\mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right]\right]\\ =&\left(\sum_{h=h^\star}^{H-2} A^{H-h-1} + A^{H-h^\star-1}\right)\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)]\\ \leqslant&3A^{H-h^\star-1}\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)],\end{aligned}$$

where the last inequality is due to $\sum_{h=h^\star}^{H-2} A^{H-h-1} = \frac{A^{H-h}-A}{A-1} \leqslant 2A^{H-h-1}$ for $A \geqslant 3$.

Therefore, we have shown that $\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)] \geqslant \frac{\omega}{3}$ for each $s^\star \in \mathcal{S}_{\mathsf{leaf}}, a^\star \in \mathcal{A}_c, h^\star \in [n+1, \frac{H}{2}]$. Taking summation over all such $(h^\star, s^\star, a^\star)$, we derive that

$$|\mathcal{S}_{\mathsf{leaf}}|\,|\mathcal{A}_c|\left(\left\lfloor\frac{H}{2}\right\rfloor - n\right)\cdot\frac{\omega}{3} \leqslant \sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}}\sum_{a^\star\in\mathcal{A}_c}\sum_{h^\star=n+1}^{\lfloor H/2\rfloor-1} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)] \leqslant T,$$

where the second inequality is because events $\{o_{h^\star}=s^\star, a_{h^\star}=a^\star\}$ are disjoint. Plugging in $|\mathcal{A}_c| = A-1, H \geqslant 4n$ and the definition of ω in (22) completes the proof of Proposition E.1. □

E.3. Proof of Proposition E.2

We first consider the case $M = M_{\theta,\mu}$. At the step $h < H$, the emission matrix $\mathbb{O}_{h;\mu}$ can be written as (up to some permutation of rows and columns)

$$\mathbb{O}_{h;\mu} = \begin{bmatrix} \frac{\mathbb{1}_{2K}+\sigma\widetilde{\mu}}{2K} & \frac{\mathbb{1}_{2K}}{2K} & 0_{2K\times\mathcal{S}_{\mathsf{tree}}} \\ 0_{\mathcal{S}_{\mathsf{tree}}\times 1} & 0_{\mathcal{S}_{\mathsf{tree}}\times 1} & I_{\mathcal{S}_{\mathsf{tree}}\times\mathcal{S}_{\mathsf{tree}}} \\ 0_{2\times 1} & 0_{2\times 1} & 0_{2\times\mathcal{S}_{\mathsf{tree}}}\end{bmatrix} \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}},$$

where $\widetilde{\mu} = [\mu; -\mu] \in \{-1,1\}^{2K}$, and $\mathbb{1} = \mathbb{1}_{2K}$ is the column vector in $\mathbb{R}^{2K}$ with all entries being one. A simple calculation shows that

$$\left[\frac{\mathbb{1}+\sigma\widetilde{\mu}}{2K}, \frac{\mathbb{1}}{2K}\right]^{\dagger\top} = \left[\frac{1}{\sigma}\widetilde{\mu}, \mathbb{1} - \frac{1}{\sigma}\widetilde{\mu}\right],$$

whose 1-norm is bounded by $\frac{2}{\sigma}+1$. Hence $\left\|\mathbb{O}_{h;\mu}^{\dagger}\right\|_1 \leqslant \frac{2}{\sigma}+1$.

Similarly, for $h=H$, $\mathbb{O}_H$ has the form (up to some permutation of rows and columns)

$$\mathbb{O}_H = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0_{1\times\mathcal{S}_{\mathsf{tree}}} \\ \frac{1}{4} & \frac{3}{4} & 0_{1\times\mathcal{S}_{\mathsf{tree}}} \\ 0_{\mathcal{S}_{\mathsf{tree}}\times 1} & 0_{\mathcal{S}_{\mathsf{tree}}\times 1} & I_{\mathcal{S}_{\mathsf{tree}}\times\mathcal{S}_{\mathsf{tree}}} \\ 0_{2K\times 1} & 0_{2K\times 1} & 0_{2K\times\mathcal{S}_{\mathsf{tree}}} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}}.$$

Notice that $\begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}$, and hence $\left\|\mathbb{O}_H^{\dagger}\right\|_1 \leqslant 2$.

Finally, by Definition 1 and noting that $\mathbb{M}_{h,1}=\mathbb{O}_h$ and taking the generalized left inverse $\mathbb{M}_{h,1}^{+}=\mathbb{O}_h^{\dagger}$ to be the pseudo-inverse for all $h\in[H]$, this gives $(\alpha_1(M))^{-1}\leqslant \max\left\{\frac{1}{\sigma}+2, 2\right\}=\frac{2}{\sigma}+1$.

We next consider the case $M=0$. In this case, $s_\oplus$ is not reachable, and hence for each step h, we can consider the generalized left inverse of $\mathbb{O}_h$ given by

$$\mathbb{O}_h^{+} := \left[\mathbb{1}(\mathbb{O}_h(o|s)>0)\right]_{(s,o)} \in \mathbb{R}^{\mathcal{S}\times\mathcal{O}},$$

with the convention that $\mathbb{1}(\mathbb{O}_h(o|s_\oplus)>0)=0$ for all $o\in\mathcal{O}$ as $\mathbb{O}_h(\cdot|s_\oplus)$ is not defined. Then it is direct to verify $\mathbb{O}_h^{+}\mathbb{O}_h\mathbf{e}_s=\mathbf{e}_s$ for all state $s\neq s_\oplus$ (because the supports $\mathrm{supp}(\mathbb{O}_h(\cdot|s))$ are disjoint by our construction). It is clear that $\left\|\mathbb{O}_h^{+}\right\|_{1\to 1}\leqslant 1$, and hence $(\alpha_1(M))^{-1}\leqslant 1$, which completes the proof. □

E.4. Proof of Lemma E.3

By definition, for any model $M\in\mathcal{M}$ and policy π,

$$\begin{aligned} V_M(\pi) = \mathbb{E}_M^{\pi}[r_H(o_H)] &= \frac{1+\varepsilon}{4}\mathbb{P}_M^{\pi}(o_H=s_0)+\mathbb{P}_M^{\pi}(o_H=\mathsf{good}) \\ &= \frac{1+\varepsilon}{4}\mathbb{P}_M^{\pi}(o_H=s_0)+\frac{3}{4}\mathbb{P}_M^{\pi}(s_H=s_\oplus)+\frac{1}{4}\mathbb{P}_M^{\pi}(s_H=s_\ominus), \end{aligned}$$

where we have used the following equality due to our construction:

$$\begin{aligned} \mathbb{P}_M^{\pi}(o_H=\mathsf{good}) =& \mathbb{P}_M(o_H=\mathsf{good}|s_H=s_\oplus)\cdot\mathbb{P}_M^{\pi}(s_H=s_\oplus)+\mathbb{P}_M(o_H=\mathsf{good}|s_H=s_\ominus)\cdot\mathbb{P}_M^{\pi}(s_H=s_\ominus) \\ =& \frac{3}{4}\mathbb{P}_M^{\pi}(s_H=s_\oplus)+\frac{1}{4}\mathbb{P}_M^{\pi}(s_H=s_\ominus). \end{aligned}$$

We next prove the result for the case $M=0$ and $M\neq 0$ separately.

Case 1: $M=0$. In this case, $s_\oplus$ is not reachable, and hence we have $V_0^{\star}=\max_\pi V_0(\pi)=\max\left\{\frac{1+\varepsilon}{4},\frac{1}{4}\right\}=\frac{1+\varepsilon}{4}$, which is attained by staying at s_0. Thus, for any policy π,

$$\begin{aligned} V_0^{\star}-V_0(\pi) =& \frac{1+\varepsilon}{4}-\frac{1+\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H=s_0)-\frac{1}{4}\mathbb{P}_0^{\pi}(s_H=s_\ominus) \\ =& \frac{1+\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H\neq s_0)-\frac{1}{4}\mathbb{P}_0^{\pi}(s_H=s_\ominus) \\ =& \frac{1}{4}(\mathbb{P}_0^{\pi}(o_H\neq s_0)-\mathbb{P}_0^{\pi}(s_H=s_\ominus))+\frac{\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H\neq s_0) \\ \geqslant& \frac{\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H\neq s_0). \end{aligned}$$

Case 2: $M=(\theta,\mu)$ for some $\theta=(h^\star,s^\star,a^\star,\mathbf{a}^\star)$. In this case, $s_\oplus$ is reachable only when $o_{h^\star}=s^\star$ and $a_{h^\star}=a^\star$, and

$$\mathbb{P}_M^{\pi}(s_H=s_\oplus)=\mathbb{P}_M^{\pi}(s_H=s_\oplus|o_{h^\star}=s^\star,a_{h^\star}=a^\star)\mathbb{P}_M^{\pi}(o_{h^\star}=s^\star,a_{h^\star}=a^\star)\leqslant\varepsilon\mathbb{P}_M^{\pi}(o_{h^\star}=s^\star,a_{h^\star}=a^\star)\leqslant\varepsilon,$$

where the equality can be attained when π is any deterministic policy that ensure $o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_{h^\star+1:H-1}=\mathbf{a}^\star$. Thus, in this case $V_M^{\star}=\max_\pi V_M(\pi)=\max\left\{\frac{1+\varepsilon}{4},\frac{3\varepsilon}{4}+\frac{1-\varepsilon}{4}\right\}=\frac{1+2\varepsilon}{4}$, and

$$V_M^{\star}-V_M(\pi) = \frac{1+2\varepsilon}{4}-\frac{1+\varepsilon}{4}\mathbb{P}_M^{\pi}(o_H=s_0)-\frac{3}{4}\mathbb{P}_M^{\pi}(s_H=s_\oplus)-\frac{1}{4}\mathbb{P}_M^{\pi}(s_H=s_\ominus)$$

$$
\begin{aligned}
&= \frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H = s_0) + \frac{1+2\varepsilon}{4}\mathbb{P}_M^\pi(o_H \neq s_0) - \frac{3}{4}\mathbb{P}_M^\pi(s_H = s_\oplus) - \frac{1}{4}\mathbb{P}_M^\pi(s_H = s_\ominus) \\
&\geqslant \frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H = s_0) + \frac{\varepsilon}{2}\mathbb{P}_M^\pi(o_H \neq s_0) - \frac{1}{2}\mathbb{P}_M^\pi(s_H = s_\oplus) \\
&\geqslant \frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H = s_0),
\end{aligned}
$$

where the first inequality is because $\mathbb{P}_M^\pi(s_H = s_\oplus) + \mathbb{P}_M^\pi(s_H = s_\ominus) \leqslant \mathbb{P}_M^\pi(o_H \neq s_0)$ by the inclusion of events. □

E.5. Proof of Lemma E.4

We first prove the following version of Lemma E.4 with an additional condition that the visitation counts are almost surely bounded under $\mathbb{P}_0^{\mathfrak{A}}$, and then prove Lemma E.4 by reducing to this case using a truncation argument.

Lemma E.5. *Suppose that algorithm $\mathfrak{A}$ (with possibly random stopping time T) satisfies $\sum_h N(E^\theta_{\mathsf{rev},h}) \leqslant \overline{N}_o$ and $N(E^\theta_{\mathsf{correct}}) \leqslant \overline{N}_r$ almost surely under $\mathbb{P}_0^{\mathfrak{A}}$, for some fixed $\overline{N}_o, \overline{N}_r$. Then*

$$
\text{either } \overline{N}_o \geqslant \frac{\delta^2\sqrt{K}}{4\varepsilon^2\sigma^2}, \quad \text{or } \overline{N}_r \geqslant \frac{\delta^2}{4\varepsilon^2},
$$

where $\delta = D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right)$.

Proof of Lemma E.5. By Lemma D.1, we have

$$
1 + \chi^2(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right] \,\|\, \mathbb{P}_0^{\mathfrak{A}}) = \mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right].
$$

To upper bound the above quantity, we invoke the following lemma, which serves a key step for bounding the above "χ^2-inner product" (Canonne, 2022, Section 3.1) between $\mathbb{P}_{\theta,\mu}/\mathbb{P}_0$ and $\mathbb{P}_{\theta,\mu'}/\mathbb{P}_0$ (proof in Appendix E.6).

Lemma E.6 (Bound on the χ^2-inner product)**.** *Under the conditions of Lemma E.5 (for a fixed θ), it holds that for any $\mu, \mu' \in \{-1,1\}^K$,*

$$
\mathbb{E}_0^{\mathfrak{A}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right] \leqslant \exp\left(\overline{N}_o\cdot\frac{C\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right| + \frac{4}{3}C\varepsilon^2\overline{N}_r\right). \tag{23}
$$

where $C := (1+\sigma)^{2H} \leqslant e$ as $\sigma \leqslant \frac{1}{2H}$.

Now we assume that Lemma E.6 holds and continue the proof of Lemma E.5. Taking expectation of (23) over $\mu, \mu' \sim \mathrm{Unif}(\{-1,+1\}^K)$, we obtain

$$
\begin{aligned}
1 + \chi^2(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right] \,\|\, \mathbb{P}_0^{\mathfrak{A}}) &= \mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right] \\
&\leqslant \mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\overline{N}_o\cdot\frac{C\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right| + \frac{4}{3}C\varepsilon^2\overline{N}_r\right)\right].
\end{aligned}
$$

Notice that μ_i, μ_i' are i.i.d. $\mathrm{Unif}(\{\pm 1\})$, and hence $\mu_1\mu_1', \cdots, \mu_K\mu_K'$ are i.i.d. $\mathrm{Unif}(\{\pm 1\})$. Then by Hoeffding's lemma, it holds that $\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(x\sum_{i=1}^K\mu_i\mu_i'\right)\right] \leqslant \exp\left(Kx^2/2\right)$ for all $x \in \mathbb{R}$, and thus by Lemma A.3, we have

$$
\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\frac{C\overline{N}_o\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right|\right)\right] \leqslant \exp\left(\max\left\{\frac{C^2\sigma^4\varepsilon^4\overline{N}_o^2}{K}, \frac{4}{3}\frac{C\sigma^2\varepsilon^2\overline{N}_o}{\sqrt{K}}\right\}\right).
$$

Therefore, combining the above inequalities with Lemma A.5, we obtain

$$
2\delta^2 = 2D_{\mathrm{TV}}\left(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right], \mathbb{P}_0^{\mathfrak{A}}\right)^2 \leqslant \log\left(1 + \chi^2(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right] \,\|\, \mathbb{P}_0^{\mathfrak{A}})\right) \leqslant \max\left\{\frac{4}{3}\frac{\overline{N}_oC\sigma^2\varepsilon^2}{\sqrt{K}}, \frac{\overline{N}_o^2C^2\sigma^4\varepsilon^4}{K}\right\} + \frac{4}{3}C\varepsilon^2\overline{N}_r.
$$

Then, we either have $\overline{N}_r \geqslant \frac{3\delta^2}{4C\varepsilon^2}$, or it holds

$$\max\left\{\frac{4}{3}\frac{\overline{N}_o C\sigma^2\varepsilon^2}{\sqrt{K}}, \frac{\overline{N}_o^2 C^2\sigma^4\varepsilon^4}{K}\right\} \geqslant \delta^2,$$

which implies that $\frac{\overline{N}_o C\sigma^2\varepsilon^2}{\sqrt{K}} \geqslant \min\left\{\frac{4}{3}, \frac{3}{4}\delta^2\right\} = \frac{3}{4}\delta^2$ (as $\delta \leqslant 1$). Using the fact that $C \leqslant e$ completes the proof of Lemma E.5. □

Proof of Lemma E.4. We perform a truncation type argument to reduce Lemma E.4 to Lemma E.5. Let us take $\overline{N}_o = \left\lceil 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[\sum_{h=h^\star}^{H-2} N(E^\theta_{\mathsf{rev},h})\right]\right\rceil$ and $\overline{N}_r = \left\lceil 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}} N(E^\theta_{\mathsf{correct}})\right\rceil$. By Markov's inequality, we have

$$\mathbb{P}_0^{\mathfrak{A}}\left(\sum_{h=h^\star}^{H-2} N(E^\theta_{\mathsf{rev},h}) \geqslant \overline{N}_o\right) \leqslant \frac{\delta}{6}, \qquad \mathbb{P}_0^{\mathfrak{A}}\left(N(E^\theta_{\mathsf{correct}}) \geqslant \overline{N}_r\right) \leqslant \frac{\delta}{6}.$$

Therefore, we can consider the following exit criterion exit for the algorithm $\mathfrak{A}$:

$$\mathsf{exit}(\tau^{(1:T')}) = \mathsf{TRUE} \quad \text{iff} \quad \sum_{t=1}^{T'}\sum_{h=h^\star}^{H-2} \mathbb{I}\left(\tau^{(t)} \in E^\theta_{\mathsf{rev},h}\right) \geqslant \overline{N}_o \text{ or } \sum_{t=1}^{T'} \mathbb{I}\left(\tau^{(t)} \in E^\theta_{\mathsf{correct}}\right) \geqslant \overline{N}_r.$$

The criterion exit induces a stopping time T_{exit}, and we have

$$\mathbb{P}_0^{\mathfrak{A}}(\exists t < T, \mathsf{exit}(\tau^{(1:t)}) = \mathsf{TRUE}) \leqslant \mathbb{P}_0^{\mathfrak{A}}\left(\sum_{h=h^\star}^{H-2} N(E^\theta_{\mathsf{rev},h}) \geqslant \overline{N}_o \text{ or } N(E^\theta_{\mathsf{correct}}) \geqslant \overline{N}_r\right) \leqslant \frac{\delta}{6} + \frac{\delta}{6} \leqslant \frac{\delta}{3}.$$

Therefore, we can consider the early stopped algorithm $\mathfrak{A}(\mathsf{exit})$ with exit criterion exit (cf. Appendix D), and by Lemma D.2 we have

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}(\mathsf{exit})}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}(\mathsf{exit})}\right]\right) \geqslant D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) - \mathbb{P}_0^{\mathfrak{A}}(\exists t < T, \mathsf{exit}(\tau^{(1:t)}) = \mathsf{TRUE}) \geqslant \frac{2\delta}{3}.$$

Notice that by our definition of exit and stopping time T_{exit}, in the execution of $\mathfrak{A}(\mathsf{exit})$, we also have

$$\sum_{t=1}^{T_{\mathsf{exit}}-1}\sum_{h=h^\star}^{H-2} \mathbb{1}\left(\tau^{(t)} \in E^\theta_{\mathsf{rev},h}\right) < \overline{N}_o, \qquad \sum_{t=1}^{T_{\mathsf{exit}}-1} \mathbb{1}\left(\tau^{(t)} \in E^\theta_{\mathsf{correct}}\right) < \overline{N}_r.$$

Therefore, algorithm $\mathfrak{A}(\mathsf{exit})$ ensures that

$$\sum_{h=h^\star}^{H-2} N(E^\theta_{\mathsf{rev},h}) = \sum_{t=1}^{T_{\mathsf{exit}}}\sum_{h=h^\star}^{H-2} \mathbb{1}\left(\tau^{(t)} \in E^\theta_{\mathsf{rev},h}\right) \leqslant \overline{N}_o + H - 1, \qquad N(E^\theta_{\mathsf{correct}}) = \sum_{t=1}^{T_{\mathsf{exit}}} \mathbb{1}\left(\tau^{(t)} \in E^\theta_{\mathsf{correct}}\right) \leqslant \overline{N}_r.$$

Applying Lemma E.5 to the algorithm $\mathfrak{A}(\mathsf{exit})$ (and $\delta' = \frac{2}{3}\delta$), we can obtain

$$\text{either } \frac{\delta^2\sqrt{K}}{9\varepsilon^2\sigma^2} \leqslant \overline{N}_o + H - 1 \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[\sum_{h=h^\star}^{H-2} N(E^\theta_{\mathsf{rev},h})\right] + H, \qquad \text{or } \frac{\delta^2}{9\varepsilon^2} \leqslant \overline{N}_r \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[N(E^\theta_{\mathsf{correct}})\right] + 1,$$

and rearranging gives the desired result. □

E.6. Proof of Lemma E.6

Throughout the proof, the parameters θ, μ, μ' are fixed.

By our discussions in Appendix D, using (16), we have

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}} \frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2} \cdot \exp\left(-\sum_{t=1}^{\mathsf{T}}\sum_{h=1}^{H} \log I(\tau_{h-1}^{(t)})\right)\right] = 1, \tag{24}$$

where for any partial trajectory τ_l up to step $l \in [H]$, $I(\tau_l)$ is defined as

$$I(\tau_l) := \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right].$$

Notice that the model $\mathbb{P}_{\theta,\mu}$ and $\mathbb{P}_0$ are different only at the transition from $s_{h^\star} = s^\star, a_{h^\star} = a^\star$ to $s_\oplus$ and the transition dynamic at state $s_\oplus$. Therefore, for any (reachable) trajectory $\tau_l = (o_1, a_1, \cdots, o_l, a_l)$, $\mathbb{P}_{\theta,\mu}(o_{l+1} = \cdot|\tau_l) \neq \mathbb{P}_0(o_{l+1} = \cdot|\tau_l)$ only if $o_{h^\star} = s^\star, a_{h^\star} = a^\star$. In other words, $I(\tau_l) = 1$ if $\tau_l \notin \{o_{h^\star} = s^\star, a_{h^\star} = a^\star\}$.

We next compute $I(\tau_l)$ for $\tau_l \in \{o_{h^\star} = s^\star, a_{h^\star} = a^\star\}$. By our construction, we have

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l) &= \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = s_\oplus)\cdot\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l)\\ &\quad + \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = s_\ominus)\cdot\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\ominus|\tau_l)\\ &= (\mathbb{O}_{l;\mu}(o|s_\oplus) - \mathbb{O}_l(o|s_\ominus))\cdot\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l) + \mathbb{O}_l(o|s_\ominus).\end{aligned} \tag{25}$$

Notice that if $\tau_l \notin E_{\mathsf{rev},l}$, then s_{l+1} must be $s_\ominus$, and hence $\mathbb{P}_{\theta,\mu}(o_{l+1} = \cdot|\tau_l) = \mathbb{O}_h(\cdot|s_\ominus) = \mathbb{P}_0(o_{l+1} = \cdot|\tau_l)$ which implies that $I(\tau_l) = 1$.

We next consider the case $\tau_l \in E_{\mathsf{rev},l}$, i.e. $a_{h^\star+1:l} = \mathbf{a}^\star_{h^\star+1:l}$:

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l) &= \mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|o_{h^\star} = s^\star, a_{h^\star} = a^\star, o_{h^\star+1:l}, a_{h^\star+1:l})\\ &= \frac{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}, s_{l+1} = s_\oplus|o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_{h^\star+1:l})}{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_{h^\star+1:l})}\\ &= \frac{\varepsilon\cdot\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\oplus, a_{h^\star+1:l})}{\varepsilon\cdot\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\oplus, a_{h^\star+1:l}) + (1-\varepsilon)\cdot\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\ominus, a_{h^\star+1:l})}\\ &= \frac{\varepsilon}{\varepsilon + (1-\varepsilon)\cdot\frac{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1}=s_\ominus, a_{h^\star+1:l})}{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1}=s_\oplus, a_{h^\star+1:l})}},\end{aligned}$$

where the third equality is because $\mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s_\oplus|o_{h^\star} = s^\star, a_{h^\star} = a^\star) = \varepsilon$. Notice that

$$\beta_{\tau_l} := \frac{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\oplus, a_{h^\star+1:l})}{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\ominus, a_{h^\star+1:l})} = \prod_{h=h^\star+1}^{l}\frac{\mathbb{O}_{h;\mu}(o_h|s_\oplus)}{\mathbb{O}_h(o_h|s_\ominus)} \leqslant (1+\sigma)^{l-h^\star},$$

where the inequality holds by our construction of $\mathbb{O}$, as long as τ_l is reachable (i.e. $o_{h^\star+1:l} \in \mathcal{O}^{l-h^\star}$). Thus, for

$$c_{\tau_l} := \mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l) = \frac{\beta_{\tau_l}}{\varepsilon\beta_{\tau_l} + 1 - \varepsilon},$$

we have $c_{\tau_l} \leqslant (1+\sigma)^H = \sqrt{C}$. Notice that by (25) and the equation above we have

$$\begin{aligned}&\text{when } l < H-1, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^+|\tau_l) = \frac{1 + c_{\tau_l}\varepsilon\sigma\mu_i}{2K}, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^-|\tau_l) = \frac{1 - c_{\tau_l}\varepsilon\sigma\mu_i}{2K} \qquad \forall i \in [K],\\ &\text{when } l = H-1, \qquad \mathbb{P}_{\theta,\mu}(o_H = \mathsf{good}|\tau_{H-1}) = \frac{1 + 2c_{\tau_{H-1}}\varepsilon}{4}, \qquad \mathbb{P}_{\theta,\mu}(o_H = \mathsf{bad}|\tau_{H-1}) = \frac{3 - 2c_{\tau_{H-1}}\varepsilon}{4}.\end{aligned}$$

On the other hand, when $l < H-1$, $\mathbb{P}_{\theta,\mu}(o_{l+1} = \cdot|\tau_l) = \mathrm{Unif}(\{o_1^+, o_1^-, \cdots, o_K^+, o_K^-\})$. Hence,

$$\begin{aligned}I(\tau_l) =& \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right]\\ =& \frac{1}{2K}\sum_{o\in\mathcal{O}_o}\frac{\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1} = o|\tau_l)}{\mathbb{P}_0(o_{l+1} = o|\tau_l)^2}\\ =& \frac{1}{2K}\sum_{i=1}^{K}(1 + c_{\tau_l}\varepsilon\sigma\mu_i)(1 + c_{\tau_l}\varepsilon\sigma\mu_i') + (1 - c_{\tau_l}\varepsilon\sigma\mu_i)(1 - c_{\tau_l}\varepsilon\sigma\mu_i')\end{aligned}$$

$$=1+\frac{c_{\tau_l}^2\varepsilon^2\sigma^2}{K}\sum_{i=1}^{K}\mu_i\mu_i' \leqslant 1+\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right|.$$

Similarly, when $l=H-1$, we can compute

$$I(\tau_{H-1})=\mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_H|\tau_{H-1})\mathbb{P}_{\theta,\mu'}(o_H|\tau_{H-1})}{\mathbb{P}_0(o_H|\tau_{H-1})^2}\middle|\tau_{H-1}\right]=1+\frac{4}{3}c_{\tau_{H-1}}^2\varepsilon^2\leqslant 1+\frac{4}{3}C\varepsilon^2.$$

Therefore, combining all these facts above, we can conclude that

$$\begin{cases} I(\tau_l)=1, & l\leqslant h^\star,\\ I(\tau_l)\leqslant 1+\mathbb{1}\left(\tau_l\in E^\theta_{\mathsf{rev},l}\right)\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right|, & h^\star<l<H-1,\\ I(\tau_{H-1})\leqslant 1+\mathbb{1}\left(\tau_{H-1}\in E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}C\varepsilon^2, & l=H-1,\end{cases}$$

where we use the fact that $E^\theta_{\mathsf{correct}}=E^\theta_{\mathsf{rev},H-1}$ by definition. Hence, using the fact $\log(1+x)\leqslant x$, we have

$$\begin{aligned}\sum_{t=1}^{\mathsf{T}}\sum_{l=0}^{H-1}\log I(\tau_l^{(t)}) &= \sum_{t=1}^{\mathsf{T}}\sum_{l=h^\star+1}^{H-1}\log I(\tau_l^{(t)})\\ &\leqslant \sum_{t=1}^{\mathsf{T}}\sum_{l=h^\star+1}^{H-2}\mathbb{1}\left(\tau_l^{(t)}\in E^\theta_{\mathsf{rev},l}\right)\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right|+\mathbb{1}\left(\tau_{H-1}^{(t)}\in E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}C\varepsilon^2\\ &= \sum_{l=h^\star+1}^{H-2}N\left(E^\theta_{\mathsf{rev},l}\right)\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right|+N\left(E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}C\varepsilon^2\\ &\leqslant \overline{N}_o\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right|+\overline{N}_r\cdot\frac{4}{3}C\varepsilon^2.\end{aligned}$$

Plugging the above inequality into (24) completes the proof of Lemma E.6. □

F. Proof of Theorem 6

We first construct a family of hard instances in Appendix F.1. We state the regret lower bound of this family of hard instances in Proposition F.1. Theorem 6 then follows from Proposition F.1 as a direct corollary. Proposition F.1 also implies a part of the PAC lower bound stated in Theorem 5.

F.1. Construction of hard instances and proof of Theorem 6

We consider the following family of m-step revealing POMDPs $\mathcal{M}$ that admits a tuple of hyperparameters $(\varepsilon,\sigma,n,m,K,H)$. All POMDPs in $\mathcal{M}$ share the state space $\mathcal{S}$, action space $\mathcal{A}$, observation space $\mathcal{O}$, and horizon length H, defined as following.

- The state space $\mathcal{S}=\mathcal{S}_{\mathsf{tree}}\bigsqcup\{s_\oplus,s_\ominus,e_\oplus,e_\ominus,\mathsf{terminal}\}$, where $\mathcal{S}_{\mathsf{tree}}$ is a binary tree with level n (so that $|\mathcal{S}_{\mathsf{tree}}|=2^n-1$). Let s_0 be the root of $\mathcal{S}_{\mathsf{tree}}$, and $\mathcal{S}_{\mathsf{leaf}}$ be the set of leaves of $\mathcal{S}_{\mathsf{tree}}$, with $|\mathcal{S}_{\mathsf{leaf}}|=2^{n-1}$.
- The observation space $\mathcal{O}=\mathcal{S}_{\mathsf{tree}}\bigsqcup\{o_1^+,o_1^-,\cdots,o_K^+,o_K^-\}\bigsqcup\{\mathsf{lock},\mathsf{good},\mathsf{bad},\mathsf{terminal}\}$. [7]
- The action space $\mathcal{A}=\{0,1,\cdots,A-1\}$.

We further define $\mathcal{A}_{\mathsf{rev}}=\{0,1,\cdots,A_1-1\},\mathcal{A}_{\mathsf{tr}}=\{A_1,\cdots,A-1\}$, with $A_1=1+\lfloor A/6\rfloor$.

Model parameters Each non-null POMDP model $M=M_{\theta,\mu}\in\mathcal{M}\backslash\{M_0\}$ is specified by two parameters (θ,μ). Here $\mu\in\{-1,+1\}^K$, and $\theta=(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)$, where

- $s^\star\in\mathcal{S}_{\mathsf{leaf}}$, $a^\star\in\mathcal{A}_c:=\{1,\cdots,A-1\}$, $a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}}$.

[7] Similarly to Appendix E, here we slightly abuse notation to reuse $\mathcal{S}_{\mathsf{tree}}$ to denote both a set of states and a corresponding set of observations, in the sense that each state $s\in\mathcal{S}_{\mathsf{tree}}\subset\mathcal{S}$ corresponds to a unique observation $o_s\in\mathcal{S}_{\mathsf{tree}}\subset\mathcal{O}$, which we also denote as s when it is clear from the context.

- $h^\star \in \mathcal{H} := \{h = n + lm : h < H, l \in \mathbb{Z}_{\geqslant 0}\}$.
- $\mathbf{a}^\star = (\mathbf{a}^\star_{h^\star+1}, \ldots, \mathbf{a}^\star_{H-1}) \in \mathcal{A}^{H-h^\star-1}$ is an action sequence indexed by $h^\star + 1, \cdots, H-1$, such that when $h \in \mathcal{H}$, we have $\mathbf{a}^\star_h \in \mathcal{A}_{\mathsf{tr}}$. We use $\mathcal{A}_{\mathsf{code},h^\star}$ to denote the set of all such $\mathbf{a}^\star$.

Our construction ensures that, only at steps $h \in \mathcal{H}$ and states $s_h \in \{s_\oplus, s_\ominus\}$, the agent can take actions in $\mathcal{A}_{\mathsf{rev}}$ and transits to $\{e_\oplus, e_\ominus\}$.

For any POMDP $M_{\theta,\mu}$, its system dynamics $\mathbb{P}_{\theta,\mu} := \mathbb{P}_{M_{\theta,\mu}}$ is defined as follows.

Emission dynamics At state $s \in \mathcal{S}_{\mathsf{tree}} \cup \{\mathsf{terminal}\}$, the agent always receives (the unique observation corresponding to) s itself as the observation.

- At state $e_\oplus$, the emission dynamic is given by
$$\mathbb{O}_\mu(o_i^+|e_\oplus) = \frac{1+\sigma\mu_i}{2K}, \qquad \mathbb{O}_\mu(o_i^-|e_\oplus) = \frac{1-\sigma\mu_i}{2K}, \qquad \forall i \in [K],$$
where we omit the subscript h because the emission distribution does not depend on h.
- At state $e_\ominus$, the observation is uniformly drawn from $\mathcal{O}_o := \{o_1^+, o_1^-, \cdots, o_K^+, o_K^-\}$, i.e. $\mathbb{O}(\cdot|e_\ominus) = \mathrm{Unif}(\mathcal{O}_o)$.
- At states $s \in \{s_\oplus, s_\ominus\}$ and steps $h \in [H-1]$, the agent always receives lock as the observation; At step H, the emission dynamics at $\{s_\oplus, s_\ominus\}$ is given by
$$\mathbb{O}_H(\mathsf{good}|s_\oplus) = \frac{3}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_\oplus) = \frac{1}{4},$$
$$\mathbb{O}_H(\mathsf{good}|s_\ominus) = \frac{1}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_\ominus) = \frac{3}{4}.$$

Transition dynamics In each episode, the agent always starts at state s_0.

- At any node $s \in \mathcal{S}_{\mathsf{tree}} \backslash \mathcal{S}_{\mathsf{leaf}}$, there are three types of available actions: $\mathsf{wait} = 0$, $\mathsf{left} = 1$ and $\mathsf{right} = 2$, such that the agent can take wait to stay at s, left to transit to the left child of s and right to transit to the right child of s.
- At any $s \in \mathcal{S}_{\mathsf{leaf}}$, the agent can take action $\mathsf{wait} = 0$ to stay at s (i.e. $\mathbb{P}(s|s, \mathsf{wait}) = 1$); otherwise, for $s \in \mathcal{S}_{\mathsf{leaf}}$, $h \in [H-1]$, $a \neq \mathsf{wait}$,
$$\mathbb{P}_{h;\theta}(s_\oplus|s,a) = \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star),$$
$$\mathbb{P}_{h;\theta}(s_\ominus|s,a) = 1 - \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star).$$
where we use subscript θ to emphasize the dependence on θ. In words, at step $h^\star$, at any leaf node taking any action, the agent will transit to one of $\{s_\oplus, s_\ominus\}$; only by taking $a^\star$ at $s^\star$, the agent can transit to state $s_\oplus$ with a small probability ε; in any other case the agent will transit to state $s_\ominus$ with probability one.
- The state $s \in \{e_\oplus, e_\ominus\}$ always transits to $\mathsf{terminal}$, regardless of the action taken.
- The $\mathsf{terminal}$ state is an absorbing state.
- At state $s_\ominus$:
  - For steps $h \in \mathcal{H}$ and $a \in \mathcal{A}_{\mathsf{rev}}$, we set $\mathbb{P}_{h;\theta}(e_\ominus|s_\ominus, a) = 1$, i.e. taking $a \in \mathcal{A}_{\mathsf{rev}}$ always transits to $e_\ominus$.
  - For steps $h \notin \mathcal{H}$ or $a \in \mathcal{A}_{\mathsf{tr}}$, we set $\mathbb{P}_{h;\theta}(s_\ominus|s_\ominus, a) = 1$, i.e. taking such action always stays at $s_\ominus$.
- At state $s_\oplus$, we only need to specify the transition dynamics for steps $h \geqslant h^\star + 1$:
  - For steps $h \in \mathcal{H}_{>h^\star} = \mathcal{H} \cap \{h > h^\star\}$ and $a \in \mathcal{A}_{\mathsf{rev}}$, we set
  $$\mathbb{P}_{h;\theta}(e_\oplus|s_\oplus, a) = \mathbb{1}(a = a^\star_{\mathsf{rev}}), \qquad \mathbb{P}_{h;\theta}(e_\ominus|s_\oplus, a) = \mathbb{1}(a \neq a^\star_{\mathsf{rev}}).$$
  In words, at steps $h \in \mathcal{H}_{>h^\star}$ and states $s_h \in \{s_\oplus, s_\ominus\}$ (corresponding to $o_h = \mathsf{lock}$), the agent can take actions $a_h \in \mathcal{A}_{\mathsf{rev}}$ to transit to $\{e_\oplus, e_\ominus\}$; but only by taking $a_h = a^\star_{\mathsf{rev}}$ "correctly" at $s_\oplus$ the agent can transit to $e_\oplus$; in any other case the agent will transit to state $e_\ominus$ with probability one. Note that $\mathcal{H} = \{h = n + lm : h < H, l \in \mathbb{Z}_{\geqslant 0}\}$, so we only allow the agent to take the reveal action $a^\star_{\mathsf{rev}}$ every m steps, which ensures that our construction is $(m+1)$-step revealing.
  - For steps $h \notin \mathcal{H}$ or $a \in \mathcal{A}_{\mathsf{tr}}$, we set
  $$\mathbb{P}_{h;\theta}(s_\oplus|s_\oplus, a) = \mathbb{1}(a = \mathbf{a}^\star_h), \qquad \mathbb{P}_{h;\theta}(s_\ominus|s_\oplus, a) = \mathbb{1}(a \neq \mathbf{a}^\star_h).$$

Reward The reward function is known (and only depends on the observation): at the first $H-1$ steps, no reward is given; at step H, we set $r_H(\mathsf{good})=1$, $r_H(\mathsf{bad})=0$, $r_H(s_0)=(1+\varepsilon)/4$, and $r_H(o)=0$ for any other $o\in\mathcal{O}$.

Reference model We use M_0 (or simply 0) to refer to the null model (reference model). The null model M_0 has transition and emission the same as any non-null model, except that the agent always arrives at $s_\ominus$ by taking any action $a\neq\mathsf{wait}$ at $s\in\mathcal{S}_{\mathsf{leaf}}$ and $h\in[H-1]$ (i.e., $\mathbb{P}_{h;M_0}(s_\ominus|s,a)=1$ for any $s\in\mathcal{S}_{\mathsf{leaf}}$, $a\in\mathcal{A}_c$, $h\in[H-1]$). In this model, $s_\oplus$ is not reachable (and so does $e_\oplus$), and hence we do not need to specify the transition and emission dynamics at $s_\oplus, e_\oplus$.

We present the expected regret lower bound and PAC-learning sample complexity lower bound of the above POMDP model class $\mathcal{M}$ in the following proposition, which we prove in Appendix F.2.

Proposition F.1. *For given $\varepsilon\in(0,0.1]$, $\sigma\in(0,1]$, $m,n\geqslant 1$, $K\geqslant 2$, $H\geqslant 8n+m+1$, the above model class $\mathcal{M}$ satisfies the following properties.*

1. $|\mathcal{S}|=2^n+4$, $|\mathcal{O}|=2^n+2K+3$, $|\mathcal{A}|=A$.
2. *For each $M\in\mathcal{M}$, M is $(m+1)$-step revealing with $\alpha_{m+1}(M)^{-1}\leqslant 1+\frac{2}{\sigma}$.*
3. $\log|\mathcal{M}|\leqslant K\log 2+H\log A+\log(SAH)$.
4. *Suppose algorithm $\mathfrak{A}$ interacts with the environment for T episodes, then*

$$\max_{M\in\mathcal{M}}\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}]\geqslant\frac{1}{120000}\min\left\{\frac{|\mathcal{S}_{\mathsf{leaf}}|\,K^{1/2}A^{m+1}H}{m\sigma^2\varepsilon^2},\frac{|\mathcal{S}_{\mathsf{leaf}}|\,A^{H/2}H}{m\varepsilon},\varepsilon T\right\},$$

   where we recall that $|\mathcal{S}_{\mathsf{leaf}}|=2^{n-1}$.

5. *Suppose algorithm $\mathfrak{A}$ interacts with the environment for T episodes and returns π^{out} such that*

$$\mathbb{P}_M^{\mathfrak{A}}\left(V_M^\star-V_M(\pi^{\mathrm{out}})<\frac{\varepsilon}{8}\right)\geqslant\frac{3}{4}.$$

   for any $M\in\mathcal{M}$, then it must hold that

$$T\geqslant\frac{1}{60000}\min\left\{\frac{|\mathcal{S}_{\mathsf{leaf}}|\,K^{1/2}A^{m+1}H}{\sigma^2\varepsilon^2},\frac{|\mathcal{S}_{\mathsf{leaf}}|\,A^{H/2}H}{\varepsilon^2}\right\}.$$

Proof of Theorem 6 We only need to suitably choose parameters when applying Proposition F.1. More specifically, given (S,O,A,H,α,m), we can let $m'=m-1$, and take $n\geqslant 1$ to be the largest integer such that $2^n\leqslant\min\{S-4,(O-5)/2\}$, and take $K=\left\lfloor\frac{O-2^n-3}{2}\right\rfloor\geqslant\frac{O-5}{4}$, $\varepsilon'=\varepsilon/8$, and $\sigma=\frac{2}{\alpha^{-1}-1}\leqslant 1$. For any fixed $\varepsilon\in(0,0.1]$, applying Proposition F.1 to the parameters $(\varepsilon,\sigma,n,m',K,H)$, we obtain a model class $\mathcal{M}_\varepsilon$ such that for any algorithm $\mathfrak{A}$,

$$\max_{M\in\mathcal{M}_\varepsilon}\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}]\geqslant c_0\min\left\{\frac{SO^{1/2}A^mH}{m\alpha^2\varepsilon^2},\frac{SA^{H/2}H}{m\varepsilon},\varepsilon T\right\},$$

where c_0 is a universal constant. We can then take the $\varepsilon\in(0,0.1]$ that maximizes the RHS of the above inequality, and applying Lemma A.1 completes the proof of Theorem 6. □

Remark F.2. The requirement $S\leqslant O$ in Theorem 6 (and Theorem 5) can actually be relaxed to $S\leqslant O^m$. The reason why we require $S\leqslant O$ in the current construction is that we directly embed $\mathcal{S}_{\mathsf{tree}}$ directly into the observation space $\mathcal{O}$, i.e. for each state $s\in\mathcal{S}_{\mathsf{tree}}$ it emits the corresponding $o_s\in\mathcal{O}$. However, when $O^m\geqslant|\mathcal{S}_{\mathsf{tree}}|\gg O$, we can alternatively take an embedding $\mathcal{S}_{\mathsf{tree}}\to\mathcal{O}^m$, i.e. for each state $s\in\mathcal{S}_{\mathsf{tree}}$ such that $s\mapsto(o_s^{(1)},\cdots,o_s^{(m)})$, it emits $o_s^{(h\bmod m)}\in\mathcal{O}$ at step h.

F.2. Proof of Proposition F.1

All propositions and lemmas stated in this section are proved in Appendix F.3-F.7.

Claim 1 follows directly by counting the number of states, observations, and actions in models in $\mathcal{M}$. Claim 3 follows as we have $|\mathcal{M}|=|\{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)\}|\times\left|\{\pm1\}^K\right|+1\leqslant HSA^H\times 2^K$. Taking logarithm yields the claim.

Claim 2 follows from this lemma, which is proved in Appendix F.3.

Lemma F.3. *For each $M \in \mathcal{M}$, it holds that $\alpha_{m+1}(M)^{-1} \leqslant \frac{2}{\sigma} + 1$.*

We now prove Claim 4 & 5. Similar to the proof of Proposition E.1, we begin by relating the learning problem to a testing problem. Recall that $\mathbb{P}_M^{\mathfrak{A}}$ is the law of $(\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(\mathsf{T})})$ induced by algorithm $\mathfrak{A}$ and model M. For any event $E \subseteq (\mathcal{O} \times \mathcal{A})^H$, we denote the visitation count of E as

$$N(E) := \sum_{t=1}^{T} \mathbb{1}(\tau^{(t)} \in E).$$

Since $N(E)$ is a function of $\tau^{(1:T)}$, we can talk about its expectation under the distribution $\mathbb{P}_M^{\mathfrak{A}}$ for any $M \in \mathcal{M}$. We first relate the expected regret to the expected visitation count of some "bad" events, giving the following lemma whose proof is contained in Appendix F.4.

Lemma F.4 (Relating regret to visitation counts)**.** *For any $M \in \mathcal{M}$ such that $M \neq 0$, it holds that*

$$\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \geqslant \frac{\varepsilon}{4}\mathbb{E}_M^{\mathfrak{A}}[N(o_H = s_0)]. \tag{26}$$

On the other hand, for the reference model 0, we have

$$\mathbb{E}_0^{\mathfrak{A}}[\mathbf{Regret}] \geqslant \frac{\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] + \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})]. \tag{27}$$

where we define $E_{\mathsf{rev}} := \{\tau : \text{for some } h \in \mathcal{H}, o_h = \mathsf{lock}, a_h \in \mathcal{A}_{\mathsf{rev}}\}$.

On the other hand, for any policy π, we have

$$V_M^\star - V_M(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H = s_0) \ \ \forall M \neq 0, \qquad \text{and} \qquad V_0^\star - V_0(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_0^\pi(o_H \neq s_0). \tag{28}$$

Therefore, we can relate the regret (or sub-optimality of the output policy) to the TV distance (under $\mu \sim \mathrm{Unif}(\{-1,+1\}^K)$ the prior distribution of parameter μ), by an argument similar to the one in Appendix E.2, giving the following lemma whose proof is contained in Appendix F.5.

Lemma F.5. *Suppose that either statement below holds for the algorithm $\mathfrak{A}$:*

(a) For any model $M \in \mathcal{M}$, $\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \leqslant T\varepsilon/32$.

(b) For any model $M \in \mathcal{M}$, the algorithm $\mathfrak{A}$ outputs a policy π^{out} such that $\mathbb{P}_M^{\mathfrak{A}}\big(V_M^\star - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\big) \geqslant \frac{3}{4}$.

Then we have

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\big]\right) \geqslant \frac{1}{2}, \qquad \forall\theta. \tag{29}$$

By our assumptions in Claim 4 (or 5), in the following we only need to consider the case that (29) holds for all θ. We will use (29) to derive lower bounds of $\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)]$ and $\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})]$, giving the following lemma whose proof is contained in Appendix F.6.

Lemma F.6. *Fix a $\theta = (h^\star, s^\star, a^\star, a_{\mathsf{rev}}^\star, \mathbf{a}^\star)$. We consider events*

$$\begin{aligned} E_{\mathsf{rev}}^\theta &:= \big\{o_{h^\star} = s^\star, a_{h^\star:h} = (a^\star, \mathbf{a}^\star_{h^\star+1:h-1}, a_{\mathsf{rev}}^\star) \text{ for some } h \in \mathcal{H}_{>h^\star}\big\}, \\ E_{\mathsf{correct}}^\theta &:= \{o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star)\}. \end{aligned}$$

Then for any algorithm $\mathfrak{A}$ with $\delta := D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) > 0$, we have

$$\text{either } \mathbb{E}_0^{\mathfrak{A}}\big[N(E_{\mathsf{rev}}^\theta)\big] \geqslant \frac{\delta^3\sqrt{K}}{18\varepsilon^2\sigma^2} - \frac{\delta}{6}, \text{ or } \mathbb{E}_0^{\mathfrak{A}}\big[N(E_{\mathsf{correct}}^\theta)\big] \geqslant \frac{\delta^3}{18\varepsilon^2} - \frac{\delta}{6}.$$

Applying Lemma F.6 for any parameter tuple $\theta = (h^\star, s^\star, a^\star, a_{\mathsf{rev}}^\star, \mathbf{a}^\star)$ with $\delta = \frac{1}{2}$, we obtain

$$\text{either } \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a_{\mathsf{rev}}^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{\sqrt{K}}{300\varepsilon^2\sigma^2}, \quad \text{or} \quad \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a_{\mathsf{rev}}^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{1}{300\varepsilon^2}, \tag{30}$$

as we choose $\varepsilon \in (0, 0.1]$.

Fix a tuple $(h^\star, s^\star, a^\star)$ such that $h^\star \in \mathcal{H}$ and $h^\star \leqslant n + m\lfloor H/10m \rfloor$, $s^\star \in \mathcal{S}_{\mathsf{leaf}}$, $a^\star \in \mathcal{A}_c$. By (30), we know that for all $\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},h^\star}$, $a^\star_{\mathsf{rev}} \in \mathcal{A}_{\mathsf{rev}}$, $\theta = (h^\star, s^\star, a^\star, a^\star_{\mathsf{rev}}, \mathbf{a}^\star)$, real constant $r \geqslant 0$, it holds that

$$
\begin{aligned}
&|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1} \cdot \mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big] + r\,|\mathcal{A}_{\mathsf{code},h^\star}| \cdot \mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big] \\
&\geqslant \frac{1}{300}\min\left\{\frac{|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{r\,|\mathcal{A}_{\mathsf{code},h^\star}|}{\varepsilon^2}\right\} \geqslant \frac{1}{300}\min\left\{\frac{|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{rA^{H/2-1}}{\varepsilon^2}\right\} =: \omega_r,
\end{aligned}
$$

where the last inequality follows from a direct calculation (see Lemma F.7). Notice that

$$
\begin{aligned}
&\sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big] \\
&= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\big[N\big(o_{h^\star} = s^\star, a_{h^\star:h} = (a^\star, \mathbf{a}^\star_{h^\star+1:h-1}, a^\star_{\mathsf{rev}}) \text{ for some } h \in \mathcal{H}_{>h^\star}\big)\big] \\
&\leqslant \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\big[N\big(o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1}), a_h = a^\star_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star}\big)\big] \\
&= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}), a_h = a^\star_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star})] \cdot \sum_{\substack{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}\\ \mathbf{a}^\star \text{ begins with } \mathbf{a}}} 1 \\
&= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}), a_h = a^\star_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star})] \cdot \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}} \\
&= \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}} \cdot \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_h \in \mathcal{A}_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star})],
\end{aligned}
$$

where the second line is due to the inclusion of events, the fourth line follows from our definition of $\mathcal{A}_{\mathsf{code},h^\star}$, and the last line is because the events $\{o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}), a_h = a^\star_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star}\}$ are disjoint and their union is simply $\{o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_h \in \mathcal{A}_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star}\}$. Similarly we have

$$
\begin{aligned}
\sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big] &= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star))] \\
&= |\mathcal{A}_{\mathsf{rev}}| \cdot \sum_{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star))] \\
&= |\mathcal{A}_{\mathsf{rev}}| \cdot \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_{h^\star+1:H-1} \in \mathcal{A}_{\mathsf{code},h^\star})].
\end{aligned}
$$

Combining all these facts, we obtain

$$
\begin{aligned}
\omega_r &\leqslant \frac{1}{|\mathcal{A}_{\mathsf{rev}}|\,|\mathcal{A}_{\mathsf{code},h^\star}|} \sum_{\substack{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\\ \mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}}} \left(|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1}\mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big] + r\,|\mathcal{A}_{\mathsf{code},h^\star}|\,\mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big]\right) \\
&\leqslant \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_h \in \mathcal{A}_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star})] + r\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star} = a^\star)]
\end{aligned}
$$

Notice that the above inequality holds for any given $s^\star \in \mathcal{S}_{\mathsf{leaf}}, a^\star \in \mathcal{A}_c, h^\star \in \mathcal{H}$ such that $h^\star \leqslant n + m\lfloor H/10m \rfloor$, and any $r \geqslant 0$. Therefore, we can take summation over all $s^\star \in \mathcal{S}_{\mathsf{leaf}}, a^\star \in \mathcal{A}_c, h^\star = n + lm \in \mathcal{H}$ with $0 \leqslant l \leqslant \lfloor H/10m \rfloor$, and obtain

$$
\begin{aligned}
&|\mathcal{S}_{\mathsf{leaf}}|\,|\mathcal{A}_c|\,(\lfloor H/10m \rfloor + 1) \cdot \min\left\{\frac{|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1}\sqrt{K}}{300\varepsilon^2\sigma^2}, \frac{rA^{H/2-1}}{300\varepsilon^2}\right\} = \sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}} \sum_{a^\star\in\mathcal{A}_c} \sum_{\substack{h^\star=n+lm:\\ 0\leqslant l\leqslant\lfloor H/10m\rfloor}} \omega_r \\
\leqslant &\sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}} \sum_{a^\star\in\mathcal{A}_c} \sum_{\substack{h^\star=n+lm:\\ 0\leqslant l\leqslant\lfloor H/10m\rfloor}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_h \in \mathcal{A}_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star})] + r\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star} = a^\star)]
\end{aligned}
$$

$\leqslant \mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})] + r\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)]$,

where the last inequality is because

$$\bigsqcup_{s^\star \in \mathcal{S}_{\mathsf{leaf}}, a^\star \in \mathcal{A}_c, h^\star \in \mathcal{H}} \{o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_h \in \mathcal{A}_{\mathsf{rev}} \text{ for some } h \in \mathcal{H}_{>h^\star}\} \subseteq \{\text{for some } h \in \mathcal{H}, o_h = \mathsf{lock}, a_h \in \mathcal{A}_{\mathsf{rev}}\} = E_{\mathsf{rev}},$$

and $\bigsqcup_{s^\star \in \mathcal{S}_{\mathsf{leaf}}, a^\star \in \mathcal{A}_c, h^\star \in \mathcal{H}} \{o_{h^\star} = s^\star, a_{h^\star} = a^\star\} \subseteq \{o_H \neq s_0\}$. Plugging in our choice $|\mathcal{A}_c| = A - 1 \geqslant \frac{2}{3}A$, $|\mathcal{A}_{\mathsf{rev}}| = 1 + \lfloor A/6 \rfloor \geqslant A/6$ and $\lfloor H/10m \rfloor + 1 \geqslant H/10m$, we conclude the proof of the following claim:

Claim: as long as (29) holds, we have

$$\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})] + r\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] \geqslant \frac{|\mathcal{S}_{\mathsf{leaf}}|\, H}{30000m} \cdot \min\left\{ \frac{A^{m+1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{rA^{H/2}}{\varepsilon^2} \right\}, \qquad \forall r \geqslant 0. \tag{31}$$

To deduce Claim 4 from the above fact, we notice that either (1) $\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] > T\varepsilon/32$ for some $M \in \mathcal{M}$, or (2) $\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \leqslant T\varepsilon/32$ for any $M \in \mathcal{M}$, and then by Lemma F.5, (29) holds, and hence we have

$$\mathbb{E}_0^{\mathfrak{A}}[\mathbf{Regret}] \geqslant \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})] + \frac{\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] \geqslant \frac{|\mathcal{S}_{\mathsf{leaf}}|\, H}{120000m} \cdot \min\left\{ \frac{A^{m+1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{A^{H/2}}{\varepsilon} \right\}$$

by setting $r = \varepsilon$ in (31). Combining these two cases, we complete the proof of Claim 4 in Proposition F.1.

Similarly, suppose that the condition in Claim 5 holds, which implies (29) (by Lemma F.5). Then we can set $r = 1$ in (31) to obtain

$$2T \geqslant \mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})] + \mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] \geqslant \frac{|\mathcal{S}_{\mathsf{leaf}}|\, H}{30000m} \cdot \min\left\{ \frac{A^{m+1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{A^{H/2}}{\varepsilon^2} \right\},$$

and hence complete the proof of Claim 4. This completes the proof of Appendix F.2. □

Lemma F.7. *As long as $|\mathcal{A}_{\mathsf{rev}}| = A_1 \leqslant 1 + \lfloor A/6 \rfloor$, we have $|\mathcal{A}_{\mathsf{code},h^\star}| \geqslant A^{H/2-1}$ for $h^\star \in \mathcal{H}$ such that $h^\star \leqslant n + m\lfloor H/10m \rfloor$.*

Proof. We denote $H_0 = \lfloor (H-n)/m \rfloor$, and assume that $h^\star = n + ml$. Recall that

$$\mathcal{A}_{\mathsf{code},h^\star} := \left\{ \mathbf{a}^\star = (\mathbf{a}^\star_{h^\star+1}, \dots, \mathbf{a}^\star_{H-1}) \in \mathcal{A}^{H-h^\star-1} : \mathbf{a}^\star_h \in \mathcal{A}_{\mathsf{tr}}, \forall h \in \mathcal{H}_{>h^\star} \right\}.$$

Hence, noticing that $|\mathcal{H}_{>h^\star}| = H_0 - l$, $|\mathcal{A}| = A$, $|\mathcal{A}_{\mathsf{tr}}| = A - A_1$, we have

$$|\mathcal{A}_{\mathsf{code},h^\star}| = A^{H-h^\star-1-(H_0-l)} \times (A - A_1)^{H_0-l}.$$

Thus, we only need to prove that

$$H - h^\star - 1 - (H_0 - l) + \frac{\log(A - A_1)}{\log A}(H_0 - l) \geqslant \frac{H}{2} - 1. \tag{32}$$

Notice that as long as $A_1 \leqslant 1 + \lfloor A/6 \rfloor$, it holds that $\frac{\log(A-A_1)}{\log A} \geqslant \frac{\log 2}{\log 3} =: w$. Using this fact and rearranging, we can see (32) holds if

$$l \leqslant \frac{\frac{H}{2} - n - (1-w)H_0}{m - 1 + w} =: l_0.$$

Now, using our assumption that $H \geqslant 10n$, we have

$$l_0 \geqslant \frac{\frac{H}{2} - n - (1-w)(H-n)}{mw} = \frac{(w-0.5)H - wn}{mw} \geqslant \frac{(w-0.5)H - 0.1wH}{mw} \geqslant \frac{H}{10m},$$

where the last inequality uses $w > \frac{5}{8}$. Therefore, as long as $l \leqslant \lfloor H/10m \rfloor$ (i.e. $h^\star \leqslant n + m\lfloor H/10m \rfloor$), we have $l \leqslant l_0$, which implies (32) and hence completes the proof. □

F.3. Proof of Lemma F.3

The idea here is similar to the proof of Proposition E.2, but as our construction is more involved, the direct description of $\mathbb{M}_{h,m+1}$ can be very complicated (even though actually only a few of its entries are non-zero). Therefore, in order to upper bound $\mathbb{M}_{h,m+1}$, we invoke the following lemmas, which will make our discussion cleaner.

Lemma F.8. *For $m \geqslant 1$, $h \in [H-m]$, $\mathbf{a} \in \mathcal{A}^m$, we consider*

$$\mathbb{M}_{h,\mathbf{a}} := [\mathbb{P}(o_{h:h+m} = \mathbf{o} | s_h = s, a_{h:h+m-1} = \mathbf{a})]_{\mathbf{o} \in \mathcal{O}^{m+1}, s \in \mathcal{S}} \in \mathbb{R}^{\mathcal{O}^{m+1} \times \mathcal{S}}.$$

Then it holds that

$$\min_{\mathbb{M}^+_{h,m+1}} \left\| \mathbb{M}^+_{h,m+1} \right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\| \mathbb{M}^+_{h,\mathbf{a}} \right\|_{1\to 1},$$

where $\min_{\mathbb{M}^+_{h,\mathbf{a}}}$ is taken over all $\mathbb{M}^+_{h,\mathbf{a}}$ such that $\mathbb{M}^+_{h,\mathbf{a}} \mathbb{M}_{h,\mathbf{a}} \mathbb{T}_{h-1} = \mathbb{T}_{h-1}$ (cf. Definition 1).

Proof of Lemma F.8. Notice that given a $\mathbf{a} \in \mathcal{A}^m$, $\mathbb{M}^+_{h,\mathbf{a}}$ such that $\mathbb{M}^+_{h,\mathbf{a}} \mathbb{M}_{h,\mathbf{a}} \mathbb{T}_{h-1} = \mathbb{T}_{h-1}$, we can construct a generalized left inverse of $\mathbb{M}_{h,m}$ as follows:

$$\mathbb{M}^+_{h,m} = \begin{bmatrix} \vdots \\ \mathbb{1}(\mathbf{a}' = \mathbf{a}) \mathbb{M}^+_{h,\mathbf{a}} \\ \vdots \end{bmatrix}_{\mathbf{a}' \in \mathcal{A}^m},$$

and clearly $\left\| \mathbb{M}^+_{h,m} \right\|_{*\to 1} \leqslant \left\| \mathbb{M}^+_{h,\mathbf{a}} \right\|_{1\to 1}$. □

In the following, for any matrix M, we write

$$\gamma(M) := \min_{M^+ : M^+ M = I} \left\| M^+ \right\|_{1\to 1}.$$

Lemma F.9. *Fix a step h and a set of states $\mathcal{S}_h$. Suppose that $\mathcal{S}_h$ contains all $s \in \mathcal{S}$ such that $\exists (s', a) \in \mathcal{S} \times \mathcal{A}$, $\mathbb{T}_{h-1}(s|s', a) > 0$. Further, suppose that $\mathcal{S}_h$ can be partitioned as $\mathcal{S}_h = \bigsqcup_{i=1}^n \mathcal{S}_h^i$, such that for each $i \neq j$, $s \in \mathcal{S}_h^i$, $s' \in \mathcal{S}_h^j$,*

$$\mathrm{supp}(\mathbb{M}_{h,\mathbf{a}}(\cdot|s)) \bigcap \mathrm{supp}(\mathbb{M}_{h,\mathbf{a}}(\cdot|s')) = \varnothing,$$

i.e. the observations emitted from different $\mathcal{S}_h^i$ are different.[8] Then it holds that

$$\min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\| \mathbb{M}^+_{h,\mathbf{a}} \right\|_{1\to 1} \leqslant \max \left\{ \gamma\big(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^1)\big), \cdots, \gamma(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^n)) \right\},$$

where

$$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}') := [\mathbb{P}(o_{h:h+m} = \mathbf{o} | s_h = s, a_{h:h+m-1} = \mathbf{a})]_{\mathbf{o} \in \mathcal{O}^{m+1}, s \in \mathcal{S}'} \in \mathbb{R}^{\mathcal{O}^{m+1} \times \mathcal{S}'}, \qquad \text{for } \mathcal{S}' \subset \mathcal{S}_h.$$

Proof of Lemma F.9. We first note that $\min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\| \mathbb{M}^+_{h,\mathbf{a}} \right\|_{1\to 1} \leqslant \gamma(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h))$, because the matrix $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)$ directly gives a generalized left inverse of $\mathbb{M}_{h,\mathbf{a}}$ (because $\mathcal{S}_h$ contains all $s \in \mathcal{S}$ such that $\exists (s', a) \in \mathcal{S} \times \mathcal{A}$, $\mathbb{T}_{h-1}(s|s', a) > 0$).

Next, as each $\mathcal{S}_h^i$ has the disjoint set of possible observation, the matrix $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)$ can be written as (up to permutation of rows and columns, and any empty entry is zero)

$$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h) = \begin{bmatrix} \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^1) & & & \\ & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^2) & & \\ & & \ddots & \\ & & & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^n) \end{bmatrix}.$$

[8] In particular, this condition is fulfilled if for each $i \neq j$, $s \in \mathcal{S}_h^i$, $s' \in \mathcal{S}_h^j$, we have $\mathrm{supp}(\mathbb{O}_h(\cdot|s)) \bigcap \mathrm{supp}(\mathbb{O}_h(\cdot|s')) = \varnothing$.

Therefore, suppose that for each i we have a left inverse $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^i)^+$ of $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^i)$, then we can form a left inverse of $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)$ as

$$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)^+ = \begin{bmatrix} \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^1)^+ & & & \\ & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^2)^+ & & \\ & & \ddots & \\ & & & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^n)^+ \end{bmatrix},$$

and hence we derive that $\gamma(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)) \leqslant \max\left\{\gamma\big(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^1)\big), \cdots, \gamma(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^n))\right\}$. □

An important observation is that, for matrix $M \in \mathbb{R}^{m\times 1}$, we have $\gamma(M) \leqslant \frac{1}{\|M\|_1}$. Thus, when the sum of entries of M equals 1, then $\gamma(M) \leqslant 1$. With the lemmas above, we now provide the proof of Lemma F.3.

Proof of Lemma F.3. We first show that the null model 0 is 1-step 1-revealing. In this model, the state $s_\oplus$ and $e_\oplus$ are not reachable, and hence for each step h, we consider the set $\mathcal{S}' = \mathcal{S}_{\mathsf{tree}} \bigsqcup \{s_\ominus, e_\ominus, \mathsf{terminal}\}$. For different states $s, s' \in \mathcal{S}'$, the support of $\mathbb{O}_h(\cdot|s)$ and $\mathbb{O}_h(\cdot|s')$ are disjoint by our construction, and hence applying Lemma F.9 gives

$$\min_{\mathbb{O}_h^+} \left\|\mathbb{O}_h^+\right\|_{1\to 1} \leqslant \max_{s\in\mathcal{S}'} \gamma(\mathbb{O}_h(s)) \leqslant 1.$$

Applying Proposition 2 completes the proof for null model 0.

We next consider the non-null model $M = M_{\theta,\mu} \in \mathcal{M}\backslash\{M_0\}$. By our construction, for $h \leqslant h^\star$, state $s_\oplus$ and $e_\oplus$ are not reachable, and hence by the same argument as in the null model, we obtain that $\min_{\mathbb{M}_{h,m+1}^+} \left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to 1} \leqslant \min_{\mathbb{O}_h^+} \left\|\mathbb{O}_h^+\right\|_{1\to 1} \leqslant 1$.

Hence, we only need to bound the quantity $\min_{\mathbb{M}_{h,m+1}^+} \left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to 1}$ for a fixed step $h > h^\star$. In this case, there exists a $l \in \mathcal{H}$ such that $h \leqslant l \leqslant h+m-1$, and we write $r = l - h + 1$. By Lemma C.1, we only need to bound $\min_{\mathbb{M}_{h,r+1}^+} \left\|\mathbb{M}_{h,r+1}^+\right\|_{*\to 1}$. Consider the action sequence $\mathbf{a} = (\mathbf{a}_{h:l-1}^\star, a_{\mathsf{rev}}^\star) \in \mathcal{A}^r$, and we partition $\mathcal{S}$ as

$$\mathcal{S} = \bigsqcup_{s\in\mathcal{S}_{\mathsf{tree}}} \{s\} \sqcup \{s_\oplus, s_\ominus\} \sqcup \{e_\oplus, e_\ominus\} \sqcup \{\mathsf{terminal}\}.$$

It is direct to verify that, in $M_{\theta,\mu}$, for states s, s' come from different subsets in the above partition, the support of $\mathbb{M}_{h,\mathbf{a}}(\cdot|s)$ and $\mathbb{M}_{h,\mathbf{a}}(\cdot|s')$ are disjoint. Then, we can apply Lemma F.8 and Lemma F.9, and obtain

$$\min_{\mathbb{M}_{h,r+1}^+} \left\|\mathbb{M}_{h,r+1}^+\right\|_{*\to 1} \leqslant \min_{\mathbb{M}_{h,\mathbf{a}}^+} \left\|\mathbb{M}_{h,\mathbf{a}}^+\right\|_{1\to 1} \leqslant \max\left\{1, \gamma(\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\})), \gamma(\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\}))\right\}.$$

Therefore, in the following we only need to consider left inverses of the matrix $\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\})$ and $\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\})$.

(1) The matrix $\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\})$. By our construction, taking $\mathbf{a}$ at $s_h = s_\oplus$ will lead to $o_{h:l} = \mathsf{lock}$ and $o_{l+1} \sim \mathbb{O}_\mu(\cdot|e_\oplus)$; taking $\mathbf{a}$ at $s_h = s_\ominus$ will lead to $o_{h:l} = \mathsf{lock}$ and $o_{l+1} \sim \mathbb{O}_\mu(\cdot|e_\ominus)$. Hence, $\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\})$ can be written as (up to permutation of rows)

$$\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K} + \sigma\widetilde{\mu}}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1}\times 2},$$

where $\widetilde{\mu} = [\mu; -\mu] \in \{-1, 1\}^{2K}$, $\mathbb{1} = \mathbb{1}_{2K}$ is the vector in $\mathbb{R}^{2K}$ with all entries being one. Similar to Proposition E.2, we can directly verify that $\gamma(\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\})) \leqslant \frac{2}{\sigma} + 1$.

(2) The matrix $\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\})$. By our construction, at $s_h = e_\oplus$, we have $o_h \sim \mathbb{O}_\mu(\cdot|e_\oplus)$ and $o_{h+1:l+1} = \mathsf{terminal}$; at $s_h = e_\ominus$, we have $o_h \sim \mathbb{O}_\mu(\cdot|e_\ominus)$ and $o_{h+1:l+1} = \mathsf{terminal}$. Thus, $\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\})$ can also be written as (up to permutation of rows)

$$\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K} + \sigma\widetilde{\mu}}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1}\times 2},$$

and hence we also have $\gamma(\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\})) \leqslant \frac{1}{\sigma} + 2$.

Combining the two cases above gives

$$\min_{\mathbb{M}^+_{h,m+1}} \left\|\mathbb{M}^+_{h,m+1}\right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,r+1}} \left\|\mathbb{M}^+_{h,r+1}\right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\|\mathbb{M}^+_{h,\mathbf{a}}\right\|_{1\to 1} \leqslant \frac{2}{\sigma} + 1,$$

and hence completes the proof of Lemma F.3. □

Remark F.10. From the proof above, it is not easy to see the POMDP $M = M_{\theta,\mu}$ is not m-step revealing for any parameters (θ, μ). Actually, for $\theta = (h^\star, s^\star, a^\star, a^\star_{\mathsf{rev}}, \mathbf{a}^\star)$, we can show that the matrix $\mathbb{M}_{h^\star+1,m}$ *does not* admit a generalized left inverse. This is because for any $\mathbf{a} \in \mathcal{A}^{m-1}$, we have

$$\mathbb{P}_{\theta,\mu}(o_{h^\star+1:h^\star+m} = \cdot | s_{h^\star+1} = s_\oplus, a_{h^\star+1:h^\star+m-1} = \mathbf{a}) = \mathbb{P}_{\theta,\mu}(o_{h^\star+1:h^\star+m} = \cdot | s_{h^\star+1} = s_\ominus, a_{h^\star+1:h^\star+m-1} = \mathbf{a}),$$

because both of the distributions are supported on the dummy observation $\mathsf{lock}^{\otimes m}$. However, it is clear that $\mathbf{e}_{s_\oplus}, \mathbf{e}_{s_\ominus} \in \mathrm{colspan}(\mathbb{T}_{h^\star})$, and hence if $\mathbb{M}_{h^\star+1,m}$ admits a generalized left inverse $\mathbb{M}^+_{h^\star+1,m}$, then $\mathbf{e}_{s_\oplus} = \mathbb{M}^+_{h^\star+1,m}\mathbb{M}_{h^\star+1,m}\mathbf{e}_{s_\oplus} = \mathbb{M}^+_{h^\star+1,m}\mathbb{M}_{h^\star+1,m}\mathbf{e}_{s_\ominus} = \mathbf{e}_{s_\ominus}$, a contradiction! Therefore, we can conclude that $\mathbb{M}_{h^\star+1,m}$ does not admit a generalized left inverse, and hence M is not m-step revealing.

F.4. Proof of Lemma F.4

In the following, we prove (26) and (27). This proof is very similar to the proof of Lemma E.3. The proof of (28) is very similar and hence omitted for succinctness.

Notice that by the definition of $\mathbf{Regret}$ and our construction of reward function, we have

$$\begin{aligned}
\mathbb{E}^{\mathfrak{A}}_M[\mathbf{Regret}] &= T \cdot V^\star_M - \mathbb{E}^{\mathfrak{A}}_M\left[\sum_{t=1}^T r_H(o_H^{(t)})\right] = T \cdot V^\star_M - \mathbb{E}^{\mathfrak{A}}_M\left[\frac{1+\varepsilon}{4} \cdot N(o_H = s_0) + N(o_H = \mathsf{good})\right] \\
&= \left(V^\star_M - \frac{1+\varepsilon}{4}\right)\mathbb{E}^{\mathfrak{A}}_M[N(o_H = s_0)] + V^\star_M \mathbb{E}^{\mathfrak{A}}_M[N(o_H \neq s_0)] - \mathbb{E}^{\mathfrak{A}}_M[N(o_H = \mathsf{good})]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}^{\mathfrak{A}}_M[N(o_H = \mathsf{good})] &= \mathbb{E}^{\mathfrak{A}}_M\left[\sum_{t=1}^T \mathbb{E}_M\Big[\mathbb{1}(o_H^{(t)} = \mathsf{good})\Big| \tau_{H-1}^{(t)}\Big]\right] \\
&= \mathbb{E}^{\mathfrak{A}}_M\left[\sum_{t=1}^T \sum_{\tau_{H-1}} \mathbb{P}_M(o_H = \mathsf{good}|\tau_{H-1}) \cdot \mathbb{1}(\tau_{H-1}^{(t)} = \tau_{H-1})\right] \\
&= \sum_{\tau_{H-1}} \mathbb{E}^{\mathfrak{A}}_M[N(\tau_{H-1})] \cdot \mathbb{P}_M(o_H = \mathsf{good}|\tau_{H-1}).
\end{aligned}$$

We prove the result for the $M \neq 0$ and the case $M = 0$ separately.

Case 1: $M = (\theta, \mu) \neq 0$. In this case, we have

$$\mathbb{P}_M(o_H = \mathsf{good}|\tau_{H-1}) = \frac{3}{4}\mathbb{P}_M(s_H = s_\oplus|\tau_{H-1}) + \frac{1}{4}\mathbb{P}_M(s_H = s_\ominus|\tau_{H-1}) \leqslant \frac{1}{4} + \frac{1}{2}\mathbb{P}_M(s_H = s_\oplus|\tau_{H-1}) \leqslant \frac{1}{4} + \frac{1}{2}\varepsilon,$$

because $\mathbb{P}_M(s_H = s_\oplus|\tau_{H-1}) \leqslant \varepsilon$ by our construction. Thus, we have shown that

$$\mathbb{E}^{\mathfrak{A}}_M[N(o_H = \mathsf{good})] \leqslant \left(\frac{1}{4} + \frac{1}{2}\varepsilon\right)\mathbb{E}^{\mathfrak{A}}_M[N(o_H \neq s_0)].$$

Notice that by this way we can also show that $V^\star_M = \frac{1+2\varepsilon}{4}$. Therefore, combining the equations above, we conclude that

$$\mathbb{E}^{\mathfrak{A}}_M[\mathbf{Regret}] \geqslant \frac{\varepsilon}{4}\mathbb{E}^{\mathfrak{A}}_M[N(o_H = s_0)].$$

Case 2: $M = 0$. In this case, $s_\oplus$ is not reachable, and hence we have

$$\mathbb{P}_0(o_H = \mathsf{good}|\tau_{H-1}) = \frac{1}{4}\mathbb{P}_0(s_H = s_\ominus|\tau_{H-1}) \leqslant \frac{1}{4}.$$

Also notice that, for any trajectory $\tau \in E_{\mathsf{rev}}$, we have $\mathbb{P}_0(o_H = \mathsf{good}|\tau_{H-1}) = 0$. Thus, we have shown that

$$\mathbb{E}_0^{\mathfrak{A}}[N(o_H = \mathsf{good})] \leqslant \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(\{o_H \neq s_0\} - E_{\mathsf{rev}})] = \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] - \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})].$$

By this way we can also show that $V_0^\star = \frac{1+\varepsilon}{4}$. Therefore, we can conclude that

$$\begin{aligned}\mathbb{E}_0^{\mathfrak{A}}[\mathbf{Regret}] &= \frac{1+\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] - \mathbb{E}_0^{\mathfrak{A}}[N(o_H = \mathsf{good})] \\ &\geqslant \frac{\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] + \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})].\end{aligned}$$

This completes the proof of Lemma F.4. □

F.5. Proof of Lemma F.5

We first consider case (a), i.e. suppose that $\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \leqslant T\varepsilon/32$ for all $M \in \mathcal{M}$. By Markov's inequality and (26) and (27), it holds that

$$\begin{aligned}&\mathbb{P}_0^{\mathfrak{A}}(N(o_H \neq s_0) \geqslant T/2) \leqslant \frac{1}{4}, \\ &\mathbb{P}_M^{\mathfrak{A}}(N(o_H = s_0) \geqslant T/2) \leqslant \frac{1}{4}, \qquad \forall M \neq 0.\end{aligned}$$

In particular, for any fixed θ, we consider the prior distribution of $M = (\theta, \mu)$ with $\mu \sim \mathrm{Unif}(\{-1,1\}^K)$, then

$$\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right](N(o_H = s_0) \geqslant T/2) \leqslant \frac{1}{4}.$$

However, we also have

$$\mathbb{P}_0^{\mathfrak{A}}(N(o_H = s_0) \geqslant T/2) = \mathbb{P}_0^{\mathfrak{A}}(N(o_H \neq s_0) \leqslant T/2) = 1 - \mathbb{P}_0^{\mathfrak{A}}(N(o_H \neq s_0) > T/2) \geqslant \frac{3}{4},$$

and then by the definition of TV distance it holds

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) \geqslant \frac{1}{2}.$$

The proof of case (b) follows from an argument which is the same as the proof of (20), and hence omitted. □

F.6. Proof of Lemma F.6

We first prove the following version of Lemma F.6 with an additional condition that the visitation counts are almost surely bounded under $\mathbb{P}_0^{\mathfrak{A}}$, and then prove Lemma F.6 by reducing to this case using a truncation argument.

To upper bound the above quantity, we invoke the following lemma, which serves a key step for bounding the above "χ^2-inner product" (Canonne, 2022, Section 3.1) between $\mathbb{P}_{\theta,\mu}/\mathbb{P}_0$ and $\mathbb{P}_{\theta,\mu'}/\mathbb{P}_0$ (proof in Appendix F.7).

Lemma F.11 (Bound on the χ^2-inner product)**.** *Suppose that algorithm $\mathfrak{A}$ (with possibly random stopping time T) satisfies $N(E_{\mathsf{rev}}^\theta) \leqslant \overline{N}_o$ and $N(E_{\mathsf{correct}}^\theta) \leqslant \overline{N}_r$ almost surely, for some fixed $\overline{N}_o, \overline{N}_r$. Then*

$$\textit{either } \overline{N}_o \geqslant \frac{3}{4}\frac{\delta^2\sqrt{K}}{\varepsilon^2\sigma^2}, \textit{ or } \overline{N}_r \geqslant \frac{3}{4}\frac{\delta^2}{\varepsilon^2},$$

where $\delta = D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right)$.

Proof of Lemma F.11. By Lemma D.1, it holds that

$$1+\chi^2\big(\mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\big]\;\|\;\mathbb{P}^{\mathfrak{A}}_0\big)=\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}^{\mathfrak{A}}_0}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right].$$

In the following lemma (proof in Appendix F.7), we bound the LHS of the equality above.

Lemma F.12. *Under the conditions of Lemma F.11, it holds that for any $\mu,\mu'\in\{-1,1\}^K$,*

$$\mathbb{E}^{\mathfrak{A}}_0\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right]\leqslant\exp\left(\overline{N}_o\cdot\frac{\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right|+\frac{4}{3}\varepsilon^2\overline{N}_r\right).\tag{33}$$

With Lemma F.12, we can take expectation of (33) over $\mu,\mu'\sim\mathrm{Unif}(\{-1,+1\}^K)$, and then

$$\begin{aligned}1+\chi^2\big(\mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\big]\;\|\;\mathbb{P}^{\mathfrak{A}}_0\big)=&\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}^{\mathfrak{A}}_0}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right]\\\leqslant&\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\overline{N}_o\cdot\frac{\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right|+\frac{4}{3}\varepsilon^2\overline{N}_r\right)\right].\end{aligned}$$

Notice that μ_i,μ'_i are i.i.d. $\mathrm{Unif}(\{\pm1\})$, and hence $\mu_1\mu'_1,\cdots,\mu_K\mu'_K$ are i.i.d. $\mathrm{Unif}(\{\pm1\})$. Then by Hoeffding's lemma, it holds that $\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(x\sum_{i=1}^K\mu_i\mu'_i\right)\right]\leqslant\exp\left(Kx^2/2\right)$ for all $x\in\mathbb{R}$, and thus by Lemma A.3, we have

$$\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\frac{\overline{N}_o\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right|\right)\right]\leqslant\exp\left(\max\left\{\frac{\sigma^4\varepsilon^4\overline{N}_o^2}{K},\frac{4}{3}\frac{\sigma^2\varepsilon^2\overline{N}_o}{\sqrt{K}}\right\}\right).$$

Therefore, combining the above inequalities with Lemma A.5, we obtain

$$2\delta^2=2D_{\mathrm{TV}}\left(\mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\big],\mathbb{P}^{\mathfrak{A}}_0\right)^2\leqslant\log\left(1+\chi^2\big(\mathbb{E}_{\mu\sim\mathsf{unif}}\big[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\big]\;\|\;\mathbb{P}^{\mathfrak{A}}_0\big)\right)\leqslant\max\left\{\frac{4}{3}\frac{\overline{N}_o\sigma^2\varepsilon^2}{\sqrt{K}},\frac{\overline{N}_o^2\sigma^4\varepsilon^4}{K}\right\}+\frac{4}{3}\varepsilon^2\overline{N}_r.$$

Then, we either have $\overline{N}_r\geqslant\frac{3\delta^2}{4\varepsilon^2}$, or it holds

$$\max\left\{\frac{4}{3}\frac{\overline{N}_o\sigma^2\varepsilon^2}{\sqrt{K}},\frac{\overline{N}_o^2\sigma^4\varepsilon^4}{K}\right\}\geqslant\delta^2,$$

which implies that $\frac{\overline{N}_o\sigma^2\varepsilon^2}{\sqrt{K}}\geqslant\min\left\{\frac{4}{3},\frac{3}{4}\delta^2\right\}=\frac{3}{4}\delta^2$ (as $\delta\leqslant1$). The proof of Lemma F.11 is completed by rearranging. □

Proof of Lemma F.6. We perform a truncation type argument to reduce Lemma F.6 to Lemma F.11, which is similar to the proof of Lemma E.4.

Let us take $\overline{N}_o=\left\lceil6\delta^{-1}\mathbb{E}^{\mathfrak{A}}_0\big[N(E^\theta_{\mathsf{rev}})\big]\right\rceil$ and $\overline{N}_r=\left\lceil6\delta^{-1}\mathbb{E}^{\mathfrak{A}}_0\big[N(E^\theta_{\mathsf{correct}})\big]\right\rceil$. By Markov's inequality, we have

$$\mathbb{P}^{\mathfrak{A}}_0\big(N(E^\theta_{\mathsf{rev}})\geqslant\overline{N}_o\big)\leqslant\frac{\delta}{6},\qquad\mathbb{P}^{\mathfrak{A}}_0\big(N(E^\theta_{\mathsf{correct}})\geqslant\overline{N}_r\big)\leqslant\frac{\delta}{6}.$$

Therefore, we can consider the following exit criterion exit for the algorithm $\mathfrak{A}$:

$$\mathsf{exit}(\tau^{(1:T')})=\mathsf{TRUE}\quad\text{iff}\quad\sum_{t=1}^{T'}\mathbb{I}\Big(\tau^{(t)}\in E^\theta_{\mathsf{rev}}\Big)\geqslant\overline{N}_o\text{ or }\sum_{t=1}^{T'}\mathbb{I}\Big(\tau^{(t)}\in E^\theta_{\mathsf{correct}}\Big)\geqslant\overline{N}_r.$$

The criterion exit induces a stopping time T_{exit}, and we have

$$\mathbb{P}^{\mathfrak{A}}_0(\exists t<T,\mathsf{exit}(\tau^{(1:t)})=\mathsf{TRUE})\leqslant\mathbb{P}^{\mathfrak{A}}_0\big(N(E^\theta_{\mathsf{rev}})\geqslant\overline{N}_o\text{ or }N(E^\theta_{\mathsf{correct}})\geqslant\overline{N}_r\big)\leqslant\frac{\delta}{6}+\frac{\delta}{6}\leqslant\frac{\delta}{3}.$$

Therefore, we can consider the early stopped algorithm $\mathfrak{A}(\mathsf{exit})$ with exit criterion exit (cf. Appendix D), and by Lemma D.2 we have

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}(\mathsf{exit})}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}(\mathsf{exit})}\right]\right) \geqslant D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) - \mathbb{P}_0^{\mathfrak{A}}(\exists t < T, \mathsf{exit}(\tau^{(1:t)}) = \mathsf{TRUE}) \geqslant \frac{2\delta}{3}.$$

Notice that by our definition of exit and stopping time T_{exit}, in the execution of $\mathfrak{A}(\mathsf{exit})$, we also have

$$\sum_{t=1}^{T_{\mathsf{exit}}-1} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{rev}}^{\theta}\right) < \overline{N}_o, \qquad \sum_{t=1}^{T_{\mathsf{exit}}-1} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{correct}}^{\theta}\right) < \overline{N}_r.$$

Therefore, algorithm $\mathfrak{A}(\mathsf{exit})$ ensures that

$$N(E_{\mathsf{rev},h}^{\theta}) = \sum_{t=1}^{T_{\mathsf{exit}}} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{rev}}^{\theta}\right) \leqslant \overline{N}_o, \qquad N(E_{\mathsf{correct}}^{\theta}) = \sum_{t=1}^{T_{\mathsf{exit}}} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{correct}}^{\theta}\right) \leqslant \overline{N}_r.$$

Applying Lemma F.11 to the algorithm $\mathfrak{A}(\mathsf{exit})$ (and $\delta' = \frac{2}{3}\delta$), we can obtain

$$\text{either } \frac{\delta^2\sqrt{K}}{3\varepsilon^2\sigma^2} \leqslant \overline{N}_o \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{rev}}^{\theta})\right] + 1, \qquad \text{or } \frac{\delta^2}{3\varepsilon^2} \leqslant \overline{N}_r \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{correct}}^{\theta})\right] + 1,$$

and rearranging gives the desired result of Lemma F.6. □

F.7. Proof of Lemma F.12

Throughout the proof, the parameters θ, μ, μ' are fixed.

By our discussion in Appendix D, using (16), we have

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}} \frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2} \cdot \exp\left(-\sum_{t=1}^{\mathsf{T}}\sum_{h=1}^{H} \log I(\tau_{h-1}^{(t)})\right)\right] = 1, \tag{34}$$

where we have defined $I(\tau_l)$ for any partial trajectory τ_l up to step $l \in [H]$ as

$$I(\tau_l) := \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right].$$

Notice that the model $\mathbb{P}_{\theta,\mu}$ and $\mathbb{P}_0$ are different only at the transition from $s_{h^\star} = s^\star, a_{h^\star} = a^\star$ to $s_\oplus$ and the dynamic at the component $\{s_\oplus, e_\oplus\}$. Therefore, for any (reachable) trajectory $\tau_l = (o_1, a_1, \cdots, o_l, a_l)$, we can consider the implication of $\mathbb{P}_{\theta,\mu}(o_{l+1} = \cdot|\tau_l) \neq \mathbb{P}_0(o_{l+1} = \cdot|\tau_l)$:
1. Clearly, $o_{h^\star} = s^\star$, $a_{h^\star} = a^\star$ (i.e. $l \geqslant h^\star + 1$ and taking action $a_{1:h^\star-1}$ from s_0 will result in $s^\star$ at step $h^\star$).
2. Either $a_{h^\star+1:l} = (\mathbf{a}_{h^\star+1:l-1}^\star, a_{\mathsf{rev}}^\star)$ for some $l \in \mathcal{H}_{>h^\star}$, or $l = H-1$ and $a_{h^\star+1:H-1} = \mathbf{a}^\star$.

Hence, for $l \in \mathcal{H}_{>h^\star}$, we define

$$E_{\mathsf{rev},l}^{\theta} := \left\{o_{h^\star} = s^\star, a_{h^\star:l} = (a^\star, \mathbf{a}_{h^\star+1:l-1}^\star, a_{\mathsf{rev}}^\star)\right\}.$$

Also recall that we define $E_{\mathsf{correct}}^{\theta} := \{o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star)\}$. Then if $\mathbb{P}_{\theta,\mu}(\cdot|\tau_l) \neq \mathbb{P}_0(\cdot|\tau_l)$, then either $l \in \mathcal{H}_{>h^\star}, \tau_l \in E_{\mathsf{rev},l}^{\theta}$, or $l = H-1, \tau_{H-1} \in E_{\mathsf{correct}}^{\theta}$. In other words, for any τ_l (that is reachable under $\mathbb{P}_0$), we have $I(\tau_l) = 1$ except for these two cases, and it remains to compute $I(\tau_l)$ for these two cases.

Case 1: $l \in \mathcal{H}_{>h^\star}, \tau_l \in E_{\mathsf{rev},l}^{\theta}$. In this case, we have

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l) &= \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = e_\oplus)\mathbb{P}_{\theta,\mu}(s_{l+1} = e_\oplus|\tau_l) + \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = e_\ominus)\mathbb{P}_{\theta,\mu}(s_{l+1} = e_\ominus|\tau_l) \\ &= (\mathbb{O}_\mu(o|e_\oplus) - \mathbb{O}(o|e_\ominus)) \cdot \mathbb{P}_{\theta,\mu}(s_{l+1} = e_\oplus|\tau_l) + \mathbb{O}(o|e_\ominus),\end{aligned}$$

where the second equality is because conditional on τ_l, we have $s_{l+1} \in \{e_\oplus, e_\ominus\}$. Now, we have

$$\mathbb{P}_{\theta,\mu}(s_{l+1} = e_\oplus|\tau_l) = \mathbb{P}_{\theta,\mu}(s_{l+1} = e_\oplus|o_{h^\star} = s^\star, a_{h^\star:l} = (a^\star, \mathbf{a}_{h^\star+1:l-1}^\star, a_{\mathsf{rev}}^\star))$$

$$= \mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s_\oplus | o_{h^\star} = s^\star, a_{h^\star} = a^\star) = \varepsilon.$$

Hence, by the definition of $\mathbb{O}_\mu(\cdot|e_\oplus)$ and $\mathbb{O}(\cdot|e_\ominus)$, we can conclude that

$$\mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^+|\tau_l) = \frac{1+\varepsilon\sigma\mu_i}{2K}, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^-|\tau_l) = \frac{1-\varepsilon\sigma\mu_i}{2K}, \qquad \forall i \in [K].$$

On the other hand, clearly $\mathbb{P}_0(o_{l+1} = \cdot|\tau_l) = \mathrm{Unif}(\{o_1^+, o_1^-, \cdots, o_K^+, o_K^-\})$. Hence, it holds that

$$\begin{aligned}
I(\tau_l) =& \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right] \\
=& \frac{1}{2K}\sum_{o\in\mathcal{O}_o}\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}=o|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}=o|\tau_l)}{\mathbb{P}_0(o_{l+1}=o|\tau_l)^2} \\
=& \frac{1}{2K}\sum_{i=1}^K (1+\varepsilon\sigma\mu_i)(1+\varepsilon\sigma\mu_i') + (1-\varepsilon\sigma\mu_i)(1-\varepsilon\sigma\mu_i') \\
=& 1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_i\mu_i' = 1 + \frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle.
\end{aligned}$$

Case 2: $l = H-1.\tau_{H-1} \in E^\theta_{\mathsf{correct}}$. In this case, the distribution $\mathbb{P}(o_H = \cdot|\tau_{H-1})$ is supported on $\{\mathsf{good}, \mathsf{bad}\}$. Similar to case 1, we have

$$\begin{aligned}
\mathbb{P}_{\theta,\mu}(o_H = \cdot|\tau_l) &= \mathbb{P}_{\theta,\mu}(o_H = \cdot|s_H = s_\oplus)\mathbb{P}_{\theta,\mu}(s_H = s_\oplus|\tau_{H-1}) + \mathbb{P}_{\theta,\mu}(o_H = \cdot|s_H = s_\ominus)\mathbb{P}_{\theta,\mu}(s_H = s_\ominus|\tau_{H-1}) \\
&= (\mathbb{O}_H(\cdot|s_\oplus) - \mathbb{O}_H(\cdot|s_\ominus))\cdot\mathbb{P}_{\theta,\mu}(s_H = s_\oplus|\tau_{H-1}) + \mathbb{O}_H(\cdot|s_\ominus),
\end{aligned}$$

where the second equality is because conditional on $\tau_{H-1} \in E^\theta_{\mathsf{correct}}$, we have $s_H \in \{s_\oplus, s_\ominus\}$. Now, we have

$$\mathbb{P}_{\theta,\mu}(s_H = s_\oplus|\tau_{H-1}) = \mathbb{P}_{\theta,\mu}(s_H = s_\oplus|o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star)) = \mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s_\oplus|o_{h^\star} = s^\star, a_{h^\star} = a^\star) = \varepsilon.$$

Hence, by the definition of $\mathbb{O}_H(\cdot|s_\oplus)$ and $\mathbb{O}_H(\cdot|s_\ominus)$, we have

$$\mathbb{P}_{\theta,\mu}(o_H = o|\tau_{H-1}) = \begin{cases} \frac{1+2\varepsilon}{4}, & o = \mathsf{good}, \\ \frac{3-2\varepsilon}{4}, & o = \mathsf{bad}. \end{cases}$$

On the other hand, clearly $\mathbb{P}_0(o_H = \mathsf{good}|\tau_{H-1}) = \frac{1}{4}, \mathbb{P}_0(o_H = \mathsf{bad}|\tau_{H-1}) = \frac{3}{4}$. Therefore, in this case, we have

$$I(\tau_{H-1}) = \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_H|\tau_{H-1})\mathbb{P}_{\theta,\mu'}(o_H|\tau_{H-1})}{\mathbb{P}_0(o_H|\tau_{H-1})^2}\middle|\tau_{H-1}\right] = \frac{1}{4}\times\left(\frac{(1+2\varepsilon)/4}{1/4}\right)^2 + \frac{3}{4}\times\left(\frac{(3-2\varepsilon)/4}{3/4}\right)^2 = 1 + \frac{4}{3}\varepsilon^2.$$

Combining the two cases above, we obtain

$$I(\tau_l) = \begin{cases} 1 + \frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle, & l \in \mathcal{H}_{>h^\star}, \tau_l \in E^\theta_{\mathsf{rev},l}, \\ 1 + \frac{4}{3}\varepsilon^2, & l = H-1, \tau_{H-1} \in E^\theta_{\mathsf{correct}}, \\ 1, & \text{otherwise}. \end{cases}$$

Hence, for each $t \in [\mathsf{T}]$,

$$\begin{aligned}
\sum_{l=0}^{H-1}\log I(\tau_l^{(t)}) &= \sum_{l=0}^{H-1}\mathbb{1}\left(l \in \mathcal{H}_{>h^\star}, \tau_l^{(t)} \in E^\theta_{\mathsf{rev},l}\right)\cdot\log\left(1 + \frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle\right) + \mathbb{1}\left(\tau_{H-1}^{(t)} \in E^\theta_{\mathsf{correct}}\right)\cdot\log\left(1+\frac{4}{3}\varepsilon^2\right) \\
&\leqslant \sum_{l\in\mathcal{H}_{>h^\star}}\mathbb{1}\left(\tau_l^{(t)} \in E^\theta_{\mathsf{rev},l}\right)\cdot\frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle + \mathbb{1}\left(\tau_{H-1}^{(t)} \in E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}\varepsilon^2 \\
&= \mathbb{1}\left(\tau_H^{(t)} \in E^\theta_{\mathsf{rev}}\right)\cdot\frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle + \mathbb{1}\left(\tau_H^{(t)} \in E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}\varepsilon^2,
\end{aligned}$$

where the last equality is because $E^\theta_{\mathsf{rev}} = \bigsqcup_{l: l \in \mathcal{H}_{>h^\star}} E^\theta_{\mathsf{rev},l}$. Taking summation over $t \in [\mathsf{T}]$, we obtain

$$\begin{aligned}\sum_{t=1}^{\mathsf{T}} \sum_{l=0}^{H-1} \log I(\tau_l^{(t)}) &\leqslant \sum_{t=1}^{\mathsf{T}} \mathbb{1}\left(\tau_H^{(t)} \in E^\theta_{\mathsf{correct}}\right) \cdot \frac{4}{3}\varepsilon^2 + \mathbb{1}\left(\tau_H^{(t)} \in E^\theta_{\mathsf{rev}}\right) \cdot \frac{\varepsilon^2\sigma^2}{K} \left\langle \mu, \mu' \right\rangle \\ &= N(E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2 + N(E^\theta_{\mathsf{rev}}) \cdot \frac{\varepsilon^2\sigma^2}{K} \left\langle \mu, \mu' \right\rangle \\ &\leqslant \overline{N}_r \cdot \frac{4}{3}\varepsilon^2 + \overline{N}_o \cdot \frac{\varepsilon^2\sigma^2}{K} \left|\left\langle \mu, \mu' \right\rangle\right|.\end{aligned}$$

Plugging the above inequality into (34) completes the proof of Lemma F.12. □

F.8. Regret calculation for hard instance in Section 5

For the hard instance presented in Section 5, we notice that any algorithm either incurs a $\Omega(\varepsilon T)$ regret, or must have successfully identified $(h^\star, s^\star, a^\star)$ within T episodes of play, which requires either at least $\Omega(SAH \times \sqrt{O}/(\sigma^2\varepsilon^2))$ episodes of taking revealing actions, each being $\Theta(1)$-suboptimal, or at least $\Omega(SAH \times A^{\Theta(H)}/\varepsilon^2)$ episodes of trying out all possible action sequences, each being $\Theta(\varepsilon)$-suboptimal. This yields a regret lower bound

$$\Omega\left(SAH \times \min\left\{\tfrac{\sqrt{O}}{\sigma^2\varepsilon^2}, \tfrac{A^{\Theta(H)}}{\varepsilon^2} \cdot \varepsilon\right\}\right) \wedge \Omega(\varepsilon T).$$

Optimizing over $\varepsilon > 0$, we obtain a $\Omega(T^{2/3})$-type regret lower bound (for $T \ll A^{\mathcal{O}(H)}$) similar as (though slightly worse rate than) Theorem 6.

G. Proof of Theorem 5

We first construct a family of hard instances in Appendix G.1. We then state the PAC lower bound of this family of hard instances in Proposition G.1. Theorem 5 then follows from combining Proposition G.1 with Proposition F.1.

G.1. Construction of hard instances and proof of Theorem 5

We consider the following family of m-step revealing POMDPs $\mathcal{M}$ that admits a tuple of hyperparameters $(\varepsilon, \sigma, n, m, K, L, H)$. All POMDPs in $\mathcal{M}$ share the state space $\mathcal{S}$, action space $\mathcal{A}$, observation space $\mathcal{O}$, and horizon length H, defined as following.

- The state space $\mathcal{S} = \mathcal{S}_{\mathsf{tree}} \bigsqcup_{j=1}^{L} \left\{s^j_\oplus, s^j_\ominus, e^j_\oplus, e^j_\ominus, \mathsf{terminal}^j\right\}$, where $\mathcal{S}_{\mathsf{tree}}$ is a binary tree with level n (so that $|\mathcal{S}_{\mathsf{tree}}| = 2^n - 1$). Let s_0 be the root of $\mathcal{S}_{\mathsf{tree}}$, and $\mathcal{S}_{\mathsf{leaf}}$ be the set of leaves of $\mathcal{S}_{\mathsf{tree}}$, with $|\mathcal{S}_{\mathsf{leaf}}| = 2^{n-1}$.
- The observation space $\mathcal{O} = \mathcal{S}_{\mathsf{tree}} \bigsqcup \left\{o_1^+, o_1^-, \cdots, o_K^+, o_K^-\right\} \bigsqcup \{\mathsf{lock}, \mathsf{good}, \mathsf{bad}\} \bigsqcup_{j=1}^{L} \left\{\mathsf{lock}^j, \mathsf{terminal}^j\right\}$.
- The action space $\mathcal{A} = \{0, 1, \cdots, A-1\}$.

We further define $\mathsf{reveal} = 0 \in \mathcal{A}$, $\mathcal{A}_c = \{1, \cdots, A-1\}$.

Model parameters Each non-null POMDP model $M = M_{\theta,\mu} \in \mathcal{M}\backslash\{M_0\}$ is specified by parameters (θ, μ), where $\mu \in \{-1, +1\}^{L \times K}$, and $\theta = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$, where

- $s^\star \in \mathcal{S}_{\mathsf{leaf}}$, $a^\star \in \mathcal{A}_c := \{1, \cdots, A-1\}$.
- $h^\star \in \mathcal{H} := \{h = n + lm : l \in \mathbb{Z}_{\geqslant 0}, h < H\}$.
- $\mathbf{a}^\star = (\mathbf{a}^\star_{h^\star+1}, \ldots, \mathbf{a}^\star_{H-1}) \in \mathcal{A}^{H-h^\star-1}$ is an action sequence indexed by $h^\star + 1, \cdots, H-1$, such that when $h \in \mathcal{H}$, we have $\mathbf{a}^\star_h \neq \mathsf{reveal}$. We use $\mathcal{A}_{\mathsf{code},h^\star}$ to denote the set of all such $\mathbf{a}^\star$.

Our construction will ensure that, only at steps $h \in \mathcal{H}$ and states $s_h \in \{s_\oplus, s_\ominus\}$, the agent can observe lock^j and take action reveal to transit to $\left\{e^j_\oplus, e^j_\ominus\right\}$.

For any POMDP $M_{\theta,\mu}$, its system dynamics $\mathbb{P}_{\theta,\mu} := \mathbb{P}_{M_{\theta,\mu}}$ is defined as follows.

Emission dynamics At state $s \in \mathcal{S}_{\mathsf{tree}} \cup \{\mathsf{terminal}\}$, the agent always receives (the unique observation corresponding to) s itself as the observation.

- At state $e_{\oplus}^{j}$, the emission dynamics is given by

$$\mathbb{O}_{\mu}(o_i^{+}|e_{\oplus}^{j}) = \frac{1+\sigma\mu_{j,i}}{2K}, \qquad \mathbb{O}_{\mu}(o_i^{-}|e_{\oplus}^{j}) = \frac{1-\sigma\mu_{j,i}}{2K}, \qquad \forall i \in [K],$$

  where we omit the subscript h because the emission distribution does not depend on h.

- At state $e_{\ominus}^{j}$, the observation is uniformly drawn from $\mathcal{O}_o := \{o_1^{+}, o_1^{-}, \cdots, o_K^{+}, o_K^{-}\}$, i.e. $\mathbb{O}(\cdot|e_{\ominus}^{j}) = \mathrm{Unif}(\mathcal{O}_o)$.

- At states $s \in \{s_{\oplus}^{j}, s_{\ominus}^{j}\}$:

  - For steps $h \in \mathcal{H}$, the agent always receives lock^{j} as the observation.
  - For steps $h \leqslant H-1$ that does not belong to $\mathcal{H}$, the agent always receives lock as the observation.
  - At step H, the emission dynamics at $\{s_{\oplus}^{j}, s_{\ominus}^{j}\}$ is given by

$$\mathbb{O}_H(\mathsf{good}|s_{\oplus}^{j}) = \frac{3}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_{\oplus}^{j}) = \frac{1}{4},$$
$$\mathbb{O}_H(\mathsf{good}|s_{\ominus}^{j}) = \frac{1}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_{\ominus}^{j}) = \frac{3}{4}.$$

Transition dynamics In each episode, the agent always starts at s_0.

- At any node $s \in \mathcal{S}_{\mathsf{tree}} \backslash \mathcal{S}_{\mathsf{leaf}}$, there are three types of available actions: $\mathsf{wait} = 0$, $\mathsf{left} = 1$ and $\mathsf{right} = 2$, such that the agent can take wait to stay at s, left to transit to the left child of s and right to transit to the right child of s.

- At any $s \in \mathcal{S}_{\mathsf{leaf}}$, the agent can take action $\mathsf{wait} = 0$ to stay at s (i.e. $\mathbb{P}(s|s, \mathsf{wait}) = 1$); otherwise, for $s \in \mathcal{S}_{\mathsf{leaf}}$, $h \in [H-1]$, $a \neq \mathsf{wait}$,

$$\mathbb{P}_{h;\theta}(s_{\oplus}^{j}|s,a) = \frac{\varepsilon}{L} \cdot \mathbb{1}(h = h^{\star}, s = s^{\star}, a = a^{\star}),$$
$$\mathbb{P}_{h;\theta}(s_{\ominus}^{j}|s,a) = \frac{1}{L} - \frac{\varepsilon}{L} \cdot \mathbb{1}(h = h^{\star}, s = s^{\star}, a = a^{\star}).$$

- The states $s \in \{e_{\oplus}^{j}, e_{\ominus}^{j}\}$ always transit to $\mathsf{terminal}^{j}$, regardless of the action taken.

- The states $\mathsf{terminal}^{1}, \cdots, \mathsf{terminal}^{L}$ are absorbing states.

- At states $s \in \{s_{\oplus}^{j}, s_{\ominus}^{j}\}$:

  - For the step $h \in \mathcal{H}$, we set

$$\mathbb{P}_{h;\theta}(e_{\oplus}^{j}|s_{\oplus}^{j}, \mathsf{reveal}) = 1, \qquad \mathbb{P}_{h;\theta}(e_{\ominus}^{j}|s_{\ominus}^{j}, \mathsf{reveal}) = 1.$$

    In words, at steps $h \in \mathcal{H}$ and states $s \in \{s_{\oplus}^{j}, s_{\ominus}^{j}\}$ (corresponding to $o = \mathsf{lock}^{j}$), the agent can take action reveal to transit to $\{e_{\oplus}^{j}, e_{\ominus}^{j}\}$, respectively. Note that $\mathcal{H} = \{h = n + lm : h < H, l \in \mathbb{Z}_{\geqslant 0}\}$, so we only allow the agent to take the reveal action reveal every m steps, which ensures that our construction is $(m+1)$-step revealing.

  - For $h \notin \mathcal{H}$ or $a \neq \mathsf{reveal}$, we set

$$\mathbb{P}_{h;\theta}(s_{\oplus}^{j}|s_{\oplus}^{j}, a) = \mathbb{1}(a = \mathbf{a}_h^{\star}), \qquad \mathbb{P}_{h;\theta}(s_{\ominus}^{j}|s_{\oplus}^{j}, a) = \mathbb{1}(a \neq \mathbf{a}_h^{\star}),$$
$$\mathbb{P}_{h;\theta}(s_{\ominus}^{j}|s_{\ominus}^{j}, a) = 1.$$

Reward The reward function is known (and only depends on the observation): at the first $H-1$ steps, no reward is given; at step H, we set $r_H(\mathsf{good}) = 1$, $r_H(\mathsf{bad}) = 0$, $r_H(s_0) = (1+\varepsilon)/4$, and $r_H(o) = 0$ for any other $o \in \mathcal{O}$.

Reference model We use M_0 (or simply 0) to refer to the null model (reference model). The null model M_0 has transition and emission the same as any non-null model, except that the agent always arrives at $s_{\ominus}^j$ (with $j \sim \mathrm{Unif}([L])$) by taking any action $a \neq \mathsf{wait}$ at $s \in \mathcal{S}_{\mathsf{leaf}}$ and $h \in [H-1]$ (i.e., $\mathbb{P}_{h;M_0}(s_{\ominus}^j | s, a) = \frac{1}{L}$ for any $s \in \mathcal{S}_{\mathsf{leaf}}$, $a \in \mathcal{A}_c$, $h \in [H-1]$). In this model, states in $\{s_{\oplus}^1, e_{\oplus}^1, \cdots, s_{\oplus}^L, e_{\oplus}^L\}$ are all not reachable, and hence we do not need to specify the transition and emission dynamics at these states.

We summarize the results of the hard instances we construct in the following proposition, which we prove in Appendix G.2.

Proposition G.1. *For given $\varepsilon \in (0, 0.1], \sigma \in (0, 1]$, $m, n \geqslant 1$, $K, L \geqslant 1$, $H \geqslant 8n + m + 1$, the above model class $\mathcal{M}$ satisfies the following properties.*

1. $|\mathcal{S}| = 2^n + 5L$, $|\mathcal{O}| = 2^n + 2K + 2L + 3$, $|\mathcal{A}| = A$.
2. *For each $M \in \mathcal{M}$, M is $(m+1)$-step revealing with $\alpha_{m+1}(M)^{-1} \leqslant 1 + \frac{2}{\sigma}$.*
3. $\log |\mathcal{M}| \leqslant LK \log 2 + H \log A + \log(SAH)$.
4. *Suppose algorithm $\mathfrak{A}$ interacts with the environment for T episodes and returns π^{out} such that*

$$\mathbb{P}_M^{\mathfrak{A}}\left(V_M^\star - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\right) \geqslant \frac{3}{4}.$$

for any $M \in \mathcal{M}$. Then it must hold that

$$T \geqslant \frac{1}{10000m} \min\left\{ \frac{|\mathcal{S}_{\mathsf{leaf}}| \sqrt{LK} A^m H}{\sigma^2 \varepsilon^2}, \frac{|\mathcal{S}_{\mathsf{leaf}}| A^{H/2} H}{\varepsilon^2} \right\}.$$

Proof of Theorem 5 We have to suitably choose parameters when applying Proposition G.1. More specifically, given (S, O, A, H, α, m), we can let $m' = m - 1$, and take $n \geqslant 1$ to be the largest integer such that $2^n \leqslant S/4$, and take $L = \lfloor (S - 2^n)/5 \rfloor$, $K = \left\lfloor \frac{O - 2^n - 2L - 3}{2} \right\rfloor \gtrsim O$ (because $O \geqslant S \geqslant 10$), $\varepsilon' = \varepsilon/8$, and $\sigma = \frac{2}{\alpha^{-1} - 1} \leqslant 1$. Applying Proposition G.1 to the parameters $(\varepsilon, \sigma, n, m', K, L, H)$, we obtain a model class $\mathcal{M}$ of m-step α-revealing POMDPs, such that if there exists an algorithm $\mathfrak{A}$ that interacts with the environment for T episodes and returns a π^{out} such that $V_M^\star - V_M(\pi^{\mathrm{out}}) < \varepsilon$ with probability at least $3/4$ for all $M \in \mathcal{M}$, then

$$T \geqslant \frac{c_0}{m} \min\left\{ \frac{S^{3/2} O^{1/2} A^{m-1} H}{\alpha^2 \varepsilon^2}, \frac{S A^{H/2} H}{\varepsilon^2} \right\},$$

where c_0 is a universal constant.

Furthermore, we can apply Proposition F.1 (claim 5) instead, and similarly obtain a model class $\mathcal{M}'$ of m-step α-revealing POMDPs, such that if there exists an algorithm $\mathfrak{A}$ that interacts with the environment for T episodes and returns a π^{out} such that $V_M^\star - V_M(\pi^{\mathrm{out}}) < \varepsilon$ with probability at least $3/4$ for all $M \in \mathcal{M}'$, then

$$T \geqslant \frac{c_0'}{m} \min\left\{ \frac{S O^{1/2} A^m H}{\alpha^2 \varepsilon^2}, \frac{S A^{H/2} H}{\varepsilon^2} \right\},$$

where c_0' is a universal constant.

Combining these two cases completes the proof of Theorem 5. □

G.2. Proof of Proposition G.1

All propositions and lemmas stated in this section are proved in Appendix G.3-G.4.

Claim 1 follows directly by counting the number of states, observations, and actions in models in $\mathcal{M}$. Claim 3 follows as we have $|\mathcal{M}| = |\{(h^\star, s^\star, a^\star, \mathbf{a}^\star)\}| \times \left|\{\pm 1\}^{L \times K}\right| + 1 \leqslant HSA^H \times 2^{LK}$. Taking logarithm yields the claim.

Claim 2 follows from this lemma, which is proved in Appendix G.3.

Lemma G.2. *For each $M \in \mathcal{M}$, it holds that $\alpha_{m+1}(M)^{-1} \leqslant \frac{2}{\sigma} + 1$.*

By our construction, we can relate the sub-optimality of the output policy to the TV distance between models (under the prior distribution of parameter $\mu \sim \mathrm{Unif}(\{-1,+1\}^{L\times K})$), by an argument similar to the one in Appendix E.2. We summarize the results in the following lemma, whose proof is omitted for succinctness.

Lemma G.3 (Relating learning to testing)**.** *In holds that*

$$V_M^\star - V_M(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H = s_0)\ \ \forall M \neq 0, \qquad \text{and} \qquad V_0^\star - V_0(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_0^\pi(o_H \neq s_0).$$

Therefore, suppose that the algorithm $\mathfrak{A}$ outputs a policy π^{out} such that $\mathbb{P}_M^{\mathfrak{A}}\big(V_M^\star - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\big) \geqslant \frac{3}{4}$ for any model $M \in \mathcal{M}$, then we have

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) \geqslant \frac{1}{2}, \qquad \forall\theta. \tag{35}$$

In the following, we use (35) to derive lower bounds of the expected visitation count of some good events, and then deduce a lower bound of T, giving the following lemma whose proof is contained in Appendix G.4.

Lemma G.4. *Fix a $\theta = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$. We consider events*

$$E_{\mathsf{reach}}^\theta := \left\{o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})\right\},$$
$$E_{\mathsf{correct}}^\theta := \left\{o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star)\right\}.$$

Then for any algorithm $\mathfrak{A}$ with $\delta := D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) > 0$, we have

$$\text{either } \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{reach}}^\theta)\right] \geqslant \frac{\delta^3\sqrt{LK}}{18\varepsilon^2\sigma^2} - \frac{\delta}{6}, \text{ or } \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{correct}}^\theta)\right] \geqslant \frac{\delta^3}{18\varepsilon^2} - \frac{\delta}{6}.$$

Applying Lemma G.4 for any parameter tuple $\theta = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$ with $\delta = \frac{1}{2}$, we obtain

$$\text{either } \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{\sqrt{LK}}{300\varepsilon^2\sigma^2}, \quad \text{or} \quad \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{1}{300\varepsilon^2}, \tag{36}$$

by our choice that $\varepsilon \in (0, 0.1]$.

Fix a tuple $(h^\star, s^\star, a^\star)$ such that $h^\star \in \mathcal{H}$ and $h^\star \leqslant n + m\lfloor H/10m\rfloor$, $s^\star \in \mathcal{S}_{\mathsf{leaf}}$, $a^\star \in \mathcal{A}_c$. By (36), we know that for all $\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},h^\star}$, it holds that

$$\begin{aligned}
&A^{m-1}\cdot\mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] + |\mathcal{A}_{\mathsf{code},h^\star}|\cdot\mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] \\
\geqslant\ &\frac{1}{300}\min\left\{\frac{A^{m-1}\sqrt{LK}}{\varepsilon^2\sigma^2}, \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{\varepsilon^2}\right\} \geqslant \frac{1}{300}\min\left\{\frac{A^{m-1}\sqrt{LK}}{\varepsilon^2\sigma^2}, \frac{A^{H/2-1}}{\varepsilon^2}\right\} =: \omega,
\end{aligned} \tag{37}$$

where the last inequality uses the fact that $|\mathcal{A}_{\mathsf{code},h^\star}| \geqslant A^{H/2-1}$ for $h^\star \leqslant n + m\lfloor H/10m\rfloor$, which follows from a direct calculation (Lemma F.7). Notice that by our definition of E_{reach},

$$\begin{aligned}
\sum_{\mathbf{a}^\star\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] &= \sum_{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})\right)\right] \\
&= \sum_{\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}))]\cdot \sum_{\substack{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}\\ \mathbf{a}^\star \text{ begins with } \mathbf{a}}} 1 \\
&= \sum_{\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}))]\cdot \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}} \\
&= \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star} = s^\star, a_{h^\star} = a^\star)]\cdot\frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}}.
\end{aligned}$$

Similarly, by our definition of E_{correct}, we have

$$\sum_{\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\Big)\Big] = \sum_{\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}^\star))]$$
$$= \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_{h^\star+1:H-1}\in\mathcal{A}_{\mathsf{code},h^\star})] \leqslant \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)].$$

Therefore, taking average of (37) over all $\mathbf{a}\in\mathcal{A}_{\mathsf{code},h^\star}$ and using the equations above, we get

$$\begin{aligned}\omega \leqslant& \frac{1}{|\mathcal{A}_{\mathsf{code},h^\star}|}\sum_{\mathbf{a}\in\mathcal{A}_{\mathsf{code},h^\star}}\Big[A^{m-1}\cdot\mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\Big)\Big] + |\mathcal{A}_{\mathsf{code},h^\star}|\cdot\mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\Big)\Big]\Big]\\ \leqslant& 2\,\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star,a_{h^\star}=a^\star)].\end{aligned}$$

Now, we have shown that $\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star,a_{h^\star}=a^\star)]\geqslant\frac{\omega}{2}$ for each $s^\star\in\mathcal{S}_{\mathsf{leaf}}, a^\star\in\mathcal{A}_c$, $h^\star\in\mathcal{H}$ such that $h^\star\leqslant n+m\lfloor H/10m\rfloor$. Taking summation over all such $(h^\star,s^\star,a^\star)$, we derive that

$$\frac{|\mathcal{S}_{\mathsf{leaf}}|\,|\mathcal{A}_c|\,(\lfloor H/10m\rfloor+1)}{600}\min\left\{\frac{A^{m-1}\sqrt{LK}}{\varepsilon^2\sigma^2},\frac{A^{H/2-1}}{\varepsilon^2}\right\} \leqslant \sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}}\sum_{a^\star\in\mathcal{A}_c}\sum_{\substack{h^\star=n+lm:\\ 0\leqslant l\leqslant\lfloor H/10m\rfloor}}\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star,a_{h^\star}=a^\star)]\leqslant T,$$

where the second inequality is because events $(\{o_{h^\star}=s^\star,a_{h^\star}=a^\star\})_{h^\star,s^\star,a^\star}$ are disjoint. Plugging in $|\mathcal{A}_c|=A-1\geqslant\frac{2}{3}A$, $\lfloor H/10m\rfloor+1\geqslant H/10m$ completes the proof of Proposition G.1. □

G.3. Proof of Lemma G.2

The proof is very similar to the proof of Lemma F.3, with only slight modification.

Case 1: We first show that the null model 0 is 1-step 1-revealing. In this model, the states in $\{s_\oplus^1,e_\oplus^1,\cdots,s_\oplus^L,e_\oplus^L\}$ are all not reachable, and hence for each step h, we consider the set $\mathcal{S}'=\mathcal{S}_{\mathsf{tree}}\sqcup\bigsqcup_{j=1}^L\{s_\ominus^j,e_\ominus^j,\mathsf{terminal}^j\}$. For different states $s,s'\in\mathcal{S}'$, the support of $\mathbb{O}_h(\cdot|s)$ and $\mathbb{O}_h(\cdot|s')$ are disjoint by our construction, and hence applying Lemma F.9 gives

$$\min_{\mathbb{O}_h^+}\|\mathbb{O}_h^+\|_{1\to1}\leqslant\max_{s\in\mathcal{S}'}\gamma(\mathbb{O}_h(s))\leqslant 1.$$

Applying Proposition 2 completes the proof for null model 0.

Case 2: We next consider the model $M=M_{\theta,\mu}\in\mathcal{M}$. By our construction, for $h\leqslant h^\star$, the states in $\{s_\oplus^1,e_\oplus^1,\cdots,s_\oplus^L,e_\oplus^L\}$ are all not reachable, and hence by the same argument as in the null model, we obtain

$$\min_{\mathbb{M}_{h,m+1}^+}\left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to1}\leqslant\min_{\mathbb{O}_h^+}\|\mathbb{O}_h^+\|_{1\to1}\leqslant 1.$$

Hence, we only need to bound the quantity $\min_{\mathbb{M}_{h,m+1}^+}\left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to1}$ for a fixed step $h>h^\star$. In this case, there exists a $l\in\mathcal{H}$ such that $h\leqslant l\leqslant h+m-1$, and we write $r=l-h+1$. By Lemma C.1, we only need to bound $\min_{\mathbb{M}_{h,r+1}^+}\left\|\mathbb{M}_{h,r+1}^+\right\|_{*\to1}$. Consider the action sequence $\mathbf{a}=(\mathbf{a}_{h:l-1}^\star,\mathsf{reveal})\in\mathcal{A}^r$, and we partition $\mathcal{S}$ as

$$\mathcal{S}=\bigsqcup_{s\in\mathcal{S}_{\mathsf{tree}}}\{s\}\sqcup\bigsqcup_{j=1}^L\{s_\oplus^j,s_\ominus^j\}\sqcup\{e_\oplus^j,e_\ominus^j\}\sqcup\{\mathsf{terminal}^j\}.$$

It is direct to verify that, in $M_{\theta,\mu}$, for states s,s' come from different subsets in the above partition, the support of $\mathbb{M}_{h,\mathbf{a}}(\cdot|s)$ and $\mathbb{M}_{h,\mathbf{a}}(\cdot|s')$ are disjoint. Then, we can apply Lemma F.8 and Lemma F.9, and obtain

$$\min_{\mathbb{M}_{h,r+1}^+}\left\|\mathbb{M}_{h,r+1}^+\right\|_{*\to1}\leqslant\min_{\mathbb{M}_{h,\mathbf{a}}^+}\left\|\mathbb{M}_{h,\mathbf{a}}^+\right\|_{1\to1}\leqslant\max_j\Big\{1,\gamma\Big(\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus^j,s_\ominus^j\})\Big),\gamma\Big(\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus^j,e_\ominus^j\})\Big)\Big\}.$$

Therefore, in the following we only need to consider left inverses of the matrix $\mathbb{M}_{h,\mathbf{a}}(\{s^j_\oplus, s^j_\ominus\})$ and $\mathbb{M}_{h,\mathbf{a}}(\{e^j_\oplus, e^j_\ominus\})$ for each $j \in [L]$.

(1) The matrix $\mathbb{M}_{h,\mathbf{a}}(\{s^j_\oplus, s^j_\ominus\})$. By our construction, taking $\mathbf{a}$ at $s_h = s^j_\oplus$ will lead to $o_{h:l-1} = \mathsf{lock}$, $o_l = \mathsf{lock}^j$ and $o_{l+1} \sim \mathbb{O}_\mu(\cdot|e^j_\oplus)$; taking $\mathbf{a}$ at $s_h = s^j_\ominus$ will lead to $o_{h:l-1} = \mathsf{lock}$, $o_l = \mathsf{lock}^j$ and $o_{l+1} \sim \mathbb{O}_\mu(\cdot|e^j_\ominus)$. Hence, $\mathbb{M}_{h,\mathbf{a}}(\{s^j_\oplus, s^j_\ominus\})$ can be written as (up to permutation of rows)

$$\mathbb{M}_{h,\mathbf{a}}(\{s^j_\oplus, s^j_\ominus\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K} + \sigma\widetilde{\mu}_j}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1} \times 2},$$

where $\widetilde{\mu}_j = [\mu_j; -\mu_j] \in \{-1, 1\}^{2K}$, $\mathbb{1} = \mathbb{1}_{2K}$ is the vector in $\mathbb{R}^{2K}$ with all entry being one. Similar to Proposition E.2, we can directly verify that $\gamma(\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\})) \leqslant \frac{2}{\sigma} + 1$.

(2) The matrix $\mathbb{M}_{h,\mathbf{a}}(\{e^j_\oplus, e^j_\ominus\})$. By our construction, at $s_h = e^j_\oplus$, we have $o_h \sim \mathbb{O}_\mu(\cdot|e^j_\oplus)$ and $o_{h+1:l+1} = \mathsf{terminal}^j$; at $s_h = e_\ominus$, we have $o_h \sim \mathbb{O}_\mu(\cdot|e^j_\ominus)$ and $o_{h+1:l+1} = \mathsf{terminal}^j$. Thus, $\mathbb{M}_{h,\mathbf{a}}(\{s^j_\oplus, s^j_\ominus\})$ can also be written as

$$\mathbb{M}_{h,\mathbf{a}}(\{e^j_\oplus, e^j_\ominus\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K} + \sigma\widetilde{\mu}_j}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1} \times 2},$$

and hence we also have $\gamma\Big(\mathbb{M}_{h,\mathbf{a}}(\{e^j_\oplus, e^j_\ominus\})\Big) \leqslant \frac{2}{\sigma} + 1$.

Combining the two cases above gives

$$\min_{\mathbb{M}^+_{h,m+1}} \left\|\mathbb{M}^+_{h,m+1}\right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,r+1}} \left\|\mathbb{M}^+_{h,r+1}\right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\|\mathbb{M}^+_{h,\mathbf{a}}\right\|_{1\to 1} \leqslant \frac{2}{\sigma} + 1,$$

and hence completes the proof of Lemma G.2. □

G.4. Proof of Lemma G.4

Similar to the proof of Lemma F.6, we only need to show the following lemma, and the proof of Lemma G.4 follows by a reduction argument (see Appendix F.6).

Lemma G.5. *Suppose that algorithm $\mathfrak{A}$ (with possibly random stopping time T) satisfies $N(E^\theta_{\mathsf{reach}}) \leqslant \overline{N}_o$ and $N(E^\theta_{\mathsf{correct}}) \leqslant \overline{N}_r$ almost surely, for some fixed $\overline{N}_o, \overline{N}_r$. Then*

$$\textit{either } \overline{N}_o \geqslant \frac{3}{4}\frac{\delta^2\sqrt{LK}}{\varepsilon^2\sigma^2}, \textit{ or } \overline{N}_r \geqslant \frac{3}{4}\frac{\delta^2}{\varepsilon^2},$$

where $\delta = D_{\mathrm{TV}}\left(\mathbb{P}^{\mathfrak{A}}_0, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}^{\mathfrak{A}}_{\theta,\mu}\right]\right)$.

Proof. Fix a $\theta = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$. Recall that we define $E^\theta_{\mathsf{reach}} := \left\{o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})\right\}$, and we further define

$$E^\theta_{\mathsf{rev}} := \left\{o_{h^\star} = s^\star, a_{h^\star:h} = (a^\star, \mathbf{a}^\star_{h^\star+1:h-1}, \mathsf{reveal}) \text{ for some } h \in \mathcal{H}_{>h^\star}\right\}.$$

For any model $M \in \mathcal{M}_\theta := \left\{M_{\theta,\mu} : \mu \in \{-1, +1\}^{L\times K}\right\} \cup \{0\}$, we consider the following "augmented" system dynamics $\overline{\mathbb{P}}_M$:
1. For each episode, after the interaction $\tau_H \sim \mathbb{P}_M$ is finished, the environment generated an extra observation $o_{H+1} = o^{\mathsf{aug}}$.
2. If $\tau_H \notin E^\theta_{\mathsf{reach}}$ or $\tau_H \in E^\theta_{\mathsf{rev}}$, then $o^{\mathsf{aug}} = \mathsf{dummy}$.
3. If $\tau_H \in E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}}$, then in $\tau_H = (o_1, a_1, \cdots, o_H, a_H)$ we have $o_{h^\star+m} = \mathsf{lock}^j$ for some $j \in [L]$, and then the environment generates o^{aug} as

$$M = M_{\theta,\mu}: \qquad \overline{\mathbb{P}}_{\theta,\mu}(o^{\mathsf{aug}} = o^+_i|\tau_H) = \frac{1 + \varepsilon\sigma\mu_{j,i}}{2K}, \qquad \overline{\mathbb{P}}_{\theta,\mu}(o^{\mathsf{aug}} = o^+_i|\tau_H) = \frac{1 - \varepsilon\sigma\mu_{j,i}}{2K}, \qquad \forall i \in [K],$$

and for $M = 0$, $\overline{\mathbb{P}}_0(o^{\mathsf{aug}} = \cdot|\tau_H) = \mathrm{Unif}(\{o^+_1, o^-_1, \cdots, o^+_K, o^-_K\})$.

Clearly, for each $M \in \mathcal{M}_\theta$, $\overline{\mathbb{P}}_M$ is still a sequential decision process. Under such construction, each policy π induces a distribution of $\overline{\tau} = (\tau_H, o^{\mathsf{aug}}) \sim \overline{\mathbb{P}}_M^\pi$, and the algorithm $\mathfrak{A}$ induce a distribution of $\overline{\tau}^{(1)}, \cdots, \overline{\tau}^{(\mathsf{T})} \sim \overline{\mathbb{P}}_0^{\mathfrak{A}}$. By data-processing inequality, we have

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) \leqslant D_{\mathrm{TV}}\left(\overline{\mathbb{P}}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\overline{\mathbb{P}}_{\theta,\mu}^{\mathfrak{A}}\right]\right).$$

Hence, by Lemma D.1, we only need to bound

$$1 + \chi^2(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\overline{\mathbb{P}}_{\theta,\mu}^{\mathfrak{A}}\right] \,\|\, \overline{\mathbb{P}}_0^{\mathfrak{A}}) = \mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}} \frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\right].$$

To upper bound the above quantity, we invoke the following lemma (proof in Appendix G.5).

Lemma G.6 (Bound on the χ^2-inner product)**.** *Under the conditions of Lemma G.5 (for a fixed θ), it holds that for any $\mu, \mu' \in \{-1, 1\}^K$,*

$$\overline{\mathbb{E}}_0^{\mathfrak{A}}\left[\prod_{t=1}^{\mathsf{T}} \frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\right] \leqslant \exp\left(\overline{N}_o \cdot \frac{\sigma^2\varepsilon^2}{LK}\left|\langle \mu, \mu'\rangle\right| + \frac{4}{3}\varepsilon^2\overline{N}_r\right). \tag{38}$$

Given Lemma G.6, the desired result follows from a standard argument (see e.g. the proof of Lemma F.11). □

G.5. Proof of Lemma G.6

We first show the following lemma, which is a single-episode version of Lemma G.6.

Lemma G.7. *For any policy π and parameter θ, μ, μ', it holds that*

$$\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2} \exp\left(-\mathbb{1}(\overline{\tau} \in E_{\mathsf{reach}}^\theta) \cdot \frac{\varepsilon^2\sigma^2}{LK}\left\langle \mu, \mu'\right\rangle - \mathbb{1}(\overline{\tau} \in E_{\mathsf{correct}}^\theta) \cdot \frac{4}{3}\varepsilon^2\right)\right] \leqslant 1. \tag{39}$$

Proof of Lemma G.7. In the following, all expectation and conditional expectation is taken with respect to $\overline{\tau} = (\tau_H, o^{\mathsf{aug}}) \sim \overline{\mathbb{P}}_0^\pi$.

Similar to the proof of Lemma F.12 (in Appendix F.7), the core of our analysis is still computing the quantity $I(\tau_l)$, defined as

$$I(\tau_l) := \mathbb{E}_{o_{l+1}\sim\overline{\mathbb{P}}_0(\tau_l)}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(o_{l+1}|\tau_l)\overline{\mathbb{P}}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\overline{\mathbb{P}}_0(o_{l+1}|\tau_l)^2}\right] = \begin{cases} \mathbb{E}\left[\dfrac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right], & l < H, \\ \mathbb{E}_{o^{\mathsf{aug}}\sim\mathbb{P}_0(\tau_H)}\left[\dfrac{\overline{\mathbb{P}}_{\theta,\mu}(o^{\mathsf{aug}}|\tau_H)\overline{\mathbb{P}}_{\theta,\mu'}(o^{\mathsf{aug}}|\tau_H)}{\overline{\mathbb{P}}_0(o^{\mathsf{aug}}|\tau_H)^2}\right], & l = H. \end{cases} \tag{40}$$

Basically, by Lemma A.2, we have

$$1 = \mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}|\tau_{h^\star+m})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}|\tau_{h^\star+m})}{\overline{\mathbb{P}}_0(\overline{\tau}|\tau_{h^\star+m})^2} \cdot \exp\left(-\sum_{l=h^\star+m}^{H} \log I(\tau_l)\right)\middle|\tau_{h^\star+m}\right]. \tag{41}$$

In the following, we first compute $I(\tau_l)$ for each (reachable) τ_l.

An important observation is that, for a trajectory $\tau_l = (o_1, a_1, \cdots, o_l, a_l)$ with $l < H$, if $\mathbb{P}_{\theta,\mu}(o_{l+1} = \cdot|\tau_l) \neq \mathbb{P}_0(o_{l+1} = \cdot|\tau_l)$, then
1. Clearly, $o_{h^\star} = s^\star$, $a_{h^\star} = a^\star$ (i.e. $l \geqslant h^\star + 1$ and taking action $a_{1:h^\star-1}$ from s_0 will result in $s^\star$ at step $h^\star$).
2. Either $a_{h^\star+1:l} = (\mathbf{a}_{h^\star+1:l-1}^\star, \mathsf{reveal})$ for some $l \in \mathcal{H}_{>h^\star}$, or $l = H-1$ and $a_{h^\star+1:H-1} = \mathbf{a}^\star$.

Therefore, for each $l \in \mathcal{H}_{>h^\star}$ we define

$$E_{\mathsf{rev},l}^\theta := \left\{o_{h^\star} = s^\star, a_{h^\star:l} = (a^\star, \mathbf{a}_{h^\star+1:l-1}^\star, \mathsf{reveal})\right\}.$$

Then, if $\mathbb{P}_{\theta,\mu}(\cdot|\tau_l) \neq \mathbb{P}_0(\cdot|\tau_l)$, either (case 1) $l \in \mathcal{H}_{>h^\star}, \tau_{H-1} \in E^\theta_{\mathsf{rev},l}$, or (case 2) $l = H-1, \tau_l \in E^\theta_{\mathsf{correct}}$, or (case 3) $l = H, \tau_H \in E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}}$.

In the following, we compute $I(\tau_l)$ for these three cases separately. We consider the events $\mathcal{L}_j := \{o_{h^\star+m} = \mathsf{lock}^j\}(j \in [L])$ to simplify our discussion.

Case 1: $l \in \mathcal{H}_{>h^\star}, \tau_l \in E^\theta_{\mathsf{rev},l}$. In this case, there exists a $j \in [L]$ such that $o_{h^\star+m} = \mathsf{lock}^j$, i.e. $\tau_l \in \mathcal{L}_j$. In other words, observing τ_l implies that $s_{h'} \in \{s^j_\oplus, s^j_\ominus\}$ for $h < h' \leqslant l$, and $s_{l+1} \in \{e^j_\oplus, e^j_\ominus\}$ (because $a_l = \mathsf{reveal}$). Therefore,

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l) &= \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = e^j_\oplus)\mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\oplus|\tau_l) + \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = e^j_\ominus)\mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\ominus|\tau_l)\\ &= \Big(\mathbb{O}_\mu(o|e^j_\oplus) - \mathbb{O}(o|e^j_\ominus)\Big)\cdot \mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\oplus|\tau_l) + \mathbb{O}(o|e^j_\ominus),\end{aligned}$$

Notice that by our construction,

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\oplus|\tau_l) &= \mathbb{P}_{\theta,\mu}(s_l = s^j_\oplus|\tau_{l-1}, o_l)\\ &= \mathbb{P}_{\theta,\mu}(s_l = s^j_\oplus|o_{h^\star} = s^\star, a_{h^\star:l-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:l-1}), \mathcal{L}_j)\\ &= \mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s^j_\oplus|o_{h^\star} = s^\star, a_{h^\star:l-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:l-1}), \mathcal{L}_j)\\ &= \mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s^j_\oplus|o_{h^\star} = s^\star, a_{h^\star:l-1} = a^\star, s_{h^\star+1} \in \{s^j_\oplus, s^j_\ominus\})\\ &= \varepsilon,\end{aligned}$$

where the first equality is because $s_{l+1} = e^j_\oplus$ if and only if $s_l = s^j_\oplus, a_l = \mathsf{reveal}$, the second inequality is because there are only lock and lock^j in $o_{h^\star+1:l}$ are the third equality is because $s_l = s^j_\oplus$ if and only if $s_{h^\star+1} = s^j_\oplus, a_{h^\star+1:l-1} = \mathbf{a}^\star_{h^\star+1:l-1}$. Combining the above equations with our definition of $\mathbb{O}_\mu(\cdot|e^j_\oplus)$ and $\mathbb{O}(\cdot|e^j_\ominus)$ gives

$$\mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^+|\tau_l) = \frac{1+\varepsilon\sigma\mu_{j,i}}{2K}, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^-|\tau_l) = \frac{1-\varepsilon\sigma\mu_{j,i}}{2K}, \qquad \forall i \in [K].$$

On the other hand, clearly $\mathbb{P}_0(o_{l+1} = \cdot|\tau_l) = \mathrm{Unif}(\{o_1^+, o_1^-, \cdots, o_K^+, o_K^-\})$. Hence, it holds that

$$I(\tau_l) = \frac{1}{2K}\sum_{o\in\mathcal{O}_o}\frac{\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1} = o|\tau_l)}{\mathbb{P}_0(o_{l+1} = o|\tau_l)^2} = 1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}, \quad \text{for any } \tau_l \in E^\theta_{\mathsf{rev},l} \cap \mathcal{L}_j. \tag{42}$$

Case 2: $l = H-1, \tau_{H-1} \in E^\theta_{\mathsf{correct}}$. In this case, by a calculation exactly the same as the proof of Lemma F.12 (Appendix F.7, case 2), we can obtain

$$I(\tau_{H-1}) = 1 + \frac{4}{3}\varepsilon^2, \quad \text{for any } \tau_{H-1} \in E^\theta_{\mathsf{correct}}. \tag{43}$$

Case 3: $l = H, \tau_H \in E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}}$. Suppose that for some $j \in [L]$, $\tau_H \in \big(E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}}\big) \cap \mathcal{L}_j$, then by our construction of $\overline{\mathbb{P}}$, we have

$$I(\tau_H) = 1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}, \quad \text{for any } \tau_H \in \big(E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}}\big) \cap \mathcal{L}_j. \tag{44}$$

Combining (42) (43) (44) together, we have shown that for any τ_H that begins with $\tau_{h^\star+m} \in E^\theta_{\mathsf{reach}} \cap \mathcal{L}_j$,

$$\begin{aligned}\sum_{l=h^\star+m}^H \log I(\tau_l) = \sum_{l\in\mathcal{H}_{>h^\star}} &\mathbb{1}\big(\tau_l \in E^\theta_{\mathsf{rev},l}\big)\cdot\log\left(1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}\right) + \mathbb{1}\big(\tau_{H-1} \in E^\theta_{\mathsf{correct}}\big)\cdot\log\left(1+\frac{4}{3}\varepsilon^2\right)\\ &+ \mathbb{1}\big(\tau_H \in E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}}\big)\cdot\log\left(1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}\right)\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{1}\big(\tau_H \in E^\theta_{\mathsf{rev}}\big) \cdot \log\left(1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}\right) + \mathbb{1}\big(\tau_H \in E^\theta_{\mathsf{correct}}\big) \cdot \log\left(1 + \frac{4}{3}\varepsilon^2\right) \\
&\quad + \mathbb{1}\big(\tau_H \in E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}}\big) \cdot \log\left(1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}\right) \\
&= \mathbb{1}\big(\tau_H \in E^\theta_{\mathsf{reach}}\big) \cdot \log\left(1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}\right) + \mathbb{1}\big(\tau_H \in E^\theta_{\mathsf{correct}}\big) \cdot \log\left(1 + \frac{4}{3}\varepsilon^2\right),
\end{aligned}
$$

where the second equality is because $E^\theta_{\mathsf{rev}} = \bigsqcup_l E^\theta_{\mathsf{rev},l}$. We also have $\sum_{l=h^\star+m}^H \log I(\tau_l) = 0$ for $\tau_{h^\star+m} \notin E^\theta_{\mathsf{reach}}$. Plugging the value of $\sum_{l=h^\star+m}^H \log I(\tau_l)$ into (41) and using the fact that $\mathbb{P}_{\theta,\mu}(\tau_{h^\star+m}) = \mathbb{P}_0(\tau_{h^\star+m})$ by our construction, we have

$$
\begin{aligned}
&\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2} \cdot \left(1 + \mathbb{1}(\tau_H \in E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2\right)^{-1} \middle| \tau_{h^\star+m}\right] \\
&= \begin{cases} 1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}, & \text{for } \tau_{h^\star+m} \in E^\theta_{\mathsf{reach}} \cap \mathcal{L}_j, \\ 1, & \text{if } \tau_{h^\star+m} \notin E^\theta_{\mathsf{reach}}. \end{cases}
\end{aligned}
$$

Notice that in $\mathbb{P}_0^\pi$, conditional on $\tau_{h^\star+m-1} \in E^\theta_{\mathsf{reach}}$, $o_{h^\star+m}$ is uniformly distributed over $\left\{\mathsf{lock}^1, \cdots, \mathsf{lock}^L\right\}$, and hence

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2} \cdot \left(1 + \mathbb{1}(\tau_H \in E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2\right)^{-1} \middle| \tau_{h^\star+m-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2} \cdot \left(1 + \mathbb{1}(\tau_H \in E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2\right)^{-1} \middle| \tau_{h^\star+m}\right] \middle| \tau_{h^\star+m-1}\right] \\
&= \frac{1}{L}\sum_{j=1}^L \left(1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_{j,i}\mu'_{j,i}\right) = 1 + \frac{\varepsilon^2\sigma^2}{LK}\left\langle \mu, \mu' \right\rangle.
\end{aligned}
$$

On the other hand, for $\tau_{h^\star+m-1} \notin E^\theta_{\mathsf{reach}}$,

$$
\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2} \cdot \left(1 + \mathbb{1}(\overline{\tau} \in E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2\right)^{-1} \middle| \tau_{h^\star+m-1}\right] = 1.
$$

Hence, taking expectation over $\tau_{h^\star+m-1}$ gives

$$
\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2} \cdot \left(1 + \mathbb{1}(\overline{\tau} \in E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2\right)^{-1} \cdot \left(1 + \mathbb{1}(\overline{\tau} \in E^\theta_{\mathsf{reach}}) \cdot \frac{\varepsilon^2\sigma^2}{LK}\left\langle \mu, \mu' \right\rangle\right)^{-1}\right] = 1.
$$

Using the fact $(1+x)^{-1} \geqslant \exp(-x)$ completes the proof. □

With Lemma G.7 proven, we continue to prove Lemma G.6. Applying Lemma G.7 to algorithm $\mathfrak{A}$, we obtain that for each $t \in [\mathsf{T}]$,

$$
\mathbb{E}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2} \cdot \exp\left(-\mathbb{1}(\overline{\tau}^{(t)} \in E^\theta_{\mathsf{reach}}) \cdot \frac{\varepsilon^2\sigma^2}{LK}\left\langle \mu, \mu' \right\rangle - \mathbb{1}(\overline{\tau}^{(t)} \in E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2\right) \middle| \overline{\tau}^{(1:t-1)}\right] \leqslant 1,
$$

where the expectation is taken over $\overline{\tau}^{(t)} \sim \overline{\mathbb{P}}_0^{\mathfrak{A}}$ conditional on $\tau^{(1:t-1)}$. Therefore, by the martingale property, it holds that

$$
\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}} \frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2} \cdot \exp\left(-\mathbb{1}(\overline{\tau}^{(t)} \in E^\theta_{\mathsf{reach}}) \cdot \frac{\varepsilon^2\sigma^2}{LK}\left\langle \mu, \mu' \right\rangle - \mathbb{1}(\overline{\tau}^{(t)} \in E^\theta_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2\right)\right] \leqslant 1.
$$

Notice that $N(E^{\theta}_{\mathsf{reach}}) = \sum_{t=1}^{T} \mathbb{1}(\overline{\tau}^{(t)} \in E^{\theta}_{\mathsf{reach}})$ and $N(E^{\theta}_{\mathsf{correct}}) = \sum_{t=1}^{T} \mathbb{1}(\overline{\tau}^{(t)} \in E^{\theta}_{\mathsf{correct}})$, and hence

$$1 \geqslant \mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})} \sim \overline{\mathbb{P}}_0^{\mathfrak{A}}} \left[\prod_{t=1}^{\mathsf{T}} \frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)}) \overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2} \times \exp\left(-N(E^{\theta}_{\mathsf{reach}}) \cdot \frac{\varepsilon^2\sigma^2}{LK} \langle \mu, \mu' \rangle - N(E^{\theta}_{\mathsf{correct}}) \cdot \frac{4}{3}\varepsilon^2 \right) \right]$$
$$\geqslant \mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})} \sim \overline{\mathbb{P}}_0^{\mathfrak{A}}} \left[\prod_{t=1}^{\mathsf{T}} \frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)}) \overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2} \times \exp\left(-\overline{N}_o \cdot \frac{\varepsilon^2\sigma^2}{LK} |\langle \mu, \mu' \rangle| - \overline{N}_r \cdot \frac{4}{3}\varepsilon^2 \right) \right].$$

Multiplying both sides by $\exp\left(\overline{N}_o \cdot \frac{\sigma^2\varepsilon^2}{LK} |\langle \mu, \mu' \rangle| + \frac{4}{3}\varepsilon^2 \overline{N}_r\right)$ completes the proof of Lemma G.6. □

H. Regret for single-step revealing POMDPs

In this section, we establish Theorem 8 on a broader class of sequential decision problems termed as *strongly B-stable PSRs*, and then deduce the guarantee for single-step revealing POMDPs as a special case. The proof is largely parallel to the analysis of PAC learning for B-stable PSRs (Chen et al., 2022a), and we follow the notations there: in the following we use θ to refer to the PSR model, and Θ to refer to the class of PSR models.

H.1. Strongly B-stable PSRs

We recall the definition of PSRs and B-stability in Appendix B. To establish $\sqrt{T}$-regret upper bound for learning PSRs, we introduce the following structural condition.

Definition H.1 (Strong B-stability). *A PSR is* strongly B-stable with parameter $\Lambda_{\mathsf{B}} \geqslant 1$ *(henceforth also* Λ_{B}-strongly-stable*) if it admits a B-representation such that for all step $h \in [H]$, policy π, $x \in \mathbb{R}^{\mathcal{U}_h}$,*

$$\sum_{\tau_{h:H}} \pi(\tau_{h:H}) \times |\mathbf{B}_{H:h}(\tau_{h:H})x| \leqslant \Lambda_{\mathsf{B}} \sum_{t_h \in \mathcal{U}_h} \pi(t_h) \times |x(t_h)|\,. \tag{45}$$

For notational simplicity, from now on we assume that for each step h, $\mathcal{U}_h = (\mathcal{O} \times \mathcal{A})^{m_h - 1} \times \mathcal{O}$ for some $m_h \in \mathbb{Z}_{\geqslant 1}$, and we define $\Omega_h := (\mathcal{O} \times \mathcal{A})^{m_h - 1}$; our results also hold for any general $\mathcal{U}_h$ using slightly more involved notation.

Proposition H.2 (Error decomposition for strongly B-stable PSRs). *Suppose that two PSR models $\theta, \bar{\theta}$ admit $\{\{\mathbf{B}^{\theta}_h(o_h, a_h)\}_{h,o_h,a_h}, \mathbf{q}^{\theta}_0\}$ and $\{\{\mathbf{B}^{\bar{\theta}}_h(o_h, a_h)\}_{h,o_h,a_h}, \mathbf{q}^{\bar{\theta}}_0\}$ as B-representation respectively. Define*

$$\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1}) := \frac{1}{2} \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left| \mathbf{B}^{\theta}_{H:h+1}(\tau_{h+1:H}) \left(\mathbf{B}^{\theta}_h(o_h, a_h) - \mathbf{B}^{\bar{\theta}}_h(o_h, a_h) \right) \mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \right|,$$
$$\mathcal{E}^{\bar{\theta}}_{\theta,0}(\pi) := \frac{1}{2} \max_{\pi' \in \Pi_0(\pi)} \sum_{\tau_{1:H}} \pi'(\tau_{1:H}) \times \left| \mathbf{B}^{\theta}_{H:1}(\tau_{1:H}) \left(\mathbf{q}^{\theta}_0 - \mathbf{q}^{\bar{\theta}}_0 \right) \right|,$$

where we define

$$\Pi_h(\pi) := \left\{ \pi' : \pi'|_{\mathcal{O} \times \mathcal{A} \times \Omega_{h+1}} = \pi|_{\mathcal{O} \times \mathcal{A} \times \Omega_{h+1}} \right\}, \qquad \Pi_0(\pi) := \left\{ \pi' : \pi'|_{\Omega_1} = \pi|_{\Omega_1} \right\},$$

i.e. $\Pi_h(\pi)$ is the set of all policy π' such that for all $(o_h, a_h, \omega_{h+1}) \in \mathcal{O} \times \mathcal{A} \times \Omega_{h+1}$, $\pi'(o_h, a_h, \omega_{h+1}|\tau_{h-1}) = \pi(o_h, a_h, \omega_{h+1}|\tau_{h-1})$.

Then the following claims hold.

1. (Performance decomposition) It holds that

$$D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}}\right) \leqslant \mathcal{E}^{\bar{\theta}}_{\theta,0}(\pi) + \sum_{h=1}^{H} \mathbb{E}^{\pi}_{\bar{\theta}}\left[\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1}) \right],$$

where for $h \in [H]$, the expectation $\mathbb{E}^{\pi}_{\bar{\theta}}$ is taking over τ_{h-1} under model $\bar{\theta}$ and policy π.

2. (Bounding errors by Hellinger distance) Suppose that θ is Λ_{B}-strong-stable and $\{\{\mathbf{B}^{\theta}_h(o_h, a_h)\}_{h,o_h,a_h}, \mathbf{q}^{\theta}_0\}$ satisfies the stability condition (45). For any step h, policy π, it holds that

$$\mathbb{E}^{\pi}_{\bar{\theta}}\left[\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1})^2 \right] \leqslant 2\Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}}\right).$$

and $(\mathcal{E}^{\bar{\theta}}_{\theta,0}(\pi))^2 \leqslant \Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}}\right)$.

Algorithm 1 OPTIMISTIC MAXIMUM LIKELIHOOD ESTIMATION (OMLE) (LIU ET AL., 2022A; CHEN ET AL., 2022A)

1: **Input:** Model class Θ, parameter $\beta > 0$.
2: **Initialize:** $\Theta^1 = \Theta$, $\mathcal{D} = \{\}$.
3: **for** iteration $k = 1, \dots, T$ **do**
4: Set $(\theta^k, \pi^k) = \arg\max_{\theta \in \Theta^k, \pi} V_\theta(\pi)$.
5: Execute π^k to collect a trajectory τ^k, and add (π^k, τ^k) into $\mathcal{D}$.
6: Update confidence set

$$\Theta^{k+1} = \left\{ \widehat{\theta} \in \Theta : \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_{\widehat{\theta}}(\tau) \geqslant \max_{\theta\in\Theta} \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_\theta(\tau) - \beta \right\}.$$

7: **end for**

H.2. Algorithms and guarantees

In this section, we state the $\sqrt{T}$-regret guarantee of the algorithm OMLE (Algorithm 1, (Liu et al., 2022a; Chen et al., 2022a)). Its proof is in presented Appendix H.4, which is adapted from the analysis of the (explorative) OMLE algorithm in Chen et al. (2022a). We also remark that the regret upper bound of OMLE in Theorem 8 can also be shown directly for single-step revealing POMDP, by strengthening the analysis in Liu et al. (2022a) using the ideas of Chen et al. (2022a).

Theorem H.3. *Suppose every $\theta \in \Theta$ is Λ_{B}-strongly stable (Definition H.1), and the true model $\theta^\star \in \Theta$ with rank $d_{\mathsf{PSR}} \leqslant d$. Then, choosing $\beta = C\log(\mathcal{N}_\Theta(1/T))/\delta)$ for some absolute constant $C > 0$, with probability at least $1-\delta$, Algorithm 1 achieves*

$$\sum_{t=1}^{T} V^\star - V_{\theta^\star}(\pi^t) \leqslant \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2 \iota \beta T}\right) \tag{46}$$

where $U_{\mathcal{T}} := \max_h |\Omega_h|$, $\iota := \log\left(1 + TdOAU_{\mathcal{T}}\Lambda_{\mathsf{B}} R_{\mathsf{B}}\right)$ with $R_{\mathsf{B}} := 1 + \max_{h,o,a} \|\mathbf{B}_h(o,a)\|_1$.

Using analysis entirely parallel to Chen et al. (2022a, Appendix G), we can show that E2D-TA (Chen et al., 2022b) and MOPS (Chen et al., 2022a, Algorithm 4) both achieve the same regret guarantees as Theorem 8.

Theorem H.4. *Suppose Θ is a PSR class with the same core test sets $\{\mathcal{U}_h\}_{h\in[H]}$, and each $\theta \in \Theta$ admits a B-representation that is Λ_{B}-strongly-stable (cf. Definition H.1), and has PSR rank $d_{\mathsf{PSR}} \leqslant d$. Then for the coefficients* dec *and* psc *introduced in Chen et al. (2022b), it holds that*

$$\overline{\mathrm{dec}}_\gamma(\Theta) \leqslant \mathcal{O}\left(\frac{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2}{\gamma}\right), \qquad \mathrm{psc}_\gamma(\Theta) \leqslant \mathcal{O}\left(\frac{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2}{\gamma}\right).$$

Therefore, we can apply Chen et al. (2022b, Theorem D.1) (for MOPS*) and Chen et al. (2022b, Theorem C.7) (for* E2D-TA*) to show that, with suitably chosen parameters,* MOPS *and* E2D-TA *both achieve a regret of*

$$\mathbf{Regret} \leqslant \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2 \log(\mathcal{N}_\Theta(1/T))/\delta) T}\right), \tag{47}$$

with probability at least $1-\delta$.

Proof of Theorem 8. To apply Theorem H.3, we first notice that Proposition C.2 readily implies that any single-step α-revealing is strongly B-stable PSR, with $\Lambda_{\mathsf{B}} \leqslant \alpha^{-1}$ and core test sets $\mathcal{U}_h = \mathcal{O}$ for all h. Therefore, applying Theorem H.3 shows that with a model class $\mathcal{M}$ of single-step α-revealing POMDPs, OMLE achieves a regret of

$$\mathbf{Regret} \leqslant \widetilde{\mathcal{O}}\left(\sqrt{\alpha^{-2} SOAH^2 \log \mathcal{N}_{\mathcal{M}}(1/T) \cdot T}\right),$$

as $U_{\mathcal{T}} = 1, d \leqslant S, \Lambda_{\mathsf{B}} \leqslant \alpha^{-1}, R_{\mathsf{B}} \leqslant \alpha^{-1}$ and $\iota = \widetilde{\mathcal{O}}(1)$. Similarly, E2D-TA and MOPS also achieve the same regret upper bound. Noticing that $\log \mathcal{N}_{\mathcal{M}}(1/T) = \widetilde{\mathcal{O}}\big(H(S^2A + SO)\big)$ (Chen et al., 2022a) completes the proof. □

H.3. Proof of Proposition H.2

Claim 1 follows from the proof of Chen et al. (2022a, Lemma D.1) directly. In the following, we show claim 2.

Fix a step $h \in [H]$. An important observation is that, by the strong Λ_{B}-stability of θ (Definition H.1), for any $\pi' \in \Pi_h(\pi)$, we have $\forall x \in \mathbb{R}^{\mathcal{U}_h}$

$$\sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^{\theta}_{H:h}(\tau_{h:H})x\right| \leqslant \Lambda_{\mathsf{B}} \sum_{t_h \in \mathcal{U}_h} \pi'(t_h|\tau_{h-1}) \times |x(t_h)| = \Lambda_{\mathsf{B}} \sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times |x(t_h)|\,, \tag{48}$$

and similarly, for $\forall x \in \mathbb{R}^{\mathcal{U}_{h+1}}$,

$$\sum_{\tau_{h+1:H}} \pi'(\tau_{h+1:H}|\tau_h) \times \left|\mathbf{B}^{\theta}_{H:h+1}(\tau_{h+1:H})x\right| \leqslant \Lambda_{\mathsf{B}} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(t_{h+1}|\tau_h) \times |x(t_{h+1})|\,. \tag{49}$$

Therefore, using use the following formula:

$$\begin{aligned}
&\left(\mathbf{B}^{\theta}_h(o_h,a_h) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\right)\mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \\
&= \mathbf{B}^{\theta}_h(o_h,a_h)\left(\mathbf{q}^{\bar{\theta}}(\tau_{h-1}) - \mathbf{q}^{\theta}(\tau_{h-1})\right) + \left(\mathbf{B}^{\theta}_h(o_h,a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right),
\end{aligned}$$

we have

$$\begin{aligned}
2\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi,\tau_{h-1}) &= \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^{\theta}_{H:h+1}(\tau_{h+1:H})\left(\mathbf{B}^{\theta}_h(o_h,a_h) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\right)\mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right| \\
&\leqslant \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^{\theta}_{H:h}(\tau_{h:H})\left(\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right)\right| \\
&\quad + \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^{\theta}_{H:h+1}(\tau_{h+1:H})\left(\mathbf{B}^{\theta}_h(o_h,a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right)\right| \\
&\leqslant \Lambda_{\mathsf{B}} \sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times \left|\mathbf{e}_{t_h}^{\top}\left(\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right)\right| \\
&\quad + \Lambda_{\mathsf{B}} \sum_{o_h,a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(o_h,a_h,t_{h+1}|\tau_{h-1}) \times \left|\mathbf{e}_{t_{h+1}}^{\top}\left(\mathbf{B}^{\theta}_h(o_h,a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right)\right|,
\end{aligned}$$

where the last inequality uses (48) and (49). Notice that $\mathbf{q}^{\theta}(\tau_{h-1}) = \left[\mathbb{P}_{\theta}(t_h|\tau_{h-1})\right]_{t_h \in \mathcal{U}_h}$, and hence

$$\begin{aligned}
&\sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times \left|\mathbf{e}_{t_h}^{\top}\left(\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right)\right| \\
&= \sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times \left|\mathbb{P}_{\theta}(t_h|\tau_{h-1}) - \mathbb{P}_{\bar{\theta}}(t_h|\tau_{h-1})\right| \\
&\leqslant D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H} = \cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H} = \cdot|\tau_{h-1})\right).
\end{aligned}$$

Also, by the definition of B-representation (cf. Definition B.3), we have

$$\left[\mathbf{B}^{\theta}_h(o,a)\mathbf{q}^{\theta}(\tau_{h-1})\right](t_{h+1}) = \mathbb{P}_{\theta}(t_{h+1}|\tau_{h-1},o,a) \times \mathbb{P}_{\theta}(o|\tau_{h-1}) = \mathbb{P}_{\theta}(o,a,t_{h+1}|\tau_{h-1}),$$

and therefore

$$\begin{aligned}
&\sum_{o_h,a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(o_h,a_h,t_{h+1}|\tau_{h-1}) \times \left|\mathbf{e}_{t_{h+1}}^{\top}\left(\mathbf{B}^{\theta}_h(o_h,a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right)\right| \\
&= \sum_{o_h,a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(o_h,a_h,t_{h+1}|\tau_{h-1}) \times \left|\mathbb{P}_{\theta}(o_h,a_h,t_{h+1}|\tau_{h-1}) - \mathbb{P}_{\bar{\theta}}(o_h,a_h,t_{h+1}|\tau_{h-1})\right| \\
&= \sum_{o_h,a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \left|\mathbb{P}^{\pi}_{\theta}(o_h,a_h,t_{h+1}|\tau_{h-1}) - \mathbb{P}^{\pi}_{\bar{\theta}}(o_h,a_h,t_{h+1}|\tau_{h-1})\right| \\
&\leqslant D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H} = \cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H} = \cdot|\tau_{h-1})\right).
\end{aligned}$$

Combining the inequalities above, we have already shown that

$$\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi,\tau_{h-1}) \leqslant \Lambda_{\mathsf{B}} D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H}=\cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H}=\cdot|\tau_{h-1})\right)$$

for any step $h \in [H]$. Therefore, we can use that fact that $D_{\mathrm{TV}} \leqslant D_{\mathrm{H}}$ and apply Lemma A.6 to obtain

$$\begin{aligned}\mathbb{E}^{\pi}_{\bar{\theta}}\left[\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi,\tau_{h-1})^2\right] \leqslant& \Lambda^2_{\mathsf{B}}\mathbb{E}^{\pi}_{\bar{\theta}}\left[D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H}=\cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H}=\cdot|\tau_{h-1})\right)^2\right] \\ \leqslant& \Lambda^2_{\mathsf{B}}\mathbb{E}^{\pi}_{\bar{\theta}}\left[D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H}=\cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H}=\cdot|\tau_{h-1})\right)\right] \leqslant 2\Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}}\right).\end{aligned}$$

A similar argument can also show that $(\mathcal{E}^{\bar{\theta}}_{\theta,0}(\pi))^2 \leqslant \Lambda^2_{\mathsf{B}} D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}}\right)^2 \leqslant \Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}}\right)$. □

H.4. Proof of Theorem H.3

The proof of Theorem H.3 uses the following fast rate guarantee for the OMLE algorithm, which is standard (e.g. Van de Geer (2000); Agarwal et al. (2020), and a simple proof can be found in (Chen et al., 2022a, Appendix E)).

Proposition H.5 (Guarantee of MLE). *Suppose that we choose $\beta \geqslant 2\log \mathcal{N}_{\Theta}(1/T) + 2\log(1/\delta) + 2$ in Algorithm 1. Then with probability at least $1-\delta$, the following holds:*

(a) For all $k \in [K]$, $\theta^\star \in \Theta^k$;

(b) For all $k \in [K]$ and any $\theta \in \Theta^k$, it holds that

$$\sum_{t=1}^{k-1} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi^t}_{\theta}, \mathbb{P}^{\pi^t}_{\theta^\star}\right) \leqslant 2\beta.$$

We next prove Theorem H.3. We adopt the definitions of $\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi,\tau_{h-1})$ as in Proposition H.2 and abbreviate $\mathcal{E}^{\star}_{k,h} = \mathcal{E}^{\theta^\star}_{\theta^k,h}$. We also condition on the success of the event in Proposition H.5.

Step 1. By Proposition H.5, it holds that $\theta^\star \in \Theta$. Therefore, $V_{\theta^k}(\pi^k) \geqslant V_\star$, and by Proposition H.2, we have

$$\begin{aligned}\sum_{t=1}^{k}\left(V_\star - V_{\theta^\star}(\pi^t)\right) \leqslant& \sum_{t=1}^{k}\left(V_{\theta^t}(\pi^t) - V_{\theta^\star}(\pi^t)\right) \leqslant \sum_{t=1}^{k} D_{\mathrm{TV}}\left(\mathbb{P}^{\pi^t}_{\theta^t}, \mathbb{P}^{\pi^t}_{\theta^\star}\right) \\ \leqslant& \sum_{t=1}^{k} 1 \wedge \left(\mathcal{E}^{\star}_{t,0}(\pi^t) + \sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\mathcal{E}^{\star}_{t,h}(\pi^t,\tau_{h-1})\right]\right) \\ \leqslant& \sum_{t=1}^{k}\left(1 \wedge \mathcal{E}^{\star}_{t,0}(\pi^t) + \sum_{h=1}^{H} 1 \wedge \mathbb{E}_{\pi^t}\left[\mathcal{E}^{\star}_{t,h}(\pi^t,\tau_{h-1})\right]\right),\end{aligned} \tag{50}$$

where the expectation $\mathbb{E}_{\pi^t}$ is taken over $\tau_{h-1} \sim \mathbb{P}^{\pi^t}_{\theta^\star}$. On the other hand, by Proposition H.2, we have

$$\mathbb{E}_{\pi^t}\left[\mathcal{E}^{\star}_{k,h}(\pi^t,\tau_{h-1})^2\right] \leqslant 2\Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi^t}_{\theta^k}, \mathbb{P}^{\pi^t}_{\theta^\star}\right), \qquad \mathcal{E}^{\star}_{k,0}(\pi^t)^2 \leqslant \Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi^t}_{\theta^k}, \mathbb{P}^{\pi^t}_{\theta^\star}\right).$$

Furthermore, by Proposition H.5 we have $\sum_{t=1}^{k-1} D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi^t}_{\theta^k}, \mathbb{P}^{\pi^t}_{\theta^\star}\right) \leqslant 2\beta$. Therefore, combining the two equations above gives

$$\sum_{t<k}\mathbb{E}_{\pi^t}[\mathcal{E}^{\star}_{k,h}(\pi^t,\tau_{h-1})^2] \leqslant 4\Lambda^2_{\mathsf{B}}\beta, \qquad \forall k \in [K], 0 \leqslant h \leqslant H. \tag{51}$$

Step 2. We would like to bridge the performance decomposition (50) and the squared B-errors bound (51) using the generalized ℓ_2-Eluder argument. We consider separately the case for $h \in [H]$ and $h = 0$.

Case 1: $h \in [H]$**.** We denote $m = m_{h+1}$ such that $\mathcal{U}_{h+1} = (\mathcal{O}\times\mathcal{A})^{m-1}\times\mathcal{O}$, $\Omega_{h+1} = (\mathcal{O}\times\mathcal{A})^{m-1}$. By definition,

$$\mathcal{E}^{\star}_{k,h}(\pi^t,\tau_{h-1}) := \frac{1}{2}\max_{\pi'\in\Pi_h(\pi^t)}\sum_{\tau_{h:H}}\pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^k_{H:h+1}(\tau_{h+1:H})\left(\mathbf{B}^k_h(o_h,a_h) - \mathbf{B}^\star_h(o_h,a_h)\right)\mathbf{q}^\star(\tau_{h-1})\right|,$$

$$=\frac{1}{2}\max_{\pi'}\sum_{\tau_{h:H}}\pi'(\tau_{h+m:H}|\tau_{h+m-1})\times\pi^t(\tau_{h:h+m-1}|\tau_{h-1})\times\left|\mathbf{B}^k_{H:h+1}(\tau_{h+1:H})\left(\mathbf{B}^k_h(o_h,a_h)-\mathbf{B}^\star_h(o_h,a_h)\right)\mathbf{q}^\star(\tau_{h-1})\right|$$

$$=\frac{1}{2}\sum_{o_h,a_h}\sum_{\omega_{h+1}\in\Omega_{h+1}}\pi^t(o_h,a_h,\omega_{h+1}|\tau_{h-1})\left\|\mathcal{B}^k_{H:h+m}\cdot\mathbf{B}^k_{h+m-1:h+1}(\omega_{h+1})\left(\mathbf{B}^k_h(o_h,a_h)-\mathbf{B}^\star_h(o_h,a_h)\right)\mathbf{q}^\star(\tau_{h-1})\right\|_\Pi,$$

where in the last equality we adopt the notation introduced in (10).

To bridge between (50) and (51), we invoke the following generalized ℓ_2-Eluder lemma, which can be obtained directly by generalizing Chen et al. (2022a, Proposition C.1 & Corollary C.2) (which correspond to the special case of the following result with $N=1$).

Lemma H.6 (Generalized ℓ_2-Eluder argument)**.** *Suppose we have a sequence of functions $\{f_{k,l}:\mathbb{R}^n\to\mathbb{R}\}_{(k,l)\in[K]\times[N]}$:*

$$f_{k,l}(x):=\max_{r\in\mathcal{R}}\sum_{j=1}^J|\langle x,y_{k,l,j,r}\rangle|,$$

which is given by the family of vectors $\{y_{k,l,j,r}\}_{(k,j,r)\in[K]\times[J]\times\mathcal{R}}\subset\mathbb{R}^n$. Further assume that there exists $L>0$ such that $f_{k,l}(x)\leqslant L\|x\|_1$.

Consider further a sequence of vector $(x_{t,l,i})_{(t,l,i)\in[K]\times[N]\times\mathcal{I}}$, satisfying the following condition

$$\sum_{t=1}^{k-1}\mathbb{E}_{i\sim q_t}\left[\left(\sum_{l=1}^N f_{k,l}(x_{t,l,i})\right)^2\right]\leqslant\beta_k,\qquad\forall k\in[K],$$

and the subspace spanned by $(x_{t,l,i})$ has dimension at most d. Then it holds that

$$\sum_{t=1}^k 1\wedge\mathbb{E}_{i\sim q_t}\left[\sum_{l=1}^N f_{t,l}(x_{t,l,i})\right]\leqslant\sqrt{4Nd\Big(k+\sum_{t=1}^k\beta_t\Big)\log\Big(1+kdL\max_i\|x_i\|_1\Big)},\qquad\forall k\in[K].$$

We have the following three preparation steps to apply Lemma H.6.

1. We define

$$x_{t,l,i}:=\pi^t(o^l_h,a^l_h,\omega^l_{h+1}|\tau^i_{h-1})\times\mathbf{q}^\star(\tau^i_{h-1})\in\mathbb{R}^{\mathcal{U}_h},$$
$$y_{k,l,j,\pi}:=\frac{1}{2}\pi(\tau^j_{h+m:H})\times\left[\mathbf{B}^k_{H:h+m}(\tau^j_{h+1:H})\mathbf{B}^k_{h+m-1:h+1}(\omega^l_{h+1})\left(\mathbf{B}^k_h(o^l_h,a^l_h)-\mathbf{B}^\star_h(o^l_h,a^l_h)\right)\right]^\top\in\mathbb{R}^{\mathcal{U}_h},$$

where $\{\tau^i_{h-1}\}_i$ is an ordering of all possible $\tau_{h-1}\in(\mathcal{O}\times\mathcal{A})^{h-1}$, $\{\tau^j_{h+m:H}=(o_{h+m},a_{h+m},\cdots,o_H,a_H)\}_{j=1}^n$ is an ordering of all possible $\tau_{h+m:H}$ (and hence $n=(OA)^{H-h-m+1}$), $\{(o^l_h,a^l_h,\omega^l_{h+1})\}_{l=1}^N$ is an ordering of $\mathcal{O}\times\mathcal{A}\times\Omega_{h+1}$ (and hence $N=OA\,|\Omega_{h+1}|\leqslant OAU_{\mathcal{T}}$), π is any policy that starts at step h. We then define

$$f_{k,l}(x)=\max_\pi\sum_j|\langle y_{k,l,j,\pi},x\rangle|,\qquad x\in\mathbb{R}^{\mathcal{U}_h}.$$

It follows from definition that

$$\mathcal{E}^\star_{k,h}(\pi^t,\tau^i_{h-1})=\sum_{l=1}^N\pi^t(o^l_h,a^l_h,\omega^l_{h+1}|\tau^i_{h-1})\times f_{k,l}(\mathbf{q}^\star(\tau^i_{h-1}))=\sum_{l=1}^N f_{k,l}(x_{t,l,i}).$$

2. By the assumption that $\theta^\star$ has PSR rank less than or equal to d, we have $\dim\operatorname{span}(x_{t,l,i})\leqslant d$. Furthermore, we have $\|x_{t,l,i}\|_1\leqslant U_A\leqslant U_{\mathcal{T}}$ by definition.

3. It remains to verify that f_k is Lipschitz with respect to 1-norm. Clearly,

$$f_{k,l}(\mathbf{q})\leqslant\frac{1}{2}\left[\|\mathcal{B}^k_{H:h}\mathbf{q}\|_\Pi+\max_{o,a}\|\mathcal{B}^k_{H:h+1}\mathbf{B}^\star_h(o,a)\mathbf{q}\|_\Pi\right]$$

$$\leqslant \frac{1}{2}\left[\Lambda_{\mathsf{B}} \left\|\mathbf{q}\right\|_1 + \Lambda_{\mathsf{B}} \max_{o,a} \left\|\mathbf{B}_h^\star(o,a)\mathbf{q}\right\|_1\right] \leqslant \frac{1}{2}\Lambda_{\mathsf{B}} R_{\mathsf{B}} \left\|\mathbf{q}\right\|_1 .$$

Hence we can take $L = \frac{1}{2}\Lambda_{\mathsf{B}} R_{\mathsf{B}}$ to ensure that $f_{k,l}(x) \leqslant L \left\|x\right\|_1$.

Therefore, applying Lemma H.6 yields

$$\sum_{t=1}^{k} 1 \wedge \mathbb{E}_{\pi^t}\left[\mathcal{E}^\star_{t,h}(\pi^t, \tau_{h-1})\right] \leqslant \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 N d\iota\beta k}\right) \leqslant \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} d\iota\beta k}\right). \tag{52}$$

This completes case 1.

Case 2: $h = 0$**.** This case follows similarly as

$$\sum_{t=1}^{k} 1 \wedge \mathcal{E}^\star_{t,0}(\pi^t) \leqslant \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} \iota\beta k}\right). \tag{53}$$

Combining these two cases, we obtain

$$\begin{aligned}\sum_{t=1}^{k}\left(V_\star - V_{\theta^\star}(\pi^t)\right) &\overset{(i)}{\leqslant} \sum_{t=1}^{k} 1 \wedge \mathcal{E}^\star_{t,0}(\pi^t) + \sum_{h=1}^{H}\sum_{t=1}^{k} 1 \wedge \mathbb{E}_{\pi^t}\left[\mathcal{E}^\star_{t,h}(\pi^t, \tau_{h-1})\right] \\ &\overset{(ii)}{\leqslant} \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} \iota\beta k}\right) + H \cdot \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} d\iota\beta k}\right) \leqslant \mathcal{O}\left(\sqrt{H^2\Lambda_{\mathsf{B}}^2 N d\iota\beta k}\right),\end{aligned}$$

where (i) used (50); (ii) used the above two cases (53) and (52). This completes the proof of Theorem H.3 □

I. Additional discussions

I.1. Impossibility of a generic sample complexity in DEC + log covering number of value/policy class

A typical guarantee of DEC theory (Foster et al., 2021; Chen et al., 2022b) asserts that for any model class $\mathcal{M}$ and policy class Π, the E2D algorithm achieves

$$\mathbb{E}[\mathbf{Regret}] \leqslant \mathcal{O}(1) \cdot \min_{\gamma>0}\left(T \cdot \mathrm{dec}^{\mathrm{H}}_\gamma(\mathcal{M}) + \gamma \log|\mathcal{M}|\right). \tag{54}$$

Foster et al. (2021) also showed that, letting $\mathrm{co}(\mathcal{M})$ denote the convex hull of $\mathcal{M}$ (the set of all mixture models of $M \in \mathcal{M}$), there is a variant of E2D that achieves

$$\mathbb{E}[\mathbf{Regret}] \leqslant \mathcal{O}(1) \cdot \min_{\gamma>0}\left(T \cdot \mathrm{dec}^{\mathrm{H}}_\gamma(\mathrm{co}(\mathcal{M})) + \gamma \log|\Pi|\right).$$

However, $\mathrm{dec}^{\mathrm{H}}_\gamma(\mathrm{co}(\mathcal{M}))$ is typically intractable large—For example, when $\mathcal{M}$ is the class of all tabular MDPs, $\mathrm{dec}^{\mathrm{H}}_\gamma(\mathrm{co}(\mathcal{M}))$ scales exponentially in S, H (Foster et al., 2022). Therefore, it is natural to ask the following

> **Question:** Is it possible to obtain a regret upper bound that replaces the term $\log|\mathcal{M}|$ in (54) by $\log|\Pi|$ or $\log|\mathcal{F}_{\mathcal{M}}|$ (where $\mathcal{F}_{\mathcal{M}}$ is a certain class of value functions induced by $\mathcal{M}$)?

The question above is of particular interest when the model class $\mathcal{M}$ itself is much larger than the value class (e.g. Q-function class), for example when $\mathcal{M}$ is a class of linear MDPs (Jin et al., 2020b) with a known feature $\phi(s,a)$ but unknown $\mu(s')$. Also, replacing $\log|\mathcal{M}|$ in (54) by $\log|\Pi|$ could be a decent improvement for specific problem classes, such as tabular MDPs in which case we can take Π to be the class of deterministic Markov policies with $\log|\Pi| = \widetilde{\mathcal{O}}(SH)$, which is smaller than $\log|\mathcal{M}| = \widetilde{\mathcal{O}}(\log\mathcal{N}_{\mathcal{M}}) = \widetilde{\mathcal{O}}\big(S^2AH\big)$ by a factor of SA.

However, our lower bounds for revealing POMDPs—specifically our hard instance construction in Appendix F—provides a (partially) negative answer to this question. For simplicity, consider the $m = 2$ case, and assume $A^H \gg \mathrm{poly}(S, O, A, \alpha^{-1}, T)$) We have the following basic facts about our model class $\mathcal{M}$.

1. The structure of $\mathcal{M}$ ensures that any possibly optimal policy is a deterministic action sequence (that does not depend on the history), and hence we can take $\Pi = \{\text{determinstic action sequences}\}$, with $\log \Pi = \widetilde{\mathcal{O}}(H)$.
2. The general results in Chen et al. (2022a) shows that as long as $\mathcal{M}$ is a subclass of 2-step α-revealing POMDPs, it holds that $\mathrm{edec}_\gamma(\mathcal{M}) \leqslant \widetilde{\mathcal{O}}(SA^2H^2\alpha^{-2}/\gamma)$—where edec is a PAC-learning analogue of the dec—which implies that $\mathrm{dec}_\gamma(\mathcal{M}) \leqslant \widetilde{\mathcal{O}}(\sqrt{SA^2H^2\alpha^{-2}/\gamma})$ (Chen et al., 2022b).
3. Proposition F.1 states that worst-case regret within family $\mathcal{M}$ for any algorithm is lower bounded by $\Omega\big((S\sqrt{O}A^2H\alpha^{-2})^{1/3}T^{2/3}\big)$.

Note that the regret lower bound involves a $\mathrm{poly}(O)$ factor, which does not appear in the upper bound for the dec. This leads to the following

> **Fact:** Without further structural assumptions for the problem, a regret upper bound of the form
>
> $$\mathbb{E}[\mathbf{Regret}] \leqslant \mathcal{O}(1) \cdot \min_{\gamma>0} \Big(T \cdot \mathrm{dec}^{\mathrm{H}}_\gamma(\mathcal{M}) + \gamma \log|\Pi| \Big) \tag{55}$$
>
> is not achievable.

The above fact is because that if (55) were achievable, then combining with the aformentioned dec upper bound would result in a regret upper bound that does not scale with $\mathrm{poly}(O)$, contradicting the lower bound.

Similarly, if we view each POMDP $M \in \mathcal{M}$ as an MDP by viewing each history τ_h as a "mega-state", then naturally the Q-function of M is given by

$$Q^\star_M(\tau_h) = \mathbb{E}^{\pi^\star_M}_M \left[\sum_{h'=1}^{H} r_{h'} \,\middle|\, \tau_h \right], \qquad \tau_h \in (\mathcal{O}\times\mathcal{A})^h, 0 \leqslant h \leqslant H,$$

where $\pi^\star_M$ is the optimal policy for M. For our family $\mathcal{M}$, it is straightforward to check that $\log \mathcal{Q}_\mathcal{M} = \widetilde{\mathcal{O}}(H)$, where $\mathcal{Q}_\mathcal{M} = \{Q^\star_M : M \in \mathcal{M}\}$. Therefore, the answer to the question above is also negative if we take the value class to be such a Q-function class.

I.2. Algorithms for hard instances of Theorem 5

We propose a brute-force algorithm $\mathfrak{A}$ to learn the class of hard instances provided in Appendix G (for proving Theorem 5), which admits a PAC sample complexity $\widetilde{\mathcal{O}}\big(S^{3/2}O^{1/2}A^mH/(\alpha^2\varepsilon^2)\big)$. Algorithm $\mathfrak{A}$ contains two stages:

1. Stage 1: For each $h \in \mathcal{H}, s \in \mathcal{S}_{\mathsf{leaf}}, a \in \mathcal{A}_c, \mathbf{a} \in \mathcal{A}^{m-1}$, the algorithm spends N_1 episodes on visiting $o_h = s$, taking actions $a_{h:h+m} = (a, \mathbf{a}, \mathsf{reveal})$, and observing (o_{h+m}, o_{h+m+1}). The observed (o_{h+m}, o_{h+m+1}) should then satisfy the joint distribution

   $$\mathbb{P}(\mathsf{lock}^j, o_i^+) = \frac{1+\sigma\varepsilon\mu_{j,i}}{2KL}, \qquad \mathbb{P}(\mathsf{lock}^j, o_i^-) = \frac{1-\sigma\varepsilon\mu_{j,i}}{2KL}, \qquad \forall (j,i) \in [L]\times[K]$$

   if $(h,s,a,\mathbf{a}) = (h^\star, s^\star, a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})$, and satisfy distribution $\mathrm{Unif}(\{\mathsf{lock}^1, \cdots, \mathsf{lock}^L\} \times \mathcal{O}_o)$ otherwise. Using the standard uniformity testing algorithm (Canonne, 2020), we can distinguish between

   $$H_0 : (h,s,a,\mathbf{a}) = (h^\star, s^\star, a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1}),$$
   $$H_1 : (h,s,a,\mathbf{a}) \neq (h^\star, s^\star, a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})$$

   with high probability using $N_1 = \widetilde{\mathcal{O}}\big(\sqrt{KL}/(\sigma^2\varepsilon^2)\big)$ samples for every fixed $(h,s,a,\mathbf{a})$. The total sample size needed in Stage 1 is thus $|\mathcal{S}_{\mathsf{leaf}}|\, H\, |\mathcal{A}_c|\, A^{m-1} \times \widetilde{\mathcal{O}}\big(\sqrt{KL}/(\sigma^2\varepsilon^2)\big)$.
2. Stage 2: Once Stage 1 is completed, the algorithm can correctly identify the parameter $(h^\star, s^\star, a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})$ (if $M \neq 0$) or find out $M = 0$. In the latter case, the algorithm can directly terminate and output the optimal policy of $M = 0$. In the former case, the algorithm needs to continue to learn the password $\mathbf{a}^\star_{h^\star+m:H-1}$:

- For each $h = h^\star + m, h^\star + 2m, \cdots$:
  - For each $\mathbf{a} \in \mathcal{A}^m$, test whether $\mathbf{a} = \mathbf{a}^\star_{h:h+m-1}$ by spending N_1 episodes on visiting $o_{h^\star} = s^\star$, taking actions $a_{h^\star:h+m} = (a, \mathbf{a}^\star_{h^\star+1:h-1}, \mathbf{a}, \mathsf{reveal})$, and observing (o_{h+m}, o_{h+m+1}).
  - By the same reason as in Stage 1 and by our choice that $N_1 = \widetilde{\mathcal{O}}\big(\sqrt{KL}/(\sigma^2\varepsilon^2)\big)$, we can learn $\mathbf{a}^\star_{h:h+m-1}$ with high probability, using the standard uniformity testing algorithm.

  Once the algorithm learns the $M = (h^\star, s^\star, a^\star, \mathbf{a}^\star)$, it terminates and outputs the optimal policy of M. The total sample size needed in Stage 2 is at most $A^m H \times \widetilde{\mathcal{O}}\big(\sqrt{KL}/(\sigma^2\varepsilon^2)\big)$ many samples.

To summarize, the brute-force algorithm $\mathfrak{A}$ we construct above can learn $\mathcal{M}$ with sample size

$$|\mathcal{S}_{\mathsf{leaf}}|\, H\, |\mathcal{A}_c|\, A^{m-1} \times \widetilde{\mathcal{O}}\left(\frac{\sqrt{KL}}{\sigma^2\varepsilon^2}\right) + A^m H \times \widetilde{\mathcal{O}}\left(\frac{\sqrt{KL}}{\sigma^2\varepsilon^2}\right) \leqslant \widetilde{\mathcal{O}}\left(\frac{S^{3/2}O^{1/2}A^m H}{\alpha^2\varepsilon^2}\right),$$

where the bound is by our choice of $\sigma, \mathcal{S}_{\mathsf{leaf}}, \mathcal{A}_c, K, L$.