Proceedings of Machine Learning Research vol 195:1–151, 2023 36th Annual Conference on Learning Theory

# Instance-Optimality in Interactive Decision Making: Toward a Non-Asymptotic Theory

**Andrew Wagenmaker** AJWAGEN@CS.WASHINGTON.EDU
*University of Washington, Seattle, WA*

**Dylan J. Foster** DYLANFOSTER@MICROSOFT.COM
*Microsoft Research, New England*

**Editors:** Gergely Neu and Lorenzo Rosasco

## Abstract

We consider the development of adaptive, instance-dependent algorithms for interactive decision making (bandits, reinforcement learning, and beyond) that, rather than only performing well in the worst case, adapt to favorable properties of real-world instances for improved performance. We aim for *instance-optimality*, a strong notion of adaptivity which asserts that, on any particular problem instance, the algorithm under consideration outperforms all consistent algorithms. Instance-optimality enjoys a rich asymptotic theory originating from the work of Lai and Robbins (1985) and Graves and Lai (1997), but *non-asymptotic* guarantees have remained elusive outside of certain special cases. Even for problems as simple as tabular reinforcement learning, existing algorithms do not attain instance-optimal performance until the number of rounds of interaction is *doubly exponential* in the number of states.

In this paper, we take the first step toward developing a non-asymptotic theory of instance-optimal decision making with general function approximation. We introduce a new complexity measure, the Allocation-Estimation Coefficient (AEC), and provide a new algorithm, $\mathsf{AE}^2$, which attains non-asymptotic instance-optimal performance at a rate controlled by the AEC. Our results recover the best known guarantees for well-studied problems such as finite-armed and linear bandits and, when specialized to tabular reinforcement learning, attain the first instance-optimal regret bounds with polynomial dependence on all problem parameters, improving over prior work exponentially. We complement these results with lower bounds that show that i) existing notions of statistical complexity are insufficient to derive non-asymptotic guarantees, and ii) under certain technical conditions, boundedness of the Allocation-Estimation Coefficient is *necessary* to learn an instance-optimal allocation of decisions in finite time.

## 1. Introduction

We consider the development of adaptive, sample-efficient algorithms for *interactive decision making*, encompassing bandit problems and reinforcement learning with general function approximation. For decision making in high-dimensional spaces with a long horizon, existing approaches (Lillicrap et al., 2015; Mnih et al., 2015; Silver et al., 2016) are sample-hungry, which presents an obstacle for real-world deployment in settings where data is scarce or high-quality simulators are not available. To overcome this challenge, algorithms should both i) flexibly incorporate users' domain knowledge, as expressed via modeling and function approximation, and ii) explore the environment in a deliberate, adaptive fashion, taking advantage of favorable structure whenever possible.

Toward achieving these goals, a major area of research aims to develop algorithms with optimal sample complexity and understand the fundamental limits for such algorithms (Russo and Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020; Du et al., 2021; Jin et al., 2021; Foster

et al., 2021), and the foundations are beginning to fall into place. In particular, focusing on *minimax regret* (that is, the best regret that can achieved for a worst-case problem instance in a given class of problems), Foster et al. (2021, 2022b, 2023) provide unified algorithm design principles and measures of statistical complexity that are both necessary and sufficient for low regret. However, minimax regret and other notions of worst-case performance are inherently pessimistic, and may not be sufficient to close the gap between theory and practice. For example, recent work has shown that algorithms that are optimal in the worst-case can be arbitrarily suboptimal on "easier" instances (Wagenmaker et al., 2022b). To overcome these challenges and develop algorithms that perform well on *every* instance, a promising approach is to develop algorithms that *adapt* to the difficulty of the problem instance under consideration.

The performance of such adaptive algorithms can be quantified through *instance-dependent* regret bounds, which become smaller (leading to low regret) when the underlying problem instance is favorable. Algorithms with such guarantees have been studied throughout the literature on bandits and reinforcement learning; basic examples include adapting to large gaps in value between alternative actions (Lai and Robbins, 1985) or low noise or variance in bandit problems (Allenberg et al., 2006; Hazan and Kale, 2011; Foster et al., 2016; Wei and Luo, 2018; Bubeck et al., 2018), and adapting to the difficulty of reaching certain states in Markov Decision Processes (Zanette and Brunskill, 2019; Simchowitz and Jamieson, 2019; Dann et al., 2021; Wagenmaker et al., 2022b).

While there are many notions of adaptivity and instance-dependence, they are generally incomparable. A stronger notion of adaptivity is *instance-optimality*, which asserts that the performance of the algorithm on a problem instance of interest exceeds that of *any consistent algorithm* (that is, any algorithm with sublinear regret for all problem instances). Instance-optimality enjoys a rich theory originating with the work of Lai and Robbins (1985) and Graves and Lai (1997), with a celebrated line of research developing sharp guarantees for the special case of finite-armed bandits (Burnetas and Katehakis, 1996; Garivier et al., 2016; Kaufmann et al., 2016; Lattimore, 2018; Garivier et al., 2019). Beyond the finite-armed bandit setting, however, development has been largely *asymptotic* in nature, and existing algorithms either:

1. achieve instance-optimality only as $T \to \infty$ (or, to the extent that they are non-asymptotic, require $T$ to be exponentially large in problem-dependent parameters) (Graves and Lai, 1997; Komiyama et al., 2015; Combes et al., 2017; Degenne et al., 2020b; Dong and Ma, 2022), or

2. achieve non-asymptotic guarantees, but require restrictive modeling assumptions such as linear function approximation (Tirinzoni et al., 2020; Kirschner et al., 2021).

Indeed, even for the simple problem of tabular (finite-state/action) reinforcement learning, existing algorithms do not attain instance-optimal performance until the number of rounds of interaction is *doubly exponential* in the number of states (Ok et al., 2018; Dong and Ma, 2022). In this paper, we address these challenges, providing algorithms that i) accommodate flexible, general-purpose function approximation, and ii) attain instance-optimality in finite time, in a sense which is itself optimal.

**Contributions.** We take the first steps toward building a non-asymptotic theory of instance-optimal decision making. We observe that asymptotic characterizations for instance-optimality:

1. reflect the regret incurred by an allocation of decisions designed to optimally distinguish the ground truth problem instance from a set of alternatives, but

2. do not capture the statistical complexity required to *learn* such an allocation.

To address this, we introduce a new complexity measure, the Allocation-Estimation Coefficient (AEC), which aims to capture the statistical complexity of learning an optimal Graves-Lai allocation. We provide a new algorithm, $\mathsf{AE}^2$, which attains non-asymptotic instance-optimal regret at a rate controlled by the AEC. We complement this result with lower bounds that show that under certain technical conditions, boundedness of the Allocation-Estimation Coefficient is not just sufficient, but *necessary* to learn an instance-optimal allocation in finite time.

Our algorithm is simple, and can be applied to any hypothesis class in a generic fashion. It recovers the best known guarantees for standard problems such as finite-armed and linear bandits and, when specialized to tabular reinforcement learning, achieves the first instance-optimal regret bounds with polynomial dependence on all problem parameters. We believe that our approach clarifies and elucidates many tradeoffs and statistical considerations left implicit in prior work, and hope that it will serve as a foundation for further development of instance-optimal algorithms.

### 1.1. Interactive Decision Making

We adopt the *Decision Making with Structured Observations* (DMSO) framework of Foster et al. (2021), which is a general setting for interactive decision making that encompasses bandit problems (structured, contextual, and so forth) and reinforcement learning with function approximation.

The DMSO framework is specified by a *decision space* $\Pi$, *reward space* $\mathcal{R} \subseteq \mathbb{R}$, and *observation space* $\mathcal{O}$. The learner is given access to a (known) *model class* $\mathcal{M} \subset (\Pi \to \triangle_{\mathcal{R}\times\mathcal{O}})$, and it is assumed there exists some true model $M^\star \in \mathcal{M}$, unknown to the learner, which represents the underlying environment. Formally, we make the following assumption.

**Assumption 1.1** (Realizability)**.** *We have that* $M^\star \in \mathcal{M}$.

The learning protocol consists of $T$ rounds. For each round $t = 1, \dots, T$:

1. The learner selects a *decision* $\pi^t \in \Pi$.
2. The learner receives a reward $r^t \in \mathcal{R}$ and observation $o^t \in \mathcal{O}$ sampled $(r^t, o^t) \sim M^\star(\pi^t)$, and observes $(r^t, o^t)$.

We can think of the model class $\mathcal{M}$ as representing the learner's prior knowledge about the decision making problem, and it allows one to appeal to estimation and function approximation. For structured bandit problems, for example, models correspond to reward distributions, and $\mathcal{M}$ encodes structure in the reward landscape. For reinforcement learning problems, models correspond to Markov decision processes (MDPs), and $\mathcal{M}$ typically encodes structure in value functions or transition probabilities. See Appendix A.6 and Appendix A.7 for concrete examples of how standard decision-making settings can be instantiated within the DMSO framework, and Foster et al. (2021) for further background. For a model $M \in \mathcal{M}$, $\mathbb{E}^{M,\pi}[\cdot]$ denotes the expectation under the process $(r, o) \sim M(\pi)$, $f^M(\pi) := \mathbb{E}^{M,\pi}[r]$ denotes the mean reward function, and $\pi_M := \arg\max_{\pi\in\Pi} f^M(\pi)$ denotes the optimal decision. When the algorithm is clear from context, $\mathbb{E}^M[\cdot]$ and $\mathbb{P}^M[\cdot]$ refer to the expectation and probability measure, respectively, induced over histories under $M$. When the context is clear, we overload notation somewhat and use $\mathbb{P}^{M,\pi}[\cdot]$ to refer to the conditional density over $\mathcal{R} \times \mathcal{O}$ induced by playing $\pi$ on $M$. We make the following assumptions.

**Assumption 1.2** (Bounded Reward Means)**.** *For all* $M \in \mathcal{M}$, $\pi \in \Pi$, *we have* $f^M(\pi) \in [0, 1]$.

**Assumption 1.3** (Unique Optimal Action). *For the ground truth model $M^\star \in \mathcal{M}$, the optimal action $\pi_{M^\star}$ is unique.*

Note that the latter assumption is standard in the literature on instance-optimality. We measure performance in terms of regret, which is given by

$$\mathbf{Reg}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right], \tag{1}$$

where $p^t$ is the learner's randomization distribution for round $t$. In addition, we define $\Delta^M(\pi) = f^M(\pi_M) - f^M(\pi)$ as the *suboptimality gap* function for model $M$ and decision $\pi$, and the *minimum suboptimality gap* as

$$\Delta^M_{\mathrm{min}} := \begin{cases} \inf_{\pi \in \Pi : \Delta^M(\pi) > 0} \Delta^M(\pi), & \pi_M \text{ is unique,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Since by assumption $\pi_{M^\star}$ is unique, we have $\Delta^{M^\star}_{\mathrm{min}} > 0$. Throughout, we replace dependence on $M^\star$ with "$\star$" when the meaning is clear from context, for example: $\Delta^\star_{\mathrm{min}} := \Delta^{M^\star}_{\mathrm{min}}$, or $f^\star(\pi) := f^{M^\star}(\pi)$.

**Further Notation.** We let $\mathcal{M}^+ = \{M : \Pi \to \triangle_{\mathcal{R}\times\mathcal{O}} \mid f^M(\pi) \in [0,1]\}$ denote the space of all possible models $M$ with rewards in $\mathcal{R}$ and $f^M(\pi) \in [0,1]$. We use $\triangle_{\mathcal{X}}$ to refer to the set of probability distributions over any $\mathcal{X}$. Throughout, we often abbreviate $\mathbb{E}_{\pi\sim p}[\cdot]$ with $\mathbb{E}_p[\cdot]$.

### 1.2. Background: Asymptotic Instance-Optimality

Our aim is to develop algorithms that are *instance-optimal* in a strong sense: for *every model* $M^\star \in \mathcal{M}$, the regret of the algorithm under $M^\star$ is at least as good as that of any *consistent* algorithm; here, an algorithm is said to be "consistent" if it ensures that $\mathbb{E}^M[\mathbf{Reg}(T)] = o(T)$ for all $M \in \mathcal{M}$. Instance-optimality is a powerful notion of performance: no algorithm—even one designed specifically with $M^\star$ in mind—can achieve lower regret on $M^\star$ without giving up consistency. For multi-armed bandits, a long line of work initiated by Lai and Robbins (1985) characterizes the instance-optimal regret as a function of the instance $M^\star$, and provides efficient algorithms that attain instance-optimality in finite time (Garivier et al., 2016; Kaufmann et al., 2016; Lattimore, 2018; Garivier et al., 2019). For the general decision making setting we consider, the forward-looking work of Graves and Lai (1997) (see Dong and Ma (2022) for a contemporary treatment) introduced a complexity measure we refer to as the *Graves-Lai Coefficient*, which asymptotically characterizes the instance-optimal performance as a function of the instance $M^\star$ and model class $\mathcal{M}$. For any class $\mathcal{M}$ and model $M \in \mathcal{M}$, the Graves-Lai Coefficient is defined as

$$\mathsf{glc}(\mathcal{M}, M) := \inf_{\eta \in \mathbb{R}^{\Pi}_+} \left\{ \sum_{\pi\in\Pi} \eta_\pi \Delta^M(\pi) \mid \forall M' \in \mathcal{M}^{\mathrm{alt}}(M) : \sum_{\pi\in\Pi} \eta_\pi D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1 \right\}, \tag{3}$$

where, for $M$ with unique optimal decision $\pi_M$, we define $\mathcal{M}^{\mathrm{alt}}(M) := \{M' \in \mathcal{M} \mid \pi_M \notin \boldsymbol{\pi}_{M'}\}$ the set of "alternative" models—the models $M' \in \mathcal{M}$ that disagree with $M$ on the optimal decision[1]—and $D_{\mathsf{KL}}(\cdot\,\|\,\cdot)$ denotes the Kullback-Leibler divergence. When $\mathcal{M}$ is clear from context, we will abbreviate $\mathsf{g}^M := \mathsf{glc}(\mathcal{M}, M)$ (and $\mathsf{g}^\star := \mathsf{glc}(\mathcal{M}, M^\star)$). We also denote any solution to Eq. (3) by $\eta^M$—note that this is not in general unique. The characterization of Graves and Lai (1997) is as follows.

1. For $M \in \mathcal{M}$ such that $\pi_M$ is not unique, we define $\mathcal{M}^{\mathrm{alt}}(M) := \{M' \in \mathcal{M} \mid \boldsymbol{\pi}_M \cap \boldsymbol{\pi}_{M'} = \varnothing\}$, and define $\mathsf{glc}(\mathcal{M}, M)$ as in Eq. (3), with respect to this $\mathcal{M}^{\mathrm{alt}}(M)$. We also define $\mathcal{M}^{\mathrm{alt}}(\pi) = \{M \in \mathcal{M} \mid \pi \notin \boldsymbol{\pi}_M\}$.

**Proposition 1.1** (Graves and Lai (1997); Dong and Ma (2022)). *For any model class $\mathcal{M}$ with $|\Pi| < \infty$, any algorithm that is consistent with respect to $\mathcal{M}$ must have*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \geq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) - o(\log(T)) \tag{4}$$

*for any $M^\star \in \mathcal{M}$, and there exists an algorithm which achieves, for all $M^\star \in \mathcal{M}$ satisfying Assumption 1.3,*[2]

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + o(\log(T)). \tag{5}$$

The interpretation of the Graves-Lai Coefficient of $M^\star$ with respect to $\mathcal{M}$, $\mathsf{glc}(\mathcal{M}, M^\star)$, is simple. It asks, if $M^\star$ is known to the learner (to be clear, $M^\star$ is not known a-priori), what is the minimum regret that must be incurred to gather enough information to rule out all possible alternatives $M' \in \mathcal{M}$ which do not have $\pi_{M^\star}$ as an optimal decision (i.e., $\pi_{M^\star} \notin \boldsymbol{\pi}_{M'}$)? In other words, it aims to *certify* that $\pi_{M^\star}$ is indeed the optimal decision while incurring the minimum regret possible.

The Graves-Lai Coefficient characterization is appealing in its simplicity, but the catch—at least when one moves beyond finite-armed bandits—is hiding in the lower-order terms, particularly for the upper bound (5). For general model classes, the best known finite-time regret bounds (Dong and Ma, 2022) take the form

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \mathrm{poly}(|\Pi|, (\Delta_{\min}^{M^\star})^{-1}) \cdot \log^{1-c}(T) \tag{6}$$

where $c > 0$ is a universal constant. While this indeed leads to instance-optimality as $T \to \infty$, the "lower-order" term in Eq. (6) scales with the size of the decision space, which is intractably large for most problems of interest. As an example, consider the problem of tabular reinforcement learning in an episodic MDP with $S$ states, $A$ actions, and horizon $H$. Here, we typically have $\mathsf{glc}(\mathcal{M}, M^\star) = \mathrm{poly}(S, A, H)$ , yet $|\Pi| = A^{HS}$. Consequently, the Graves-Lai Coefficient does not become the dominant term in (6) until $\geq \exp(\exp(S))$. That is, for realistic time horizons, asymptotic instance-optimality does not tell the full story.

**Learning an optimal allocation.** Given knowledge of an optimal Graves-Lai allocation $\eta^{M^\star}$ solving Eq. (3), a learner could simply take actions as specified by $\eta^{M^\star}$, and would achieve the instance-optimal rate given in Proposition 1.1. However, this is typically infeasible, as the optimal allocation itself depends strongly upon the ground truth model $M^\star$, and is therefore unknown to the learner. In light of this challenge, the approach taken by essentially all existing algorithms (Burnetas and Katehakis, 1996; Graves and Lai, 1997; Magureanu et al., 2014; Komiyama et al., 2015; Lattimore and Szepesvari, 2017; Combes et al., 2017; Hao et al., 2019, 2020; Van Parys and Golrezaei, 2020; Degenne et al., 2020b; Tirinzoni et al., 2020; Kirschner et al., 2021; Dong and Ma, 2022) is to first learn an estimate for a Graves-Lai allocation $\eta^{M^\star}$, and then take actions as specified by this estimate. In addition to being natural, this approach is *necessary* in a certain (weak) sense: for any algorithm that achieves instance-optimality, the expected decision frequencies must converge to an approximately optimal allocation as $T$ grows (cf. Lemma B.1).[3]

The presence of the lower-order term scaling with $|\Pi|$ in Eq. (5) (and in similar regret bounds from most existing work) reflects the sample complexity required to learn an optimal Graves-Lai allocation

2. To be precise, rather than scaling directly with $\mathsf{glc}(\mathcal{M}, M^\star)$, the upper bound given by Dong and Ma (2022) scales with a quantity $\mathsf{glc}_T(\mathcal{M}, M^\star)$ such that $\mathsf{glc}_T(\mathcal{M}, M^\star) \to_T \mathsf{glc}(\mathcal{M}, M^\star)$.
3. The connection between instance-optimal regret and learning an optimal allocation has many subtleties; we refer ahead to Appendix B for extensive discussion.

through uniform exploration. Specifically, one can estimate an allocation by uniformly exploring the decision space to gather data, and then solving an empirical approximation to the Graves-Lai program (3). Naive exploration of this type inevitably results in $\Omega(|\Pi|)$ sample complexity, and it is natural to ask whether a more deliberate exploration strategy, perhaps by exploiting the structure of $\mathcal{M}$, could lead to better finite-time regret bounds. For the setting of linear bandits, where $\Pi \subseteq \mathbb{R}^d$ and the mean reward function $\pi \mapsto f^M(\pi)$ is linear, this is indeed the case: a recent line of work (Tirinzoni et al., 2020; Kirschner et al., 2021) provides regret bounds of the form

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \mathrm{poly}(d, (\Delta_{\min}^{M^\star})^{-1}) \cdot \log^{1-c}(T).$$

This bound replaces the size of the decision space in the lower-order term by the dimension $d$, reflecting the fact that there are only $d$ "effective" directions in which exploration is required. While this is an encouraging start, the techniques used in these works are specialized to linear bandits, and it is unclear how to generalize them beyond this setting.

### 1.3. A Motivating Example

As discussed in the prequel, for finite-armed bandits and linear bandits, it is possible to achieve instance-optimal regret bounds where the lower-order terms scale with the number of actions $A$, or dimension $d$, respectively. Extrapolating, one might be tempted to ask whether we can always learn a near-optimal allocation with sample complexity no larger than, say, the minimax rate for $\mathcal{M}$. The starting point for our work is to recognize that in general, the answer is no: existing notions of statistical complexity are insufficient to capture the complexity of learning the Graves-Lai allocation in finite time, as illustrated in the following simple example.

**Example 1.1** (Searching for an informative arm)**.** Let $A, N \geq 2$ and $\beta \in (0,1)$ be parameters, and consider the class $\mathcal{M}$ of all models defined as follows. First, $\Pi = [A] \cup \{\pi_i^\circ\}_{i\in[N]}$; decisions in $[A]$ are "bandit arms", and decisions in $\{\pi_i^\circ\}_{i\in[N]}$ are "informative" (or, revealing) arms. Each model $M$ has a unique optimal decision $\pi_M$, and the following structure, with $\mathcal{O} = [A] \cup \{\perp\}$.

- For each bandit arm $k \in [A]$ we have $r \sim \mathcal{N}(f^M(k), 1)$ for $f^M \in [0,1]$. There are no observations, i.e. $o = \perp$ almost surely.
- All informative arms $\pi_k^\circ$ give 0 reward almost surely. There exists a unique informative arm $\pi_M^\circ \in \{\pi_i^\circ\}_{i\in[N]}$ associated with $M$, so that if we play any $\pi_k^\circ$, we receive an observation

$$o \sim \begin{cases} \mathrm{Unif}([A]), & \pi_k^\circ \neq \pi_M^\circ, \\ \beta\mathbb{I}_{\pi_M} + (1-\beta)\mathrm{Unif}([A]), & \pi_k^\circ = \pi_M^\circ. \end{cases}$$

We take $\mathcal{M}$ to consist of all possible models with this structure. The interpretation here is as follows. If one were to ignore the revealing arms $\{\pi_i^\circ\}_{i\in[N]}$, this would be a standard finite-armed bandit problem. In particular, if we were to consider a model $M^\star$ with $f^{M^\star}(\pi) = \frac{1}{2} + \Delta\mathbb{I}\{\pi = i\}$ for $i \in [A]$, a standard calculation would yield $\mathsf{g}^{M^\star} \propto \frac{A}{\Delta}$. However, the presence of the informative arms makes the problem substantially easier. With $\beta = 9/10$ (for concreteness), one can see that for the model $M$, pulling the informative arm $\pi_{M^\star}^\circ$ will give $o = \pi_{M^\star}$ with probability at least $9/10$, meaning that we can identify that $\pi_{M^\star}$ is optimal with high probability by pulling $\pi_{M^\star}^\circ$ a constant number of times. It follows that the optimal allocation is to ignore the bandit arms and set

$\eta^{M^\star}(\pi) \propto \mathbb{I}\{\pi = \pi^\circ_{M^\star}\}$. This gives $\mathsf{g}^{M^\star} \leq O(1)$, which is substantially better than $\mathsf{g}^{M^\star} \propto \frac{A}{\Delta}$ if $\Delta$ is small or $A$ is large.

If one only is only concerned with asymptotic rates, this is the end of the story, but for non-asymptotic rates, we need to consider the amount of exploration required to *learn the optimal allocation*. In particular, in order to identify the informative arm $\pi^\circ_{M^\star}$, which is necessary to learn the optimal allocation, it is clear that in the worst case, any algorithm needs to try all of the revealing arms, leading to $\mathbb{E}[\mathbf{Reg}(T)] = \Omega(N)$. While the complexity of learning the optimal allocation is washed away by an asymptotic analysis with $T \to \infty$, it cannot be ignored for finite $T$. In addition, the $\Omega(N)$ factor cannot be explained away by standard complexity measures. As we have seen, $\sup_{M \in \mathcal{M}} \mathsf{g}^M = O(1)$, so the Graves-Lai Coefficient is not sufficient to explain it. Furthermore, the minimax rate for this problem is always bounded by $O(\sqrt{AT})$, which does not scale with $N$; yet $\Omega(A)$ sample complexity does not suffice to learn an optimal allocation. In addition, it can be shown that existing complexity measures such as the Decision-Estimation Coefficient (Foster et al., 2021) and information ratio (Russo and Van Roy, 2018) also do not scale with $N$. ◁

Example 1.1 shows that if we want to achieve instance-optimality in finite time, new notions of problem complexity for the class $\mathcal{M}$ are required, motivating the following central questions:

1. Can we develop algorithms for general model classes $\mathcal{M}$ that achieve non-asymptotic instance-optimal regret bounds of the form
$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \mathrm{comp}(\mathcal{M}) \cdot \log^{1-c}(T), \tag{7}$$
where $\mathrm{comp}(\mathcal{M})$ is a complexity measure that reflects the intrinsic difficulty of exploring in order to learn a Graves-Lai optimal allocation for $\mathcal{M}$?
2. Can we understand when the presence of such lower-order terms is *necessary*?

### 1.4. Organization

The remainder of the paper is organized as follows. In Section 2 we introduce a novel complexity measure, the Allocation-Estimation Coefficient, which captures the complexity of learning a Graves-Lai optimal allocation (Section 2.1), present our main upper and lower bounds (Section 2.2), and instantiate our bounds on several examples (Section 2.3). In Section 3 we present an overview of our main algorithm, $\mathsf{AE}^2$, and in Section 4 offer directions for future work. Due to space constraints, results in the main body are presented informally—see Part I of the appendix for full statements.

## 2. Overview of Results

To capture the statistical complexity of learning an optimal Graves-Lai allocation in finite time, we provide a new complexity measure, the *Allocation-Estimation Coefficient* (AEC).

### 2.1. The Allocation-Estimation Coefficient

For a model $M \in \mathcal{M}$ and parameter $\varepsilon \in [0, 1]$, we define

$$\Lambda(M; \varepsilon) = \Bigg\{ \lambda \in \triangle_\Pi \mid \exists \mathsf{n} \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{\pi \sim \lambda}[\Delta^M(\pi)] \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}, \\ \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi \sim \lambda}\big[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)\big] \geq \frac{1-\varepsilon}{\mathsf{n}} \Bigg\} \tag{8}$$

the set of (normalized) allocations $\lambda \in \triangle_\Pi$ which are $\varepsilon$-optimal for the Graves-Lai program $\mathsf{glc}(\mathcal{M}, M)$ in Eq. (3)—both in terms of achieving the optimal objective value and satisfying the information constraint. In addition, for a distribution $\lambda \in \triangle_\Pi$, we define

$$\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda) = \{M \in \mathcal{M} \mid \lambda \in \Lambda(M; \varepsilon)\}. \tag{9}$$

Informally, $\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ represents the set of models for which the (normalized allocation) $\lambda \in \triangle_\Pi$ is $\varepsilon$-optimal for the Graves-Lai program $\mathsf{glc}(\mathcal{M}, M)$.

For a *reference model* $\overline{M} : \Pi \to \triangle_{\mathcal{R}\times\mathcal{O}}$ (not necessarily in $\mathcal{M}$) and parameter $\varepsilon > 0$, the Allocation-Estimation Coefficient is given by

$$\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) = \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\}, \tag{10}$$

where we adopt the convention that the value is 0 if $\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda) = \mathcal{M}$. In addition, letting $\mathrm{co}(\mathcal{M})$ denote the convex hull for $\mathcal{M}$, we define $\mathsf{aec}_\varepsilon(\mathcal{M}) := \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M})$.

The Allocation-Estimation Coefficient is a game between a min-player choosing $\lambda, \omega \in \triangle_\Pi$ and a max-player choosing a model $M \in \mathcal{M}$ (with the restriction that $M \notin \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$). The distribution $\lambda \in \triangle_\Pi$ represents a normalized Graves-Lai allocation, while $\omega \in \triangle_\Pi$ is an *exploration* distribution used to gather information. The reference model $\overline{M}$ should be interpreted as a guess for the underlying $M^\star \in \mathcal{M}$. When $\lambda \in \triangle_\Pi$ is fixed, $\inf_{\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} (\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right])^{-1}$ represents the time required to gather enough information to distinguish between the reference model $\overline{M}$ and all alternative models $M \notin \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ for which $\lambda$ is not an $\varepsilon$-optimal Graves-Lai allocation—provided that we explore optimally by minimizing over $\omega \in \triangle_\Pi$. For intuition, consider the case when $\overline{M} \in \mathcal{M}$. In this case, $\lambda$ *must* be chosen so that $\overline{M} \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ (i.e., $\lambda$ must be a Graves-Lai optimal allocation for $\overline{M}$), as otherwise the value of the AEC will be infinite, since $\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, \overline{M}(\pi)\right)\right] = 0$. Therefore, in such cases, the AEC reflects the *difficulty of distinguishing $\overline{M}$ from models that have different Graves-Lai optimal allocations*. Such models might have the *same optimal decision $\pi_{\overline{M}}$ as $\overline{M}$* (cf. Example 1.1) but, if our goal is to play a Graves-Lai optimal allocation for $\overline{M}$, we must still distinguish $\overline{M}$ from such models.

The Allocation-Estimation Coefficient plays a natural role for deriving both upper and lower bounds on the time required to learn an optimal allocation. For lower bounds, the significance of the AEC is somewhat immediate: it precisely quantifies the time required to acquire enough information to learn an $\varepsilon$-optimal allocation for the *best possible exploration strategy*, and thus leads to a lower bound on time required to learn such an allocation for any algorithm. Notably, the AEC serves as a lower bound for *all possible model classes* $\mathcal{M}$, and hence may be thought of as an intrinsic structural property of the class $\mathcal{M}$. For upper bounds, the Allocation-Estimation Coefficient acts as a mechanism to drive exploration; see Section 3 for further explanation.

**Generalized Allocation-Estimation Coefficient.** For certain results, we make use of the following, slightly more general variant of the AEC. For a reference model $\overline{M} : \Pi \to \triangle_{\mathcal{R}\times\mathcal{O}}$ and *subset of models* $\mathcal{M}_0 \subseteq \mathcal{M}$, we define

$$\mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0, \overline{M}) = \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}_0\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\}, \tag{11}$$

where we adopt the convention that the value is 0 if $\mathcal{M}_0 \backslash \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda) = \varnothing$. Here $\mathcal{M}_0$ denotes the set we take the supremum over, while $\mathcal{M}$ denotes the set that $\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ is defined with respect to (i.e., the set with respect to which the Graves-Lai allocation is defined). When $\mathcal{M}_0 = \mathcal{M}$, we recover the AEC as defined in Eq. (10): $\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) = \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}, \overline{M})$.

### 2.2. Main Results

Building on the intuition above, our main results show that boundedness of the Allocation-Estimation Coefficient is sufficient to achieve instance-optimality in finite time, and is also necessary in order to learn a near-optimal allocation. Formal statements of our upper bounds are given in Appendix A and formal statements for our lower bounds in Appendix B.

**Upper Bound.** Our upper bounds are based on a new algorithm, $\mathsf{AE}^2$ *(Allocation Estimation via Adaptive Exploration)*, achieves instance-optimality by using the Allocation-Estimation Coefficient to drive exploration.

**Theorem 2.1** (Upper Bound—Informal Version of Theorem A.1)**.** *For any model class $\mathcal{M}$ satisfying certain regularity conditions, the $\mathsf{AE}^2$ algorithm ensures that for all $\varepsilon > 0$, $M^\star \in \mathcal{M}$, and $T \in \mathbb{N}$:*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon) \cdot \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \widetilde{O}^+\big(\mathsf{aec}_{\varepsilon/12}(\mathcal{M}) + \mathsf{aec}_{\varepsilon/12}^{1/2}(\mathcal{M}) \cdot \log^{1/2}(T)\big),$$

*where $\widetilde{O}^+(\cdot)$ suppresses polynomial dependence on $\varepsilon^{-1}$, the log-covering number of $\mathcal{M}$, $\sup_{M \in \mathcal{M}} 1/\Delta_{\min}^M$, $\log\log T$, and several other measures of the regularity for the class $\mathcal{M}$.*

Theorem 2.1 shows that it is therefore possible to achieve instance-optimality in finite time with lower-order terms scaling (primarily) as the cost of learning the optimal allocation, as captured by the AEC. For multi-armed bandits with $\Pi = [A]$, we have $\mathsf{aec}_\varepsilon(\mathcal{M}) = \widetilde{O}^+(\mathrm{poly}(A))$, and for linear bandits with $\Pi \subseteq \mathbb{R}^d$, we have $\mathsf{aec}_\varepsilon(\mathcal{M}) = \widetilde{O}^+(\mathrm{poly}(d))$, so that the regret bound in Theorem 2.1 enjoys similar scaling as existing non-asymptotic approaches (Tirinzoni et al., 2020; Kirschner et al., 2021). For tabular reinforcement learning, we have $\mathsf{aec}_\varepsilon(\mathcal{M}) = \widetilde{O}^+(\mathrm{poly}(H, S, A))$, which leads to exponential improvement over prior work (Dong and Ma, 2022). Finally, for the instance in Example 1.1, in cases when $N \gg A, 1/\Delta_{\min}$, $\mathsf{aec}_\varepsilon(\mathcal{M}) = O(N)$, so $\mathsf{aec}_\varepsilon(\mathcal{M})$ captures the intuitive difficulty of learning a Graves-Lai allocation in this setting.

**Remark 2.1** (Technical Conditions)**.** *The technical conditions under which Theorem 2.1 is proven are relatively mild, and include certain smoothness of the KL divergences, sub-Gaussian tail behavior for log-likelihood ratios, and bounded covering number for $\mathcal{M}$ with standard parametric growth (note that $\mathcal{M}$ may be infinite), all of which can be shown to hold for standard classes. In addition, we requires that the amount of information that can be gained by playing the optimal decision for $M^\star$ is bounded (see Appendix A.1 for precise statements of our conditions).*

**Remark 2.2** (Asymptotic Performance)**.** *Asymptotically as $T \to \infty$, the regret bound given in Theorem 2.1 scales as $(1+\varepsilon) \cdot \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log T$, which is a factor of $(1+\varepsilon)$ off of the lower bound. For all standard classes, $\mathsf{aec}_{\varepsilon/12}(\mathcal{M})$ scales polynomially in $1/\varepsilon$ so, to obtain an optimal leading-order constant, it suffices to choose $\varepsilon = 1/\log^a T$, for small enough $a > 0$.*

**Adapting to Minimum Gap.** Note that the lower-order term given in Theorem 2.1 scales with $\sup_{M \in \mathcal{M}} 1/\Delta_{\min}^{M}$, the minimum gap of the entire model class. In Appendix A.5, we give a refinement of the $\mathsf{AE}^2$ algorithm ($\mathsf{AE}^2_\star$) which attains an improved regret bound which replaces the term $\sup_{M \in \mathcal{M}} 1/\Delta_{\min}^{M}$ with $1/\Delta_{\min}^{\star}$, the minimum gap of the underlying model; notably $\mathsf{AE}^2_\star$ requires no prior knowledge of $\Delta_{\min}^{\star}$ (i.e., it is able to adapt to the minimum gap of the underlying model). In addition, rather than scaling with $\mathsf{aec}_{\varepsilon/12}(\mathcal{M})$, the lower-order term now scales with $\mathsf{aec}^{\mathcal{M}}_{\varepsilon/12}(\mathcal{M}^\star)$ for a subset $\mathcal{M}^\star \subset \mathcal{M}$ which, informally, restricts to models in $\mathcal{M}$ for which the minimum gap is at least $\Delta_{\min}^{\star}$.

**Theorem 2.2** (Upper Bound—Informal Version of Theorem A.2)**.** *For any model class $\mathcal{M}$ satisfying certain regularity conditions, the $\mathsf{AE}^2_\star$ algorithm ensures that for all $\varepsilon > 0$, $M^\star \in \mathcal{M}$, and $T \in \mathbb{N}$:*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon) \cdot \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \widetilde{O}^{+}\big((\mathsf{aec}^{\mathcal{M}}_{\varepsilon/12}(\mathcal{M}^\star))^3 + \log^{6/7}(T)\big), \tag{12}$$

*where $\widetilde{O}^{+}(\cdot)$ suppresses polynomial dependence on $\varepsilon^{-1}$, the log-covering number of $\mathcal{M}$, $1/\Delta_{\min}^{\star}$, $\log\log T$, and several other measures of the regularity for the class $\mathcal{M}$.*

**Lower Bound.** To provide lower bounds, we adopt a novel minimax framework which asks, for the model class $\mathcal{M}$ under consideration, what is the least value of $T \in \mathbb{N}$ for which it is possible to learn an $\varepsilon$-optimal Graves-Lai allocation for any model in $\mathcal{M}$. To state our result, we introduce the following notation, defined with respect to any $\overline{M} \in \mathcal{M}^{+}$:

$$\mathcal{M}^{\mathrm{opt}}(\overline{M}) = \big\{M \in \mathcal{M} \mid \boldsymbol{\pi}_{M} \subseteq \boldsymbol{\pi}_{\overline{M}}, \;\; D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) = 0 \;\; \forall \pi \in \boldsymbol{\pi}_{\overline{M}}\big\}.$$

The set $\mathcal{M}^{\mathrm{opt}}(\overline{M})$ represents the set of models where 1) the optimal decision coincides with that of $\overline{M}$ and 2) $\overline{M}$ and $M \in \mathcal{M}$ cannot be distinguished by playing the optimal decision.

Our main lower bound provides a sort of converse to the upper bound in Theorem 2.1.

**Theorem 2.3** (Lower Bound—Informal Version of Theorem B.2)**.** *For any model class $\mathcal{M}$ and $\varepsilon > 0$, it holds that unless*

$$\log(T) \geq \sup_{\overline{M} \in \mathcal{M}^{+}} \widetilde{\Omega}^{+}\big(\mathsf{aec}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^{\mathrm{opt}}(\overline{M}), \overline{M})\big), \tag{13}$$

*no algorithm can simultaneously achieve the following for all instances $M \in \mathcal{M}$:*

1. *attain Graves-Lai optimality on $M$ within a constant factor (i.e., ensure $\mathbb{E}^{M}[\mathbf{Reg}(T)] \leq 2 \cdot \mathsf{glc}(\mathcal{M}, M)\log(T)$).*
2. *discover an $\varepsilon$-optimal allocation for $M$ (i.e., find $\lambda$ with $M \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)$) with probability greater than $\widetilde{\Omega}^{+}(1)$.*

*Here, $\widetilde{\Omega}^{+}(\cdot)$ hides polynomial dependence on regularity parameters of $\mathcal{M}$.*

Observe that the Graves-Lai Coefficient becomes the dominant term in the upper bound Theorem 2.1 as soon as $\log(T) \geq \widetilde{\Omega}^{+}(\mathsf{aec}_{\varepsilon/12}(\mathcal{M}))$. The lower bound (13) shows that for any algorithm that aims to estimate the Graves-Lai allocation (in particular, $\mathsf{AE}^2$), such scaling is necessary, and therefore the lower-order term in Theorem 2.1 is in some sense unimprovable. To the best of our knowledge, this is the first general approach to quantifying the lower-order terms necessary in order to achieve instance-optimality. We make several remarks on the lower bound.

**Remark 2.3** (Scaling in AEC)**.** *Our upper and lower bounds scale with a slightly different version of the AEC, as the lower bound restricts the AEC to* $\mathcal{M}^{\mathrm{opt}}(\overline{M})$*. In Appendix B, we show an additional lower bound that scales directly with* $\mathsf{aec}_\varepsilon(\mathcal{M})$*, matching our upper bound, but which only provides a lower bound on* $T$ *rather than* $\log(T)$ *(see Theorem B.1).*

**Remark 2.4** (Asymptotic Optimality and Learning Optimal Allocations)**.** *Theorem 2.3 gives a lower bound on the time needed to learn a near-optimal Graves-Lai allocation, but does not directly imply that it is* necessary *that an asymptotically optimal algorithm learn such an allocation. As we have noted, the allocations played by any asymptotically optimal algorithm must converge to an optimal allocation* in expectation*. However, showing that this convergence is necessary with even constant probability (the condition under which Theorem 2.3 is proved) is rather subtle. As we show in Appendix B (Theorem B.3), if one assumes that, in addition to being asymptotically optimal in expectation, the algorithm under consideration also has regret with appropriately bounded second moment, then if* $\mathbb{E}^M[\mathbf{Reg}(T)] \le (1+\varepsilon)\cdot \mathsf{glc}(\mathcal{M}, M)\log(T)$ *for all* $M \in \mathcal{M}$*, it is indeed necessary that a burn-in time analogous to Eq. (13) is satisfied.*

Together, our upper and lower bounds represent an initial step toward building a sharp non-asymptotic theory of instance-optimality, and lead to a number of new conceptual insights. Our results open the door for further-development, and to this end we highlight a number of opportunities for improvement (Appendix B.4), as well as open problems (Section 4).

### 2.3. Concrete Examples

We next present several examples illustrating our upper and lower bounds. All results in this section are informal—see Appendices A.3, A.6, A.7 and B.3 for formal results and additional examples.

**Example 2.1** (Searching for an Informative Arm (revisited))**.** We return to the example of Section 1.3. Some calculation shows that, for the choice of $\mathcal{M}$ in Example 1.1, as long as $\beta$ is constant and $N \ge A/\Delta^2$, we have

$$\Omega(N) \le \sup_{\overline{M}\in\mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}^{\mathrm{opt}}(\overline{M}), \overline{M}) \quad \text{and} \quad \mathsf{aec}_\varepsilon(\mathcal{M}) \le O(N).$$

Theorem 2.1 then implies that $\mathsf{AE}^2$ has regret on Example 1.1 of

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\cdot \mathsf{g}^\star \log(T) + N \cdot \mathrm{poly}(A, \tfrac{1}{\Delta}, \tfrac{1}{\varepsilon}, \log N, \log\log T)\cdot \log^{1/2}(T).$$

Furthermore, Theorem 2.3 shows that a scaling of $\log(T) \ge \widetilde{\Omega}^+(N)$ is necessary for any algorithm to learn a Graves-Lai optimal allocation. It follows that, on this example, the AEC reflects a notion of problem difficulty not captured by any existing complexity measure, matching our intuitive understanding of what the correct scaling should be. We remark that the scaling $\log(T) \ge \widetilde{\Omega}^+(N)$ is natural (as compared to $T \ge \widetilde{\Omega}^+(N)$) since, if an algorithm is instance-optimal as required by Theorem 2.3, it can allocate at most $O^+(\log(T))$ pulls to suboptimal decisions. To pull every informative arm (each of which is suboptimal) while achieving instance-optimality, it follows that we must have $\log(T) \ge \widetilde{\Omega}^+(N)$. ◁

**Example 2.2** (Tabular Reinforcement Learning)**.** Consider the setting of tabular reinforcement learning. Here we take $M$ to be a (tabular) episodic Markov Decision Processes (MDP) with $S$

states, $A$ actions, horizon $H$, probability transition kernels $\{P_h^M\}_{h=1}^H$, and Gaussian rewards; see Appendix A.7 for a full definition of this setting. Let $\mathcal{M}$ denote the set of all such tabular MDPs which, for each state-action-state triple $(s, a, s')$ and $h \in [H]$, have $P_h^M(s' \mid s, a) \geq P_{\min} > 0$; that is, each transition can occur with some minimum probability. Then it can be shown that:

$$\mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq \mathrm{poly}\Big(S, A, H, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \log \tfrac{1}{P_{\min}}\Big).$$

This implies that for any tabular MDP in $\mathcal{M}$, the $\mathsf{AE}^2_\star$ algorithm has regret bounded as:

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon) \cdot \mathsf{g}^\star \log(T) + \mathrm{poly}(S, A, H, \tfrac{1}{\Delta^\star_{\min}}, \tfrac{1}{\varepsilon}, \log \tfrac{1}{P_{\min}}, \log\log T) \cdot \log^{1/2}(T).$$

To the best of our knowledge, this is the first regret bound in the setting of tabular reinforcement learning which is instance-optimal with lower-order terms scaling only polynomially in problem parameters, an exponential improvement over past work (Ok et al., 2018; Dong and Ma, 2022). Furthermore, one can also show that $\sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}^{\mathrm{opt}}(\overline{M}), \overline{M}) \geq \widetilde{\Omega}\Big(\frac{1}{\varepsilon^2} \cdot \frac{SA}{(\Delta^\star_{\min})^2}\Big)$, so that our lower bound, Theorem 2.3, implies that a burn-in time scaling polynomially in $S$, $A$, $\frac{1}{\varepsilon}$, and $\frac{1}{\Delta^\star_{\min}}$ is necessary to learn an $\varepsilon$-optimal Graves-Lai allocation for every model in $\mathcal{M}$.

We remark that the prior work of Dong and Ma (2022) does not require that $P_h^M(s' \mid s, a) \geq P_{\min}$ as we do, yet their bound scales *polynomially* in the inverse probability of observing the trajectory that occurs with minimum non-zero probability (the work of Ok et al. (2018) only holds for ergodic MDPs, itself a very strong assumption). Our finite-time results are therefore, in general, significantly stronger, scaling only logarithmically in $P_{\min}$. Removing the $P_{\min}$ assumption while still achieving reasonable lower-order terms is an interesting direction for future work. ◁

**Example 2.3** (Linear Bandits)**.** Consider the setting of linear bandits in $d$ dimensions with unit-variance Gaussian noise. Let $\mathcal{M}$ denote the set of all linear bandit models defined with respect to some arm set $\mathcal{X} \subseteq \mathbb{R}^d$ and parameter set $\Theta \subseteq \mathbb{R}^d$. Concretely, each model $M \in \mathcal{M}$ takes the form

$$M(\pi) = \mathcal{N}(\langle \theta, x_\pi \rangle, 1),$$

for some $\theta \in \Theta$, where $x_\pi \in \mathcal{X}$ is an embedding of $\pi$. Let $\Delta^\star_{\min}$ denote the minimum gap of $M^\star$ (which is unknown to the algorithm). Then it can be shown that

$$\mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq \mathrm{poly}\Big(d, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}\Big)$$

which implies that the $\mathsf{AE}^2_\star$ algorithm has regret bounded as

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon) \cdot \mathsf{g}^\star \log(T) + \mathrm{poly}(d, \tfrac{1}{\Delta^\star_{\min}}, \tfrac{1}{\varepsilon}, \log\log T) \cdot \log^{6/7}(T). \tag{14}$$

We remark that the scaling of Eq. (14) matches the state-of-the-art instance-optimal bounds for linear bandits (in that all have polynomial lower-order terms—our polynomial dependence on $d$ is slightly worse as our upper bound on the AEC is somewhat coarse) (Tirinzoni et al., 2020; Kirschner et al., 2021). Notably, it is a simple corollary of a much more general result, while prior work relies on specialized algorithms tailored to linear bandits. ◁

In Appendices A.3, A.6, A.7 and B.3 we formalize these examples and present additional examples, including structured bandits with bounded eluder dimension and finite-action contextual bandits. In all cases, we obtain lower-order terms scaling only polynomially with problem parameters, and in each setting either match the best-known existing bound, or are the first to provide any meaningful finite-time bounds.

**Algorithm 1** Allocation Estimation via Adaptive Exploration ($\mathsf{AE}^2$, Informal)

1: **input:** optimality tolerance $\varepsilon$, model class $\mathcal{M}$.
2: Initialize $s \leftarrow 1$ and $q \leftarrow$ class-dependent quantity.
3: Compute $\xi^1 \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
4: **for** $t = 1, 2, 3, \ldots$ **do**
5: **if** $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then**
6: Play $\pi_{\widehat{M}^s}$. // Exploit
7: **else** // Explore
8: Set $p^s \leftarrow q\lambda^s + (1-q)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \operatorname*{arg\,min}_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda)} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\widehat{M}(\pi)\,\|\,M(\pi)\right)\right]\right]}. \tag{15}$$

9: Draw $\pi^s \sim p^s$ and observe reward $r^s$ and observation $o^s$.
10: Compute $\widehat{M}^{s+1} = \mathbb{E}_{\widehat{M}\sim\xi^{s+1}}[\widehat{M}]$ for $\xi^{s+1} \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{(\pi^i, r^i, o^i)\}_{i=1}^s)$, $s \leftarrow s+1$.

## 3. Algorithm Overview

Finally, we present our algorithm, $\mathsf{AE}^2$, in Algorithm 1. $\mathsf{AE}^2$ relies on an *online estimation oracle*, denoted by $\mathbf{Alg}_{\mathsf{KL}}$, which at every step $s$, given access to data $\left\{(\pi^i, r^i, o^i)\right\}_{i=1}^{s-1}$ with $\pi^i \sim p^i$ and $(r^i, o^i) \sim M^\star(\pi^i)$ returns a randomized estimate $\xi^i = \mathbf{Alg}_{\mathsf{KL}}\left(\left\{(\pi^i, r^i, o^i)\right\}_{i=1}^{s-1}\right) \in \triangle_{\mathcal{M}}$ with the goal of approximating $M^\star$ (Foster and Rakhlin, 2020; Foster et al., 2021). The estimates produced by $\mathbf{Alg}_{\mathsf{KL}}$ must ensure the total KL estimation error is bounded:

$$\mathbf{Est}_{\mathsf{KL}}(s) := \textstyle\sum_{i=1}^s \mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}\left(\widehat{M}(\pi)\,\|\,M^\star(\pi)\right)]] \lesssim O(\log s). \tag{16}$$

We show that, under the regularity conditions required by Theorem 2.1, such a guarantee can be achieved, with $O(\cdot)$ hiding the log-covering number of $\mathcal{M}$.

$\mathsf{AE}^2$ alternates between *exploit* steps and *explore* steps, tracking the number of explore steps that have been performed with a counter $s \in \mathbb{N}$. For each step $t \in \mathbb{N}$, the algorithm makes use of an estimator $\widehat{M}^s = \mathbb{E}_{\widehat{M}\sim\xi^s}[\widehat{M}]$, where $\xi^s = \mathbf{Alg}_{\mathsf{KL}}\left(\left\{(\pi^i, r^i, o^i)\right\}_{i=1}^{s-1}\right)$ is computed by calling the estimation oracle with data gathered at previous explore steps. Given the estimator, the algorithm first checks whether it has enough information to guarantee that the greedy decision is optimal, in which case it exploits (Line 6); otherwise it explores. In explore steps, the key component is the choice of the exploration distributions $\lambda^s$ and $\omega^s$ in Eq. (15), which mimics the Allocation-Estimation Coefficient program. To understand the role of the Allocation-Estimation Coefficient here, we consider two cases. In the first case, if $\lambda^s$ is an $\varepsilon$-optimal Graves-Lai allocation for $M^\star$ (that is, $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s)$), then playing $\lambda^s$ will optimize the tradeoff between minimizing regret and collecting information, and will therefore match the optimal performance prescribed by the Graves-Lai Coefficient.

In the second case, if $\lambda^s$ is not an $\varepsilon$-optimal Graves-Lai allocation for $M^\star$, we have $M^\star \notin \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s)$, so $\omega^s$ will place mass on actions that ensure $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega}[D_{\mathsf{KL}}\left(\widehat{M}(\pi)\,\|\,M^\star(\pi)\right)]]$ is large. Since $p^s$ plays $\omega^s$ with constant probability, the quantity $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\left(\widehat{M}(\pi)\,\|\,M^\star(\pi)\right)]]$ will also be large. If our estimator is consistent and Eq. (16) holds, this can only happen a small number of times without violating Eq. (16); at most logarithmic in the number of exploration rounds. As such,

we can show that $\lambda^s$ must be a near-optimal Graves-Lai allocation for $M^\star$ on all but a logarithmic number of exploration rounds, and that $\mathsf{AE}^2$ achieves the optimal rate on such rounds, yielding the optimal performance rate of Theorem 2.1. Critically, rather than exploring in a naive fashion (e.g., by sampling decisions uniformly), $\mathsf{AE}^2$ explores specifically with the goal of learning a Graves-Lai optimal allocation for $M^\star$ and adapts to the structure of $\mathcal{M}$ to perform this exploration efficiently.

We emphasize the simplicity of $\mathsf{AE}^2$. While most existing instance-optimal algorithms are quite complicated even for basic settings, $\mathsf{AE}^2$ relies on a few simple components yet is far more general than existing approaches and performs comparably or better. See Appendix A.2 for a full description.

## 4. Discussion

Our work initiates the systematic study of non-asymptotic instance-optimality in interactive decision making. We close by highlighting a number of interesting open problems and future directions raised by our work. On the technical side:

- Our upper bounds depend on a number of different problem-dependent parameters, such as the minimum gap in the lower-order terms. Can we improve the dependence on these parameters, or understand to what extend they are necessary?
- Our lower bounds concern the problem of learning a near-optimal allocation. Like the upper bounds, these results are likely loose in terms of dependence on various problem parameters, and new techniques will be required to tighten them. Furthermore, it remains to develop a complete understanding of the connections between this problem and the problem of minimizing regret. While our results show that "well-behaved" algorithms which achieve instance-optimal regret must pay a burn-in proportional to the cost of learning a near-optimal allocation, it remains unclear if this is truly necessary for algorithms which only have optimal expected regret (but, for example, could exhibit heavy-tailed behavior).
- While our algorithm achieves the instance-optimal rate, its regret could scale linearly over shorter time horizons, until it has learned a near-optimal allocation. Can we develop "best-of-both-worlds" algorithms that achieve the same instance-optimal guarantees of $\mathsf{AE}^2$, yet also achieves the minimax-optimal rate (for example, a $\mathcal{O}(\sqrt{T})$-style guarantee) over shorter time horizons?
- The focus of this work is primarily on regret minimization, yet the challenge of learning the optimal allocation also arises in the PAC setting. Does the AEC extend to the PAC setting, and can algorithms be developed in the PAC setting which achieve the instance-optimal rate in the leading-order term, while scaling with an AEC-like quantity in the lower-order term?

More broadly, it will be interesting to explore whether our framework and algorithm design ideas can be used to develop practical and computationally efficient algorithms.

### Acknowledgements

The authors would like to thank Johannes Kirschner for helpful discussions. The work of AW was supported in part by NSF TRIPODS 62-2945 and NSF HDR 62-0221. A portion of this work was completed while AW was an intern at Microsoft Research, and while visiting the Simons Institute for the Theory of Computing.

## References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled markov chains: Finite parameter space. Technical report, MICHIGAN UNIV ANN ARBOR COMMUNICATIONS AND SIGNAL PROCESSING LAB, 1988.

Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes. In *International Conference on Machine Learning*, pages 7459–7468. PMLR, 2021.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34:25852–25864, 2021.

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. *Hannan Consistency in On-Line Learning in Case of Unbounded Losses Under Partial Monitoring*, pages 229–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-46650-5. doi: 10.1007/11894841_20. URL http://dx.doi.org/10.1007/11894841_20.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pages 111–127. PMLR, 2018.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

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

Lijie Chen, Jian Li, and Mingda Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110. PMLR, 2017.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1761–1769, 2017.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, 2008.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020a.

Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, pages 2443–2452. PMLR, 2020b.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

Kefan Dong and Tengyu Ma. Asymptotic instance-optimal algorithms for interactive decision making. *arXiv preprint arXiv:2206.02326*, 2022.

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

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Dylan J Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning (ICML)*, 2020.

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

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *Conference on Learning Theory (COLT)*, 2020.

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

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

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

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

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pages 784–792, 2016.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. *arXiv preprint arXiv:1910.06996*, 2019.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

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

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

Kwang-Sung Jun and Chicheng Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. *Advances in Neural Information Processing Systems*, 33:6366–6376, 2020.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Julian Katz-Samuels, Lalit Jain, Kevin G Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33:10371–10382, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. *Advances in Neural Information Processing Systems*, 28, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR, 2014.

Teodor V Marinov, Mehryar Mohri, and Julian Zimmert. Stochastic online learning with feedback graphs: Finite-time and asymptotic optimality. *arXiv preprint arXiv:2206.10022*, 2022a.

Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Open problem: Finite-time instance dependent optimality for stochastic online learning with feedback graphs. In *Conference on Learning Theory*, pages 5644–5649. PMLR, 2022b.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

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

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

Ohad Shamir. A variant of azuma's inequality for martingales with subgaussian tails. *arXiv preprint arXiv:1110.2392*, 2011.

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

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pages 1794–1834. PMLR, 2017.

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

Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33:1417–1427, 2020.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*, 2021.

Andrea Tirinzoni, Aymen Al-Marjani, and Emilie Kaufmann. Near instance-optimal pac reinforcement learning for deterministic mdps. *arXiv preprint arXiv:2203.09251*, 2022.

Bart PG Van Parys and Negin Golrezaei. Optimal learning for structured bandits. *arXiv preprint arXiv:2007.07302*, 2020.

Andrew Wagenmaker and Kevin Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *arXiv preprint arXiv:2207.02575*, 2022.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. *arXiv preprint arXiv:2211.04974*, 2022.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *International Conference on Machine Learning*, pages 10641–10652. PMLR, 2021.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022a.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pages 358–418. PMLR, 2022b.

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

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

Yuhong Yang and Andrew R Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 1998.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

# Contents

# Part I
# Main Results

Part I is organized as follows. Appendix A presents our algorithm and main upper bounds, as well as examples. Appendix B presents complementary lower bounds. In Appendix C we review additional related work. We conclude with discussion of open problems and future directions in Section 4. Proofs are deferred to the appendix.

**Additional notation.** For an integer $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, \dots, n\}$. For a set $\mathcal{Z}$, we let $\triangle_{\mathcal{Z}}$ denote the set of all probability distributions over $\mathcal{Z}$. We adopt standard big-oh notation, and write $f = \widetilde{O}(g)$ to denote that $f = O(g \cdot \max\{1, \mathrm{polylog}(g)\})$. We use $\lesssim$ only in informal statements to emphasize the most notable elements of an inequality. We will let $\mathrm{lin}(\cdot)$ denote a function multi-linear and poly-logarithmic in its arguments. For a decision $\pi \in \Pi$, we use $\mathbb{I}_\pi \in \triangle_\Pi$ to denote the delta distribution which places probability mass 1 on $\pi$.

Define the Kullback-Leibler divergence by

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

## Appendix A. The AE$^2$ Algorithm: Regret Bounds and Examples

This section presents our main algorithm and regret bounds. We begin by introducing the most basic variant of our algorithm, AE$^2$, and using it to provide instance-optimal regret bounds for simple settings (Appendices A.2 and A.3); with preliminaries in Appendix A.1. We then give a refined variant of the algorithm, AE$^2_\star$, which adapts to the minimum gap $\Delta^\star_{\min}$ and leads to regret bounds under relaxed regularity conditions (Appendix A.4 and Appendix A.5). We use this variant to provide applications to structured and contextual bandits (Appendix A.6) and tabular reinforcement learning (Appendix A.7). We conclude with an overview of our analysis in Appendix A.8. For all results in this section, we assume that Assumptions 1.1 to 1.3 hold.

### A.1. Regularity Conditions

To present our algorithm and results, we first introduce several regularity conditions for the model class $\mathcal{M}$.

**Likelihood ratios.** We next make two assumptions concerning smoothness of KL divergences and behavior of log-likelihood ratios.

**Assumption A.1** (Smooth KL)**.** *There exists $L_{\mathsf{KL}} > 0$ such that for all $M, M', M'' \in \mathcal{M}$ and $\pi \in \Pi$,*

$$\left|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) - D_{\mathsf{KL}}\big(M'(\pi) \,\|\, M''(\pi)\big)\right| \le L_{\mathsf{KL}} \sqrt{D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.$$

**Assumption A.2** (Sub-Gaussian Log-Likelihood)**.** *There exists $V_{\mathcal{M}} > 0$ such that for all $M, M', M'' \in \mathcal{M}$ and $\pi \in \Pi$,*

$$\mathbb{P}_{(r,o)\sim M(\pi)}\left[\left|\log \frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M'',\pi}(r,o)} - \mathbb{E}_{(r',o')\sim M(\pi)}\left[\log \frac{\mathbb{P}^{M',\pi}(r',o')}{\mathbb{P}^{M'',\pi}(r',o')}\right]\right| \ge x\right] \le 2\exp(-x^2/V^2_{\mathcal{M}})$$

*for all $x \ge 0$.*

Assumption A.1 and Assumption A.2 facilitate finite-sample estimation guarantees with respect to the KL divergence. Both assumptions are met by standard problem classes, including general structured bandit problems with Gaussian noise. Existing works that consider general model classes make similar assumptions (Dong and Ma, 2022).

**Estimation.** To provide estimation guarantees that accommodate infinite classes $\mathcal{M}$, we assume certain covering properties. We will consider the following notion of a cover.

**Definition A.1** ($(\rho, \mu)$-Cover)**.** *We say that a set $\mathcal{M}_{\mathrm{cov}} \subseteq \mathcal{M}$ is a $(\rho, \mu)$-cover of $\mathcal{M}$ if there exists some event $\mathcal{E}$ such that:*[4]

1. $\sup_{M \in \mathcal{M}} \sup_{\pi \in \Pi} \mathbb{P}^{M,\pi}(\mathcal{E}^c) \leq \mu$.
2. *For each $M \in \mathcal{M}$, there exists some $M' \in \mathcal{M}_{\mathrm{cov}}$ such that*

$$\left|\log \mathbb{P}^{M,\pi}(r, o) - \log \mathbb{P}^{M',\pi}(r, o)\right| \leq \rho$$

*for all $(r, o) \in \mathcal{R} \times \mathcal{O}$ with $\sup_{M'' \in \mathcal{M}} \mathbb{P}^{M'',\pi}(r, o \mid \mathcal{E}) > 0$.*

*We denote the size of the smallest such cover by $\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho, \mu)$.*

Definition A.1 states that the log-likelihoods are "covered" under some good event $\mathcal{E}$ which occurs with high probability: for any model in the class, we can find some model in the cover with log-likelihoods that are "close" on $\mathcal{E}$. We assume that the covering number for the model class $\mathcal{M}$ is bounded, and has reasonable ("parametric") growth.

**Assumption A.3** (Bounded Covering Number)**.** *For some parameters $d_{\mathsf{cov}} \geq 1, C_{\mathsf{cov}} \geq 1$, we have*

$$\log \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho, \mu) \leq d_{\mathsf{cov}} \cdot \log\left(\frac{C_{\mathsf{cov}}}{\rho\mu}\right).$$

Note that the rate of growth of the covering number required by Assumption A.3 is the standard rate of growth for parametric (e.g., linear) classes. Our results easily extend to accommodate general growth rates, but we adopt Assumption A.3 because it suffices for all of the examples we will consider, and simplifies presentation.

**Information content of optimal decisions.** As noted in the introduction, the Graves-Lai Coefficient $\mathsf{g}^\star = \mathsf{glc}(\mathcal{M}, M^\star)$ can be thought of as the minimal regret needed to distinguish $M^\star$ from all possible models with different optimal decisions. As playing the optimal decision, $\pi_\star$, incurs no regret, any allocation $\eta$ which is optimal for the Graves-Lai program, Eq. (3), will still be optimal if we increase the number of plays of $\pi_\star$ arbitrarily. As we are interested in finite-time behavior in this work, it is undesirable to consider allocations for which the number of pulls of optimal decisions are arbitrarily large. Instead, we would like to consider allocations which play optimal decisions only as long as they still provides useful information about models in the alternate set. The following definition gives a formal quantification of this.

4. Note that we require that $\mathcal{M}_{\mathrm{cov}} \subseteq \mathcal{M}$, i.e. that $\mathcal{M}_{\mathrm{cov}}$ is a *proper* cover.

**Definition A.2** (Information Content of Optimal Decision). *Fix $\epsilon \in (0, 1/2]$. For a model $M \in \mathcal{M}$, we define $\mathsf{n}_\varepsilon^M > 0$ as the minimum value such that, for any allocation $\eta \in \mathbb{R}_+^\Pi$ satisfying*

$$(1+\varepsilon)\mathsf{g}^M \geq \sum_{\pi\in\Pi} \eta(\pi)\Delta^M(\pi) \quad \textit{and} \quad \inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi\in\Pi} \eta(\pi) D_{\mathsf{KL}}\big(M(\pi)\,\|\,M'(\pi)\big) \geq 1-\varepsilon,$$

*we have*

$$\inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi\in\Pi, \pi\notin\boldsymbol{\pi}_M} \eta(\pi) D_{\mathsf{KL}}\big(M(\pi)\,\|\,M'(\pi)\big) + \sum_{\pi\in\boldsymbol{\pi}_M} \mathsf{n}_\varepsilon^M D_{\mathsf{KL}}\big(M(\pi)\,\|\,M'(\pi)\big) \geq 1-2\varepsilon.$$

*We denote $\mathsf{n}_\varepsilon^{\mathcal{M}} := \sup_{M\in\mathcal{M}} \mathsf{n}_\varepsilon^M$.*

Intuitively, any allocation which is $\varepsilon$-optimal for the Graves-Lai program Eq. (3) need not play any optimal decision $\pi_M \in \boldsymbol{\pi}_M$ more than $\mathsf{n}_\varepsilon^M$ times. Therefore, for model $M$, $\mathsf{n}_\varepsilon^M$ can be thought of as a quantification of the extent to which playing optimal decisions provides useful information—no additional useful information can be acquired on models in the alternate set $\mathcal{M}^{\mathrm{alt}}(M)$ by playing optimal decisions more than $\mathsf{n}_\varepsilon^M$ times. As we will see, $\mathsf{n}_\varepsilon^M$ is bounded polynomially in problem parameters for many classes of interest.

**Uniformly regular classes.** We refer to a class as *uniformly regular* if $\mathsf{n}_\varepsilon^{\mathcal{M}} < \infty$, and the following assumption on the minimum gaps holds.

**Assumption A.4** (Lower-Bounded Minimum Gap). *We have $\inf_{M\in\mathcal{M}} \Delta_{\min}^M > 0$. We denote by $\Delta_{\min} > 0$ a (known) lower bound on $\inf_{M\in\mathcal{M}} \Delta_{\min}^M$.*

Note that Assumption A.4 implies that for all $M \in \mathcal{M}$, $\pi_M$ is unique. For the results concerning the most basic version of our algorithm, $\mathsf{AE}^2$ (Appendices A.2 and A.3), we assume for expositional purposes that the class $\mathcal{M}$ is uniformly regular. Our more general algorithm, $\mathsf{AE}^2_\star$ (Appendix A.5), achieves guarantees similar to those of $\mathsf{AE}^2$, but without uniform regularity. In particular, $\mathsf{AE}^2_\star$ replaces dependence on $\Delta_{\min}$ with the minimum gap $\Delta_{\min}^\star := \Delta_{\min}^{M^\star}$ for the true model, and replaces dependence on $\mathsf{n}_\varepsilon^{\mathcal{M}}$ with $\mathsf{n}_\varepsilon^\star := \mathsf{n}_\varepsilon^{M^\star}$. We note, however, that in cases where a lower bound $\Delta$ on the minimum gap $\Delta_{\min}^\star$ of the true model is known a-priori, Assumption A.4 can be satisfied by restricting the model class to models with minimum gap at least $\Delta$.

### A.2. The $\mathsf{AE}^2$ Algorithm

We now present the most basic variant of our main algorithm, $\mathsf{AE}^2$ (Algorithm 2). This will serve as the starting point for the most general version of our algorithm, $\mathsf{AE}^2_\star$ (Appendix A.5). To describe the algorithm, we first introduce the primitive of an *online estimation oracle* (Foster and Rakhlin, 2020; Foster et al., 2021).

**Estimation oracles.** Algorithm 2 makes use of an *online estimation oracle*, denoted by $\mathbf{Alg}_{\mathsf{KL}}$, which is an algorithm that, given knowledge of the class $\mathcal{M}$, estimates the underlying model $M^\star \in \mathcal{M}$ from data in a sequential fashion. When invoked at step $s \in \mathbb{N}$ with the data $(\pi^1, r^1, o^1), \ldots, (\pi^{s-1}, r^{s-1}, o^{s-1})$ observed so far, the estimation oracle builds an estimate

$$\widehat{M}^s = \mathbf{Alg}_{\mathsf{KL}}\Big(\big\{(\pi^i, r^i, o^i)\big\}_{i=1}^{s-1}\Big)$$

**Algorithm 2** Allocation Estimation via Adaptive Exploration ($\mathsf{AE}^2$)

**input:** optimality tolerance $\varepsilon$, model class $\mathcal{M}$.
Initialize $s \leftarrow 1$, $\mathsf{n}_{\max} \leftarrow \mathsf{n}_{\max}(\mathcal{M}, \varepsilon/6)$, and $q \leftarrow \frac{4\mathsf{n}_{\max}+\varepsilon \underline{\mathsf{g}}^{\mathcal{M}}}{4\mathsf{n}_{\max}+2\varepsilon \underline{\mathsf{g}}^{\mathcal{M}}}$ for $\underline{\mathsf{g}}^{\mathcal{M}} := \inf_{M\in\mathcal{M}:\mathsf{g}^M>0} \mathsf{g}^M$.
Compute $\xi^1 \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
**for** $t = 1, 2, 3, \ldots$ **do**
  **if** $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then**
    Play $\pi_{\widehat{M}^s}$. // Exploit
  **else** // Explore
    Set $p^s \leftarrow q\lambda^s + (1-q)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \arg\min_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda;\mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\widehat{M}(\pi)\,\|\,M(\pi)\right)\right]\right]}. \tag{18}$$

    Draw $\pi^s \sim p^s$ and observe reward $r^s$ and observation $o^s$.
    Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{(\pi^i, r^i, o^i)\}_{i=1}^s)$ and $\widehat{M}^{s+1} = \mathbb{E}_{\widehat{M}\sim\xi^{s+1}}[\widehat{M}]$ .
    $s \leftarrow s+1$.

which aims to approximate the true model $M^\star$. Following (Foster et al., 2021; Chen et al., 2022; Foster et al., 2022a), we make use of *randomized* estimation oracles that, at each step produces $\xi^s = \mathbf{Alg}_{\mathsf{KL}}\left(\left\{(\pi^i, r^i, o^i)\right\}_{i=1}^{s-1}\right)$, where $\xi^s \in \triangle_{\mathcal{M}}$ is a randomization distribution, and draw $\widehat{M} \sim \xi^s$. We measure the oracle's performance in terms of cumulative estimation error, defined as follows.

**Definition A.3** (Cumulative Estimation Error)**.** *Consider the process where, for each round $i \in \mathbb{N}$, given $(\pi^1, r^1, o^1), \ldots, (\pi^{i-1}, r^{i-1}, o^{i-1})$ with $\pi^i \sim p^i$ and $(r^i, o^i) \sim M^\star(\pi^i)$, the estimation oracle returns $\xi^i = \mathbf{Alg}_{\mathsf{KL}}\left((\pi^1, r^1, o^1), \ldots, (\pi^{i-1}, r^{i-1}, o^{i-1})\right)$. For any $s \in \mathbb{N}$, we define the oracle's cumulative KL estimation error under this process as:*

$$\mathbf{Est}_{\mathsf{KL}}(s) := \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\mathbb{E}_{\pi\sim p^i}\left[D_{\mathsf{KL}}\left(M^\star(\pi)\,\|\,\widehat{M}(\pi)\right)\right]\right].$$

Algorithm 2 can be invoked with any off-the-shelf algorithm for estimation, but our main results make use of the fact that under Assumption A.2 and Assumption A.3, there exists an estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$ (Algorithm 7 in Appendix F.3) which ensures that with probability at least $1-\delta$, for all $s \in \mathbb{N}$:

$$\mathbf{Est}_{\mathsf{KL}}(s) \lesssim V_{\mathcal{M}} \cdot d_{\mathsf{cov}} \cdot \log^{3/2}\left(\frac{C_{\mathsf{cov}} \cdot s}{\delta}\right). \tag{17}$$

That is, the estimation oracle ensures that the KL divergence between the true model $M^\star$ and the estimates returned scales at most poly-logarithmically in the exploration horizon. Note that on its own, this guarantee does not necessarily imply that $\widehat{M}^s = \mathbb{E}_{M\sim\xi^s}[M] \to M^\star$—low online estimation error only requires that $\widehat{M}^s$ is on average close to $M^\star$ on the decisions we have actually played.

**Algorithm overview.** We now present a formal overview of our algorithm $\mathsf{AE}^2$. We restate it here for convenience in Algorithm 2. The algorithm alternates between *exploit* steps and *explore* steps, tracking the number of explore steps that have been performed with a counter $s \in \mathbb{N}$. For each step $t \in \mathbb{N}$, the algorithm makes use of an estimator $\widehat{M}^s = \mathbb{E}_{\widehat{M}\sim\xi^s}[\widehat{M}]$, where $\xi^s = \mathbf{Alg}_{\mathsf{KL}}\big(\{(\pi^i, r^i, o^i)\}_{i=1}^{s-1}\big)$ is computed by calling the estimation oracle with data gathered at previous explore steps. Given the estimator, $\mathsf{AE}^2$ performs a test based on likelihood ratios (Line 5) to check whether it has collected enough information to rule out all models for which $\pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ is not an optimal decision. If so, it *exploits*, and plays $\pi_{\widehat{M}^s}$ (Line 6), as in this case $\pi_{\widehat{M}^s} = \pi_\star$ with high probability. If the test fails, the algorithm must gather more information to eliminate alternatives, and it *explores* (Line 8). The key component of the explore phase is the choice of the exploration distribution in Eq. (18), which is based on the Allocation-Estimation Coefficient program, but incorporates some small modifications: 1) First, $\widehat{M}$ is randomized according to the distribution $\xi^s$, 2) Second, the set $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)$ is replaced with a smaller set $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda; \mathsf{n}_{\max})$, which requires that $\lambda$ obeys certain normalization constraints; this is detailed below. Using the distributions $\lambda^s$ (representing a normalized allocation) and $\omega^s$ (representing an exploration distribution) returned in Eq. (18), the algorithm computes a mixture $p^s = q\lambda^s + (1-q)\omega^s$, where $q \in (0,1)$ is a carefully chosen parameter, and plays $\pi^s \sim p^s$ (Line 9). The reward and observation $(r^s, o^s)$ that result from playing $\pi^s$ are then used to update the estimation oracle for subsequent rounds (Line 10).

To understand the intuition behind the explore phase and why the Allocation-Estimation Coefficient plays a useful role here, we can consider two cases. In the first case, if $\lambda^s$ is an $\varepsilon$-optimal Graves-Lai allocation for $M^\star$ (that is, $M^\star \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$), then playing $\lambda^s$ will optimize the tradeoff between minimizing regret on $M^\star$ and collecting information that allows one to distinguish $M^\star$ from $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, and will therefore match the optimal performance prescribed by the Graves-Lai Coefficient, incurring regret scaling as $\mathsf{g}^\star$.

In the second case, if $\lambda^s$ is not an $\varepsilon$-optimal Graves-Lai allocation for $M^\star$, we have $M^\star \notin \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$, so by the definition of the AEC (Eq. (18)), $\omega^s$ will place mass on actions that ensure $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega}[D_{\mathsf{KL}}\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]$ is large; exactly how large this quantity is will be quantified by the value of the AEC. Since $p^s$ plays $\omega^s$ with constant probability, the quantity

$$\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]]$$

will also be large, but if the estimation oracle is consistent in the sense of Definition A.3, this can only happen a small number of times. In particular, if Eq. (17) holds, the number of times in which we encounter this second case is at most *logarithmic* in the number of exploration rounds. As such, we can show that $\lambda^s$ must be a near-optimal Graves-Lai allocation for $M^\star$ on all but a logarithmic number of exploration rounds, and that $\mathsf{AE}^2$ achieves the optimal rate on such rounds.

Critically, rather than exploring in a naive fashion (e.g., by sampling decisions uniformly), $\mathsf{AE}^2$ explores only to the extent necessary to learn a Graves-Lai allocation for $M^\star$. There may exist instances $M \neq M^\star$ which differ significantly from $M^\star$ but have a similar Graves-Lai allocations—$\mathsf{AE}^2$ will make no effort to distinguish such instances since, as long as it knows that one of these instances is correct, it can simply play their shared Graves-Lai allocation. This notion of exploration, which is targeted toward distinguishing instances that have different Graves-Lai allocations, is precisely the notion captured by the Allocation-Estimation Coefficient.

**Normalization factor for allocations.** As noted in Appendix A.1, while an optimal Graves-Lai allocation may place an arbitrarily large number of pulls on an optimal decision, for finite-time

guarantees it is useful to restrict to allocations which place only finite mass on optimal decisions. To this end, $\mathsf{AE}^2$ restricts the optimization problem based on the AEC in Eq. (18) to only consider normalized allocations $\lambda$ for which the *normalization factor* is at most

$$\mathsf{n}_{\max}(\mathcal{M}, \varepsilon) := \frac{64}{\Delta_{\min}^2} \cdot \left( \frac{1}{\varepsilon} + V_{\mathcal{M}} \mathsf{n}_{\varepsilon}^{\mathcal{M}} \right) \cdot \max_{M \in \mathcal{M}} \mathsf{g}^M. \tag{19}$$

where the normalization factor refers to the value $\mathsf{n}$ in the definition of $\Lambda(M; \varepsilon)$ (see Eq. (8)). In particular, to enforce this restriction, the optimization problem in Eq. (18) restricts the max-player to $M \in \mathcal{M} \backslash \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda; \mathsf{n}_{\max})$, where $\mathsf{n}_{\max} := \mathsf{n}_{\max}(\mathcal{M}, \varepsilon/6)$ and $\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda; \mathsf{n}_{\max})$ is defined identically to $\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda)$ in Eq. (9), but with $\mathsf{n}$ restricted to $\mathsf{n} \le \mathsf{n}_{\max}$. As we show in Lemma E.4, for $\mathsf{n}_{\max}(\mathcal{M}, \varepsilon)$ defined as in Eq. (19), the optimal value in Eq. (18) can be bounded by the AEC.

**Computational efficiency.** The primary computational burden in $\mathsf{AE}^2$ lies in solving the optimization problem (18) to compute the exploration distributions. In general there is little hope of solving this efficiently (i.e., in time sublinear in $|\Pi|$ and $|\mathcal{M}|$)—indeed, in some cases it may be that to even determine whether $M \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)$ will require enumerating the model class $\mathcal{M}$. However, for nicely structured problems, we anticipate that this program can be solved, or at least approximated, efficiently. As the focus of our work is primarily statistical, we leave further exploration as to when the algorithm can be implemented efficiently to future work.

**Simplicity.** We emphasize the simplicity of $\mathsf{AE}^2$. Most existing algorithms which achieve instance-optimality are quite complex, even in specialized settings such as linear bandits. In contrast, $\mathsf{AE}^2$ is very simple and intuitive, and relies only on three basic components: an explore-exploit test, an estimation oracle, and a single optimization to compute the exploration distributions. Despite its simplicity, as we show, $\mathsf{AE}^2$ obtains comparable or better performance over existing approaches.

**Relation to existing approaches.** At a very high level, $\mathsf{AE}^2$ bears some similarity to the E2D algorithm of Foster et al. (2021), which achieves the *minimax* optimal rate for general classes $\mathcal{M}$ in the DMSO framework. Both algorithms rely on online estimation algorithms, and both solve min-max programs based on the output of the estimator to determine which allocations to play. However, the algorithm design and analysis principles for the two algorithms, and in particular the motivation for the min-max programs they solve, differ significantly.

### A.3. $\mathsf{AE}^2$ Algorithm: Regret Bound for Uniformly Regular Classes

We present upper bounds for $\mathsf{AE}^2$ in the setting where our class $\mathcal{M}$ is uniformly regular: Assumption A.4 holds and $\mathsf{n}_{\varepsilon}^{\mathcal{M}} < \infty$; these assumptions are relaxed by the $\mathsf{AE}^2_\star$ algorithm in the sequel. To state the regret bound for $\mathsf{AE}^2$ in the tightest form possible, we introduce the following variant of the Allocation-Estimation Coefficient, which incorporates randomized estimators $\xi \in \triangle_{\mathcal{M}}$:

$$\overline{\mathsf{aec}}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}_0, \xi) := \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \backslash \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim \omega}[D_{\mathsf{KL}}(\overline{M}(\pi) \,\|\, M(\pi))]]}, \tag{20}$$

with $\overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}, \xi) := \overline{\mathsf{aec}}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}, \xi)$ and $\overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}) := \sup_{\xi \in \triangle_{\mathcal{M}}} \overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}, \xi)$. Note that one can always bound $\overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}) \le \mathsf{aec}_{\varepsilon}(\mathcal{M})$ due to the convexity of the KL divergence. In fact, these definitions are equivalent up to dependence on problem-dependent parameters in Appendix A.1 (indeed, our

lower bounds in Appendix B scale with the latter quantity), but the former can be simpler to bound for some of the examples we consider.

Our main theorem concerning the performance of $\mathsf{AE}^2$ is as follows.

**Theorem A.1** (Regret Bound for $\mathsf{AE}^2$)**.** *For any $\varepsilon \in (0, 1/2]$, there exists a choice for the estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$ such that for all $T \in \mathbb{N}$, under Assumptions A.1 to A.4 and if $\mathsf{g}^\star > 0$, the expected regret of $\mathsf{AE}^2$ is bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T) + C_{\mathsf{low}} \cdot \log^{1/2}(T), \tag{21}$$

*where*

$$C_{\mathsf{aec}} := c \cdot \frac{V_{\mathcal{M}}^2 d_{\mathsf{cov}} \log(C_{\mathsf{cov}}) \cdot \max_{M \in \mathcal{M}} \mathsf{g}^M}{\varepsilon \Delta_{\min}^3} \cdot \left(\varepsilon^{-1} + V_{\mathcal{M}} \mathsf{n}^{\mathcal{M}}_{\varepsilon/6}\right) \cdot \log(C_{\mathsf{low}}),$$

*for a universal numerical constant $c > 0$, and $C_{\mathsf{low}}$ is a lower-order constant given by*

$$C_{\mathsf{low}} := \mathrm{lin}\left(\max_{M \in \mathcal{M}} \mathsf{g}^M, \mathsf{aec}^{1/2}_{\varepsilon/12}(\mathcal{M}), \tfrac{1}{\varepsilon^2}, \tfrac{1}{\Delta_{\min}^3}, \mathsf{n}^{\mathcal{M}}_{\varepsilon/6}, L^2_{\mathsf{KL}}, V_{\mathcal{M}}^{13/2}, d_{\mathsf{cov}}, \log(C_{\mathsf{cov}}), \log\log T\right),$$

*where* $\mathrm{lin}(\cdot)$ *denotes a function multi-linear and poly-logarithmic in its arguments.*

We prove Theorem A.1 in Appendix F.1, and give a proof sketch in Appendix A.8. Theorem A.1 shows that $\mathsf{AE}^2$ achieves the asymptotically optimal Graves-Lai rate for $M^\star$, as given in Proposition 1.1, up to a $(1+\varepsilon)$ approximation factor. In more detail, if we label the terms in Eq. (21) as

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le \underbrace{(1+\varepsilon)\mathsf{g}^\star \cdot \log(T)}_{\text{(I)}} + \underbrace{\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T)}_{\text{(II)}} + \underbrace{C_{\mathsf{low}} \cdot \log^{1/2}(T)}_{\text{(III)}},$$

the regret bound can be seen to consist of:

- The *leading-order term* $\text{(I)} = (1+\varepsilon)\mathsf{g}^\star \cdot \log(T)$. This is the only term that scales linearly with $\log(T)$, as a consequence we have $\lim_{T\to\infty} \frac{\mathbb{E}^{M^\star}[\mathbf{Reg}(T)]}{\log(T)} \le (1+\varepsilon)\mathsf{g}^\star$, which matches the instance-optimal rate given in Proposition 1.1 up to a factor of $(1+\varepsilon)$.
- A lower-order term $\text{(II)} = \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T)$, which is polylogarithmic in $\log(T)$, and scales with $\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})$, as well as regularity parameters from Appendix A.1.
- A second lower-order term $\text{(III)} = C_{\mathsf{low}} \cdot \log^{1/2} T$. This term scales with $\log^{1/2}(T) = o(\log(T))$ and, like the term (II), scales with the AEC and regularity parameters from Appendix A.1. Compared to (II), this term has worse dependence on $\log(T)$, but enjoys sublinear $\overline{\mathsf{aec}}^{1/2}_{\varepsilon/12}(\mathcal{M})$ scaling with the AEC.

Critically, both of the $o(\log T)$ lower-order terms above do not scale with (often exponentially large) terms such as $|\Pi|$ or $|\mathcal{M}|$ found in prior work, and instead scale principally with $\overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M})$, which, as

we will show in Appendix B, is unavoidable in a certain sense. In particular, note that once $T$ is large enough that

$$\log(T) \geq \widetilde{\Omega}^{+}(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})),$$

the leading-order term $(\mathrm{I}) = (1+\varepsilon)\mathsf{g}^\star \cdot \log(T)$ term in Theorem A.1 will dominate the regret. This is precisely the time horizon given by the lower bound in Theorem 2.3, which is necessary for an algorithm to learn a near-optimal allocation for the Graves-Lai program. We offer a more thorough comparison of Theorem A.1 with our lower bounds in Appendix B.4. Below, we discuss the lower-order terms and asymptotic performance in greater detail.

**Remark A.1** (Additional Lower-Order Terms)**.** *The lower-order terms in Theorem A.1 depend on the model class $\mathcal{M}$ through the regularity, covering, and smoothness assumptions, as well as the minimum gap (Appendix A.1). For many of the examples we consider, the Allocation-Estimation Coefficient will dominate these other terms, yet there may exist classes where this is not the case. Resolving the optimal dependence on these problem-dependent parameters in the lower-order terms, as well as understanding when these parameters are necessary, remains an interesting direction for future work.*

*In addition, let us mention that while both lower-order terms scale with $o(\log T)$ (note that the scaling is no larger than $O(\sqrt{\log T} \cdot \operatorname{polylog} \log T)$), it is not clear what the optimal dependence on $T$ should be for the lower-order terms. For example, one might hope to replace the dependence on $\log^{1/2}(T)$ with $\log^{a}(T)$ for some constant $a < 1/2$, or even with $\operatorname{polylog}(\log(T))$. Precisely characterizing the optimal $\log(T)$ scaling for lower-order terms remains an interesting open question. To this end, we remark that Jun and Zhang (2020) show that in some cases, an $\Omega(\log \log T)$ term is indeed necessary.*

**Remark A.2** (Asymptotic Performance)**.** *Asymptotically, as $T \to \infty$, the regret of $\mathsf{AE}^2$ scales with $(1+\varepsilon)\mathsf{g}^\star \cdot \log(T)$, which is a factor of $(1+\varepsilon)$ off from the asymptotic lower bound in Proposition 1.1. For any fixed $T \in \mathbb{N}$ of interest, as long as $\mathsf{aec}_\varepsilon(\mathcal{M}) = \operatorname{poly}(\varepsilon^{-1})$, one can obtain an asymptotic constant of $1$ by choosing $\varepsilon = 1/\log^{a}(T)$ for a sufficiently small constant $a > 0$. For example, when $|\Pi| < \infty$, it is always possible to bound $\mathsf{aec}_\varepsilon(\mathcal{M}) \lesssim \operatorname{poly}(|\Pi|)/\varepsilon^4$ (see Proposition A.1 below), so choosing $\varepsilon$ as above ensures that the lower-order terms scale $o(\log T)$, while the leading-order term scales as $\mathsf{g}^\star \cdot \log(T)$ asymptotically.*

### A.3.1. EXAMPLE: SEARCHING FOR AN INFORMATIVE ARM

We next provide an example of a uniformly regular class in order to illustrate a case where Theorem A.1 holds. In particular, we revisit the *informative arm* setting described in the introduction (Example 1.1). Recall that we exhibited a model class for which the complexity of learning the Graves-Lai allocation is not governed by existing complexity measures, and can be larger than the minimax optimal rate for learning with $\mathcal{M}$. In what follows, we show that on this example the Allocation-Estimation Coefficient correctly adapts to the complexity of this model class. We emphasize that the main applications of our results, which take advantage of the more general $\mathsf{AE}^2_\star$ algorithm, will be given in Appendix A.6 and Appendix A.7, and we also present additional examples of uniformly regular classes in Appendix G.2.1.

**Example A.1** (Searching for an Informative Arm (revisited))**.** Let $\mathcal{M}$ denote the model class constructed in Example 1.1, with parameters $A, N \geq 5$ and $\beta \in [4/A, 9/10]$. We additionally discretize

the space so that, for each $M \in \mathcal{M}$ and $\pi \in [A]$, we have $f^M(\pi) \in \{0, \Delta_{\min}, 2\Delta_{\min}, \ldots, \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}\}$,[5] and furthermore restrict $\mathcal{M}$ so that it does not include instances with multiple optimal arms. For this class, one can show that Assumptions A.1 to A.4 hold with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq O(\log A)$ and $d_{\mathsf{cov}} = O(A), C_{\mathsf{cov}} = O(N)$ (see Appendix G.4). Furthermore, we can bound $\mathsf{n}_\varepsilon^{\mathcal{M}} \leq \frac{2}{\Delta_{\min}^2}$, and

$$\mathsf{aec}_\varepsilon(\mathcal{M}) \leq \frac{64N}{\beta^2} + \frac{16A}{\Delta_{\min}^2}.$$

As a result, for this class, $\mathsf{AE}^2$ has expected regret bounded as

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + N \cdot \mathrm{poly}(A, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta_{\min}}, \log N, \log\log T) \cdot \log^{1/2}(T).$$

Note that here the only term that scales linearly with $N$ is the Allocation-Estimation Coefficient—every other class-dependent term appearing in the regret bound scales at most logarithmically in $N$. We are particularly interested in situations where the cost of finding the correct informative arm is much larger than any existing complexity measures for the problem: that is, when $\beta$ is constant and $N \gg A, \frac{1}{\Delta_{\min}}$. In this case, we have $\mathsf{aec}_\varepsilon(\mathcal{M}) \leq O(N)$, and the Allocation-Estimation Coefficient correctly captures the intuitive complexity of learning the optimal allocation. In particular, the dependence on $N$ reflects the fact that we need to test each informative arm at least once. Furthermore, as we show in Example B.1, we can lower bound $\mathsf{aec}_\varepsilon(\mathcal{M}) \geq \Omega(N)$ as well, so in the regime where $N \gg A, \frac{1}{\Delta_{\min}}$, the AEC is the dominant lower-order term. ◁

See Appendix G.2 for the proof of this example.

### A.4. The $\mathsf{AE}^2_\star$ Algorithm

While it may be reasonable to assume that the minimum gap of $M^\star$, $\Delta^\star_{\min}$, is bounded away from 0, and that the amount of useful information playing $\pi_\star$ provides is also bounded on $M^\star$, assuming that this is true for every model in the model class (as in the prequel) is a significantly stronger assumption. For example, if we let $\mathcal{M}$ denote the space of all multi-armed bandits with means in $[0, 1]$, the only possible value of $\Delta_{\min}$ is 0, as we can always find some instance with minimum gap arbitrarily close to 0. In this section, we dispense with the uniform regularity assumption: we relax Assumption A.4, and additionally prove that it suffices if only $\mathsf{n}^\star_\varepsilon := \mathsf{n}^{M^\star}_\varepsilon$ (as opposed to $\mathsf{n}^{\mathcal{M}}_\varepsilon$) is bounded.

Our main algorithm for this section, $\mathsf{AE}^2_\star$, is given in Algorithm 3. It is very similar to $\mathsf{AE}^2$ but to remove the requirement of uniform regularity, the algorithm avoids solving Eq. (18) over the entire model class $\mathcal{M}$, and instead solves it over a carefully restricted model class. For $x, y > 0$, define

$$\mathcal{M}_{x,y} := \{M \in \mathcal{M} \ : \ \Delta^M_{\min} \geq x, \mathsf{n}^M_\varepsilon \leq y\}. \tag{23}$$

$\mathsf{AE}^2_\star$ breaks its explore rounds into doubling epochs. For each epoch $\ell$, Eq. (22) in Algorithm 3 solves an AEC-like optimization problem over a restricted class $\mathcal{M}^\ell \subseteq \mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}}$, which is chosen in Line 9 to explicitly ensure that the value of the optimization problem in Eq. (22), is bounded; this is guaranteed by the definition of $\Delta^\ell$ in Line 8. Similar to $\mathsf{AE}^2$, the value of the optimization

5. The discretization is required to satisfy the uniform regularity assumption—we include it here to provide a concrete example of a uniformly regular class. However, it can be shown that, without this discretization assumption, Theorem A.2 applies to the original formulation given in Example 1.1 with the AEC again scaling as $O(N)$.

**Algorithm 3** Adaptive Exploration for Allocation Estimation for classes without uniform regularity ($\mathsf{AE}^2_\star$)

**input:** Optimality tolerance $\varepsilon$, estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$, growth parameters $\alpha_q$, $\alpha_n$, $\alpha_{\mathcal{M}} \geq 0$.
$s \leftarrow 1$, $\ell \leftarrow 1$, $\delta \leftarrow \frac{\varepsilon}{4+2\varepsilon}$, $q^s \leftarrow 1 - s^{-\alpha_q}$, $\mathsf{n}^s \leftarrow s^{\alpha_{\mathsf{n}}}$.
Compute $\xi^1 \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
**for** $t = 1, 2, 3, \ldots$ **do**
  **if** $s \geq 2^\ell$ **then** // Form active set and cover
    $\ell \leftarrow \ell + 1$.
    $\Delta^\ell \leftarrow \arg\min_{\Delta \geq 0} \Delta \quad \text{s.t.} \quad \mathsf{aec}^{\mathcal{M}}_{\delta/2}(\mathcal{M}_{\Delta,\frac{1}{\Delta}}) \leq s^{\alpha_{\mathcal{M}}}$.
    $\mathcal{M}^\ell \leftarrow \mathcal{M}_{\Delta^\ell,\frac{1}{\Delta^\ell}} \cap \left\{M \in \mathcal{M} \;:\; \mathsf{n}^M_\varepsilon + \frac{1}{\Delta^M_{\min}} + \frac{4\mathsf{g}^M}{\Delta^M_{\min}} + \frac{2\mathsf{n}^M_\varepsilon}{\mathsf{g}^M} + \frac{4}{\Delta^M_{\min}\delta} \leq \sqrt{\mathsf{n}^s}\right\}$.
    $\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}^\ell$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
  **if** $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then**
    Play $\pi_{\widehat{M}^s}$. // Exploit
  **else** // Explore
    Set $p^s \leftarrow q^s\lambda^s + (1-q^s)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \arg\min_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}^\ell\setminus\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]]}. \tag{22}$$

    Draw $\pi^s \sim p^s$, observe $r^s, o^s$, set $\mathfrak{D}^\ell \leftarrow \mathfrak{D}^\ell \cup \{(\pi^s, r^s, o^s)\}$.
    Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\mathfrak{D}^\ell, \mathcal{M}^\ell_{\mathrm{cov}})$ and $\widehat{M}^{s+1} = \mathbb{E}_{M\sim\xi^{s+1}}[M]$.
    $s \leftarrow s + 1$.

problem in Eq. (22) quantifies how much information we are gaining about the Graves-Lai allocation of $M^\star$, and the regret of the explore phase can be bounded in terms of the value of this optimization. By restricting $\mathcal{M}^\ell$ so that the value of Eq. (22) is always bounded, we can therefore ensure that the regret during the exploration phase is bounded.

Intuitively, this restriction of $\mathcal{M}^\ell$ reduces the space of models we must distinguish $M^\star$ from in order to identify its Graves-Lai allocation: rather than distinguishing $M^\star$ from all models in $\mathcal{M}$, we must only distinguish it from models in $\mathcal{M}^\ell$, which could be significantly easier. The caveat is that, since we do not know the value of $\mathsf{n}^\star_\varepsilon$ or $\Delta^\star_{\min}$, $M^\star$ may not always be in $\mathcal{M}^\ell$. In such cases, little can be said about the exploration phase—we are not able to provide any meaningful guarantees on how much information $\lambda^s$ and $\omega^s$ acquire about $M^\star$. To mitigate this, as $s$ increases we gradually relax the criteria for inclusion in $\mathcal{M}^\ell$, ensuring that for large enough $s$, $M^\star$ will be in $\mathcal{M}^\ell$. In particular, one can show that the number of exploration rounds needed to guarantee $M^\star \in \mathcal{M}^\ell$ scales with $\mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star)$, for

$$\mathcal{M}^\star := \left\{M \in \mathcal{M} \;:\; \Delta^M_{\min} \geq \Delta_\star, \mathsf{n}^M_{\varepsilon/36} \leq 1/\Delta_\star\right\} \quad \text{for} \quad \Delta_\star := \min\{\Delta^\star_{\min}, 1/\mathsf{n}^\star_{\varepsilon/36}\}. \tag{24}$$

That is, $\mathcal{M}^\star$ is the restriction of $\mathcal{M}$ to models with gap at least $\min\{\Delta^\star_{\min}, 1/\mathsf{n}^\star_{\varepsilon/36}\}$ (implying all models in $M^\star$ have a unique optimal decision), and for which the information content of the optimal decision is at most $\max\{1/\Delta^\star_{\min}, \mathsf{n}^\star_{\varepsilon/36}\}$.

**Estimation oracle.** While $\mathsf{AE}^2$ simply requires that the estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$ returns randomized estimators supported on $\triangle_{\mathcal{M}}$, for $\mathsf{AE}^2_\star$, we wish to ensure that the estimators produced are instead only supported on $\mathcal{M}^\ell$. To this end, we restrict the estimator to $\mathcal{M}^\ell_{\mathrm{cov}}$, the $(\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}^\ell$. We denote the resulting estimation oracle with $\mathbf{Alg}_{\mathsf{KL}}(\mathfrak{D}^\ell, \mathcal{M}^\ell_{\mathrm{cov}})$, where the first argument represents the set of available observations, and the second argument the set over which the estimation oracle must return an estimate.

**Computational efficiency of $\mathsf{AE}^2_\star$.** Similar to $\mathsf{AE}^2$, it is not clear how to solve the main optimization required by $\mathsf{AE}^2_\star$, Eq. (22), in general. In addition, unlike $\mathsf{AE}^2$, $\mathsf{AE}^2_\star$ maintains a version space of models, $\mathcal{M}^\ell$, which could increase the computational burden further. We emphasize that the focus of this work is primarily statistical, and leave addressing the computational challenges for future work.

### A.5. $\mathsf{AE}^2_\star$ Algorithm: Regret Bound without Uniform Regularity

The following theorem provides the main guarantee for $\mathsf{AE}^2_\star$.

**Theorem A.2** (Regret Bound for $\mathsf{AE}^2_\star$)**.** *For any $\varepsilon \in (0, 1/2]$, if Assumptions A.1 to A.3 hold and $\mathsf{g}^\star > 0$, $\mathsf{AE}^2_\star$ (Algorithm 3) ensures that for all $T \in \mathbb{N}$, the expected regret is bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \left(\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon/12}(\mathcal{M}^\star)\right)^3 \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T) + C_{\mathsf{low}} \cdot \log^{6/7}(T),$$

*where*

$$C_{\mathsf{aec}} := \widetilde{O}\left(\frac{V^3_{\mathcal{M}}(V_{\mathcal{M}} + L_{\mathsf{KL}}) \cdot d_{\mathsf{cov}} \log(C_{\mathsf{cov}})}{\varepsilon \Delta^\star_{\min}}\right),$$

*and $C_{\mathsf{low}}$ is a lower-order constant given by*

$$C_{\mathsf{low}} := \mathrm{poly}\left(\mathsf{g}^\star, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/6}, \tfrac{1}{\varepsilon}, V_{\mathcal{M}}, L_{\mathsf{KL}}, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, \log\log T\right).$$

The proof of Theorem A.2 is given in Appendix F.2. As Theorem A.2 illustrates, at the expense of a slightly larger polynomial dependence on the Allocation-Estimation Coefficient, and slightly larger lower-order terms, we can obtain near instance-optimal regret—matching the instance-optimal lower bound given in Proposition 1.1 up to a factor of $(1+\varepsilon)$—without requiring any assumption on the minimum gap, or boundedness of $\mathsf{n}^{\mathcal{M}}_\varepsilon$. Rather than scaling with the minimum gap for the entire class, $\Delta_{\min}$, Theorem A.2 scales only with the minimum gap of the ground truth model, $\Delta^\star_{\min}$, which could be substantially larger than $\Delta_{\min}$. An additional advantage of Theorem A.2 is that it scales with $\mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star)$ as opposed to $\mathsf{aec}_\varepsilon(\mathcal{M})$; for the examples we consider in the sequel, the former quantity enjoys better dependence on problem-dependent parameters. For example, we show in Appendix B that for standard classes, $\mathsf{aec}_\varepsilon(\mathcal{M})$ can scale with the minimum gap amongst all models in the class. On the other hand $\mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star)$ typically scales with $\Delta^\star_{\min}$.

We emphasize that $\mathsf{AE}^2_\star$ *requires no prior knowledge of* $\Delta^\star_{\min}$ *or* $\mathsf{n}^\star_\varepsilon$—it is able to adapt to the minimum gap and regularity of the underlying model.

**Remark A.3** (Dependence on $\mathsf{n}^\star_\varepsilon$)**.** *As we show in the following examples, $\mathsf{n}^\star_\varepsilon$ is typically bounded polynomially in standard problem parameters, though in practice this needs to be verified for each problem instance. We remark that some scaling in terms of $\mathsf{n}^\star_\varepsilon$ seems unavoidable—if there is a significant amount of information to be gained playing the optimal decision, any algorithm which is nearly instance-optimal will play the optimal decision at least $\mathsf{n}^\star_\varepsilon$ times, and therefore the "effective horizon" to eliminate alternate instances scales with $\mathsf{n}^\star_\varepsilon$. As we are the first to formalize this notion of how informative the optimal decision is, we believe more research in this direction is required.*

## A.6. Application: Structured and Contextual Bandits

We now instantiate Theorem A.2 to give regret bounds for $\mathsf{AE}^2_\star$ in standard settings of interest, bounding the Allocation-Estimation Coefficient for each setting. We begin by focusing on structured bandit settings and contextual bandits, then turn to tabular reinforcement learning in the sequel. We recall that, to map bandit problems to the DMSO framework, we take the decision space $\Pi$ to be the set of "arms", the observation space $\mathcal{O} = \{\varnothing\}$, and the reward space $\mathcal{R}$ to be the rewards from the bandit (while we do not explicitly include the rewards in the observation space, we assume they are observed). We defer proofs for all examples to Appendix G.

### A.6.1. THE UNIFORM EXPLORATION COEFFICIENT

For the main examples we consider, we proceed by first bounding the Allocation-Estimation Coefficient in terms of another, somewhat simpler parameter we refer to as the *uniform exploration coefficient*.

**Definition A.4** (Uniform Exploration Coefficient)**.** *For a randomized estimator $\xi \in \triangle_{\mathcal{M}}$, we define the uniform exploration coefficient with respect to $\xi$ at scale $\varepsilon > 0$ as the value of the following program:*

$$C^{\xi}_{\mathrm{exp}}(\varepsilon) := \min_{C\in\mathbb{R}_+, p\in\triangle_\Pi} \left\{ C \;\middle|\; \forall M, M' \in \mathcal{M} \;:\; \begin{array}{c} \max_{M''\in\{M,M'\}} \mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p}[D_{\mathsf{KL}}(\overline{M}(\pi) \,\|\, M''(\pi))]] \le 1/C \\ \Longrightarrow \max_{p'\in\triangle_\Pi} \mathbb{E}_{\pi\sim p'}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \le \varepsilon \end{array} \right\}.$$

*We define $p^{\xi}_{\mathrm{exp}}(\varepsilon)$ as any minimizing distribution for this program, and let*

$$C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) := \sup_{\xi\in\triangle_{\mathcal{M}}} C^{\xi}_{\mathrm{exp}}(\varepsilon)$$

*denote the uniform exploration constant for class $\mathcal{M}$.*

Intuitively, the uniform exploration coefficient characterizes the extent to which it is possible to explore by uniformly covering the decision space. In particular, one can always choose $p$ to be uniform over $\Pi$, which gives $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \lesssim |\Pi|/\varepsilon$, but in cases where information is shared between actions, the parameter is significantly smaller, as we will show for familiar examples below. For example, in the case of linear bandits with dimension $d$, we have $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \le \widetilde{O}(\frac{d\cdot\log 1/\varepsilon}{\varepsilon})$.

The following result shows that the Allocation-Estimation Coefficient can be bounded in terms of the uniform exploration coefficient.

**Proposition A.1** (Informal)**.** *For $\mathcal{M}_0 \subseteq \mathcal{M}$, we can bound $\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}_0) \le C_{\mathrm{exp}}(\mathcal{M}_0, \delta)$ for any*

$$\sqrt{\delta} \le \min_{M\in\mathcal{M}_0} \min\left\{ \min\left\{ \frac{1}{81 L_{\mathsf{KL}}}, \frac{\Delta^M_{\min}}{34 V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M/\Delta^M_{\min} + \mathsf{n}^M_{\varepsilon/36}}, \frac{\Delta^M_{\min}}{3} \right\}.$$

The full statement of Proposition A.1 is given in Lemma E.6. Using Proposition A.1, we obtain guarantees for $\mathsf{AE}^2_\star$ on several familiar classes, beginning with several bandit settings. We remark that Proposition A.1 is not in general tight—it simply shows that the Allocation-Estimation Coefficient is bounded by a simple, general, and interpretable notion of how easily a class can be explored. As, such the bounds on the Allocation-Estimation Coefficient in the following examples can almost certainly be improved using more specialized tools.

### A.6.2. FINITE-ARMED BANDITS

We first consider the simplest bandit setting: multi-armed bandits with finite arms.

**Example A.2** (Finite-Armed Bandit)**.** Fix $A > 0$, and consider the class of finite-armed bandits with $A$ arms and unit-variance Gaussian noise:

$$\mathcal{M} = \left\{M(\pi) = \mathcal{N}(f^M(\pi), 1) \mid f^M \in [0,1]^A\right\}.$$

It is straightforward to verify that Assumptions A.1 to A.3 hold with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$ and $d_{\mathsf{cov}} = O(A), C_{\mathsf{cov}} = O(1)$, and it can also be shown that, as long as $f^{M^\star}(\pi_\star) < 1$, we can bound $\mathsf{n}^\star_\varepsilon \leq c \cdot \frac{A^2}{\varepsilon(\Delta^\star_{\min})^4}$. In addition, we can bound $C_{\exp}(\mathcal{M}^\star, \varepsilon) \leq 4A/\varepsilon$, so Proposition A.1 gives the following result.

**Proposition A.2.** *For the finite-armed bandit problem with $A$ actions, there exists a universal constant $c > 0$ such that*

$$\overline{\mathsf{aec}}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star) \leq c \cdot \frac{A^{15}}{\varepsilon^8(\Delta^\star_{\min})^{24}}. \tag{25}$$

We immediately obtain the following corollary to Theorem A.2.

**Corollary A.1.** *For finite-armed bandits with Gaussian noise, as long as $f^{M^\star}(\pi_\star) < 1$, $\mathsf{AE}^2_\star$ has regret bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(A, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \log\log T) \cdot \log^{6/7}(T).$$

With more refined analyses, various works have achieved instance-optimal regret bounds for finite-armed bandits with tighter lower-order terms than Corollary A.1 (Garivier et al., 2016; Kaufmann et al., 2016; Lattimore, 2018; Garivier et al., 2019). We emphasize that Corollary A.1 is a special case of a much more general result. In particular, we proved the bound on the Allocation-Estimation Coefficient, Eq. (25), using tools which hold for general classes (e.g. Proposition A.1). An analysis of $\mathsf{AE}^2_\star$ specialized to finite-armed bandits would likely yield a tighter result. ◁

### A.6.3. STRUCTURED BANDITS

Many bandit problems exhibit richer structure than the multi-armed bandit setting, and the study of these settings has been the focus of much of the recent work on instance-optimal learning. We next consider one such setting, that of structured bandits with bounded *eluder dimension* (Russo and Van Roy, 2013).

**Example A.3** (Structured Bandits with Bounded Eluder Dimension)**.** Consider a bandit problem with unit-variance Gaussian noise but where the means are now given by a *general function class* $\mathcal{F}$ mapping from $\Pi$ to $[0,1]$:

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(f(\pi), 1) \mid f \in \mathcal{F}\}. \tag{26}$$

For such general settings, we might hope to capture the complexity of learning in terms of generalized notions of dimension for $\mathcal{F}$. We consider one such notion here: the eluder dimension.

**Definition A.5** (Eluder Dimension (Russo and Van Roy, 2013)). *Let $\check{d}_{\mathrm{E}}(\mathcal{F}, \varepsilon')$ denote the length of the longest sequence of actions $\{\pi_1, \dots, \pi_d\}$ such that, for each $n \leq d$, there exist functions $f, f' \in \mathcal{F}$ with $\sqrt{\sum_{i=1}^{n-1}(f(\pi_i) - f'(\pi_i))^2} \leq \varepsilon'$ but $f(\pi_n) - f'(\pi_n) > \varepsilon'$. The* eluder dimension *of function class $\mathcal{F}$ at scale $\varepsilon$ is then defined as $d_{\mathrm{E}}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \geq \varepsilon} \check{d}_{\mathrm{E}}(\mathcal{F}, \varepsilon') \vee 1$.*

The eluder dimension can be thought of as quantifying how easily a function class can be "explored": evaluating a pair of functions on $d_{\mathrm{E}}(\mathcal{F}, \varepsilon)$ points allows one to determine whether they are nearly identical over the entire space. It is known to be bounded for many standard classes—for example, for linear function classes with dimension $d$, $d_{\mathrm{E}}(\mathcal{F}, \varepsilon) \leq \widetilde{O}(d \cdot \log 1/\varepsilon)$—and is also closely related to the *disagreement coefficient* (Foster et al., 2020). Furthermore, it can be shown to be a sufficient condition for bounded (worst-case) regret in general bandit problems (Russo and Van Roy, 2013). The following result shows that the eluder dimension bounds the Allocation-Estimation Coefficient.

**Proposition A.3.** *For the structured bandit class $\mathcal{M}$ considered in (26), we have*

$$C_{\mathrm{exp}}(\mathcal{M}^\star, \delta) \leq \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)}{\delta}.$$

*This implies that*

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)}{\delta} \quad \textit{for scale} \quad \delta = c \cdot \frac{\varepsilon^2 \Delta_\star^8}{d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta_\star)^2} \quad \textit{and} \quad \Delta_\star := \min\Big\{\Delta^\star_{\min}, 1/\mathsf{n}^\star_{\varepsilon/36}\Big\},$$

*where $c > 0$ is a universal constant.*

Proposition A.3 highlights the ability of the Allocation-Estimation Coefficient to adapt to the inherent complexity of "exploring" for the model class under consideration. We henceforth abbreviate $d_{\mathrm{E}} := d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)$.

It is straightforward to show that Assumptions A.1 and A.2 are met in this setting with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$ (see Appendix G.2). Furthermore, Assumption A.3 can be shown to hold with $d_{\mathsf{cov}}$ scaling with the covering number of $\mathcal{F}$ in the distance $d(f, f') = \sup_{\pi \in \Pi} |f(\pi) - f'(\pi)|$ and $C_{\mathsf{cov}} = O(1)$. In general, it must be shown that $\mathsf{n}^\star_\varepsilon$ is bounded for each $M^\star$ and class of interest; as we show in the following examples, it is bounded for standard structured bandit settings such as linear bandits. We have the following corollary to Theorem A.2.

**Corollary A.2.** *In the structured bandit setting with bounded eluder dimension considered above, $\mathsf{AE}^2_\star$ has regret bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(d_{\mathrm{E}}, d_{\mathsf{cov}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log T) \cdot \log^{6/7}(T).$$

To the best of our knowledge, Corollary A.2 is the first result to show that it is possible to obtain the instance-optimal rate in classes with bounded eluder dimension, with lower-order terms scaling only polynomially in the eluder dimension. More generally, Corollary A.2 illustrates that $\mathsf{AE}^2_\star$ can adapt to the structural properties of the given model class, and achieve regret scaling with existing notions of intrinsic dimension. ◁

We next consider two examples of structured bandits where it is known that the eluder dimension is bounded: linear bandits and generalized linear models. While these results are immediate given Corollary A.2, the additional structure present in these settings allows us to obtain somewhat more explicit results.

**Example A.4** (Linear Bandits)**.** Consider the class of linear bandits with unit-variance Gaussian noise defined as

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(\langle \theta, x_\pi \rangle, 1) \mid \theta \in \Theta\},$$

where $\Theta \subseteq \mathbb{R}^d$ is some convex set with $\ell_2$ diameter $O(1)$ and $\mathcal{X} := \{x_\pi : \pi \in \Pi\} \subseteq \mathbb{R}^d$ is the arm set, which we assume has $\|x_\pi\|_2 \leq 1$ for all $\pi \in \Pi$. As in Example A.3, Assumptions A.1 and A.2 are met in this setting with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$; furthermore, Assumption A.3 is also met with $d_{\mathsf{cov}} = O(d)$ and $C_{\mathsf{cov}} = O(1)$. We then have the following bound on the AEC.

**Proposition A.4.** *For the linear bandit class $\mathcal{M}$ defined above, we have*

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq c \cdot \frac{d^3}{\varepsilon^2} \cdot \left( \frac{1}{\Delta^\star_{\min}} + \mathsf{n}^\star_{\varepsilon/36} \right)^8 \cdot \mathrm{polylog}\left( d, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36} \right)$$

*for a universal constant $c > 0$.*

Using Proposition A.4, we obtain the following corollary to Theorem A.2.

**Corollary A.3.** *In the linear bandit setting defined above, $\mathsf{AE}^2_\star$ has regret bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(d, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log T) \cdot \log^{6/7}(T).$$

Corollary A.4 has lower-order terms scaling similarly to the best known lower-order terms in the linear bandit setting (Tirinzoni et al., 2020; Kirschner et al., 2021). These works, however, develop algorithms which are specialized to the linear bandit setting, while Corollary A.4 is the instantiation of a more general result designed for arbitrary general decision-making settings.

The following result shows that under general conditions, we can bound the parameter $\mathsf{n}^\star_\varepsilon$ for linear bandits.

**Proposition A.5** (Informal)**.** *For linear bandits satisfying certain regularity conditions, $\mathsf{n}^\star_\varepsilon$ is bounded by a polynomial function of problem parameters and a geometry-dependent term scaling with the structure of $\mathcal{X}$ and $\Theta$.*

The full statement of Proposition A.5 is given in Proposition G.2. The regularity condition for Proposition A.5 requires primarily that $\theta^\star$ lies sufficiently far within the interior of $\Theta$ (see Appendix G.2.4 for further details). We remark that the guarantees given in both Tirinzoni et al. (2020) and Kirschner et al. (2021) scale with geometric parameters very similar to $\mathsf{n}^\star_\varepsilon$.

◁

**Example A.5** (Generalized Linear Models)**.** In the generalized linear model setting, we take the model class to be

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(g(\langle \theta, x_\pi \rangle), 1) \mid \theta \in \Theta\},$$

where $\Theta$ and $\mathcal{X}$ are as in Example A.4, and $g(\cdot)$ is a known link function which is increasing and Lipschitz, but potentially nonlinear. Let $g_{\max}$ and $g_{\min}$ denote upper and lower bounds on the derivative of $g$, respectively:

$$g_{\max} := \max_{\theta\in\Theta, x\in\mathrm{co}(\mathcal{X})} g'(\langle\theta, x\rangle) \quad \text{and} \quad g_{\min} := \min_{\theta\in\Theta, x\in\mathrm{co}(\mathcal{X})} g'(\langle\theta, x\rangle).$$

As in the linear bandit setting, we can show that Assumptions A.1 and A.2 are both met with $L_{\mathsf{KL}}, V_{\mathcal{M}} \le 4$, and that Assumption A.3 is also met with $d_{\mathsf{cov}} = O(d)$ and $C_{\mathsf{cov}} = O(g_{\max})$. Furthermore, under the same conditions as for linear bandits, $\mathsf{n}^\star_\varepsilon$ can be bounded for generalized linear models exactly as for linear bandits, but with an additional scaling of $(\frac{g_{\max}}{g_{\min}})^2$. We then have the following.

**Proposition A.6.** *For the generalized linear model class $\mathcal{M}$ defined above, we have*

$$\overline{\mathsf{aec}}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star) \le c\cdot\frac{d^3 g_{\max}^3}{\varepsilon^2 g_{\min}^3}\cdot\left(\frac{1}{\Delta^\star_{\min}} + \mathsf{n}^\star_{\varepsilon/36}\right)^8 \cdot \mathrm{polylog}\left(d, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36}\right)$$

*for a universal constant $c > 0$.*

Using Proposition A.4, we obtain the following corollary to Theorem A.2.

**Corollary A.4.** *In the generalized linear model setting defined above, $\mathsf{AE}^2_\star$ has regret bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star\cdot\log(T) + \mathrm{poly}(d, \tfrac{g_{\max}}{g_{\min}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log T)\cdot\log^{6/7}(T).$$

To the best of our knowledge, this is the first result to obtain finite-time instance-optimality for generalized linear models with lower-order terms polynomial in problem parameters. ◁

### A.6.4. CONTEXTUAL BANDITS

The previous examples illustrate that $\mathsf{AE}^2_\star$ is able to learn efficiently in a variety of structured bandit settings. We now show that it leads to new guarantees for finite-action contextual bandits with general function approximation.

**Example A.6** (Contextual Bandits with Finitely Many Arms)**.** Consider the contextual bandit setting with context set $\mathcal{X}$ (which could be arbitrarily large) and action set $\mathcal{A}$ such that $A := |\mathcal{A}| < \infty$. Let $p_{\mathcal{X}}$ denote the context distribution, which we assume is known to the learner. The learning protocol is then, for step $t = 1, 2, 3, \ldots$:

1. Environment samples context $x^t \sim p_{\mathcal{X}}$.
2. Learner chooses action $a^t \in \mathcal{A}$, receives reward $r_t$.

We assume that $r^t = f^\star(x^t, a^t) + w^t$ for $w^t \sim \mathcal{N}(0,1)$, for some $f^\star : \mathcal{X}\times\mathcal{A} \to [0,1]$. We assume as well that the learner is given access to a set of functions $\mathcal{F}$ such that $f^\star \in \mathcal{F}$.

To view this setting as a special case of the DMSO framework, we take the decision space to be the set $\Pi = (\mathcal{X} \to \mathcal{A})$ of all policies mapping from $\mathcal{X}$ to $\mathcal{A}$, and take $\mathcal{O} = \mathcal{X}$ as the observation space. The learner's decision at round $t$ is a policy $\pi^t$, and they receive a reward-observation pair $(r^t, o^t) = (r^t, x^t)$ under the process $x^t \sim p_{\mathcal{X}}$, $r^t \sim \mathcal{N}(f^\star(x^t, \pi^t(x^t)), 1)$. The model class $\mathcal{M}$ is the set of all instances of this form for $f^\star \in \mathcal{F}$.

The following result shows that the Allocation-Estimation Coefficient is be bounded by the number of actions $A$, and is *independent* of the size of the context space. See Appendix G.3 for a proof.

**Proposition A.7.** *For the contextual bandit setting, we can bound*

$$C_{\mathrm{exp}}(\mathcal{M}^\star, \varepsilon) \le \frac{4A}{\varepsilon},$$

*which implies that*

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \le c \cdot \frac{A^3}{\varepsilon^2} \cdot \left( \frac{1}{\Delta^\star_{\mathrm{min}}} + \mathsf{n}^\star_{\varepsilon/36} \right)^8,$$

*for a universal constant $c > 0$.*

As in the cases of bandits with bounded eluder dimension, $\mathsf{n}^\star_\varepsilon$ must be bounded for each $M^\star$ and class $\mathcal{F}$ of interest. It is straightforward to show, however, that Assumptions A.1 and A.2 are met in this setting with $L_{\mathsf{KL}}, V_{\mathcal{M}} \le 4$, and, furthermore, that Assumption A.3 is also met with $d_{\mathsf{cov}}$ scaling as the covering number of $\mathcal{F}$ in the distance $d(f, f') = \sup_{x \in \mathcal{X}, a \in \mathcal{A}} |f(x,a) - f'(x,a)|$, and $C_{\mathsf{cov}} = O(1)$. We then have the following corollary.

**Corollary A.5.** *In the finit-action contextual bandit setting considered above, $\mathsf{AE}^2_\star$ has regret bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(A, d_{\mathsf{cov}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\mathrm{min}}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log T) \cdot \log^{6/7}(T).$$

To the best of our knowledge, Corollary A.5 is the first instance-optimal guarantee in the contextual bandit setting with general function approximation that obtains lower-order term scaling polynomially in problem parameters. Notably, the lower-order term scales independently of the size of the context space, $|\mathcal{X}|$. We anticipate that extending this result to contextual bandit settings that have large action spaces, but which exhibit additional structure allowing for efficient exploration (e.g., linearity), will be straightforward. ◁

### A.7. Application: Tabular Reinforcement Learning

As a final application of our results, we turn to the setting of episodic tabular reinforcement learning.

**Episodic Markov decision processes.** Recall that episodic reinforcement learning is a special case of the DMSO framework in which each model $M \in \mathcal{M}$ is an episodic Markov Decision Process (MDP) given by the tuple $M = (\mathcal{S}, \mathcal{A}, H, \{P^M_h\}_{h=1}^H, \{R^M_h\}_{h=1}^H, s_1)$. Here $\mathcal{S}$ is a set of states, $\mathcal{A}$ a set of actions, $H$ the horizon, $P^M_h : \mathcal{S} \times \mathcal{A} \to \triangle_{\mathcal{S}}$ the probability transition kernel at step $h$, $R^M_h : \mathcal{S} \times \mathcal{A} \to \triangle_{\mathbb{R}}$ the reward distribution at step $h$, and $s_1$ a deterministic initial state, which we take to be fixed across models. We assume that $R^M_h(s,a)$ is unit-variance Gaussian, and that $\mathbb{E}_{r_h \sim R^M_h(s,a)}[r_h] \in [0, 1/H]$.[6]

The decision space $\Pi$ consists of non-stationary policies $\pi = (\pi_1, \dots, \pi_H)$, where $\pi_h : \mathcal{S} \to \mathcal{A}$. For a fixed policy $\pi$, an episode proceeds in an MDP $M$ proceeds as follows. First, beginning from the initial state $s_1$, we take action $a_1 \sim \pi_1(s_1)$, receive reward $r_1 \sim R^M_1(s_1, a_1)$, and transition to $s_2 \sim P^M_1(\cdot \mid s_1, a_1)$. This continues for $H$ steps at which point the episode terminates and the process repeats. We define $f^M(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^H r_h\right]$ as the expected reward achieved over the entire episode under this process.

6. This assumption only serves to ensure that $f^M(\pi) \in [0,1]$, in line with the convention for the rest of the paper. Our results continue to hold up to $\mathrm{poly}(H)$ factors if $\mathbb{E}_{r_h \sim R^M_h(s,a)}[r_h] \in [0,1]$.

Each round $t \in [T]$ in the DMSO framework corresponds to an episode in the underlying MDP $M^\star$. At each round, the learner selects a policy $\pi^t$, and receives reward $r^t = \sum_{h=1}^H r_h^t$ and $o^t = (s_1^t, a_1^t, r_1^t, \ldots, s_H^t, a_H^t, r_H^t)$, where $(s_1^t, a_1^t, r_1^t, \ldots, s_H^t, a_H^t, r_H^t)$ is the trajectory that results from executing $\pi^t$ in $M^\star$ for a single episode.

**Tabular model class.** In the tabular RL setting, it is assumed that $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$ are both finite, and we take $\Pi$ to be the set of all deterministic policies. In addition to assuming that $\mathcal{M}$ consists of tabular MDPs, we restrict to the following subclass:

$$\mathcal{M}_{\mathrm{tab}}(P_{\min}) := \left\{ M = (\mathcal{S}, \mathcal{A}, H, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, s_1) \ : \ \min_{s,a,s',h} P_h^M(s' \mid s, a) \geq P_{\min} \right\}. \tag{27}$$

While the assumption that $\min_{s,a,s',h} P_h^M(s' \mid s, a) \geq P_{\min}$ may be seen as restrictive, the guarantees we provide scale only with $\log \frac{1}{P_{\min}}$, so $P_{\min}$ can be taken to be extremely small without affecting the result significantly.

Note that when our results are specialized to reinforcement learning, $\Delta_{\min}^\star$ denotes the gap between the performance of the optimal policy, and the next-best *deterministic* policy. This quantity can be lower bounded in terms of other standard quantities including gaps in the rewards at each state and the transition probabilities.

Toward instantiating Theorem A.2 in this tabular RL setting, we first provide a bound on the Allocation-Estimation Coefficient, which we establish by first bounding the Uniform Exploration Coefficient.

**Proposition A.8.** *For $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$, we can bound*[7]

$$C_{\mathrm{exp}}^{\mathsf{H}}(\mathcal{M}^\star, \varepsilon) \leq c \cdot \frac{SAH^2 \cdot \log^2 H}{\varepsilon^2}$$

*for a universal constant $c > 0$, which implies that*

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq c \cdot \frac{S^5 A^5 H^{14} \cdot \log^{10} H}{\varepsilon^4} \cdot \left( \frac{1}{\Delta_{\min}^\star} + \mathsf{n}_{\varepsilon/36}^\star \right)^{24} \cdot \log^4 \frac{1}{P_{\min}}$$

*for a universal constant $c > 0$.*

Next, it can be shown that Assumptions A.1 to A.3 hold for $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$ with constants (see Appendix G.5):

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = O(H \cdot \log 1/P_{\min}), \quad d_{\mathsf{cov}} = O(S^2 AH), \quad C_{\mathsf{cov}} = O(H/P_{\min}).$$

We then obtain the following corollary to Theorem A.2.

**Corollary A.6.** *For $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$ and $\varepsilon$, $\mathsf{AE}_\star^2$ has regret bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(S, A, H, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta_{\min}^\star}, \log \tfrac{1}{P_{\min}}, \mathsf{n}_{\varepsilon/36}^\star, \log\log(T)) \cdot \log^{6/7}(T).$$

7. Here $C_{\mathrm{exp}}^{\mathsf{H}}$ denote the uniform exploration coefficient as defined in Definition A.4, but with $D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$ replaced with $D_{\mathsf{H}}^2(\cdot, \cdot)$. To prove Proposition A.8, we show that a variant of Proposition A.1 still holds with this alternate definition of $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon)$.

To our knowledge, Corollary A.6 is the first guarantee for tabular RL that achieves the instance-optimal rate while obtaining lower-order terms that scale only polynomially in problem parameters. As noted previously, existing approaches to instance-optimal regret in tabular RL (Ok et al., 2018; Dong and Ma, 2022) have lower-order terms that scale exponentially in problem parameters, and as a result are truly asymptotic in nature.

While Corollary A.6 is stated in terms of $\mathsf{n}^\star_\varepsilon$ for the sake of generality, as we show in Appendix G.5, if $M^\star$ has rewards that are sufficiently small (in particular, if it satisfies $\mathbb{E}_{r_h \sim R_h^{M^\star}(s,a)}[r_h] < 1/H^2$ for all $(s,a,h)$), then we can bound

$$\mathsf{n}^\star_\varepsilon \le c \cdot \frac{\mathsf{g}^\star}{\Delta^\star_{\min}} \cdot \left(1 + \frac{\mathsf{g}^\star}{\varepsilon(\Delta^\star_{\min})^2}\right).$$

In this case, Corollary A.6 scales polynomially in all standard problem parameters.

Let us remark that the prior work of Dong and Ma (2022) does not require that $P_h^M(s' \mid s,a) \ge P_{\min}$ as we do (the work of Ok et al. (2018) only holds for ergodic MDPs, itself a very strong assumption). However, the lower-order term obtained in Dong and Ma (2022) scales polynomially in the inverse probability of observing the trajectory that occurs with minimum (non-zero) probability. In general, this will scale exponentially in $H$, and inversely with the probability of the transition with minimum (non-zero) probability occurring, that is $\min_{s,a,s',h:P_h^M(s'|s,a)>0} P_h^M(s'|s,a)$. Thus, while we must impose the stronger condition that all transitions occur with some probability $P_{\min}$, our bounds only scale logarithmically in this quantity, and polynomially in $S$, $A$, and $H$, a significant improvement over Dong and Ma (2022). Understanding whether it is possible to remove the additional restrictions we impose while still obtaining reasonable finite-time performance is an interesting direction for future work.

As far as we are aware, there is no prior work on instance-optimal algorithms for RL settings with general function approximation. While we have only instantiated Theorem A.2 and $\mathsf{AE}^2_\star$ in the tabular RL setting, the tools we have developed can also be applied to RL with general model classes. Exploring the application of $\mathsf{AE}^2_\star$ to, for example, bilinear classes (Du et al., 2021) is an exciting avenue for future work.

### A.8. Overview of Analysis

To close this section we briefly sketch the proof of the regret bound for $\mathsf{AE}^2$ (Theorem A.1); the proof of the regret bound for $\mathsf{AE}^2_\star$ (Theorem A.2) builds on these ideas, but is slightly more involved. See Appendix F for full proofs.

Let us refer to the *exploit phase* as the subset of rounds $t$ in which Line 6 of $\mathsf{AE}^2$ is reached, and refer to the *explore phase* as the subset of rounds in which Line 8 is reached. We focus on bounding the regret in the explore phase—it can be shown (Lemma F.1) that in the exploit phase, where the if statement on Line 6 is true, $\pi_{\widehat{M}^s} = \pi_\star$ for all but $O(\log\log T)$ rounds, so that the regret incurred in this phase is at most $O(\log\log T)$.

Let $s_T$ denote the total number of rounds in the explore phase up to time $T$. Fix an explore round $s \in [s_T]$. We bound the regret $\Delta^\star(p^s)$ by considering three cases.

**Case 1:** $M^\star \in \mathcal{M} \backslash \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$**.** In this case, $\lambda^s$ is not an optimal (normalized) allocation for $M^\star$, but we can use the AEC to argue that the information gained by the algorithm is large. In

particular, since $p^s$ plays $\omega^s$ with probability at least $1-q$, we can bound

$$\begin{aligned}
&\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]} \\
&\leq \frac{1}{1-q}\cdot\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]} \\
&\overset{(a)}{\leq} \frac{1}{1-q}\cdot\min_{\lambda,\omega\in\triangle_\Pi}\sup_{M\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda;\mathsf{n}_{\max})}\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]} \\
&\lesssim \frac{1}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})
\end{aligned}$$

as long as $\mathsf{n}_{\max}$ is chosen appropriately; here, $(a)$ follows because $M^\star\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda^s;\mathsf{n}_{\max})$ by assumption in this case, and by the choice of $\lambda^s$ and $\omega^s$ given in Eq. (18). Rearranging this gives

$$1\lesssim\frac{1}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]].$$

This reflects that, when $M^\star\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s;\mathsf{n}_{\max})$, our choice of $p^s$ ensures that $\widehat{M}\sim\xi^s$ and $M^\star$ can be distinguished, with the amount of information gained lower bounded by $O(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})^{-1})$

Adding and subtracting $\frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]$ to $\Delta^\star(p^s)$, and using that $\Delta^\star(p^s)\leq 1$ always, we then have that the instantaneous regret in this case is bounded by

$$\begin{aligned}
\Delta^\star(p^s) &= \Delta^\star(p^s)-\frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\
&\quad+\frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\
&\leq -1+\frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]].
\end{aligned}$$

Summing over $s$, it follows that the total regret in this case is bounded by

$$\sum_{s=1}^{s_T}\Delta^\star(p^s)\cdot\mathbb{I}\{s\text{ in Case 1}\}\leq\frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)-\sum_{s=1}^{s_T}\mathbb{I}\{s\text{ in Case 1}\}.$$

Furthermore, since regret is always non-negative—that is, $\Delta^\star(p)\geq 0$ for all $p$—rearranging this inequality leads to a bound on the total number of times this case can occur:

$$\sum_{s=1}^{s_T}\mathbb{I}\{s\text{ in Case 1}\}\leq\frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbf{Est}_{\mathsf{KL}}(s_T).$$

Critically, as given in (17), $\mathbf{Est}_{\mathsf{KL}}(s_T)$ scales at most poly-logarithmically in $s_T$. Thus, as long as $s_T$ is at most $O(\log T)$, the total regret incurred in this case (as well as the total number of times this case can occur), will be at most $O(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\log\log T)$.

**Case 2: $M^\star \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$ and $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$.** In this case, we have that $\lambda^s$ is a Graves-Lai optimal allocation for $M^\star$. Thus, it follows that

$$\Delta^\star(\lambda^s) \le (1+\varepsilon/6)\mathsf{g}^\star/\mathsf{n}^\star \quad \text{and} \quad \inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{\pi\sim\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \ge (1-\varepsilon/6)/\mathsf{n}^\star$$

for some $\mathsf{n}^\star \le \mathsf{n}_{\max}$. This implies that, for any $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, we can bound

$$\begin{aligned}
\Delta^\star(p^s) &= \Delta^\star(p^s) - (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&\lesssim (1+\varepsilon/6)\mathsf{g}^\star/\mathsf{n}^\star - (1+\varepsilon)(1-\varepsilon/6)\mathsf{g}^\star/\mathsf{n}^\star + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&\lesssim -\varepsilon\mathsf{g}^\star/\mathsf{n}^\star + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&\le -\varepsilon\mathsf{g}^\star/\mathsf{n}_{\max} + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]
\end{aligned}$$

Since this bound holds uniformly for all $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, it follows that the total regret in this case can be bounded as

$$\begin{aligned}
\sum_{s=1}^{s_T} \Delta^\star(p^s) \cdot \mathbb{I}\{s \text{ in Case 2}\} \lesssim (1+\varepsilon)\mathsf{g}^\star \cdot \sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\
- \frac{\varepsilon\mathsf{g}^\star}{\mathsf{n}_{\max}} \sum_{s=1}^{s_T} \mathbb{I}\{s \text{ in Case 2}\}.
\end{aligned}$$

To bound this, the key observation is that, if we explore at round $s$, then it must be the case that, for all $\pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$, there exists some $M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$ such that $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}] \le \log(T\log T)$. Using Assumption A.1 and Assumption A.2 to move from $\widehat{M} \sim \xi^i$ to $M^\star$ and to relate the observed log-likelihood ratios to the KL divergence, we can furthermore show that

$$\begin{aligned}
&\inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} \sum_{s=1}^{s_T} \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\
&\quad\lesssim \inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} \sum_{s=1}^{s_T} \mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M,\pi^s}(r^s,o^s)}\right] + \sqrt{s_T} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T) \\
&\quad\lesssim \log(T\log T) + \sqrt{s_T} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T).
\end{aligned}$$

This allows us to bound

$$\begin{aligned}
(1+\varepsilon)\mathsf{g}^\star \cdot \sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\
\lesssim (1+\varepsilon)\mathsf{g}^\star \cdot \log T + \mathsf{g}^\star\sqrt{s_T} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T).
\end{aligned}$$

Thus, as long as $s_T = O(\log T)$, we can bound the total regret incurred in Case by $(1+\varepsilon)\mathsf{g}^\star \cdot \log T + o(\log T)$. Using that regret is always lower bounded by 0 in the same fashion as Case 1, we can further use this to bound the total number of times that Case 2 occurs by

$$\sum_{s=1}^{s_T} \mathbb{I}\{s \text{ in Case 2}\} \le \frac{\mathsf{n}_{\max}}{\varepsilon\mathsf{g}^\star}\Big(\mathsf{g}^\star \cdot \log T + \mathsf{g}^\star\sqrt{s_T} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T)\Big).$$

The intuition for this case is that, since we are playing a Graves-Lai allocation for $M^\star$, the regret will scale with $\mathsf{g}^\star$, the instance-optimal rate, and, furthermore, the allocation will allow us to distinguish $M^\star$ from alternatives $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$. Using that the total estimation error is bounded, and that we only enter the explore phase if there exists some $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ that we cannot distinguish from $M^\star$, this ultimately implies that the total number of times this phase occurs, and therefore the total regret incurred by this phase, is bounded.

**Case 3: $M^\star \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$ and $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$.** In this case, we bound $\Delta^\star(p^s)$ by adding and subtracting $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]$ in the same fashion as Case 1. Since $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$, it can be shown that $\xi^s$ must place $\Omega(\Delta_{\min})$ probability mass on $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, allowing us to lower bound $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]$ using the same reasoning as in Case 2. In total, we can show that the regret in this case is bounded by $O(\frac{\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T))$, and the number of times this case can occur is at most $O(\frac{\mathsf{n}_{\max}}{\varepsilon\Delta_{\min}} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T))$.

**Concluding the Proof.** Combining all three cases, we have shown that the regret of the explore phase is bounded by

$$(1+\varepsilon)\mathsf{g}^\star \cdot \log T + \widetilde{O}\left(\frac{1}{1-q} \cdot \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot \mathbf{Est}_{\mathsf{KL}}(s_T) + \mathsf{g}^\star\sqrt{s_T} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T) + \frac{\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T)\right).$$

Furthermore, using our bounds on the number of times each case can occur, one can show that $s_T = O(\log T)$. Since $\mathbf{Est}_{\mathsf{KL}}(s_T)$ is at most polylogarithmic in $s_T$, it follows that the regret is bounded as

$$(1+\varepsilon)\mathsf{g}^\star \cdot \log T + \widetilde{O}\left(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) + \overline{\mathsf{aec}}^{1/2}_{\varepsilon/12}(\mathcal{M}) \cdot \log^{1/2} T\right),$$

as stated in Theorem A.1.

## Appendix B. Lower Bounds for Learning the Optimal Allocation

The AE$^2$ algorithm achieves instance-optimal regret by explicitly learning an $\varepsilon$-optimal Graves-Lai allocation for the underlying model $M^\star$. In this section, we introduce an abstract formulation for the problem of learning an optimal allocation (Appendix B.1), and provide lower bounds which show that the Allocation-Estimation Coefficient is a fundamental limit for this task (Appendix B.2). We then present several examples illustrating lower bounds on the Allocation-Estimation Coefficient (Appendix B.3), and discuss how our lower bounds relate to the problem of minimizing regret (Appendix B.4).

**Additional Notation.** Throughout this section we will also make use of the following definition:

$$\Lambda(M; \varepsilon, \mathsf{n}_{\max}) := \Bigg\{\lambda \in \triangle_\Pi \;:\; \exists \mathsf{n} \in (0, \mathsf{n}_{\max}] \text{ s.t. } \Delta^M(\lambda) \le \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}, \\ \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi\sim\lambda}\big[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)\big] \ge \frac{1-\varepsilon}{\mathsf{n}}\Bigg\}. \tag{28}$$

That is, $\Lambda(M; \varepsilon, \mathsf{n}_{\max})$ denotes the set of normalized allocations that are Graves-Lai optimal for $M$ with tolerance $\varepsilon$, and have normalization factor at most $\mathsf{n}_{\max}$. Unless otherwise stated, the results in this section do not make use of Assumption 1.3 or Assumption A.4.

### B.1. Learning the Optimal Allocation: Minimax Formulation

We consider the following protocol, which captures the task of learning an optimal Graves-Lai allocation for an unknown model $M^\star$.

- For $t = 1, \dots, T$, sample $\pi^t \sim p^t$ and observe $(r^t, o^t)$.
- Based on the entire history $\mathcal{H}^T = (\pi^1, r^1, o^1), \dots, (\pi^T, r^T, o^T)$, output a normalized allocation $\widehat{\lambda} \in \triangle_\Pi$. The allocation may be randomized according to a distribution $q \in \triangle_{\triangle_\Pi}$.

We formalize an *algorithm* for this task as a pair $\mathbb{A} = (p, q)$, where $q(\cdot \mid \mathcal{H}^T)$ is the distribution over $\widehat{\lambda}$ given the history, and $p = \{p^t\}_{t\in[T]}$ is a sequence of exploration distributions of the form $p^t(\cdot \mid \mathcal{H}^{t-1})$. We let $\mathbb{P}^{M,\mathbb{A}}(\cdot)$ denote the law of $\mathcal{H}^T$ when $M$ is the underlying model and $\mathbb{A}$ is the algorithm, and let $\mathbb{E}^{M,\mathbb{A}}[\cdot]$ denote the corresponding expectation. The goal of the algorithm is to ensure that $\widehat{\lambda}$ is an $\varepsilon$-optimal allocation for $M^\star$ with high probability, i.e.

$$\mathbb{P}^{M^\star,\mathbb{A}}\Big(\widehat{\lambda} \in \Lambda(M^\star; \varepsilon, \mathsf{n}_{\max})\Big) \geq 1 - \delta$$

for some failure probability $\delta > 0$ and normalization factor $\mathsf{n}_{\max} > 0$

Intuitively, learning an optimal Graves-Lai allocation is closely related to achieving instance-optimal regret, but there are some subtle technical differences which we discuss in detail in the sequel. We study the former task because we find it to be more amenable to non-asymptotic lower bounds, and because it captures the behavior of "natural" algorithms such as $\mathsf{AE}^2$ and essentially every existing asymptotically optimal algorithm we are aware of.

**Minimax framework.** To provide lower bounds on the complexity of learning an optimal Graves-Lai allocation, we consider a minimax framework. Our main quantity of interest will be:

$$T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) = \inf_{\mathbb{A}} \inf\Big\{T \in \mathbb{N} \mid \mathbb{P}^{M,\mathbb{A}}\Big(\widehat{\lambda} \in \Lambda(M; \varepsilon, \mathsf{n}_{\max})\Big) \geq 1 - \delta,\ \forall M \in \mathcal{M}\Big\}. \quad (29)$$

This represents the earliest time $T \in \mathbb{N}$ for which there exists an algorithm that learns an $\varepsilon$-optimal allocation with probability at least $1 - \delta$, and does so uniformly for all $M \in \mathcal{M}$. Recall that for our upper bounds (Theorem A.1), the Allocation-Estimation Coefficient gives a bound on the lower-order terms in the regret (reflecting the time required to learn an $\varepsilon$-optimal allocation) that holds *uniformly* for all models in the class $\mathcal{M}$. To understand the optimality of uniform bounds of this type, a minimax framework is natural. This framework also naturally complements recent non-asymptotic algorithms for linear models such as (Tirinzoni et al., 2020; Kirschner et al., 2021), where the complexity of exploration is captured by problem-dependent quantities such as the feature dimension, which are bounded uniformly for all models in the class. Nonetheless, exploring other notions of optimality (for example, instance-dependent complexity) for learning the Graves-Lai allocation is an interesting direction for future research.

Note that the quantity (29) does not place any constraint on the regret of the algorithm under consideration. It will also be useful to consider the notion

$$T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta, R)$$
$$= \inf_{\mathbb{A}} \inf\Big\{T \in \mathbb{N} \mid \mathbb{P}^{M,\mathbb{A}}\Big(\widehat{\lambda} \in \Lambda(M; \varepsilon, \mathsf{n}_{\max})\Big) \geq 1 - \delta,\ \mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq R \cdot \log(T),\ \ \forall M \in \mathcal{M}\Big\}, \quad (30)$$

which captures the minimax complexity of learning the Graves-Lai allocation, subject to the constraint that the algorithm achieves logarithmic regret throughout the learning process. This notion is particularly well suited to complement the upper bounds achieved by algorithms such as $\mathsf{AE}^2$.

We remark that the restriction to allocations with normalization factor no more than $\mathsf{n}_{\max}$ in the definitions above is natural for several reasons. First, without such a restriction, the returned allocation can place an arbitrarily small amount of mass on informative actions, and an arbitrarily large amount of mass on the optimal action, since any Graves-Lai optimal allocation is still Graves-Lai optimal if the amount of mass on the optimal action is increased arbitrarily. While technically Graves-Lai optimal, such allocations do not reflect an allocation one could play over a finite time horizon in order to certify the optimal decision $\pi_\star$—as any algorithm with finite-time guarantees must do—and therefore do not reflect the cost such an algorithm must pay to learn an allocation. Second, given knowledge of the optimal action, any Graves-Lai optimal allocation which has a normalization factor larger than $\mathsf{n}_{\max}$ can be transformed into a Graves-Lai optimal allocation with normalization factor $\mathsf{n}_{\max}$ by adjusting the mass on the optimal action, assuming $\mathsf{n}_{\max}$ is taken to be sufficiently large (in particular, as large as $\mathsf{n}_{\max}(\mathcal{M}, \varepsilon)$; cf. Eq. (19)). Therefore, if we restrict our attention to the task to learning the Graves-Lai allocation for a subclass of models which agree on the optimal action (as we do in Theorem B.2), any lower bound for learning an allocation with normalization $\mathsf{n}_{\max}$ also applies to learning an unrestricted allocation, for large enough $\mathsf{n}_{\max}$. Finally, $\mathsf{AE}^2$ itself plays allocations with bounded normalization, which as we note in Appendix A.2, does not affect the optimality of its performance—allocations with bounded normalization are always sufficient.

### B.2. Main Result

We state two lower bounds. The first scales with the version of the Allocation-Estimation Coefficient appearing in our upper bound (Theorem A.1), but leads to a lower bound on $T^{\mathrm{gl}}$, while the second lower bound scales with the AEC for a restriction $\mathcal{M}$, but is exponentially stronger in the sense that it provides a similar lower bound on $\log(T^{\mathrm{gl}})$. We remark briefly that the regularity conditions of Appendix A.1 are not required to hold here, unless otherwise stated.

**Theorem B.1** (Main lower bound—weak variant)**.** *Let $\varepsilon > 0$, $\mathsf{n}_{\max} > 0$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given, and set $\delta := \frac{\varepsilon}{2} \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}\}$. Unless*

$$T > \frac{\delta}{8} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}),$$

*any algorithm must have, for some $M \in \mathcal{M}_0$:*

$$\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \mathsf{n}_{\max})\right] \geq \frac{\delta}{6}.$$

Stated equivalently, Theorem B.1 implies that for any $\varepsilon > 0$, if we set $\delta = c \cdot \varepsilon \cdot \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}$ for sufficiently small numerical constant $c$, then

$$T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) \geq \delta \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}(\mathcal{M}_0, \overline{M}).$$

**Remark B.1** (Choice of $\mathcal{M}_0$)**.** *Theorem B.1 is stated with respect to an arbitrary subset, $\mathcal{M}_0$, of $\mathcal{M}$. While one could simply choose $\mathcal{M}_0 \leftarrow \mathcal{M}$, in some cases it is advantageous to choose $\mathcal{M}_0 \subsetneq \mathcal{M}$. In particular, note that our lower bound scale with $\inf_{M \in \mathcal{M}_0} \mathsf{g}^M$. For classes $\mathcal{M}$ where $\inf_{M \in \mathcal{M}} \mathsf{g}^M = 0$—which will be the case, for example, if $\mathcal{M}$ corresponds to a model class where reward means form a compact set—it is advantageous to restrict $\mathcal{M}_0$ so that it corresponds only to instances with $\mathsf{g}^M > 0$.*

*We remark as well that the restriction of the lower bound to $\mathcal{M}_0$ is somewhat analogous to our upper bound Theorem A.2, which provides guarantees in terms of the AEC of a restriction of $\mathcal{M}$, $\mathcal{M}^\star$. We may therefore choose $\mathcal{M}_0 \leftarrow \mathcal{M}^\star$ to obtain lower bounds matching the scaling of Theorem A.2. In the following section we provide several examples of how $\mathcal{M}_0$ can be chosen to yield intuitive lower bounds.*

*Lastly, let us mention that $\mathcal{M}_0$ may also be chosen to yield lower bounds that have a more instance-dependent flavor. For example, given some instance $M^\star$, we could choose $\mathcal{M}_0$ to correspond to all instances identical to $M^\star$ up to a permutation of the decisions, in which case Theorem B.1 is will yield lower bounds on the performance of any algorithm on a permutation of $M^\star$, rather than over the entire class.*

**Remark B.2.** *The lower bound in Theorem B.1, for $\mathcal{M}_0 = \mathcal{M}$, scales with the quantity*

$$\sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \sup_{\xi \in \triangle_{\mathcal{M}}} \overline{\mathsf{aec}}_\varepsilon(\mathcal{M}, \xi) = \overline{\mathsf{aec}}_\varepsilon(\mathcal{M}),$$

*which at first glance might appear to be larger than the version of the AEC appearing in our upper bounds. However, there is no contradiction, because these quantities can be shown to be equivalent (up to problem-dependent parameters) under the assumptions with which Theorem A.1 is proven.*

Our second lower bound yields a lower bound on $\log(T^{\mathrm{gl}})$ as opposed to $T^{\mathrm{gl}}$—a significantly stronger result—but scales with the AEC for a restricted class. To state the result, for $\overline{M} \in \mathcal{M}^+$ and $\mathcal{M}_0 \subseteq \mathcal{M}$, define

$$\mathcal{M}_0^{\mathrm{opt}}(\overline{M}) = \left\{ M \in \mathcal{M}_0 \mid \boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}, \;\; D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) = 0 \;\; \forall \pi \in \boldsymbol{\pi}_{\overline{M}} \right\}.$$

This represents the set of models $M$ where 1) the optimal decisions for $M$ are also optimal for $\overline{M}$ and 2) playing the an optimal decision reveals no information that can distinguish $M$ and $\overline{M}$. Our second lower bound scales with the AEC for $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, and is restricted to algorithms with low regret.

**Theorem B.2** (Main lower bound—strong variant)**.** *Let $\varepsilon > 0$, $\mathsf{n}_{\max} > 0$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given, and define $\delta = \frac{\varepsilon}{2} \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}\}$. Unless*

$$\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \geq \Omega(\delta^2) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

*there is no algorithm that simultaneously ensures that*

1. $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2 \cdot \mathsf{g}^M \log(T), \;\; \forall M \in \mathcal{M}_0.$
2. $\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \mathsf{n}_{\max})\right] \leq \frac{\delta}{12}, \;\; \forall M \in \mathcal{M}_0.$

Equivalently, Theorem B.2 implies that for any $\varepsilon > 0$, if we set $\delta = c \cdot \varepsilon \cdot \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}$ for a sufficiently small numerical constant $c$, then for $R := 2 \sup_{M \in \mathcal{M}_0} \mathsf{g}^M$,

$$\log(T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \geq \frac{\delta^2 \cdot \inf_{M \in \mathcal{M}_0} \Delta^M_{\min}}{R} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}).$$

### B.3. Examples

As we discuss in the sequel, the dependence on the Allocation-Estimation Coefficient in our lower bounds (particularly Theorem B.2) qualitatively matches the dependence on the AEC in our main upper bounds, Theorems A.1 and A.2. We defer a detailed discussion comparing our upper and lower bounds for a moment, and make matters concrete by considering three examples: the informative arm example from the introduction (Example 1.1), multi-armed bandits, and tabular reinforcement learning. We defer proofs for all examples to Appendix H.4.

**Example B.1** (Searching for an Informative Arm (revisited))**.** Let $\mathcal{M}$ be the class constructed in Example 1.1 with parameters $N, A \in \mathbb{N}$ and $\beta \in (0, 1/2)$. Let $\mathcal{M}_0$ denote the restriction of $\mathcal{M}$ to models with $\Delta^M_{\min} \geq \Delta$ for some $\Delta \in (0, 1/6)$ (so that $\pi_M$ is unique), and $f^M(\pi_M) < 1$. As long as $N \geq 16$ and $\beta \log(1 + \beta A) \geq 2\Delta/(A-1)$, we have

$$\sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}) \geq \frac{N}{4\beta}. \tag{31}$$

For this construction, we have $\Delta^M_{\min} \geq \Delta$ and $\Omega(1) \leq \mathsf{g}^M \leq O(\beta^{-1})$ for all $M \in \mathcal{M}_0$. Thus, Theorem B.2 implies that any algorithm with $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for all $M \in \mathcal{M}_0$ must fail to learn an $\varepsilon$-optimal allocation with probability $\delta = \Omega(\varepsilon/\mathsf{n}_{\max})$ unless

$$\sup_{M \in \mathcal{M}} \frac{\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T) \gtrsim \frac{\varepsilon^2}{\mathsf{n}^2_{\max}} \cdot \frac{N}{\beta},$$

which implies that $\log(T^{\mathrm{gl}}(\mathcal{M}_0; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \varepsilon^2/\mathsf{n}^2_{\max} \cdot N$ for $R = 2 \sup_{M \in \mathcal{M}} \mathsf{g}^M = O(\beta^{-1})$. Any Graves-Lai allocation need only take at most $O(\beta^{-1})$ pulls to eliminate alternative instances, so an appropriate choice of $\mathsf{n}_{\max}$ is $O(\beta^{-1})$, yielding $\log(T^{\mathrm{gl}}(\mathcal{M}_0; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \beta^2\varepsilon^2 \cdot N$. While the dependence on the parameter $\mathsf{n}_{\max}, \varepsilon, \Delta > 0$ here is certainly loose, this formalizes the intuition sketched in the introduction: any algorithm that learns an optimal allocation must explore $\Omega(N)$ times, yet any algorithm that achieves near-instance-optimal regret $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T) \lesssim \beta^{-1} \log(T)$ can play a sub-optimal decision no more than roughly $\beta^{-1} \log(T)/\Delta$ times, leading to the constraint that

$$\log(T) \gtrsim N.$$

We can also apply Theorem B.1 to show that any algorithm must fail to learn an $\varepsilon$-optimal allocation with probability $\delta = \Omega(\varepsilon/\mathsf{n}_{\max})$ unless $T \gtrsim \varepsilon/\mathsf{n}_{\max} \cdot \frac{N}{\beta}$. ◁

**Example B.2** (Finite-Armed Bandit)**.** Let $A \geq 6$ and $\Delta \in (0, 1/2)$ be given and set $\Pi = [A]$. Let $\mathcal{M}$ be the set of all multi-armed bandit instances with Gaussian noise:

$$\mathcal{M} = \left\{ M(\pi) = \mathcal{N}(f^M(\pi), 1/2) \mid f^M \in [0,1]^A \right\} \tag{32}$$

and let $\mathcal{M}_0 = \{M \in \mathcal{M} \ : \ |\boldsymbol{\pi}_M| = 1, \Delta^M(\pi) \in [\Delta, 2\Delta] \text{ for } \pi \neq \pi_M, f^M(\pi_M) < 1\}$ denote the set of all bandit instances where suboptimal arms have gaps on order $\Delta$. Then for all $\varepsilon \in (0, 1/32)$,

$$\sup_{\overline{M} \in \mathcal{M}} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}) \geq c \cdot \varepsilon^{-2} \cdot \frac{A}{\Delta^2}, \tag{33}$$

where $c > 0$ is an absolute constant.

It can be shown that, by the construction of $\mathcal{M}_0$, $\mathsf{g}^M = \Theta(A/\Delta)$ for all $M \in \mathcal{M}_0$, so $\delta \propto \varepsilon \cdot \min\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\}$. Theorem B.1 then implies that any algorithm must fail to learn an $\varepsilon$-optimal allocation with probability $\delta = \Omega(\varepsilon \cdot \min\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\})$ unless

$$T \gtrsim \min\left\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\right\} \cdot \frac{A}{\varepsilon \Delta^2}.$$

Any Graves-Lai allocation need only take $O(\frac{A}{\Delta^2})$ pulls to eliminate all alternate instances, so a reasonable choice of $\mathsf{n}_{\max}$ is therefore $O(\frac{A}{\Delta^2})$. With this choice of $\mathsf{n}_{\max}$, we have $T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) \gtrsim \frac{A}{\varepsilon \Delta}$. We remark that the scaling on all parameters here is natural. Intuitively, we would expect that we need to pull each arm at least once to learn a near-optimal allocation, yielding an $\Omega(A)$ scaling. In addition, note that for multi-armed bandits, the optimal allocation places mass $\propto \frac{1}{\Delta^2}$ on arms with gap $\Delta$. To correctly estimate this proportion requires an accurate estimate of $\Delta$, which becomes increasingly difficult as $\Delta$ becomes smaller, yielding an $\Omega(\frac{1}{\Delta})$ scaling. Finally, as we decrease $\varepsilon$, we require that the returned allocation becomes closer to a truly optimal allocation, and we would therefore expect an $\Omega(\frac{1}{\varepsilon})$ scaling.

Note that the result derived by applying Theorem B.1 above only gives a lower bound on $T$. To obtain a lower bound on $\log(T)$, we combine Theorem B.2 and Eq. (33) with $\mathsf{n}_{\max} = O(\frac{A}{\Delta^2})$ as above, which implies that any algorithm with $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for all $M \in \mathcal{M}_0$ must fail to learn an $\varepsilon$-optimal allocation with probability $\delta = \Omega(\varepsilon \cdot \min\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\})$ unless

$$\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \gtrsim A,$$

or equivalently $\log(T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \frac{A}{\sup_{M \in \mathcal{M}_0} \mathsf{g}^M / \Delta_{\min}^M}$ for $R = 2 \sup_{M \in \mathcal{M}} \mathsf{g}^M$. To see why such scaling is natural, note that any algorithm which has $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for each instance $M$ can afford to explore (that is, play a suboptimal decision) at most $2\frac{\mathsf{g}^M}{\Delta_{\min}^M} \log(T)$ times, or their regret could exceed $2\mathsf{g}^M \log(T)$. However, no algorithm has any hope of learning an optimal allocation unless they play every arm at least once, so taking at least $A$ pulls of suboptimal arms seems unavoidable, and we therefore would expect that we must have $2\frac{\mathsf{g}^M}{\Delta_{\min}^M} \log(T) \gtrsim A$, which is precisely the necessary scaling shown here.

We make two remarks on this $\log(T)$ lower bound. First, note that due to the presence of the $\delta^2$ term in Eq. (33) (which we believe to be loose), the lower bound we derive by applying Theorem B.2 does not scale with the parameter $\varepsilon^{-1}$, as one might hope. Second, as noted, for $M \in \mathcal{M}_0$ for $\mathcal{M}_0$ chosen as in Example B.2, we have $\mathsf{g}^M = \Omega(A/\Delta)$, in which cases the dependence on $A$ cancels, and the lower bound becomes trivial. Note that this is somewhat to be expected. Theorem B.2 will give a trivial lower bound whenever $\sup_{M \in \mathcal{M}_0} \mathsf{g}^M$ is much larger than the AEC. Recall that in our

upper bound, Theorem A.2, the leading order $\mathsf{g}^\star \cdot \log T$ term will dominate the lower-order terms once

$$\mathsf{g}^\star \cdot \log T \gtrsim \Omega(\mathsf{aec}_\varepsilon(\mathcal{M})).$$

Therefore, if $\mathsf{g}^\star$ is much larger than the AEC, Theorem A.2 simply gives an upper bound scaling as approximately $\mathsf{g}^\star \cdot \log T$, as long as $\log T = \Omega(1)$. The interpretation in this setting is that the complexity of learning the Graves-Lai allocation is dominated by the regret incurred by *playing* the Graves-Lai allocation, which we know is necessary from Proposition 1.1, and therefore we would not expect lower-order terms to be significant components of the regret, as reflected by Theorem A.2. However, as we have already shown, In settings such as Example B.1 where this is not the case and the AEC is much larger than $\mathsf{g}^\star$, Theorem B.2 will give a non-trivial lower bound which reflects the difficulty of learning the optimal allocation. ◁

**Example B.3** (Tabular Reinforcement Learning)**.** Let $S, A, H \in \mathbb{N}$ and $\Delta \in (0, 1/2)$ be given, and assume that $SA \geq 24$ and $H \geq \log_2(S/2)$. Let $\mathcal{M}$ be the set of all tabular MDPs with 1) $|\mathcal{S}| = S$, $|\mathcal{A}| = A$ and horizon $H$, and 2) Gaussian rewards with variance $\sigma^2 = 1/2$ (cf. Appendix A.7). Let $\mathcal{M}_0$ be the result of restricting $\mathcal{M}$ in the same fashion as Example B.2. Then for all $\varepsilon \in (0, 1/32)$,

$$\sup_{\overline{M} \in \mathcal{M}} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}) \geq c \cdot \varepsilon^{-2} \cdot \frac{SA}{\Delta^2}, \tag{34}$$

where $c > 0$ is an absolute constant.

It can be shown that, by the construction of $\mathcal{M}_0$, $\mathsf{g}^M \geq \Omega(SA/\Delta)$ for all $M \in \mathcal{M}_0$. Thus, analogous to the multi-armed bandit example, Theorem B.1 and Eq. (34) imply that any algorithm must fail to learn an $\varepsilon$-optimal allocation with probability $\delta = \Omega(\varepsilon \cdot \min\{1, \frac{SA}{\Delta \mathsf{n}_{\max}}\})$ unless

$$T \gtrsim \min\left\{1, \frac{SA}{\Delta \mathsf{n}_{\max}}\right\} \cdot \frac{SA}{\varepsilon \Delta^2}.$$

Choosing $\mathsf{n}_{\max} = O(\frac{SA}{\Delta^2})$ gives $T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) \gtrsim \frac{SA}{\varepsilon \Delta}$. With this same choice of $\mathsf{n}_{\max}$, Theorem B.2 implies that any algorithm with $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for all $M \in \mathcal{M}_0$ must fail to learn an $\varepsilon$-optimal allocation with probability $\delta = \Omega(\varepsilon\Delta)$ unless

$$\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \gtrsim SA,$$

or equivalently $\log(T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \frac{SA}{\sup_{M \in \mathcal{M}_0} \mathsf{g}^M / \Delta_{\min}^M}$ for $R = 2 \sup_{M \in \mathcal{M}_0} \mathsf{g}^M$.

◁

### B.4. Discussion and Interpretation

Our lower bounds show that the Allocation-Estimation Coefficient serves as a fundamental limit on the sample complexity required to learn an approximate Graves-Lai allocation. In particular, they capture phenomena such as the necessity of searching for an informative arm in Example 1.1 that are missed by purely asymptotic analyses. To the best of our knowledge, our lower bounds represent the first attempt to systematically understand the sample complexity of learning the Graves-Lai allocation in a general decision making framework. As such, they are somewhat coarse (in particular, the dependence on parameters such as $\varepsilon$, $\mathsf{n}_{\max}$, and $\Delta_{\min}^M$ is almost certainly loose), and they are best thought of as a starting point for further research. In what follows, we provide additional interpretation of the results, and highlight some of the most interesting remaining questions.

**Regret versus learning the optimal allocation.** Theorems B.1 and B.2 lower bound the sample complexity required to learn an $\varepsilon$-optimal Graves-Lai allocation. Intuitively, this task is closely related to achieving instance-optimal regret. Our analysis of $\mathsf{AE}^2$ shows that it is *sufficient*, and many prior works aim to directly estimate the optimal allocation as well. However, it is unclear to what extent learning the optimal allocation is *necessary* to achieve instance-optimal regret.

In more detail, it is quite straightforward to show that if an algorithm achieves instance-optimal regret, its empirical frequencies act as an optimal allocation *in expectation*.

**Lemma B.1.** *Let $\varepsilon \in (0, 2)$, and suppose that Assumption A.4 holds. Fix $T \in \mathbb{N}$ and consider an algorithm $\mathbb{A}$ such that for all $M \in \mathcal{M}$,*

$$\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^M \cdot \log(T).$$

*For each $M \in \mathcal{M}$, define $\eta^M \in \mathbb{R}_+^{\Pi}$ via $\eta^M(\pi) = \mathbb{E}^{M, \mathbb{A}}\left[\frac{T(\pi)}{\log(T)}\right]$, where $T(\pi)$ denotes the number of pulls of decision $\pi$, and define $\lambda^M = \eta^M / \|\eta^M\|_1$. Then if*

$$\log(T) \ge \frac{6}{\varepsilon} \log\left( \sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \right),$$

*we have that for all $M \in \mathcal{M}$,*

$$\lambda^M \in \Lambda(M; \varepsilon). \tag{35}$$

This result gives a guarantee on the expected frequencies of any instance-optimal algorithm, but does not give any guarantee for the realized frequencies. As such, without further assumptions on the algorithm under consideration, it is unclear whether instance-optimal regret implies that it is possible to learn an optimal allocation with high or even *constant* probability. We cannot currently rule out the existence of pathological algorithms for which $\eta^M$ is optimal in expectation, yet the empirical arm frequencies deviate from the mean with moderate probability. Nonetheless, if one is willing to make stronger assumptions on the algorithm under consideration—in particular, that the *second moment* of regret is controlled—then it is possible to derive lower bounds on regret directly.

**Theorem B.3** (Simplified version of Theorem H.3)**.** *Let the time horizon $T \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$ be given, and suppose that Assumptions A.2 and A.4 hold. Suppose there exists an algorithm $\mathbb{A}$ with the property that for all $M \in \mathcal{M}$: 1) $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^M \log(T)$, 2) $\sqrt{\mathbb{E}^{M, \mathbb{A}}[(\mathbf{Reg}(T))^2]} \le 2\mathsf{g}^M \log(T)$, and 3) for all $\pi \in \Pi$, if $\mathbb{E}^{M, \mathbb{A}}[T(\pi)] \neq 0$, then $\mathbb{E}^{M, \mathbb{A}}[T(\pi)] \ge 1$. Then if we define $\delta = \varepsilon \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{3\mathsf{g}^M/\Delta_{\min}^M + \mathsf{n}_\varepsilon^{\mathcal{M}}}\}$, it must be the case that*

$$\log^3(T) \ge \frac{\delta^2}{C} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{4\varepsilon}^{\mathcal{M}}(\mathcal{M}^{\mathrm{opt}}(\overline{M}), \overline{M}).$$

*for $C \le O\Big((\sup_{M \in \mathcal{M}} \frac{\mathsf{g}^M}{\Delta_{\min}^M})^4 \cdot \frac{V_{\mathcal{M}}^2 \log(\delta^{-1})}{\varepsilon^2}\Big)$.*

See Appendix H.5 for a full statement and details. The idea behind the proof is to 1) show (via robust mean estimation) that any instance-optimal algorithm with well-behaved tails can be used to estimate the optimal allocation with high probability (with a small blowup in time horizon), and then 2) appeal to Theorem B.2. More work is required to understand whether 1) we can prove lower bounds on regret directly, and 2) whether it is possible to show that low regret and learning the optimal allocation are equivalent in a stronger sense.

**Comparing upper and lower bounds.** Keeping the differences between regret minimization and learning the optimal allocation in mind, let us highlight that the lower bound on $T$ provided by Theorem B.2 seems to qualitatively match the upper bound from Theorems A.1 and A.2. In particular, ignoring problem-dependent parameters and polylogarithmic factors, the upper bound Theorem A.1 scales, for every model $M \in \mathcal{M}$, as

$$\mathbb{E}^M[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^M \log(T) + \widetilde{O}^+(\mathsf{aec}_\varepsilon(\mathcal{M})).$$

In order for this bound to simplify to, say,

$$\mathbb{E}^M[\mathbf{Reg}(T)] \leq (1+2\varepsilon)\mathsf{g}^M \log(T),$$

we need

$$\mathsf{g}^M \cdot \log(T) \geq \widetilde{\Omega}^+(1) \cdot \frac{\mathsf{aec}_\varepsilon(\mathcal{M})}{\varepsilon},$$

which has similar scaling to the lower bound

$$\log(T) \gtrsim \widetilde{\Omega}^+(1) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})$$

from Theorem B.2. As discussed in the prequel, the former result is concerned with regret, while the latter considers the task of learning the optimal allocation, but the scaling $\log(T) \gtrsim \mathsf{aec}_\varepsilon(\mathcal{M})$ seems to be fundamental for both. Of course, beyond the gap between regret and learning the allocation, there is still much room to improve the dependence on problem-dependent parameters in both results.

**Comparing Theorem B.1 and Theorem B.2.** Theorem B.1 and Theorem B.2 exhibit an interesting dichotomy: Theorem B.1 places no constraints on the regret of the algorithm under consideration, and gives a lower of the form $T \gtrsim \widetilde{\Omega}^+(1) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0, \overline{M})$, while Theorem B.2 gives a lower bound of the form $\log(T) \gtrsim \widetilde{\Omega}^+(1) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})$, or equivalently $T \gtrsim \exp(\widetilde{\Omega}^+(1) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}))$; the latter lower bound is exponentially stronger, with the caveat that 1) the class $\mathcal{M}_0$ is replaced with the subclass $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$ and 2) Theorem B.2 assumes that the algorithm achieve nearly-instance optimal regret for every model in $\mathcal{M}_0$ ($\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}] \leq 2 \cdot \mathsf{g}^M \log(T)$). In what follows, we argue that this tradeoff is fundamental.

- First, let us consider the role of the assumption $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}] \leq 2 \cdot \mathsf{g}^M \log(T)$. Without this assumption, the lower bound from Theorem B.1 is qualitatively tight: if the algorithm explores optimally for every round $t \in [T]$, it gains roughly $(\sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0, \overline{M}))^{-1}$ units of information per round, which is sufficient to identify an optimal allocation as soon as $T \gtrsim \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0, \overline{M})$.
- On the other hand, if the require that $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}] \leq 2 \cdot \mathsf{g}^M \log(T)$, then for each $M \in \mathcal{M}_0$, the algorithm can afford to explore (i.e., play a non-optimal action) at most $2 \cdot \frac{\mathsf{g}^M}{\Delta_{\min}^M} \log(T)$ times. This changes the "effective" time horizon for exploration to $T' = 2 \cdot \frac{\mathsf{g}^M}{\Delta_{\min}^M} \log(T)$, but only if we restrict to models for which playing an optimal decision gives no information. This is precisely what the subclass $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$ captures: models $M \in \mathcal{M}_0$ for which decisions that are optimal for $\overline{M}$ lead to no information. Combining these insights leads to the lower bound $T' \approx \frac{\mathsf{g}^M}{\Delta_{\min}^M} \log(T) \gtrsim \Omega(1) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})$ in Theorem B.2.

We remark in passing that the definition of $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, which places the constraint that $D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) = 0, \;\; \forall \pi \in \boldsymbol{\pi}_{\overline{M}}$ is somewhat coarse. We expect that both Theorem B.2 and Theorem A.1/Theorem A.2 can be improved to scale with

$$\mathcal{M}_0^{\mathrm{opt}}(\overline{M}; \alpha) = \big\{M \in \mathcal{M}_0 \mid \boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}, \;\; D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) \le \alpha^2 \;\; \forall \pi \in \boldsymbol{\pi}_{\overline{M}}\big\}$$

for $\alpha \approx 1/\sqrt{T}$; the intuition is that we get $\Omega(T)$ rounds worth of information on $\boldsymbol{\pi}_M$ for "free", which facilitates accurate estimation.

**Minimax versus instance-dependent lower bounds.** As mentioned in the prequel, our lower bounds have a (constrained) minimax flavor. Specifically, Theorem B.2 shows that if $T$ is not sufficiently large, then for any algorithm, there must exist a "worst-case" model $M \in \mathcal{M}$ for which the algorithm either 1) fails to achieve (approximately) instance-optimal regret or 2) fails to learn an $\varepsilon$-optimal Graves-Lai allocation. While this is quite different from a classical minimax analysis, and certainly is closely connected to instance-optimality, an interesting direction for future work is to develop a fully instance-dependent understanding of the complexity of learning Graves-Lai allocations.

## Appendix C. Additional Related Work

In this section, we discuss further related work not already covered in detail.

**Asymptotic guarantees for general decision making.** For the general decision making framework we consider, which allows for arbitrary model classes and subsumes structured bandits and reinforcement, the only prior works we are aware of that achieve the instance-optimal lower bound from Graves and Lai (1997) are Komiyama et al. (2015), which restricts to finite observation spaces, and Dong and Ma (2022), which restricts to finite decision spaces; these works do not provide non-asymptotic guarantees.

Many works provide purely asymptotic instance-optimality guarantees for more specialized settings, including multi-armed bandits (Lai and Robbins, 1985; Garivier et al., 2016; Lattimore, 2018; Garivier et al., 2019), linear bandits (Lattimore and Szepesvari, 2017; Hao et al., 2019, 2020), and general structured bandits (Burnetas and Katehakis, 1996; Magureanu et al., 2014; Combes et al., 2017; Van Parys and Golrezaei, 2020; Degenne et al., 2020b). Some of these works do provide non-asymptotic bounds on regret, but these results generally have lower-order terms that scale linearly in $|\Pi|$, which renders them vacuous until $\log(T) \gtrsim |\Pi|$; we consider such results to be asymptotic in spirit. Along these lines, it is worth discussing Jun and Zhang (2020), which provides non-asymptotic guarantees for structured bandits in which the lower-order terms scale with a quantity $K_\psi$ that aims to capture the number of "effective arms". While this quantity can improve over $|\Pi|$ in certain situations, it is not clear whether it is well behaved for standard classes of interest (e.g., linear bandits).

**Non-asymptotic guarantees for linear bandits.** For linear bandits, a number of recent works provide non-asymptotic instance-optimal regret bounds in which lower order terms scale only with the dimension $d$ rather than the number of decisions $|\Pi|$ (Tirinzoni et al., 2020; Kirschner et al., 2021). These results take advantage of the specialized geometric structure of the linear bandit setting (e.g., existence of optimal design) for exploration, and cannot be directly adapted to general function approximation, but our results can be viewed as generalizing these guarantees.

**Reinforcement learning.** For reinforcement learning, a number of works—mostly focusing on tabular settings or linear function approximation—provides non-asymptotic guarantees that are instance-dependent, but not necessarily instance-optimal (Simchowitz and Jamieson, 2019; Al Marjani and Proutiere, 2021; Dann et al., 2021; Al Marjani et al., 2021; Wagenmaker et al., 2022b; Wagenmaker and Jamieson, 2022; Wagenmaker and Pacchiano, 2022). For instance-optimality, the results we are aware of are the classical work of Agrawal et al. (1988) and very recent work Dong and Ma (2022), which provides asymptotic guarantees for finite-horizon tabular RL, Ok et al. (2018), which provides asymptotic guarantees for an infinite-horizon setting under ergodicity assumptions, and Tirinzoni et al. (2022) which provides PAC guarantees for deterministic MDPs (though we note that the guarantee of Tirinzoni et al. (2022) is also achieved, up to $H$ factors, by Wagenmaker and Jamieson (2022)).

**Instance-optimal PAC guarantees.** Our discussion has largely centered on regret, which is the focus of our work. For the PAC setting, where the goal is to identify the optimal decision (or a near-optimal decision) as quickly as possible, a number of recent works have employed similar techniques to derive instance-optimal algorithms for settings such as multi-armed and structured bandits (Kaufmann et al., 2016; Garivier and Kaufmann, 2016; Russo, 2016; Degenne and Koolen, 2019; Degenne et al., 2019, 2020a). While many of these works are asymptotic in nature, in specialized settings such as multi-armed bandits (Jamieson et al., 2014), linear bandits (Fiez et al., 2019; Katz-Samuels et al., 2020), and linear dynamical systems (Wagenmaker et al., 2021), recent work has shown that the optimal rates are achievable in finite-time.

**Complexity of learning the Graves-Lai allocation.** The Allocation-Estimation Coefficient aims to capture the sample complexity required to learn an $\varepsilon$-optimal Graves-Lai allocation. To the best of our knowledge, our work is the first to study the complexity of learning the allocation with general function approximation, but a small body of work has studied the complexity in simple settings such as top-$k$ bandits (Simchowitz et al., 2017; Chen et al., 2017), and graph bandits (Marinov et al., 2022a,b).

**Minimax regret.** While the focus of this work has been on instance-optimality, a large body of work exists on *minimax* optimality, where the goal is to perform optimally on the *hardest* instance within a class. This line of work has established worst-case optimal (or nearly optimal) rates in settings such as multi-armed bandits (Auer et al., 2002; Audibert and Bubeck, 2009), linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011), tabular reinforcement learning (Dann et al., 2019; Zhang et al., 2021), and reinforcement learning with function approximation (Zhou et al., 2021; Du et al., 2021). The recent line of work Foster et al. (2021, 2022b, 2023) shows that, in the interactive decision-making setting considered in this work, the minimax-optimal rates are governed by a quantity known as the *Decision-Estimation Coefficient*. While our work takes inspiration and bears some similarity with this work, we remark that the techniques necessary to establish instance-optimality are significantly more intricate. It is also worth stating that it is always possible to bound the DEC by the AEC; the converse, however, is not true.

# Part II

# Proofs

## Appendix D. Additional Notation

| Mathematical Notation | Definition |
|---|---|
| $D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$ | KL divergence |
| $D_{\mathsf{H}}(\cdot, \cdot)$ | Hellinger distance |
| $D_{\mathsf{TV}}(\cdot, \cdot)$ | Total variation distance |
| $D(\cdot \,\|\, \cdot)$ | General divergence |
| $\triangle_{\mathcal{X}}$ | Set of probability distributions over $\mathcal{X}$ |
| **DMSO Notation** | |
| $M$ | Model |
| $M^\star$ | Ground truth model |
| $\mathcal{M}$ | Set of models |
| $\mathcal{M}_0$ | Arbitrary subset of $\mathcal{M}$ |
| $\pi, \Pi$ | Decision $\pi$, set of all decisions $\Pi$ |
| $r, \mathcal{R}$ | Reward $r$, set of all rewards $\mathcal{R}$ |
| $o, \mathcal{O}$ | Observation $o$, set of all observations $\mathcal{O}$ |
| $\mathbb{E}^{M,\pi}[\cdot], \mathbb{P}^{M,\pi}[\cdot]$ | Expectation and distribution of $(r, o) \sim M(\pi)$ |
| $\mathbb{E}^{M}[\cdot], \mathbb{P}^{M}[\cdot]$ | Expectation and distribution induced over histories on $M$ |
| $f^M(\pi)$ | Expected reward playing $\pi$ on $M$, $f^M(\pi) = \mathbb{E}^{M,\pi}[r]$ |
| $\pi_M$ | Optimal decision of model $M$, $\pi_M \in \arg\max_{\pi \in \Pi} f^M(\pi)$ |
| $\boldsymbol{\pi}_M$ | Set of optimal decisions of model $M$ |
| $\Delta^M(\pi)$ | Gap of decision $\pi$ on model $M$, $\Delta^M(\pi) = f^M(\pi_M) - f^M(\pi)$ |
| $\Delta^M_{\min}$ | Minimum gap on model $M$ (see Eq. (2)) |
| $\mathcal{M}^+$ | Set of all possible models, $\mathcal{M}^+ = \{M : \Pi \to \triangle_{\mathcal{R}\times\mathcal{O}} \mid f^M(\pi) \in [0,1]\}$ |
| $\mathbf{Reg}(T)$ | Regret after $T$ rounds (see Eq. (1)) |
| $L_{\mathsf{KL}}$ | Lipschitz constant of KL divergence (see Assumption A.1) |
| $V_{\mathcal{M}}$ | Sub-gaussian parameter of log-likehood ratio (see Assumption A.2) |
| $\mathsf{N}_{\mathsf{cov}}(\mathcal{M}, \rho, \mu)$ | $(\rho, \mu)$ covering number of $\mathcal{M}$ (see Definition A.1) |
| $d_{\mathsf{cov}}, C_{\mathsf{cov}}$ | Bounds on covering number (see Assumption A.3) |
| $\mathsf{n}^M_\varepsilon$ | Information content of optimal decision on $M$ with tolerance $\varepsilon$ (see Definition A.2) |
| $\mathsf{n}^{\mathcal{M}}_\varepsilon$ | Maximum information content of optimal decision on $\mathcal{M}$, $\mathsf{n}^{\mathcal{M}}_\varepsilon = \sup_{M\in\mathcal{M}} \mathsf{n}^M_\varepsilon$ |
| $\mathcal{M}_{x,y}$ | $\mathcal{M}_{x,y} = \{M \in \mathcal{M} \,:\, \Delta^M_{\min} \geq x, \mathsf{n}^M_\varepsilon \leq y\}$ |
| $\Delta_\star$ | $\Delta_\star = \min\{\Delta^\star_{\min}, 1/\mathsf{n}^\star_{\varepsilon/36}\}$ |
| $\mathcal{M}^\star$ | Restriction of $\mathcal{M}$ induced by $M^\star$ (see Eq. (24)) |

| **Graves-Lai Notation** | |
|---|---|
| $\mathsf{glc}(\mathcal{M}, M)$ | Graves-Lai Coefficient for class $\mathcal{M}$, model $M$ (see Eq. (3)) |
| $\mathsf{g}^M, \mathsf{g}^\star$ | Graves-Lai Coefficient for model $M$, $M^\star$; $\mathsf{g}^M = \mathsf{glc}(\mathcal{M}, M)$, $\mathsf{g}^\star = \mathsf{glc}(\mathcal{M}, M^\star)$ |
| $\underline{\mathsf{g}}^{\mathcal{M}}$ | Minimum non-zero Graves-Lai Coefficient on $\mathcal{M}$, $\underline{\mathsf{g}}^{\mathcal{M}} := \min_{M \in \mathcal{M}: \mathsf{g}^M > 0} \mathsf{g}^M$ |
| $\mathcal{M}^{\mathrm{alt}}(M)$ | Alternate set for model $M$, $\mathcal{M}^{\mathrm{alt}}(M) = \{M' \in \mathcal{M} \mid \boldsymbol{\pi}_M \cap \boldsymbol{\pi}_{M'} = \varnothing\}$ |
| $\mathcal{M}^\star_{\mathrm{alt}}$ | Alternate set for $M^\star$, $\mathcal{M}^\star_{\mathrm{alt}} = \mathcal{M}^{\mathrm{alt}}(M^\star)$ |
| $\eta$ | Allocation, $\eta \in \mathbb{R}^{\Pi}_+$ |
| $\lambda$ | Normalized allocation, $\lambda \in \triangle_\Pi$ |
| $\omega$ | Exploration distribution, $\omega \in \triangle_\Pi$ |
| $\Lambda(M; \varepsilon)$ | Set of $\varepsilon$-optimal normalized Graves-Lai allocations for model $M$ (see Eq. (8)) |
| $\Lambda(M; \varepsilon, \mathsf{n}_{\max})$ | Set of $\varepsilon$-optimal normalized Graves-Lai allocations for model $M$ with normalization factor at most $\mathsf{n}_{\max}$ (see Eq. (28)) |
| $\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)$ | Models for which $\lambda$ is an $\varepsilon$-optimal Graves-Lai allocation (see Eq. (9)) |
| $\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda; \mathsf{n}_{\max})$ | Models for which $\lambda$ is an $\varepsilon$-optimal Graves-Lai allocation with normalization factor at most $\mathsf{n}_{\max}$ (see Eq. (37)) |
| $I^M(\eta; \mathcal{M})$ | Information content of $\eta$ on $M$ with respect to $\mathcal{M}$ (see Eq. (36)) |
| $I^M(\eta)$ | $I^M(\eta) = I^M(\eta; \mathcal{M})$ |
| $T^{\mathrm{gl}}(\mathcal{M}_0; \varepsilon, \mathsf{n}_{\max}, \delta)$ | Minimum time to learn Graves-Lai allocation over $\mathcal{M}_0$ (see Eq. (29)) |
| **AEC Notation** | |
| $\mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}_0, \overline{M})$ | AEC with tolerance $\varepsilon$, model set $\mathcal{M}_0$, reference model $\overline{M}$ (see Eq. (10)) |
| $\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M})$ | $\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) = \mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}, \overline{M})$ |
| $\mathsf{aec}_\varepsilon(\mathcal{M})$ | $\mathsf{aec}_\varepsilon(\mathcal{M}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M})$ |
| $\overline{\mathsf{aec}}^{\mathcal{M}}_\varepsilon(\mathcal{M}_0, \xi)$ | AEC with randomized estimator $\xi$ (see Eq. (20)) |
| $\overline{\mathsf{aec}}_\varepsilon(\mathcal{M}, \xi)$ | $\overline{\mathsf{aec}}_\varepsilon(\mathcal{M}, \xi) = \overline{\mathsf{aec}}^{\mathcal{M}}_\varepsilon(\mathcal{M}, \xi)$ |
| $\overline{\mathsf{aec}}_\varepsilon(\mathcal{M})$ | $\overline{\mathsf{aec}}_\varepsilon(\mathcal{M}) = \sup_{\xi \in \triangle_{\mathcal{M}}} \overline{\mathsf{aec}}_\varepsilon(\mathcal{M}, \xi)$ |
| $\mathsf{aec}^{\mathsf{D}}_\varepsilon(\mathcal{M}, \overline{M})$ | AEC defined with respect to general divergence (see Eq. (38)) |
| $\overline{\mathsf{aec}}^{\mathsf{D}}_\varepsilon(\mathcal{M}, \xi)$ | AEC with randomized estimator, general divergence (see Eq. (39)) |
| **Uniform Exploration Notation** | |
| $C^\xi_{\mathrm{exp}}(\varepsilon)$ | Uniform exploration coefficient with respect to $\xi$ at scale $\varepsilon$ (see Definition A.4) |
| $p^\xi_{\mathrm{exp}}(\varepsilon)$ | Uniform exploration distribution with respect to $\xi$ at scale $\varepsilon$ |
| $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon)$ | Uniform exploration coefficient for class $\mathcal{M}$ at scale $\varepsilon$ |
| $C^{\mathsf{D},\xi}_{\mathrm{exp}}(\varepsilon)$ | Uniform exploration coefficient, general divergence (see Definition E.1) |
| $p^{\mathsf{D},\xi}_{\mathrm{exp}}(\varepsilon)$ | Uniform exploration distribution, general divergence |
| $C^{\mathsf{D}}_{\mathrm{exp}}(\mathcal{M}, \varepsilon)$ | Uniform exploration coefficient for class $\mathcal{M}$, general divergence |
| **Estimation Notation** | |
| $\mathbf{Alg}_{\mathsf{KL}}$ | Estimation oracle |
| $\mathbf{Est}_{\mathsf{KL}}(s)$ | Cumulative KL estimation error (see Definition A.3) |
| $\mathbf{Est}_{\mathsf{D}}(s)$ | Cumulative estimation error with respect to divergence $D$ (see Eq. (40)) |
| $\widehat{\mathbf{Est}}_{\mathsf{D}}(s)$ | Cumulative estimation error with arguments flipped (see Eq. (41)) |

**Divergences.** For probability distributions $\mathbb{P}$ and $\mathbb{Q}$ over a measurable space $(\Omega, \mathscr{F})$ with a common dominating measure, we define the total variation distance as

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

and (squared) Hellinger distance as

$$D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) = \int \left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}}\right)^2.$$

**Interactive decision making.** We formalize probability spaces in the same fashion as Foster et al. (2021, 2022b). Decisions are associated with a measurable space $(\Pi, \mathscr{P})$, rewards are associated with the space $(\mathcal{R}, \mathscr{R})$, and observations are associated with the space $(\mathcal{O}, \mathscr{O})$. The history after round $t$ is denoted by $\mathcal{H}^t = (\pi^1, r^1, o^1), \ldots, (\pi^t, r^t, o^t)$. We define

$$\Omega^t = \prod_{i=1}^{t} (\Pi \times \mathcal{R} \times \mathcal{O}), \quad \text{and} \quad \mathscr{F}^t = \bigotimes_{i=1}^{t} (\mathscr{P} \otimes \mathscr{R} \otimes \mathscr{O})$$

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

When the algorithm is clear from context, we let $\mathbb{P}^M$ denote the law it induces on $\mathcal{H}^T$ when $M : \Pi \to \triangle_{\mathcal{R} \times \mathcal{O}}$ is the underlying model, and let $\mathbb{E}^M[\cdot]$ denote the corresponding expectation. We will also overload notation somewhat and let $\mathbb{P}^{M,\pi}$ the density of $(r, o) \sim M(\pi)$.

**Notation for complexity measures and allocations.** We let $T(\pi)$ denote the number of times decision $\pi$ is taken up to time $T$. For $\eta \in \mathbb{R}_+^{\Pi}$, we define

$$I^M(\eta\,; \mathcal{M}) = \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \in \Pi} \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big), \tag{36}$$

so that we can write $\mathsf{glc}(\mathcal{M}, M^\star) = \inf_{\eta \in \mathbb{R}_+^{\Pi}} \big\{ \sum_{\pi \in \Pi} \eta(\pi) \Delta^{M^\star}(\pi) \mid I^{M^\star}(\eta; \mathcal{M}) \geq 1 \big\}$. We abbreviate $I^M(\eta) = I^M(\eta; \mathcal{M})$ whenever the class $\mathcal{M}$ is clear from context. We will occasionally overload notation and write $\Delta^M(\eta) = \sum_{\pi \in \Pi} \eta(\pi) \Delta^M(\pi)$ for $\eta \in \mathbb{R}_+^{\Pi}$. We also let $\mathcal{M}^\star_{\mathrm{alt}} := \mathcal{M}^{\mathrm{alt}}(M^\star)$ and

$$\mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda; \mathsf{n}_{\max}) := \{M \in \mathcal{M} \;:\; \lambda \in \Lambda(M; \varepsilon, \mathsf{n}_{\max})\}. \tag{37}$$

Recall the definition

$$\Lambda(M; \varepsilon) = \Bigg\{ \lambda \in \triangle_{\Pi} \mid \exists \mathsf{n} \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{\pi \sim \lambda}[\Delta^M(\pi)] \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}, \\ \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi \sim \lambda}\big[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)\big] \geq \frac{1-\varepsilon}{\mathsf{n}} \Bigg\}.$$

For a given $\lambda \in \Lambda(M; \varepsilon)$, we refer to the $\mathsf{n} \in \mathbb{R}_+$ which realizes

$$\mathbb{E}_{\pi \sim \lambda}[\Delta^M(\pi)] \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}} \quad \text{and} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi \sim \lambda}\big[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)\big] \geq \frac{1-\varepsilon}{\mathsf{n}}$$

as the *normalization factor* of $\lambda$.

**General divergences.** While in the main text we have focused on results that hold for the KL divergence, throughout the appendix we will consider more general divergences $D(\cdot \,\|\, \cdot)$. In particular, rather than fixing the divergence in Eq. (18) to the KL divergence, we utilize divergence $D$. We also perform online estimation with respect to $D$ rather than with respect to the KL divergence. While $D$ could be an arbitrary non-negative function, we make the following assumptions on it.

First, we replace Assumption A.1 with the following more general assumption.

**Assumption D.1.** *For all $M, M', M'' \in \mathcal{M}$, and $\pi \in \Pi$, there exists some $L_{\mathsf{KL}}$ such that*

$$\left|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) - D_{\mathsf{KL}}\big(M'(\pi) \,\|\, M''(\pi)\big)\right| \le L_{\mathsf{KL}}\sqrt{D\big(M(\pi) \,\|\, M'(\pi)\big)}.$$

Note that, by Jensen's inequality, Assumption D.1 immediately implies that, for $\xi \in \triangle(\mathcal{M})$,

$$\left|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) - \mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)]\right| \le L_{\mathsf{KL}}\sqrt{\mathbb{E}_{\overline{M}\sim\xi}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]}.$$

We in addition make the following assumption, which we note is met for the KL divergence.

**Assumption D.2.** *For all $M, M' \in \mathrm{co}(\mathcal{M})$, and $\pi$, we have*

$$D_{\mathsf{H}}^2\big(M(\pi), M'(\pi)\big) \le D\big(M(\pi) \,\|\, M'(\pi)\big).$$

*Furthermore, $D(\cdot \,\|\, \cdot)$ is convex in its second argument.*

A direct consequence of this assumption is that, when rewards are observed and bounded in $[0, 1]$, we have

$$|f^M(\pi) - f^{M'}(\pi)| \le \sqrt{D\big(M(\pi) \,\|\, M'(\pi)\big)}.$$

Throughout the appendix, we assume that Assumption D.1 and Assumption D.2 hold for our divergence $D$.

We also generalize the definition of the Allocation-Estimation Coefficient to account for general divergences as

$$\mathsf{aec}_\varepsilon^{\mathsf{D}}(\mathcal{M}, \overline{M}) := \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\pi\sim\omega}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]} \tag{38}$$

and

$$\overline{\mathsf{aec}}_\varepsilon^{\mathsf{D}}(\mathcal{M}; \xi) := \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim\omega}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]}. \tag{39}$$

Other variants of the AEC are generalized similarly with a superscript $\mathsf{D}$. Our guarantees will also depend on a notion of estimation error for general divergences, given by

$$\mathbf{Est}_{\mathsf{D}}(s) := \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]. \tag{40}$$

It will also be convenient to work with the following notion of estimation error:

$$\widehat{\mathbf{Est}}_{\mathsf{D}}(s) := \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]. \tag{41}$$

## Appendix E. Technical Tools

### E.1. Online Learning

In this section, we state online estimation guarantees for variants of the Tempered Aggregation algorithm of Chen et al. (2022). Throughout the section we abbreviate $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}^{t-1}, p^t, \xi^t]$. We recall that $\mathbb{P}^{M,\pi}(r, o)$ denotes the density over rewards and observations $(r, o)$ under $M(\pi)$

**Algorithm 4** Tempered Aggregation

**input:** Finite class $\mathcal{M}$.
Initialize $\xi^1 \leftarrow \mathrm{Unif}(\mathcal{M})$.
**for** $t = 1, 2, 3, \ldots$ **do**
  Receive $(\pi^t, r^t, o^t)$.
  Update estimator:

$$\xi^{t+1}(M) \propto \xi^t(M) \cdot \exp\left(\frac{1}{2}\log \mathbb{P}^{M,\pi^t}(r^t, o^t)\right) \quad \forall M \in \mathcal{M}.$$

**Proposition E.1** (Tempered Aggregation, Finite Class Setting). *Assume $|\mathcal{M}| \leq \infty$ and $M^\star \in \mathcal{M}$. Then Algorithm 4 produces estimates $(\xi^t)_{t=1}^T$ which satisfy, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \leq 2\log\frac{|\mathcal{M}|}{\delta}.$$

**Proof of Proposition E.1.** We follow closely the proof of Theorem C.1 of Chen et al. (2022). Define the random variable

$$A^t := -\log \mathbb{E}_{M\sim\xi^t}\left[\exp\left(\beta \log \frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}\right)\right].$$

We have

$$\begin{aligned}
\mathbb{E}_{t-1}[\exp(-A^t)] &= \mathbb{E}_{t-1}\left[\mathbb{E}_{M\sim\xi^t}\left[\exp\left(\frac{1}{2}\log\frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}\right)\right]\right] \\
&= \sum_{M\in\mathcal{M}} \xi^t(M)\mathbb{E}_{t-1}\left[\exp\left(\frac{1}{2}\log\frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}\right)\right] \\
&= \sum_{M\in\mathcal{M}} \xi^t(M)\mathbb{E}_{t-1}\left[\mathbb{E}_{o\sim M^\star(\pi^t)}\left[\sqrt{\frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}}\right]\right] \\
&= \sum_{M\in\mathcal{M}} \xi^t(M)\cdot\left(1 - \frac{1}{2}\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]\right)
\end{aligned}$$

where the last equality holds by the definition of the Hellinger distance. This implies that

$$1 - \mathbb{E}_{t-1}[\exp(-A^t)] = \frac{1}{2}\mathbb{E}_{M\sim\xi^t}\left[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]\right].$$

By Lemma A.4 of Foster et al. (2021), we have that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T} A^t + \log\frac{1}{\delta} &\geq \sum_{t=1}^{T} -\log \mathbb{E}_{t-1}[\exp(-A^t)] \\ &\geq \sum_{t=1}^{T} \big(1 - \mathbb{E}_{t-1}[\exp(-A^t)]\big) \\ &= \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}\left[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]\right].\end{aligned}$$

We turn to upper bounding $\sum_{t=1}^{T} A^t$. Following the proof of Theorem C.1 of Chen et al. (2022), we have

$$\sum_{t=1}^{T} A^t = -\log\left(\sum_{M\in\mathcal{M}} \xi^1(M)\exp\left(\sum_{t=1}^{T}\frac{1}{2}\log\frac{\mathbb{P}^{M,\pi^t}(r^t,o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}\right)\right).$$

Since $M^\star \in \mathcal{M}$, we can then bound

$$\sum_{t=1}^{T} A^t \leq -\log\left(\xi^1(M^\star)\exp\left(\sum_{t=1}^{T}\frac{1}{2}\log\frac{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}\right)\right) = \log|\mathcal{M}|.$$

Combining these expressions gives the result. □

**Proposition E.2** (Tempered Aggregation, Infinite Class Setting)**.** *Let $\mathcal{M}_{\mathrm{cov}}$ denote a $(\rho,\mu)$-cover of $\mathcal{M}$ with covering number $\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho,\mu)$. If we apply Algorithm 4 to $\mathcal{M}_{\mathrm{cov}}$, we have that whenever $M^\star \in \mathcal{M}$, with probability at least $1-\delta-T\mu$,*

$$\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \leq 2\log\frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho,\mu)}{\delta} + T\rho.$$

**Proof of Proposition E.2.** Defining $A^t$ as in Proposition E.1, the first part of the proof is identical to that of Proposition E.1, and we conclude that, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} A^t + \log\frac{1}{\delta} \geq \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]$$

and

$$\sum_{t=1}^{T} A^t = -\log\left(\sum_{M\in\mathcal{M}_{\mathrm{cov}}} \xi^1(M)\exp\left(\sum_{t=1}^{T}\frac{1}{2}\log\frac{\mathbb{P}^{M,\pi^t}(r^t,o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}\right)\right).$$

Let $\mathcal{E}$ denote the event associated with $\mathcal{M}_{\mathrm{cov}}$, as defined in Definition A.1, which satisfies $\sup_{M'\in\mathcal{M}}\sup_{\pi}\mathbb{P}^{M'}(\mathcal{E}\mid\pi)\leq\mu$. Let $\widetilde{M}\in\mathcal{M}_{\mathrm{cov}}$ denote the element in the cover which satisfies

$$\left|\log\frac{\mathbb{P}^{M^\star,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}\right| = \left|\log\mathbb{P}^{M^\star,\pi}(r,o) - \log\mathbb{P}^{\widetilde{M},\pi}(r,o)\right| \leq \rho,$$

for all $(r, o, \pi)$ with $\sup_{M' \in \mathcal{M}} \mathbb{P}^{M', \pi}(r, o \mid \mathcal{E}) > 0$. We then have

$$\sum_{t=1}^{T} \log \frac{\mathbb{P}^{M, \pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)} = \sum_{t=1}^{T} \left( \log \frac{\mathbb{P}^{M, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)} + \log \frac{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)} \right)$$

so we can bound

$$\sum_{t=1}^{T} A^t \le \log |\mathcal{M}_{\mathrm{cov}}| + \frac{1}{2} \sum_{t=1}^{T} \log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)}$$

which gives that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{M \sim \xi^t}[\mathbb{E}_{\pi \sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \le 2 \log |\mathcal{M}_{\mathrm{cov}}| + 2 \log \frac{1}{\delta} + \sum_{t=1}^{T} \log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)}.$$

Let $\mathcal{E}_t$ denote the event that $\mathcal{E}$ occurs at step $t$. Denote the event $\mathcal{E}_1 := \cap_{t=1}^{T} \mathcal{E}_t$ and

$$\mathcal{E}_2 := \left\{ \sum_{t=1}^{T} \mathbb{E}_{M \sim \xi^t}[\mathbb{E}_{\pi \sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \le 2 \log |\mathcal{M}_{\mathrm{cov}}| + 2 \log \frac{1}{\delta} + T\rho \right\}.$$

Then

$$\mathbb{P}^{M^\star}[\mathcal{E}_2] \le \mathbb{P}^{M^\star}[\mathcal{E}_2 \cap \mathcal{E}_1] + \mathbb{P}^{M^\star}[\mathcal{E}_1^c].$$

By definition of $\mathcal{E}_t$ and a union bound we have $\mathbb{P}^{M^\star}[\mathcal{E}_1^c] \le T\mu$. Furthermore, on the event $\mathcal{E}_1$ we can bound, for each $t \le T$,

$$\log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)} \le \rho.$$

Thus, it follows that on $\mathcal{E}_1$, $\sum_{t=1}^{T} \log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)} \le T\rho$, which implies that $\mathbb{P}^{M^\star}[\mathcal{E}_2 \cap \mathcal{E}_1] \le \delta$. The result follows.

□

## E.2. Properties of Graves-Lai Program

In this section, we establish some basic properties of the Graves-Lai coefficient $\mathsf{g}^M \equiv \mathsf{glc}(\mathcal{M}, M)$. Through out the section, we omit dependence on the class $\mathcal{M}$ for various quantities of interest whenever it is clear from context. Throughout, we will use the fact that whenever $\Delta^M_{\min} > 0$, $\pi_M$ is unique.

### E.2.1. BASIC PROPERTIES OF GRAVES-LAI PROGRAM

**Lemma E.1.** *For any $M \in \mathcal{M}$ and $\mathsf{n} > 0$, we have*

$$\frac{\mathsf{g}^M}{\mathsf{n}} \le \inf_{\lambda \in \triangle_\Pi} \left\{ \Delta^M(\lambda) : I^M(\lambda) \ge \frac{1}{\mathsf{n}} \right\}.$$

**Proof of Lemma E.1.** Assume the contrary. Then there exists some $\widetilde{\lambda}$ such that

$$\Delta^M(\widetilde{\lambda}) < \mathsf{g}^M/\mathsf{n} \quad \text{and} \quad I^M(\widetilde{\lambda}) \geq 1/\mathsf{n}.$$

However, since both $\Delta^M(\lambda)$ and $I^M(\lambda)$ are linear in rescaling of $\lambda$, this implies that

$$\Delta^M(\mathsf{n}\widetilde{\lambda}) < \mathsf{g}^M \quad \text{and} \quad I^M(\mathsf{n}\widetilde{\lambda}) \geq 1.$$

By definition we have

$$\mathsf{g}^M = \inf_{\eta \in \mathbb{R}_+^\Pi} \Delta^M(\eta) \quad \text{s.t.} \quad I^M(\eta) \geq 1.$$

This is a contradiction, so the desired conclusion follows. □

**Lemma E.2.** *For any $M \in \mathcal{M}^+$ with $\Delta^M_{\min} > 0$, we can bound*

$$\mathsf{g}^M \leq C^\xi_{\exp}(\tfrac{1}{4}(\Delta^M_{\min})^2)$$

*for $\xi = \mathbb{I}_M$.*

**Proof of Lemma E.2.** By definition we have

$$\begin{aligned}
\mathsf{g}^M &= \inf_{\eta \in \mathbb{R}_+^\Pi} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\text{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1 \\
&\leq \inf_{\eta \in \mathbb{R}_+^\Pi} \|\eta\|_1 \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\text{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1.
\end{aligned}$$

Let $\xi \in \triangle_{\mathcal{M}}$ denote the distribution with $\xi(M) = 1$. Let and let $p_{\exp} := p^\xi_{\exp}(\varepsilon)$ denote the uniform exploration distribution defined with respect to $\xi$, and $C^\xi_{\exp}(\varepsilon)$ the corresponding uniform exploration coefficient, for some $\varepsilon$ to be chosen.

Consider some $M' \in \mathcal{M}^{\text{alt}}(M)$. Since $\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)]] = \mathbb{E}_{\pi \sim p}[D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi))]$ for all $M''$ and $D_{\mathsf{KL}}(M(\pi) \,\|\, M(\pi)) = 0$ for all $\pi$, it follows from the definition of the uniform exploration coefficient that

$$\mathbb{E}_{p_{\exp}}[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] \leq 1/C^\xi_{\exp}(\varepsilon) \implies \sup_{p \in \triangle_\Pi} \mathbb{E}_p[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] \leq \varepsilon$$

or, alternatively,

$$\sup_{p \in \triangle_\Pi} \mathbb{E}_p[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] > \varepsilon \implies \mathbb{E}_{p_{\exp}}[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] > 1/C^\xi_{\exp}(\varepsilon).$$

If $M' \in \mathcal{M}^{\text{alt}}(M)$, then it follows that $\pi_M \notin \boldsymbol{\pi}_{M'}$. Take some $\pi_{M'} \in \boldsymbol{\pi}_{M'}$. By definition we have $f^M(\pi_M) \geq f^M(\pi_{M'}) + \Delta^M_{\min}$ and $f^{M'}(\pi_{M'}) \geq f^{M'}(\pi_M)$. Thus,

$$\begin{aligned}
\Delta^M_{\min} &\leq f^M(\pi_M) - f^M(\pi_{M'}) + f^{M'}(\pi_{M'}) - f^{M'}(\pi_M) \\
&\leq |f^M(\pi_M) - f^{M'}(\pi_M)| + |f^{M'}(\pi_{M'}) - f^M(\pi_{M'})| \\
&\leq \sqrt{D\big(M(\pi_M) \,\|\, M'(\pi_M)\big)} + \sqrt{D\big(M(\pi_{M'}) \,\|\, M'(\pi_{M'})\big)}.
\end{aligned}$$

This implies that there exists some $\pi$ such that $D\big(M(\pi) \,\|\, M'(\pi)\big) \geq (\Delta^M_{\min}/2)^2$, so we can lower bound

$$\sup_{p \in \triangle_\Pi} \mathbb{E}_p[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] \geq \frac{1}{4}(\Delta^M_{\min})^2.$$

Thus, setting $\varepsilon = \frac{1}{4}(\Delta^M_{\min})^2$, we have that

$$\mathbb{E}_{p_{\mathrm{exp}}}[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] > 1/C^\xi_{\mathrm{exp}}(\tfrac{1}{4}(\Delta^M_{\min})^2).$$

It follows that the allocation $\eta = C^\xi_{\mathrm{exp}}(\frac{1}{4}(\Delta^M_{\min})^2) \cdot p_{\mathrm{exp}}$ realizes

$$\inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) = \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} C^\xi_{\mathrm{exp}}(\tfrac{1}{4}(\Delta^M_{\min})^2) \cdot \mathbb{E}_{p^M_{\mathrm{exp}}}[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] \geq 1,$$

which proves the result. □

**Lemma E.3.** *Assume* $\mathsf{g}^M > 0, \Delta^M_{\min} > 0$, *and* $\mathsf{n}^M_{1/4} < \infty$. *Then it must be the case that*

$$\mathsf{g}^M \geq \Delta^M_{\min} \cdot \frac{1}{\max_{M' \in \mathcal{M}, \pi \in \Pi} D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.$$

**Proof of Lemma E.3.** Recall that

$$\mathsf{g}^M = \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1.$$

By the definition of $\mathsf{n}^M_{1/4}$, for any allocation $\eta \in \mathbb{R}^\Pi_+$ satisfying $\inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 3/4$, the allocation $\widetilde{\eta}$ defined as $\widetilde{\eta}(\pi) = \eta(\pi)$ for $\pi \neq \pi_M$, and $\widetilde{\eta}(\pi_M) = \mathsf{n}^M_{1/2}$ satisfies

$$\inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \widetilde{\eta}(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1/2,$$

and furthermore $\Delta^M(\eta) = \Delta^M(\widetilde{\eta})$. It follows that

$$\begin{aligned}
\mathsf{g}^M &\geq \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 3/4 \\
&\geq \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1/2,\ \eta(\pi_M) \leq \mathsf{n}^M_{1/4}.
\end{aligned}$$

If for all $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ we have $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$, this implies that we can distinguish $M$ from $M'$ by playing only $\pi_M$. Furthermore, by what we have just shown, this can be achieved by playing $\pi_M$ at most $2\mathsf{n}^M_{1/4}$ times. It follows that, if $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$ for all $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, then $\mathsf{g}^M = 0$. Thus, if $\mathsf{g}^M > 0$, there must exist some $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ such that $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) = 0$.

We then have

$$\mathsf{g}^M \geq \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \sum_{\pi \neq \pi_M} \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1$$

$$
\begin{aligned}
&= \inf_{\eta \in \mathbb{R}^{\Pi}_{+}, \eta(\pi_M)=0} \Delta^M(\eta) \quad \text{s.t.} \quad \sum_{\pi \neq \pi_M} \eta(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq 1 \\
&\geq \inf_{\eta \in \mathbb{R}^{\Pi}_{+}, \eta(\pi_M)=0} \Delta^M_{\min} \cdot \|\eta\|_1 \quad \text{s.t.} \quad \max_{\pi} D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \cdot \|\eta\|_1 \geq 1 \\
&= \Delta^M_{\min} \cdot \frac{1}{\max_\pi D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.
\end{aligned}
$$

The result follows. □

### E.2.2. PROPERTIES OF THE INFORMATION CONTENT OF OPTIMAL DECISIONS

**Lemma E.4.** *Let $\varepsilon \in [0, 1/2)$ and $\bar{\mathsf{n}} > 0$ be given. We can bound, for any function $g(\omega, M)$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ with $\inf_{M \in \mathcal{M}_0} \Delta^M_{\min} > 0$,*

$$
\inf_{\omega, \lambda \in \triangle_\Pi} \sup_{M \in \mathcal{M}_0 \setminus \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda; \bar{\mathsf{n}})} g(\omega, M) \leq \inf_{\omega, \lambda \in \triangle_\Pi} \sup_{M \in \mathcal{M}_0 \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)} g(\omega, M)
$$

*as long as*

$$
\bar{\mathsf{n}} \geq \max_{M \in \mathcal{M}_0} \max\left\{ \mathsf{n}^M_\varepsilon, \frac{4\mathsf{g}^M}{\Delta^M_{\min}}, \frac{2\mathsf{g}^M}{\zeta \Delta^M_{\min}} \right\},
$$

*where*

$$
\zeta := \min_{M \in \mathcal{M}_0 : \mathsf{g}^M > 0} \min\left\{ \frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}^M_\varepsilon}, \frac{\Delta^M_{\min} \varepsilon}{4} \right\}.
$$

**Proof of Lemma E.4.** Note that each of these expressions only depend on $\lambda$ through $\mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda; \bar{\mathsf{n}})$ and $\mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)$, respectively. To prove the result, it therefore suffices to show that, for every $\lambda \in \triangle_\Pi$, there exists $\lambda' \in \triangle_\Pi$ such that $\mathcal{M}_0 \cap \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda) \subseteq \mathcal{M}_0 \cap \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda'; \bar{\mathsf{n}})$.

Fix $\lambda \in \triangle_\Pi$. Consider $M \in \mathcal{M}_0 \cap \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)$. By definition we have that there exists some $\mathsf{n} > 0$ such that

$$
\Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \quad \text{and} \quad I^M(\lambda) \geq (1-\varepsilon)/\mathsf{n},
$$

where here $I^M(\lambda) = I^M(\lambda; \mathcal{M})$. We consider two cases.

**Case 1: $\lambda = \mathbb{I}_\pi$ for some $\pi \in \Pi$.** First, suppose that $\pi \neq \pi_M$. Then we have

$$
\Delta^M_{\min} \leq \Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies \mathsf{n} \leq (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}.
$$

It follows that as long as $\bar{\mathsf{n}} \geq (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}$, then $M \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda; \bar{\mathsf{n}})$.

Now, suppose that $\pi = \pi_M$. In this case, by the definition of $\mathsf{n}^M_\varepsilon$, we immediately have that it suffices to take $\mathsf{n} = \mathsf{n}^M_\varepsilon$. Thus, for $\lambda = \mathbb{I}_\pi$, we have $\mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda) = \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda; \bar{\mathsf{n}})$ as long as

$$
\bar{\mathsf{n}} \geq \max\{\mathsf{n}^M_\varepsilon, (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}\}.
$$

**Case 2a: $\lambda \neq \mathbb{I}_\pi$ for any $\pi \in \Pi$.** Fix some $\zeta \in (0, 1/2)$ to be chosen. Suppose that there exists some $\pi'$ such that $\lambda(\pi') \geq 1 - \zeta$, and note that there can exist at most one such $\pi'$. Define $\lambda'$ as

$$\lambda'(\pi') = 1 - \zeta, \quad \lambda'(\pi) = \frac{\zeta}{1 - \lambda(\pi')}\lambda(\pi), \quad \forall \pi \neq \pi'$$

and note that $\lambda' \in \triangle_\Pi$. Our goal is to show that $\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda) \subseteq \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda'; \bar{\mathsf{n}})$.

Suppose that $\pi_M = \pi'$. Then

$$\Delta^M(\lambda') = \frac{\zeta}{1 - \lambda(\pi')} \cdot \Delta^M(\lambda) \leq \frac{\zeta}{1 - \lambda(\pi')} \cdot \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}.$$

Denote $\mathsf{n}' := (\frac{\zeta}{1-\lambda(\pi')} \cdot \frac{1}{\mathsf{n}})^{-1}$. Since $I^M(\lambda) \geq (1-\varepsilon)/\mathsf{n}$, we have $I^M(\mathsf{n}\lambda) \geq 1 - \varepsilon$. Then, by the definition of $\mathsf{n}^M_\varepsilon$, we have

$$\inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \neq \pi_M} \mathsf{n}\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) + \mathsf{n}^M_\varepsilon D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \geq 1 - 2\varepsilon.$$

However, note that

$$\begin{aligned}
I^M(\mathsf{n}'\lambda') &= \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \neq \pi_M} \frac{\zeta \mathsf{n}'}{1 - \lambda(\pi_M)} \lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) + \mathsf{n}'(1-\zeta) D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \\
&= \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \neq \pi_M} \mathsf{n}\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) + \frac{(1-\zeta)(1-\lambda(\pi'))\mathsf{n}}{\zeta} \cdot D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \\
&\overset{(a)}{\geq} \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \neq \pi_M} \mathsf{n}\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) + \mathsf{n}^M_\varepsilon D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \\
&\overset{(b)}{\geq} 1 - 2\varepsilon
\end{aligned}$$

where $(a)$ follows as long as

$$\frac{(1-\zeta)(1-\lambda(\pi'))\mathsf{n}}{\zeta} \geq \mathsf{n}^M_\varepsilon, \tag{42}$$

and $(b)$ follows from what we have just shown. Rearranging, we have that Eq. (42) is equivalent to

$$\frac{(1-\lambda(\pi'))\mathsf{n}}{(1-\lambda(\pi'))\mathsf{n} + \mathsf{n}^M_\varepsilon} \geq \zeta.$$

Note that by Lemma E.1 and the definition of $\lambda$, we have

$$\frac{(1-\varepsilon)\mathsf{g}^M}{\mathsf{n}} \leq \inf_{\widetilde{\lambda} \in \triangle_\Pi} \left\{ \Delta^M(\widetilde{\lambda}) : I^M(\widetilde{\lambda}) \geq \frac{1-\varepsilon}{\mathsf{n}} \right\} \leq \Delta^M(\lambda),$$

which implies that

$$(1-\varepsilon)\mathsf{g}^M \leq \Delta^M(\lambda) \cdot \mathsf{n} \leq (1 - \lambda(\pi_M)) \cdot \mathsf{n}.$$

As the function $\frac{x}{x+\mathsf{n}\varepsilon^M}$ is increasing in $x$, a sufficient choice of $\zeta$ for this $M$ is then

$$\min\{\frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}\varepsilon^M}, 3/8\} \geq \zeta$$

Thus, for such a $\zeta$, we have that $M \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda'; \mathsf{n}')$, which implies that $M \in \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda'; \mathsf{n}')$. Note that $(1-\lambda(\pi'))\mathsf{n} \leq (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}$ in this case, so we have that $M \in \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda'; \bar{\mathsf{n}})$ as long as

$$\bar{\mathsf{n}} \geq \frac{2\mathsf{g}^M}{\zeta\Delta^M_{\min}}.$$

Consider now the case where $\pi_M \neq \pi'$. In this case, defining $\lambda'$ as before, we can bound

$$\Delta^M(\lambda') \leq \zeta + (1-\zeta)\Delta^M(\pi') \leq \zeta + \lambda(\pi')\Delta^M(\pi') \leq \zeta + \Delta^M(\lambda) \leq \zeta + (1+\varepsilon)\mathsf{g}^M/\mathsf{n}.$$

Furthermore,

$$\begin{aligned}
I^M(\lambda') &= \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \neq \pi'} \frac{\zeta}{1-\lambda(\pi')}\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) + (1-\zeta) D_{\mathsf{KL}}\big(M(\pi') \,\|\, M'(\pi')\big) \\
&\geq \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \neq \pi'} (1-\zeta)\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) + (1-\zeta)\lambda(\pi') D_{\mathsf{KL}}\big(M(\pi') \,\|\, M'(\pi')\big) \\
&= (1-\zeta) I^M(\lambda) \\
&\geq (1-\zeta)(1-\varepsilon)/\mathsf{n}.
\end{aligned}$$

Since $\pi' \neq \pi_M$, we can lower bound $\Delta^M(\lambda) \geq (1-\zeta)\Delta^M_{\min}$, so

$$(1-\zeta)\Delta^M_{\min} \leq \Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies \mathsf{n} \leq \frac{(1+\varepsilon)\mathsf{g}^M}{(1-\zeta)\Delta^M_{\min}} \leq \frac{4\mathsf{g}^M}{\Delta^M_{\min}}$$

We can therefore bound $\zeta \leq \varepsilon\mathsf{g}^M/\mathsf{n}$ as long as

$$\zeta \leq \frac{\Delta^M_{\min} \cdot \varepsilon}{4}.$$

Consider $\zeta$ that satisfies this inequality. Then $\Delta^M(\lambda') \leq (1+2\varepsilon)\mathsf{g}^M/\mathsf{n}$. We can also lower bound $(1-\zeta)(1-\varepsilon) \geq (1-2\varepsilon)$ as long as $\zeta \leq \frac{\varepsilon}{1-\varepsilon} \leq \varepsilon$. Thus, as long as

$$\zeta \leq \min\{1, \frac{\Delta^M_{\min}}{4}\} \cdot \varepsilon \quad \text{and} \quad \bar{\mathsf{n}} \geq \frac{4\mathsf{g}^M}{\Delta^M_{\min}},$$

we have that $M \in \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda'; \bar{\mathsf{n}})$.

**Case 2b: $\lambda \neq \mathbb{I}_\pi$ for any $\pi \in \Pi$.** Finally, it remains to handle the case then there does not exist $\pi'$ such that $\lambda(\pi') \geq 1-\zeta$. In this case, we can always lower bound

$$\zeta\Delta^M_{\min} \leq \Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies \mathsf{n} \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\zeta\Delta^M_{\min}},$$

so as long as $\bar{\mathsf{n}} \geq \frac{(1+\varepsilon)\mathsf{g}^M}{\zeta\Delta^M_{\min}}$, we have $M \in \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda; \bar{\mathsf{n}})$.

Since we are in the regime where $\lambda \neq \mathbb{I}_\pi$, it must be the case that if $M \in \mathcal{M}_0 \cap \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)$, then $\mathsf{g}^M > 0$. Thus, a sufficient choice of $\zeta$ is

$$\zeta = \min_{M \in \mathcal{M}_0 : \mathsf{g}^M > 0} \min\left\{ \frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}\varepsilon^M}, \frac{\Delta^M_{\min}\varepsilon}{4} \right\}.$$

**Concluding the Proof.** To show the result, we need that $\bar{\mathsf{n}}$ is large enough for each $M \in \mathcal{M}_0$. The argument above shows that it suffices to take

$$\bar{\mathsf{n}} \geq \max_{M \in \mathcal{M}_0} \max\left\{\mathsf{n}_\varepsilon^M, \frac{4\mathsf{g}^M}{\Delta_{\min}^M}, \frac{2\mathsf{g}^M}{\zeta \Delta_{\min}^M}\right\}$$

for

$$\zeta := \min_{M \in \mathcal{M}_0 : \mathsf{g}^M > 0} \min\left\{\frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}_\varepsilon^M}, \frac{\Delta_{\min}^M \varepsilon}{4}\right\}.$$

This proves the result.

□

**Lemma E.5.** *For every $M \in \mathcal{M}$ with $\Delta_{\min}^M > 0$, there exists some $\lambda \in \Lambda(M; \varepsilon)$ with normalization factor $\mathsf{n}$ satisfying*

$$\mathsf{n} \leq \mathsf{g}^M / \Delta_{\min}^M + \mathsf{n}_\varepsilon^M,$$

*i.e., we have $M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda; \mathsf{n})$.*

**Proof of Lemma E.5.** Consider some allocation $\eta \in \mathbb{R}_+^\Pi$ such that

$$\Delta^M(\eta) \leq \mathsf{g}^M \quad \text{and} \quad I^M(\eta) \geq 1. \tag{43}$$

Let $\eta'$ denote the allocation satisfying $\eta'(\pi) = \eta(\pi)$ for $\pi \neq \pi_M$, and $\eta'(\pi_M) = 0$. Note that $\Delta^M(\eta) \geq \Delta_{\min}^M \|\eta'\|_1$, which implies that

$$\|\eta'\|_1 \leq \mathsf{g}^M / \Delta_{\min}^M.$$

Let $\eta''$ denote the allocation satisfying $\eta''(\pi) = \eta(\pi)$ for $\pi \neq \pi_M$ and $\eta''(\pi_M) = \mathsf{n}_\varepsilon^M$. Then by definition of $\mathsf{n}_\varepsilon^M$, and since $\eta$ satisfies Eq. (43), we have $I^M(\eta'') \geq 1 - \varepsilon$. Furthermore, it is straightforward to see that $\Delta^M(\eta'') \leq \mathsf{g}^M$. This implies that $\eta''/\|\eta''\|_1 \in \Lambda(M; \varepsilon)$ with normalization factor $\|\eta''\|_1$. However, we can bound

$$\|\eta''\|_1 = \|\eta'\|_1 + \mathsf{n}_\varepsilon^M \leq \mathsf{g}^M / \Delta_{\min}^M + \mathsf{n}_\varepsilon^M.$$

This proves the result. □

### E.2.3. BOUNDING ALLOCATION-ESTIMATION COEFFICIENT VIA UNIFORM EXPLORATION COEFFICIENT

In this section we prove a generalized version of Proposition A.1. In particular, rather than specializing to the KL divergence, we consider a general divergence $D$. We define the uniform exploration coefficient with respect to $D$ as follows.

**Definition E.1** (Uniform Exploration Coefficient, General Divergences)**.** *For a randomized estimator $\xi \in \triangle_{\mathcal{M}}$ and divergence $D(\cdot \,\|\, \cdot)$, we define the uniform exploration coefficient with respect to $\xi$ at scale $\varepsilon > 0$ as the value of the following program:*

$$C_{\mathrm{exp}}^{\mathsf{D},\xi}(\varepsilon) := \min_{C \in \mathbb{R}_+, p \in \triangle_\Pi} \left\{ C \;\middle|\; \forall M, M' \in \mathcal{M} \;:\; \begin{array}{c} \max_{M'' \in \{M, M'\}} \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p}[D\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)]] \le 1/C \\ \implies \max_{p' \in \triangle_\Pi} \mathbb{E}_{\pi \sim p'}[D\big(M(\pi) \,\|\, M'(\pi)\big)] \le \varepsilon \end{array} \right\}.$$

*We define $p_{\mathrm{exp}}^{\mathsf{D},\xi}(\varepsilon)$ as the minimizing distribution for this program, and let*

$$C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}, \varepsilon) := \sup_{\xi \in \triangle_{\mathcal{M}}} C_{\mathrm{exp}}^{\mathsf{D},\xi}(\varepsilon)$$

*denote the uniform exploration constant for class $\mathcal{M}$.*

**Lemma E.6** (Formal version of Proposition A.1)**.** *Let $\varepsilon \in [0, 1/2)$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given, and assume that $\inf_{M \in \mathcal{M}_0} \mathsf{g}^M > 0, \inf_{M \in \mathcal{M}_0} \Delta_{\min}^M > 0$, $\sup_{M \in \mathcal{M}_0} \mathsf{n}_{1/4}^M < \infty$, and Assumptions A.2, D.1 and D.2 hold. Then for any $\xi \in \triangle_{\mathcal{M}_0}$, we can bound*

$$\min_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}_0 \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\omega}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} \le C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}_0, \delta)$$

*for any $\delta > 0$ satisfying*

$$\sqrt{\delta} \le \min_{M \in \mathcal{M}_0} \min\left\{ \min\left\{ \frac{1}{81 L_{\mathsf{KL}}}, \frac{\Delta_{\min}^M}{34 V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M/\Delta_{\min}^M + \mathsf{n}_{\varepsilon/36}^M}, \frac{\Delta_{\min}^M}{3} \right\}.$$

**Proof of Lemma E.6.** Let $\delta > 0$ be some tolerance to be chosen. Let $\widetilde{M}$ denote some $M \in \mathcal{M}_0$ such that $\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \le 1/C_{\mathrm{exp}}$, where we abbreviate $C_{\mathrm{exp}} := C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}_0, \delta)$ and $p_{\mathrm{exp}} := p_{\mathrm{exp}}^{\mathsf{D},\xi}(\delta)$ is a distribution that achieves the value of $C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}_0, \delta)$ for $\xi$; if such an $\widetilde{M}$ does not exist, we let $\widetilde{M} = \arg\min_{M \in \mathcal{M}_0} \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]$. Let $\varepsilon' > 0$ be some value to be chosen, and let $\widetilde{\lambda} \in \Lambda(\widetilde{M}; \varepsilon')$ denote the allocation in $\Lambda(\widetilde{M}; \varepsilon')$ with smallest normalizing factor $\mathsf{n}$. Let $\widetilde{\mathsf{n}}$ denote the value of this normalizing factor, then:

$$\Delta^{\widetilde{M}}(\widetilde{\lambda}) \le (1 + \varepsilon')\mathsf{g}^{\widetilde{M}}/\widetilde{\mathsf{n}} \quad \text{and} \quad I^{\widetilde{M}}(\widetilde{\lambda}) \ge (1 - \varepsilon')/\widetilde{\mathsf{n}}.$$

We can bound:

$$\begin{aligned} &\min_{\lambda \in \triangle_\Pi} \min_{\omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)^c \cap \mathcal{M}_0} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\omega}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} \\ &\quad \le \sup_{M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\widetilde{\lambda})^c \cap \mathcal{M}_0} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} \\ &\quad \le \sup_{M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\widetilde{\lambda})^c \cap \mathcal{M}_0} \frac{\mathbb{I}\{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \le 1/C_{\mathrm{exp}}\}}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} + C_{\mathrm{exp}}, \end{aligned} \tag{44}$$

where here we take $\frac{0}{0} = 0$. If $\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, \widetilde{M}(\pi)\big)]] > 1/C_{\mathrm{exp}}$, then by definition of $\widetilde{M}$, for all $M \in \mathcal{M}_0$ we have

$$\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] > 1/C_{\mathrm{exp}},$$

so we can simply bound Eq. (44) $\leq C_{\mathrm{exp}}$. Going forward, we assume this is not the case, so that $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi)\,\|\,\widetilde{M}(\pi)\big)]] \leq 1/C_{\mathrm{exp}}$. Our goal is to show that, for small enough $\delta$, $\widetilde{\lambda} \in \Lambda(M;\varepsilon)$ for every other $M \in \mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi)\,\|\,\widetilde{M}(\pi)\big)]] \leq 1/C_{\mathrm{exp}}$, so that $M \in \mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\widetilde{\lambda})$. This will imply that there does not exist $M \in \mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\widetilde{\lambda})^c \cap \mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi)\,\|\,\widetilde{M}(\pi)\big)]] \leq 1/C_{\mathrm{exp}}$, which further implies that Eq. (44) $\leq C_{\mathrm{exp}}$.

Fix any $M \in \mathcal{M}_0$. We note that by the definition of $p_{\mathrm{exp}}$, if $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi)\,\|\,M(\pi)\big)]] \leq 1/C_{\mathrm{exp}}$, then

$$\sup_{p\in\triangle_{\Pi}} \mathbb{E}_{p}[D\big(\widetilde{M}(\pi)\,\|\,M(\pi)\big)] \leq \delta,$$

This implies in particular that, for each $\pi$, $D\big(\widetilde{M}(\pi)\,\|\,M(\pi)\big) \leq \delta$.

**Step 1: $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi)\,\|\,M(\pi)\big)]] \leq 1/C_{\mathrm{exp}}$ implies $\pi_M = \pi_{\widetilde{M}}$.** As noted, if

$$\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi)\,\|\,M(\pi)\big)]] \leq 1/C_{\mathrm{exp}},$$

we have $D\big(\widetilde{M}(\pi)\,\|\,M(\pi)\big) \leq \delta$ for all $\pi$. Assume that $\pi_M \neq \pi_{\widetilde{M}}$ (note that since $\inf_{M\in\mathcal{M}_0}\Delta_{\mathrm{min}}^{M} > 0$ by assumption, all $M \in \mathcal{M}_0$ have unique optimal). By definition we have $f^{\widetilde{M}}(\pi_{\widetilde{M}}) \geq f^{\widetilde{M}}(\pi_M) + \Delta_{\mathrm{min}}^{\widetilde{M}}$ and $f^{M}(\pi_M) \geq f^{M}(\pi_{\widetilde{M}})$. Thus,

$$\begin{aligned}
\Delta_{\mathrm{min}}^{\widetilde{M}} &\leq f^{\widetilde{M}}(\pi_{\widetilde{M}}) - f^{\widetilde{M}}(\pi_M) + f^{M}(\pi_M) - f^{M}(\pi_{\widetilde{M}}) \\
&\leq |f^{\widetilde{M}}(\pi_{\widetilde{M}}) - f^{M}(\pi_{\widetilde{M}})| + |f^{M}(\pi_M) - f^{\widetilde{M}}(\pi_M)| \\
&\leq \sqrt{D\big(\widetilde{M}(\pi_{\widetilde{M}})\,\|\,M(\pi_{\widetilde{M}})\big)} + \sqrt{D\big(\widetilde{M}(\pi_M)\,\|\,M(\pi_M)\big)}.
\end{aligned}$$

This implies that there exists some $\pi$ such that $D\big(\widetilde{M}(\pi)\,\|\,M(\pi)\big) \geq (\Delta_{\mathrm{min}}^{\widetilde{M}}/2)^2$. Assuming

$$\delta \leq \min_{M\in\mathcal{M}_0} (\Delta_{\mathrm{min}}^{\widetilde{M}}/3)^2, \tag{45}$$

this is a contradiction. Thus, it must be the case that $\pi_M = \pi_{\widetilde{M}}$, as long as Eq. (45) is satisfied.

**Step 2: $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi)\,\|\,M(\pi)\big)]] \leq 1/C_{\mathrm{exp}}$ implies $\widetilde{\lambda} \in \Lambda(M;\epsilon)$.** Under Assumption D.2, we can bound, $\forall \pi \in \Pi$,

$$|f^{\widetilde{M}}(\pi) - f^{M}(\pi)| \leq \sqrt{D\big(\widetilde{M}(\pi)\,\|\,M(\pi)\big)} \leq \sqrt{\delta}.$$

This implies that, for any $\lambda \in \triangle_{\Pi}$,

$$|\Delta^{\widetilde{M}}(\lambda) - \Delta^{M}(\lambda)| \leq |f^{\widetilde{M}}(\pi_M) - f^{M}(\pi_M)| + \sum_{\pi} \lambda_\pi |f^{\widetilde{M}}(\pi) - f^{M}(\pi)| \leq 4\sqrt{\delta}.$$

In addition, under Assumption D.1, we have

$$\begin{aligned}
D_{\mathsf{KL}}\big(M(\pi)\,\|\,M'(\pi)\big) &\geq D_{\mathsf{KL}}\big(\widetilde{M}(\pi)\,\|\,M'(\pi)\big) - L_{\mathsf{KL}}\sqrt{D\big(\widetilde{M}(\pi)\,\|\,M(\pi)\big)} \\
&\geq D_{\mathsf{KL}}\big(\widetilde{M}(\pi)\,\|\,M'(\pi)\big) - 2L_{\mathsf{KL}}\sqrt{\delta}.
\end{aligned}$$

This implies, for any $\lambda \in \triangle_\Pi$,

$$\begin{aligned} I^M(\lambda) &= \inf_{M' \in \mathcal{M}^{\text{alt}}(M)} \sum_\pi \lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \\ &\geq \inf_{M' \in \mathcal{M}^{\text{alt}}(M)} \sum_\pi \lambda(\pi) D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - 2L_{\mathsf{KL}}\sqrt{\delta} \\ &= I^{\widetilde{M}}(\lambda) - 2L_{\mathsf{KL}}\sqrt{\delta} \end{aligned}$$

where the final equality uses that, given what we have already shown, $\pi_M = \pi_{\widetilde{M}}$, so that $\mathcal{M}^{\text{alt}}(M) = \mathcal{M}^{\text{alt}}(\widetilde{M})$. Repeating the calculation in the other direction, we get that $|I^M(\lambda) - I^{\widetilde{M}}(\lambda)| \leq 2L_{\mathsf{KL}}\sqrt{\delta}$.

We next relate $\mathsf{g}^M$ to $\mathsf{g}^{\widetilde{M}}$. By definition we have

$$(1+\varepsilon')\mathsf{g}^M/\widetilde{\mathsf{n}} \geq \inf_{\lambda \in \triangle_\Pi} \Delta^M(\lambda) \quad \text{s.t.} \quad I^M(\lambda) \geq (1-\varepsilon')/\widetilde{\mathsf{n}}.$$

Applying our perturbation bounds we can lower bound this as

$$\begin{aligned} &\geq \inf_{\lambda \in \triangle_\Pi} \Delta^{\widetilde{M}}(\lambda) - 4\sqrt{\delta} \quad \text{s.t.} \quad I^{\widetilde{M}}(\lambda) \geq (1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta} \\ &\geq \frac{\mathsf{g}^{\widetilde{M}}}{((1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta})^{-1}} - 4\sqrt{\delta} \end{aligned}$$

where the last inequality follows from Lemma E.1. This implies that

$$\mathsf{g}^{\widetilde{M}} \leq ((1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta})^{-1}(1+\varepsilon') \cdot \frac{\mathsf{g}^M}{\widetilde{\mathsf{n}}} + 4((1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta})^{-1}\sqrt{\delta}. \tag{46}$$

Assuming that

$$\sqrt{\delta} \leq \frac{\varepsilon' - 2(\varepsilon')^2}{2(1+2\varepsilon')L_{\mathsf{KL}}\widetilde{\mathsf{n}}},$$

some algebra shows that

$$\text{Eq. (46)} \leq (1+2\varepsilon')^2\mathsf{g}^M + 4(1+2\varepsilon')\widetilde{\mathsf{n}}\sqrt{\delta}.$$

Now we can bound

$$\begin{aligned} \Delta^M(\widetilde{\lambda}) &\leq \Delta^{\widetilde{M}}(\widetilde{\lambda}) + 4\sqrt{\delta} \\ &\leq (1+\varepsilon')\mathsf{g}^{\widetilde{M}}/\widetilde{\mathsf{n}} + 4\sqrt{\delta} \\ &\leq (1+2\varepsilon')^3\mathsf{g}^M/\widetilde{\mathsf{n}} + 4(1+2\varepsilon')^2\sqrt{\delta} + 4\sqrt{\delta} \end{aligned}$$

and

$$I^M(\widetilde{\lambda}) \geq I^{\widetilde{M}}(\widetilde{\lambda}) - 2L_{\mathsf{KL}}\sqrt{\delta} \geq (1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta}.$$

If $\varepsilon'$ and $\delta$ are small enough so that

$$(1+2\varepsilon')^3 \leq 1+\varepsilon/2 \quad \text{and} \quad 4(1+2\varepsilon')^2\sqrt{\delta} + 4\sqrt{\delta} \leq \varepsilon\mathsf{g}^M/2\widetilde{\mathsf{n}}$$

and

$$1-\varepsilon' \geq 1-\varepsilon/2 \quad \text{and} \quad 2L_{\mathsf{KL}}\sqrt{\delta} \leq \varepsilon/2\widetilde{\mathsf{n}},$$

then $\Delta^M(\widetilde{\lambda}) \leq (1+\varepsilon)\mathsf{g}^M/\widetilde{\mathsf{n}}$ and $I^M(\widetilde{\lambda}) \geq (1-\varepsilon)/\widetilde{\mathsf{n}}$, which implies that $\widetilde{\lambda} \in \Lambda(M;\varepsilon)$ with scaling factor $\widetilde{\mathsf{n}}$.

**Step 3: Condition on $\delta$.** Altogether, we have assumed that $\delta$ satisfies Eq. (45) and, that for some $M \in \mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \le 1/C_{\mathrm{exp}}$, we have

$$\sqrt{\delta} \le \frac{\varepsilon' - 2(\varepsilon')^2}{2(1+2\varepsilon')L_{\mathsf{KL}}\widetilde{\mathsf{n}}}, \quad 4(1+2\varepsilon')^2\sqrt{\delta} + 4\sqrt{\delta} \le \varepsilon \mathsf{g}^M/2\widetilde{\mathsf{n}}, \quad 2L_{\mathsf{KL}}\sqrt{\delta} \le \varepsilon/2\widetilde{\mathsf{n}} \tag{47}$$

and

$$(1+2\varepsilon')^3 \le 1 + \varepsilon/2, \quad 1 - \varepsilon' \ge 1 - \varepsilon/2.$$

Some algebra shows that, as long as $\varepsilon \le 1$, it suffices to take $\varepsilon' = \varepsilon/36$ to satisfy the latter two conditions. Furthermore, some calculation shows that a sufficient condition for Eq. (47) to be met is that

$$\sqrt{\delta} \le \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\mathsf{g}^M}{17}\right\} \cdot \frac{\varepsilon}{\widetilde{\mathsf{n}}}.$$

By Lemma E.5 and our choice of $\widetilde{\mathsf{n}}$, we can bound

$$\widetilde{\mathsf{n}} \le 2\mathsf{g}^{\widetilde{M}}/\Delta^{\widetilde{M}}_{\mathrm{min}} + \mathsf{n}^{\widetilde{M}}_{\varepsilon/36}$$

so it suffices that we take

$$\sqrt{\delta} \le \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\mathsf{g}^M}{17}\right\} \cdot \varepsilon\left(\frac{2\mathsf{g}^{\widetilde{M}}}{\Delta^{\widetilde{M}}_{\mathrm{min}}} + \mathsf{n}^{\widetilde{M}}_{\varepsilon}\right)^{-1}.$$

As $\widetilde{M}$ was chosen to an arbitrary model in $\mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, \widetilde{M}(\pi)\big)]] \le 1/C_{\mathrm{exp}}$, we take it to minimize $\mathsf{g}^{\widetilde{M}}$ over this constraint. It suffices then that

$$\sqrt{\delta} \le \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\mathsf{g}^{\widetilde{M}}}{17}\right\} \cdot \varepsilon\left(\frac{2\mathsf{g}^{\widetilde{M}}}{\Delta^{\widetilde{M}}_{\mathrm{min}}} + \mathsf{n}^{\widetilde{M}}_{\varepsilon/36}\right)^{-1}.$$

Finally, by Lemma F.13 (under Assumption A.2) and Lemma E.3, we can lower bound $\mathsf{g}^{\widetilde{M}} \ge \Delta^{\widetilde{M}}_{\mathrm{min}}/2V_{\mathcal{M}}$. Combining this condition with Eq. (45) gives the result.

□

## Appendix F. Proofs from Appendix A

**Organization of Appendix F.** In this section we prove the main results from Appendix A. We consider a slightly generalized version of the setting in Appendix A, where we allow for divergences other than just the KL divergence, as described below. This section is organized as follows.

- First, in Appendix F.1, we give the proof of our main result, Theorem A.1. We break this proof into two principle components: bounding the regret of $\mathsf{AE}^2$ in the exploit phase (Section F.1.1), and explore phase (Section F.1.2). The key results in this section are Lemma F.3, which formalizes the key algorithm intuition given in Appendix A, showing that exploring via the AEC yields low regret, and Lemma F.4, which shows that, to enter the explore phase, the total "information gain" must be bounded as $O(\log T)$, which ultimately yields the optimal leading-order scaling. We combine these results with our estimation guarantees in Appendix F.1.3, where we give the proof of Theorem A.1.

- In Appendix F.2, we extend AE$^2$ and Theorem A.1 to the case where we assume no lower bound on the minimum gap over the model class, first presenting our main algorithm in this setting, Algorithm 6, and then giving a proof of Theorem A.2. The structure of this section is similar to Appendix F.1—the primary difference being a slightly different argument to handle the need to adapt to the minimum gap of the ground truth instance.

- Finally, in Appendix F.3 we present our estimation routine with covering, and prove that it achieves low estimation error, and in Appendix F.4 we provide the proofs of miscellaneous results used throughout Appendix F.

### F.1. Regret Bound for Uniformly Regular Classes (Theorem A.1)

In this section we prove Theorem F.1, which generalizes Theorem A.1. To do so, we analyze Algorithm 5, wich generalizes Algorithm 2 to allow for general divergences.

Throughout, we define

$$\underline{\mathrm{g}}^{\mathcal{M}} := \min_{M \in \mathcal{M}: \mathrm{g}^M > 0} \mathrm{g}^M.$$

**Algorithm 5** Allocation Estimation via Adaptive Exploration (AE$^2$, general divergences)

1: **input:** Optimality tolerance $\delta$, model class $\mathcal{M}$, estimation oracle $\mathbf{Alg}_D$.
2: Initialize $s \leftarrow 1$, $\varepsilon \leftarrow \frac{\delta}{4+2\delta}$, $\mathsf{n}_{\max} \leftarrow \mathsf{n}_{\max}(\mathcal{M}, \varepsilon)$, $q \leftarrow \frac{4\mathsf{n}_{\max}+\delta\underline{\mathrm{g}}^{\mathcal{M}}}{4\mathsf{n}_{\max}+2\delta\underline{\mathrm{g}}^{\mathcal{M}}}$.
3: Compute $\xi^1 \leftarrow \mathbf{Alg}_D(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
4: **for** $t = 1, 2, 3, \ldots$ **do**
5: **if** $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then**
6: Play $\pi_{\widehat{M}^s}$. // Exploit
7: **else** // Explore
8: Set $p^s \leftarrow q\lambda^s + (1-q)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \arg\min_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{\pi\sim\omega}\left[D\left(\widehat{M}(\pi)\,\|\,M(\pi)\right)\right]\right]}. \tag{48}$$

9: Draw $\pi^s \sim p^s$, observe $r^s, o^s$.
10: Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_D(\{(\pi^i, r^i, o^i)\}_{i=1}^s)$ and let $\widehat{M}^s = \mathbb{E}_{M\sim\xi^s}[M]$.
11: $s \leftarrow s+1$.

#### F.1.1. BOUNDING REGRET OF EXPLOIT PHASE

We refer to the *exploit phase* as the subset of rounds $t$ in which Line 6 is reached, and refer to the *explore phase* as the subset of rounds in which Line 8 is reached.

**Lemma F.1.** *The total expected regret incurred by the exploit phase of Algorithm 5 is bounded by* $2\log\log T + 3$.

**Proof of Lemma F.1.** Let $\mathcal{E}_t$ denote the event that we exploit at round $t$, and that $\pi_{\widehat{M}^s} \neq \pi_\star$. Then, since we can incur suboptimality of at most 1 at each round, the total expected regret incurred by the exploit phase is bounded by

$$\sum_{t=1}^{T} \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}].$$

Let $\widetilde{\mathcal{E}}_t$ denote the event

$$\widetilde{\mathcal{E}}_t := \left\{ \forall s \geq 1 \ : \ \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M^\star,\pi^i}(r^i,o^i)}\right] < \log(t\log t) \right\}.$$

By Lemma F.2, we have $\mathbb{P}^{M^\star}[\mathbb{I}\{\widetilde{\mathcal{E}}_t^c\}] \leq \frac{1}{t\log t}$, and we can bound

$$\mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}] \leq \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t \cap \widetilde{\mathcal{E}}_t\}] + \mathbb{E}^{M^\star}[\mathbb{I}\{\widetilde{\mathcal{E}}_t^c\}] \leq \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t \cap \widetilde{\mathcal{E}}_t\}] + \frac{1}{t\log t}.$$

Let $s_t$ denote the exploration round at round $t$. If we exploit at round $t$, this implies that for all $M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, we have

$$\sum_{i=1}^{s_t-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t). \tag{49}$$

If $\pi_{\widehat{M}^s} \neq \pi_\star$, then $M^\star \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, so Eq. (49) must hold for $M \leftarrow M^\star$. This contradicts $\widetilde{\mathcal{E}}_t$, however, so $\mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t \cap \widetilde{\mathcal{E}}_t\}] = 0$. Thus, $\mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}] \leq \frac{1}{t\log t}$, so

$$\sum_{t=1}^{T} \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}] \leq 3 + \sum_{t=3}^{T} \frac{1}{t\log t} \leq 3 + 2\int_e^T \frac{1}{t\log t}\mathrm{d}t = 3 + 2\log\log T.$$

□

**Lemma F.2.** *For $\{(r^i,o^i,\xi^i)\}_{i=1}^s$ generated as in Algorithm 5, we have that*

$$\mathbb{P}^{M^\star}\left[\exists s \geq 1 \ : \ \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M^\star,\pi^i}(r^i,o^i)}\right] \geq \varepsilon\right] \leq e^{-\varepsilon}.$$

**Proof of Lemma F.2.** Denote

$$X_s := \exp\left(\sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M^\star,\pi^i}(r^i,o^i)}\right]\right).$$

We first show that $X_s$ is a supermartingale. Letting $\mathcal{F}_{s-1}$ denote the filtration up to $s-1$, we have

$$\mathbb{E}^{M^\star}[X_s \mid \mathcal{F}_{s-1}] = \exp\left(\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M^\star,\pi^i}(r^i,o^i)}\right]\right) \cdot \mathbb{E}^{M^\star}\left[\exp\left(\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)}\right]\right) \mid \mathcal{F}_{s-1}\right]$$

$$
\begin{aligned}
&= X_{s-1} \cdot \mathbb{E}^{M^\star}\left[\exp\left(\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)}\right]\right) \mid \mathcal{F}_{s-1}\right] \\
&\overset{(a)}{\leq} X_{s-1} \cdot \mathbb{E}^{M^\star}\left[\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\exp\left(\log\frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)}\right)\right] \mid \mathcal{F}_{s-1}\right] \\
&= X_{s-1} \cdot \mathbb{E}^{M^\star}\left[\frac{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)]}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)} \mid \mathcal{F}_{s-1}\right] \\
&= X_{s-1}
\end{aligned}
$$

where $(a)$ holds by Jensen's inequality, and the final equality holds since $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{P}^{\widehat{M},\pi}(\cdot,\cdot)]$ is a valid distribution over $\mathcal{R}\times\mathcal{O}$. Thus, $X_s$ is a supermartingale. Ville's Maximal Inequality then immediately gives that

$$
\mathbb{P}^{M^\star}[\exists s \geq 1 \ : \ X_s \geq e^{\varepsilon}] \leq \frac{\mathbb{E}^{M^\star}[X_1]}{e^{\varepsilon}}.
$$

To complete the proof, using the same calculation as above, we bound

$$
\mathbb{E}^{M^\star}[X_1] = \mathbb{E}^{M^\star}\left[\exp\left(\mathbb{E}_{\widehat{M}\sim\xi^1}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^1}(r^1,o^1)}{\mathbb{P}^{M^\star,\pi^1}(r^1,o^1)}\right]\right)\right] \leq \mathbb{E}^{M^\star}\left[\mathbb{E}_{\widehat{M}\sim\xi^1}\left[\exp\left(\log\frac{\mathbb{P}^{\widehat{M},\pi^1}(r^1,o^1)}{\mathbb{P}^{M^\star,\pi^1}(r^1,o^1)}\right)\right]\right] \leq 1.
$$

□

### F.1.2. BOUNDING REGRET OF EXPLORE PHASE

**Lemma F.3** (Main Explore Phase Regret Bound)**.** *Let $s_T$ denote the total number of exploration rounds (which is a random variable), and assume that $\delta \in [0, 1/2)$. Then running Algorithm 5, if $\mathsf{g}^\star > 0$, we can bound*

$$
\begin{aligned}
\mathbb{E}[s_T] \leq{}& \frac{24\mathsf{n}_{\max}^2 + 8\mathsf{n}_{\max}\underline{\mathsf{g}}^{\mathcal{M}}}{(\delta\underline{\mathsf{g}}^{\mathcal{M}})^2} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot \mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{12\mathsf{n}_{\max}}{\delta\Delta_{\min}} \cdot \mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] \\
&+ \frac{6\mathsf{n}_{\max}}{\delta} \cdot \mathbb{E}\left[\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\right]
\end{aligned}
$$

*and the regret during exploration rounds is bounded as*

$$
\begin{aligned}
\mathbb{E}\left[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)\right] \leq{}& \frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot \mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] \\
&+ (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}\left[\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\right].
\end{aligned}
$$

**Proof of Lemma F.3.** The expected regret during exploration can be written as $\mathbb{E}[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)] = \mathbb{E}[\sum_{s=1}^{s_T}\mathbb{E}_{p^s}[\Delta^\star(\pi)]]$. By definition, for exploration rounds, we have $p^s \leftarrow q\lambda^s + (1-q)\omega^s$. For each $s \leq s_T$, we consider three cases to bound the instantaneous expected regret, $\mathbb{E}_{p^s}[\Delta^\star(\pi)] = \Delta^\star(p^s)$.

**Case 1:** $M^\star \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})$**.** Denote such rounds as $\mathcal{S}^1_{\mathrm{exp}}$. Write

$$\Delta^\star(p^s) = \Big[\Delta^\star(p^s) - \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\Big[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\Big]\Big] + \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\Big[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\Big]$$

for

$$\gamma^s := \frac{1+\delta}{1-q}\cdot\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}. \tag{50}$$

In this case we have that

$$\begin{aligned}
&\gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\\
&\quad= \frac{1+\delta}{1-q}\cdot\frac{1}{\mathbb{E}_{\omega^s}[\mathbb{E}_{\widehat{M}\sim\xi^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\\
&\quad\geq \frac{1+\delta}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\\
&\quad= 1+\delta.
\end{aligned}$$

Thus, since the suboptimality gap is always bounded by 1, we can bound

$$\Delta^\star(p^s) - \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\Big[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\Big] \leq 1 - \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\Big[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\Big] \leq -\delta.$$

So,

$$\Delta^\star(p^s) \leq -\delta + \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\Big[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\Big].$$

**Case 2:** $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max}), \pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^2_{\mathrm{exp}}$, and write

$$\Delta^\star(p^s) = [\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star\cdot\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]] + (1+\delta)\mathsf{g}^\star\cdot\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]$$

for any $M \in \mathcal{M}^\star_{\mathrm{alt}}$. In this case, since $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})$, we have that $\lambda^s \in \Lambda(M^\star;\varepsilon)$. This then implies that

$$\Delta^\star(\lambda^s) \leq (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star \quad\text{and}\quad \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] \geq (1-\varepsilon)/\mathsf{n}^\star$$

for some $\mathsf{n}^\star \leq \mathsf{n}_{\max}$. Since $M \in \mathcal{M}^\star_{\mathrm{alt}}$, it follows that $\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] \geq (1-\varepsilon)/\mathsf{n}^\star$. Thus,

$$\begin{aligned}
\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] &\leq q\left[\Delta^\star(\lambda^s) - (1+\delta)\mathsf{g}^\star\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\right] + 1 - q\\
&\leq q\left[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star\right] + 1 - q\\
&= q\left[2\varepsilon - \delta(1-\varepsilon)\right]\cdot\frac{\mathsf{g}^\star}{\mathsf{n}^\star} + 1 - q\\
&\overset{(a)}{=} -\frac{q\delta}{2}\frac{\mathsf{g}^\star}{\mathsf{n}^\star} + 1 - q\\
&\leq -\frac{q\delta}{2}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}} + 1 - q
\end{aligned}$$

$$\overset{(b)}{\leq} -\frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}},$$

where $(a)$ follows from our choice of $\varepsilon = \frac{\delta/2}{2+\delta}$ and setting of $\mathsf{n}^\star$, and $(b)$ follows from our setting of $q = \frac{4\mathsf{n}_{\max}+\delta\underline{\mathsf{g}}^{\mathcal{M}}}{4\mathsf{n}_{\max}+2\delta\underline{\mathsf{g}}^{\mathcal{M}}}$, since $\mathsf{g}^\star > 0$, and some algebra. Thus,

$$\Delta^\star(p^s) \leq (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] - \frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}}.$$

As this holds for every $M \in \mathcal{M}^\star_{\mathrm{alt}}$, we therefore have

$$\Delta^\star(p^s) \leq \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] - \frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}}.$$

**Case 3:** $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}_{\max}), \pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^3_{\mathrm{exp}}$, and write

$$\begin{aligned}\Delta^\star(p^s) = \left[\Delta^\star(p^s) - \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]\right] \\ + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]].\end{aligned}$$

Since $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}_{\max})$, we have that $\lambda^s \in \Lambda(M^\star; \varepsilon)$. This then implies that for any $M \in \mathcal{M}^\star_{\mathrm{alt}}$:

$$\Delta^\star(\lambda^s) \leq (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star \quad \text{and} \quad \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \geq (1-\varepsilon)/\mathsf{n}^\star$$

for some $\mathsf{n}^\star \leq \mathsf{n}_{\max}$. By Lemma F.9, since $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$, we can lower bound $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{I}\{\widehat{M} \in \mathcal{M}^\star_{\mathrm{alt}}\}] \geq \frac{1}{2}\Delta_{\min}$. Thus, we have

$$\begin{aligned}\frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] &\geq \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big) \cdot \mathbb{I}\{\widehat{M} \in \mathcal{M}^\star_{\mathrm{alt}}\}]] \\ &\geq \frac{(1+\delta)(1-\varepsilon)\mathsf{g}^\star}{\mathsf{n}^\star}.\end{aligned}$$

This implies that

$$\begin{aligned}\Delta^\star(p^s) - \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] &\leq q[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star] + 1 - q \\ &\leq -\frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}},\end{aligned}$$

where the final inequality follows by the same argument as in Case 2. Thus,

$$\Delta^\star(p^s) \leq \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] - \frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}}.$$

**Completing the Proof.** In total we have

$$\begin{aligned}\mathbb{E}\Big[\sum_{s=1}^{s_T}\Delta^\star(p^s)\Big] &\le \mathbb{E}\Big[-\delta|\mathcal{S}^1_{\mathrm{exp}}| - \frac{\delta\mathsf{g}^\star}{4\mathsf{n}_{\max}}|\mathcal{S}^2_{\mathrm{exp}}\cup\mathcal{S}^3_{\mathrm{exp}}| + \sum_{s\in\mathcal{S}^1_{\mathrm{exp}}}\gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\\ &\quad+(1+\delta)\mathsf{g}^\star\sum_{s\in\mathcal{S}^2_{\mathrm{exp}}}\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\\ &\quad+\frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\sum_{s\in\mathcal{S}^3_{\mathrm{exp}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\Big]\\ &\le \mathbb{E}\Big[-\frac{\delta\mathsf{g}^\star}{4\mathsf{n}_{\max}}\cdot s_T+\frac{8\mathsf{n}_{\max}+2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)+\frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)\\ &\quad+(1+\delta)\mathsf{g}^\star\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\Big]\end{aligned}$$

where the last inequality follows from Lemma F.10, which bounds

$$\gamma^s\le\frac{8\mathsf{n}_{\max}+2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}),$$

and also using that for any $s\in\mathcal{S}^2_{\mathrm{exp}}$ we have $\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}$. Upper bounding $-\frac{\delta\mathsf{g}^\star}{4\mathsf{n}_{\max}}\cdot s_T\le 0$ proves the second claim in the lemma statement.

For the first claim, as regret is always nonnegative, it follows that

$$\begin{aligned}0\le\mathbb{E}\Big[&-\frac{\delta\mathsf{g}^\star}{4\mathsf{n}_{\max}}\cdot s_T+\frac{8\mathsf{n}_{\max}+2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)+\frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)\\ &+(1+\delta)\mathsf{g}^\star\cdot\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\Big]\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}[s_T]\le\frac{4\mathsf{n}_{\max}}{\delta\mathsf{g}^\star}\cdot\mathbb{E}\Big[&\frac{8\mathsf{n}_{\max}+2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)+\frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)\\ &+(1+\delta)\mathsf{g}^\star\cdot\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\Big].\end{aligned}$$

□

**Lemma F.4.** *When running both Algorithm 5 and Algorithm 6, for all $\alpha\in[0,1)$, we have*

$$\begin{aligned}&\mathbb{E}\Big[\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)}s^\alpha\cdot\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\Big]\\ &\le\mathbb{E}[s_T^\alpha]\log T+\mathbb{E}\Big[V_{\mathcal{M}}s_T^{1/2+\alpha}\sqrt{1344d_{\mathsf{cov}}\cdot\log(128C_{\mathsf{cov}}s_T)}\Big]\end{aligned}$$

$$+ (V_{\mathcal{M}} + L_{\mathsf{KL}})\left(4\frac{s_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{s_T} s^{1/2+\alpha/2} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\right)\Bigg]$$
$$+ \mathbb{E}[s_T^\alpha]\log\log T + 7V_{\mathcal{M}}.$$

**Proof of Lemma F.4.** Let $\widetilde{s}_T$ denote the final exploration round for which $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$. Then, upper bounding the KL divergence by $2V_{\mathcal{M}}$ via Lemma F.13, we have

$$\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} s^\alpha \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}$$
$$\le \widetilde{s}_T^\alpha \sum_{s=1}^{\widetilde{s}_T-1} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] + 2V_{\mathcal{M}} s_T^\alpha.$$

Under Assumption D.1 and via Jensen's inequality and AM-GM, we have, for any $\alpha_s > 0$ and $M \in \mathcal{M}$,

$$\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] \le \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]] + L_{\mathsf{KL}}\sqrt{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}$$
$$\le \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]] + L_{\mathsf{KL}}\left(\frac{1}{\alpha_s} + \alpha_s\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right),$$

We now wish to bound

$$\mathbb{E}\left[\widetilde{s}_T^\alpha \sum_{s=1}^{\widetilde{s}_T-1} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]\right].$$

Let $\mathcal{M}^j_{\mathrm{cov}}$ denote an $(\rho_j,\mu_j)$-cover of $\mathcal{M}^\star_{\mathrm{alt}}$ and $\mathcal{E}_j^s$ the corresponding event at step $s \le 2^j$, for some $\rho_j$ and $\mu_j$ to be chosen. Let $\mathcal{E}_j := \cap_{s\le 2^j}\mathcal{E}_j^s$. By definition $\mathbb{P}_{M^\star}[\mathcal{E}_j^s] \ge 1-\mu_j$, so $\mathbb{P}_{M^\star}[\mathcal{E}_j] \ge 1 - 2^j\mu_j$. Define an event

$$A_j := \Bigg\{\forall s \le 2^j, M\in\mathcal{M}^j_{\mathrm{cov}} \;:\; (s+1)^\alpha \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{p^i}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]$$
$$\le (s+1)^\alpha \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] + V_{\mathcal{M}}(s+1)^{\frac{1}{2}+\alpha}\sqrt{56\log\frac{2^j\mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}},\rho_j,\mu_j)}{\delta_j}}$$
$$+ V_{\mathcal{M}}\cdot\left(4\frac{(s+1)^{\frac{1+\alpha}{2}}}{1-\alpha} + \sum_{i=1}^{s} i^{\frac{1+\alpha}{2}}\cdot\mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right)\Bigg\}$$

sfor some $\delta_j$ to be chosen. By invoking Lemma F.12 with $\beta_i = i^{1/2+\alpha/2}/(s+1)^\alpha$ and a union bound, we have $\mathbb{P}[A_j] \ge 1-\delta_j$ (while Lemma F.12 does not contain the $(s+1)^\alpha$ term, the bound in the expression for $A_j$ simply gives the bound from Lemma F.12 multiplied through with $(s+1)^\alpha$, which is non-random, so this is admissible). We can decompose

$$\mathbb{E}\left[\widetilde{s}_T^\alpha \sum_{s=1}^{\widetilde{s}_T-1} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]\right]$$

$$\leq \sum_{j=1}^{\lceil \log T \rceil} \mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j \cap \mathcal{E}_j\}\right]$$
$$+ \sum_{j=1}^{\lceil \log T \rceil} \mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j^c \cup \mathcal{E}_j^c\}\right].$$

We bound these terms separately. We can bound the second term as

$$\sum_{j=1}^{\lceil \log T \rceil} \mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j^c \cup \mathcal{E}_j^c\}\right]$$
$$\leq \sum_{j=1}^{\lceil \log T \rceil} 2V_{\mathcal{M}} 2^{2j} (\mathbb{P}[A_j^c] + \mathbb{P}[\mathcal{E}_j^c])$$
$$\leq \sum_{j=1}^{\lceil \log T \rceil} 2V_{\mathcal{M}} 2^{2j} (\delta_j + 2^j \mu_j)$$

where the first inequality follows by bounding Lemma F.13, and $\widetilde{s}_T \leq 2^j$, and the second follows by our bound on the probability of $A_j$. Letting $\delta_j = \frac{1}{2^{3j}}, \mu_j = \frac{1}{2^{4j}}$, this term can be bounded by $4V_{\mathcal{M}}$. We turn now to the first term. Fix $j \in [\lceil \log T \rceil]$. By definition of the event $A_j$, and since $\widetilde{s}_T \leq 2^j$, plugging in our choice of $\delta_j$ we can bound, for any $M \in \mathcal{M}^j_{\mathrm{cov}}$,

$$\mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j \cap \mathcal{E}_j\}\right]$$
$$\leq \mathbb{E}\left[\left(\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^s}(r^s, o^s)}{\mathbb{P}^{M, \pi^s}(r^s, o^s)}\right] + V_{\mathcal{M}} \widetilde{s}_T^{1/2+\alpha} \sqrt{56 \log(2^{3j} \mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}}, \rho_j, \mu_j))}\right.\right.$$
$$\left.\left. + V_{\mathcal{M}} \cdot \left(4 \frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]\right)\right) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), \mathcal{E}_j\}\right]$$
$$\leq \mathbb{E}\left[\left(\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^s}(r^s, o^s)}{\mathbb{P}^{M, \pi^s}(r^s, o^s)}\right] + V_{\mathcal{M}} \widetilde{s}_T^{1/2+\alpha} \sqrt{168 \log(8 \widetilde{s}_T \mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}}, \rho_j, \tfrac{\widetilde{s}_T^{-4}}{16}))}\right.\right.$$
$$\left.\left. + V_{\mathcal{M}} \cdot \left(4 \frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]\right)\right) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), \mathcal{E}_j\}\right] \tag{51}$$

where the second inequality follows since, if $\widetilde{s}_T \in [2^{j-1}, 2^j)$, then we can bound $2^j \leq 2\widetilde{s}_T$.

Since $\widetilde{s}_T$ is an exploration round by definition, we know that for all $\pi_{\widehat{M}^{\widetilde{s}_T}} \in \boldsymbol{\pi}_{\widehat{M}^{\widetilde{s}_T}}$, there exists some $M' \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^{\widetilde{s}_T}})$ such that

$$\sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^s}(r^s, o^s)}{\mathbb{P}^{M', \pi^s}(r^s, o^s)}\right] \leq \log(T \log T).$$

By assumption we have $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^{\widetilde{s}_T}}$, so it follows that there exists some (random) $M' \in \mathcal{M}^\star_{\mathrm{alt}}$ such that the above inequality holds. Let $M'' \in \mathcal{M}^j_{\mathrm{cov}}$ denote the model in $\mathcal{M}^j_{\mathrm{cov}}$ such that

$$\left|\log \frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M'',\pi}(r,o)}\right| = \left|\log \mathbb{P}^{M',\pi}(r,o) - \log \mathbb{P}^{M'',\pi}(r,o)\right| \le \rho_j,$$

for all $(r,o,\pi)$ for which $\sup_{\widetilde{M}\in\mathcal{M}} \mathbb{P}^{\widetilde{M},\pi}(r,o \mid \mathcal{E}_j) > 0$, which is guaranteed to exist by Definition A.1. Note that by definition of $\mathcal{M}^j_{\mathrm{cov}}$, we have $M'' \in \mathcal{M}^\star_{\mathrm{alt}}$. We then have

$$\begin{aligned}
&\widetilde{s}_T^{\alpha} \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T-1} \mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M,\pi^s}(r^s,o^s)}\right] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1},2^j), \mathcal{E}_j\} \\
&\quad \le \widetilde{s}_T^{\alpha} \sum_{s=1}^{\widetilde{s}_T-1} \mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M'',\pi^s}(r^s,o^s)}\right] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1},2^j), \mathcal{E}_j\} \\
&\quad = \widetilde{s}_T^{\alpha} \sum_{s=1}^{\widetilde{s}_T-1} \left[\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M',\pi^s}(r^s,o^s)}\right] + \log \frac{\mathbb{P}^{M',\pi^s}(r^s,o^s)}{\mathbb{P}^{M'',\pi^s}(r^s,o^s)}\right] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1},2^j), \mathcal{E}_j\} \\
&\quad \le \widetilde{s}_T^{\alpha} \log(T\log T) + \rho_j \cdot \widetilde{s}_T^{1+\alpha},
\end{aligned}$$

where the inequality holds since on $\mathcal{E}_j$, we can ensure that $\log \frac{\mathbb{P}^{M',\pi^s}(r^s,o^s)}{\mathbb{P}^{M'',\pi^s}(r^s,o^s)} \le \rho_j$. Therefore, choosing $\rho_j = 2^{-3j}$, we have

$$\begin{aligned}
\text{Eq. (51)} &\le \mathbb{E}\Bigg[\Bigg(\widetilde{s}_T^{\alpha}\log(T\log T) + \rho_j \cdot \widetilde{s}_T^{1+\alpha} + V_{\mathcal{M}}\widetilde{s}_T^{1/2+\alpha}\sqrt{168\widetilde{s}_T \log(8\mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}},\rho_j,\tfrac{\widetilde{s}_T^{-4}}{16}))} \\
&\qquad + V_{\mathcal{M}} \cdot \left(4\frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right)\Bigg) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1},2^j)\}\Bigg] \\
&\le \mathbb{E}\Bigg[\Bigg(\widetilde{s}_T^{\alpha}\log(T\log T) + 2^{-j} + V_{\mathcal{M}}\widetilde{s}_T^{\alpha+1/2}\sqrt{168\log(8\widetilde{s}_T\mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}},\tfrac{\widetilde{s}_T^{-3}}{8},\tfrac{\widetilde{s}_T^{-4}}{16}))} \\
&\qquad + V_{\mathcal{M}} \cdot \left(4\frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right)\Bigg) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1},2^j)\}\Bigg].
\end{aligned}$$

As this holds for each $j$, and the events $\{\widetilde{s}_T \in [2^{j-1},2^j)\}$ are disjoint, we can sum over $j$ to bound

$$\begin{aligned}
&\sum_{j=1}^{\lceil\log T\rceil} \mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T-1} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1},2^j), A_j \cap \mathcal{E}_j\}\right] \\
&\quad \le \mathbb{E}\Bigg[\widetilde{s}_T^{\alpha}\log(T\log T) + 1 + V_{\mathcal{M}}\widetilde{s}_T^{\alpha+1/2}\sqrt{168\log(8\widetilde{s}_T\mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}},\tfrac{\widetilde{s}_T^{-3}}{8},\tfrac{\widetilde{s}_T^{-4}}{16}))} \\
&\qquad + V_{\mathcal{M}} \cdot \left(4\frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right)\Bigg].
\end{aligned}$$

Finally, under Assumption A.3 and Lemma F.15 we can bound

$$\begin{aligned}\log(8\widetilde{s}_T\mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}}, \tfrac{\widetilde{s}_T^{-3}}{8}, \tfrac{\widetilde{s}_T^{-4}}{16})) &\le \log(8\widetilde{s}_T) + \log \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tfrac{\widetilde{s}_T^{-3}}{8}, \tfrac{\widetilde{s}_T^{-4}}{16}) \\ &\le \log(8\widetilde{s}_T) + d_{\mathsf{cov}} \cdot \log\big(128 C_{\mathsf{cov}} \widetilde{s}_T^7\big) \\ &\le 8 d_{\mathsf{cov}} \cdot \log(128 C_{\mathsf{cov}} \widetilde{s}_T).\end{aligned}$$

The result follows. □

### F.1.3. COMPLETING THE PROOF

**Theorem F.1** (Full version of Theorem A.1)**.** *For any $\delta \le 1/2$, if we set $D(\cdot \,\|\, \cdot) = D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$ and instantiate $\mathbf{Alg}_D$ with Algorithm 7, under Assumptions 1.1 to 1.3 and A.1 to A.4 and if $\mathsf{g}^\star > 0$, the expected regret of Algorithm 5 is bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\delta)\mathsf{g}^\star \cdot \log T + C_{\mathsf{aec}} \cdot \overline{\mathsf{aec}}_{\varepsilon/2}(\mathcal{M}) \cdot \log^{3/2}(\log T) + \mathrm{lin}\left(C_{\mathsf{low}}, \sqrt{\log T}, \log^{3/2}(\log T)\right),$$

*for*

$$\begin{aligned}C_{\mathsf{low}} &:= \mathrm{lin}\left(\mathsf{g}^\star, \max_{M\in\mathcal{M}} \mathsf{g}^M, V_{\mathcal{M}}^{13/2}, L_{\mathsf{KL}}^2, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, \frac{1}{\delta^2}, \frac{1}{\Delta_{\min}^3}, \mathsf{n}^{\mathcal{M}}_{\delta/6}, \sqrt{\mathsf{aec}_{\varepsilon/2}(\mathcal{M})}\right) \\ C_{\mathsf{aec}} &:= c \cdot \frac{V_{\mathcal{M}}^2 d_{\mathsf{cov}} \log(C_{\mathsf{cov}}) \cdot \max_{M\in\mathcal{M}} \mathsf{g}^M \cdot (\delta^{-1} + V_{\mathcal{M}} \mathsf{n}^{\mathcal{M}}_{\delta/6})}{\delta \Delta_{\min}^3} \cdot \log(C_{\mathsf{low}})\end{aligned}$$

*and where $\mathrm{lin}(\cdot)$ denotes a function linear and poly-logarithmic in its arguments and $c > 0$ is a universal constant.*

**Proof of Theorem F.1.** Letting $\mathcal{T}_{\mathrm{exploit}}$ denote the exploitation rounds, we can bound the total expected regret as

$$\sum_{t=1}^T \mathbb{E}[\Delta^\star(\pi^t)] = \mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{exploit}}} \Delta^\star(\pi^t)\right] + \mathbb{E}\left[\sum_{s=1}^{s_T} \Delta^\star(\pi^s)\right] \le 2\log\log T + 3 + \mathbb{E}\left[\sum_{s=1}^{s_T} \Delta^\star(\pi^s)\right],$$

where the inequality follows from Lemma F.1. It remains to bound the regret in the exploration rounds. By Lemma F.3, we can bound this as

$$\begin{aligned}\mathbb{E}\left[\sum_{s=1}^{s_T} \Delta^\star(\pi^s)\right] \le \frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta \underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot \mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] \\ + (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}\left[\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\right].\end{aligned}$$

Applying Lemma F.4 with $\alpha = 0$, we can bound

$$\mathbb{E}\left[\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\right] \le \log T + \mathbb{E}\left[V_{\mathcal{M}} \sqrt{1344 d_{\mathsf{cov}} s_T \cdot \log(128 C_{\mathsf{cov}} s_T)}\right.$$

$$
\begin{aligned}
&+ (V_{\mathcal{M}} + L_{\mathsf{KL}})\left(4\sqrt{s_T} + \sum_{s=1}^{s_T} \sqrt{s} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\right)\Bigg] + \log\log T + 7V_{\mathcal{M}} \\
&\le \log T + V_{\mathcal{M}}\sqrt{1344 d_{\mathsf{cov}} \mathbb{E}[s_T] \cdot \log(128 C_{\mathsf{cov}} \mathbb{E}[s_T])} \\
&\quad + (V_{\mathcal{M}} + L_{\mathsf{KL}})\left(4\sqrt{\mathbb{E}[s_T]} + \mathbb{E}\left[\sum_{s=1}^{s_T} \sqrt{s} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\right]\right) + \log\log T + 7V_{\mathcal{M}}
\end{aligned} \tag{52}
$$

where the last inequality follows from Lemma F.16—applied with $\alpha = \beta = 1/2$, $a = 128C_{\mathsf{cov}}$—and the concavity of $\sqrt{x}$. Note that the condition of Lemma F.16 is met here since we assume $C_{\mathsf{cov}} \ge 1$. It follows from Lemma F.7 that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{s_T} \sqrt{s} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\right] \\
&\le \mathbb{E}\Big[(2 + 6V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}} s_T) + 1) \cdot 5\sqrt{2 s_T \log(2 s_T)}\Big] \\
&\qquad + \mathbb{E}[32(1 + V_{\mathcal{M}}) \log(s_T)] + 8 \\
&\le (2 + 6V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}} \mathbb{E}[s_T]) + 1) \cdot 5\sqrt{2\mathbb{E}[s_T] \log(2\mathbb{E}[s_T])} \\
&\qquad + 32(1 + V_{\mathcal{M}}) \log \mathbb{E}[s_T] + 8 \\
&= O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \sqrt{\mathbb{E}[s_T] \log(\mathbb{E}[s_T])} + V_{\mathcal{M}} \sqrt{\mathbb{E}[s_T]}\Big),
\end{aligned} \tag{53}
$$

where the last inequality follows from Lemma F.16. Similarly, by Lemma F.8, we can bound both $\mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)]$ and $\mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)]$ as

$$
\begin{aligned}
\mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)], \mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] &\le \mathbb{E}\Big[(2 + 5V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}} s_T) + 1) \cdot \sqrt{\log(2 s_T)}\Big] \\
&\qquad + \mathbb{E}[32(1 + V_{\mathcal{M}}) \log(s_T)] + 8 \\
&\le (2 + 5V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}} \mathbb{E}[s_T]) + 1) \cdot \sqrt{\log(2\mathbb{E}[s_T])} \\
&\qquad + 32(1 + V_{\mathcal{M}}) \log(\mathbb{E}[s_T]) + 8 \\
&= O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}} \log(\mathbb{E}[s_T])\Big),
\end{aligned} \tag{54}
$$

where the second inequality follows from Lemma F.16 and Jensen's inequality, which lets us pass the expectation through. Together this gives a regret bound of

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{s_T} \Delta^\star(\pi^s)\right] \\
&\le (1+\delta)\mathsf{g}^\star \cdot \log T + (1+\delta)\mathsf{g}^\star V_{\mathcal{M}} \sqrt{1344 d_{\mathsf{cov}} \mathbb{E}[s_T] \cdot \log(128 C_{\mathsf{cov}} \mathbb{E}[s_T])} + (1+\delta)\mathsf{g}^\star(\log\log T + 7V_{\mathcal{M}}) \\
&\quad + (1+\delta)\mathsf{g}^\star(V_{\mathcal{M}} + L_{\mathsf{KL}}) \cdot O\Big(\sqrt{\mathbb{E}[s_T]} + V_{\mathcal{M}} d_{\mathsf{cov}} \cdot \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \sqrt{\mathbb{E}[s_T] \log(\mathbb{E}[s_T])} + V_{\mathcal{M}} \log \mathbb{E}[s_T]\Big) \\
&\quad + \frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta \underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}} \log(\mathbb{E}[s_T])\Big)
\end{aligned}
$$

$$+ \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}}\mathbb{E}[s_T]) \sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}} \log(\mathbb{E}[s_T])\Big)$$

By Lemma F.3 we can bound the total number of exploration rounds by

$$\begin{aligned}
\mathbb{E}[s_T] &\le \frac{24\mathsf{n}_{\max}^2 + 8\mathsf{n}_{\max}\underline{\mathsf{g}}^{\mathcal{M}}}{(\delta\underline{\mathsf{g}}^{\mathcal{M}})^2} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot \mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{12\mathsf{n}_{\max}}{\delta\Delta_{\min}} \cdot \mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] \\
&\quad + \frac{6\mathsf{n}_{\max}}{\delta} \cdot \mathbb{E}\left[\sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\right] \\
&\le \frac{24\mathsf{n}_{\max}^2 + 8\mathsf{n}_{\max}\underline{\mathsf{g}}^{\mathcal{M}}}{(\delta\underline{\mathsf{g}}^{\mathcal{M}})^2} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}}\mathbb{E}[s_T]) \sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}} \log(\mathbb{E}[s_T])\Big) \\
&\quad + \frac{12\mathsf{n}_{\max}}{\delta\Delta_{\min}} \cdot O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}}\mathbb{E}[s_T]) \sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}} \log(\mathbb{E}[s_T])\Big) \\
&\quad + \frac{6\mathsf{n}_{\max}}{\delta} \cdot (\log T + V_{\mathcal{M}} \sqrt{1344 d_{\mathsf{cov}} \mathbb{E}[s_T] \cdot \log(128 C_{\mathsf{cov}}\mathbb{E}[s_T])} + \log\log T + 7V_{\mathcal{M}}) \\
&\quad + \frac{6\mathsf{n}_{\max}}{\delta}(V_{\mathcal{M}} + L_{\mathsf{KL}})\Big(4\sqrt{\mathbb{E}[s_T]} + O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}}\mathbb{E}[s_T]) \sqrt{\mathbb{E}[s_T]\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\sqrt{\mathbb{E}[s_T]}\Big)\Big)
\end{aligned}$$

where the second inequality follows from Equations (52) to (54). Using Lemma F.17 and bounding $\underline{\mathsf{g}}^{\mathcal{M}} \le \mathsf{n}_{\max}$, we can solve this for $\mathbb{E}[s_T]$ to get

$$\begin{aligned}
\mathbb{E}[s_T] \le \widetilde{O}\Bigg( &\frac{\mathsf{n}_{\max}}{\delta} \cdot \log T + \frac{V_{\mathcal{M}} d_{\mathsf{cov}} \log C_{\mathsf{cov}} \cdot \mathsf{n}_{\max}^2}{(\delta\underline{\mathsf{g}}^{\mathcal{M}})^2} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) + \frac{\mathsf{n}_{\max} V_{\mathcal{M}} d_{\mathsf{cov}} \log C_{\mathsf{cov}}}{\delta\Delta_{\min}} \\
&+ \frac{\mathsf{n}_{\max}^2 (V_{\mathcal{M}} + L_{\mathsf{KL}})^2 (V_{\mathcal{M}}^2 d_{\mathsf{cov}}^2 \log^2 C_{\mathsf{cov}} + V_{\mathcal{M}}^2)}{\delta^2}\Bigg).
\end{aligned}$$

Finally, by Lemma F.13 and Lemma E.3 we can lower bound $\underline{\mathsf{g}}^{\mathcal{M}} \ge \Delta_{\min}/2V_{\mathcal{M}}$. Plugging this into the above expression and using that

$$\mathsf{n}_{\max} = \mathsf{n}_{\max}(\mathcal{M}, \delta/6) := \frac{32}{\Delta_{\min}^2} \cdot \left(\frac{6}{\delta} + 2V_{\mathcal{M}} \mathsf{n}^{\mathcal{M}}_{\delta/6}\right) \cdot \max_{M \in \mathcal{M}} \mathsf{g}^M,$$

gives the result, after simplifying.

□

### F.2. Regret Bound without Uniform Regularity (Theorem A.2)

In this section, we prove Theorem A.2 (as well as a more general result, Theorem F.2), which gives regret bounds for a variant of $\mathsf{AE}^2$, Algorithm 6, $\mathsf{AE}^2_\star$, which does not require uniform regularity, and adapts to the gap $\Delta^\star_{\min} := \Delta^{M^\star}_{\min}$ for the underlying model $M^\star$. Algorithm 6 is a slightly more general version of Algorithm 3, incorporating general divergences $D$.

Throughout this section, we let $s^\star$ denote the first exploration round $s$ such that $M^\star \in \mathcal{M}^\ell$ in Algorithm 3. Note that $s^\star$ is a deterministic quantity (though, the first round $t$ in which $s = s^\star$ is not deterministic).

We remark briefly that, to bound Eq. (55) by a restriction of the AEC, it is critical that our estimator produced at each exploration round $s$, $\xi^s$, is only supported on $\mathcal{M}^\ell$. To accomplish this, we explicitly generate a cover of $\mathcal{M}^\ell$, $\mathcal{M}^\ell_{\mathrm{cov}}$, and run an estimation procedure on this cover. As Algorithm 7, the estimation procedure used to prove Theorem A.1, directly covers all of $\mathcal{M}$, to prove Theorem A.2 we do not appeal directly to this algorithm, yet we note that the covering and estimation procedure employed by $\mathsf{AE}^2_\star$ are essentially identical to that in Algorithm 7, modulo the choice of which set is being covered.

**Algorithm 6** Adaptive Exploration for Allocation Estimation for classes without uniform regularity ($\mathsf{AE}^2_\star$)

**input:** Optimality tolerance $\delta$, divergence $D$, estimation oracle $\mathbf{Alg}_D$, growth parameters $\alpha_q$, $\alpha_n$, $\alpha_{\mathcal{M}}$.
$s \leftarrow 1$, $\ell \leftarrow 1$, $\varepsilon \leftarrow \frac{\delta}{4+2\delta}$.
$q^s \leftarrow 1 - s^{-\alpha_q}$, $\mathsf{n}^s \leftarrow s^{\alpha_{\mathsf{n}}}$.
Compute $\xi^1 \leftarrow \mathbf{Alg}_D(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
**for** $t = 1, 2, 3, \ldots$ **do**
  **if** $s \geq 2^\ell$ **then** // Form active set and cover
    $\ell \leftarrow \ell + 1$.
    $\Delta^\ell \leftarrow \arg\min_{\Delta\geq 0} \Delta \quad \text{s.t.} \quad \overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}_{\Delta,\frac{1}{\Delta}}) \leq s^{\alpha_{\mathcal{M}}}$.
    $\mathcal{M}^\ell \leftarrow \mathcal{M}_{\Delta^\ell,\frac{1}{\Delta^\ell}} \cap \left\{M \in \mathcal{M} \;:\; \mathsf{n}^M_\varepsilon + \frac{1}{\Delta^M_{\min}} + \frac{4\mathsf{g}^M}{\Delta^M_{\min}} + \frac{2\mathsf{n}^M_\varepsilon}{\mathsf{g}^M} + \frac{4}{\Delta^M_{\min}\varepsilon} \leq \sqrt{\mathsf{n}^s}\right\}$.
    $\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}^\ell$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
  **if** $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then**
    Play $\pi_{\widehat{M}^s}$. // Exploit
  **else** // Explore
    Set $p^s \leftarrow q^s\lambda^s + (1-q^s)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \arg\min_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}^\ell\setminus\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega}[D(\widehat{M}(\pi)\,\|\,M(\pi))]]}. \tag{55}$$

    Draw $\pi^s \sim p^s$, observe $r^s, o^s$, set $\mathfrak{D}^\ell \leftarrow \mathfrak{D}^\ell \cup \{(\pi^s, r^s, o^s)\}$.
    Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_D(\mathfrak{D}^\ell, \mathcal{M}^\ell_{\mathrm{cov}})$ and $\widehat{M}^{s+1} = \mathbb{E}_{M\sim\xi^{s+1}}[M]$.
    $s \leftarrow s + 1$.

### F.2.1. BOUNDING REGRET OF EXPLORE PHASE

**Lemma F.5.** *We have*

$$s^\star \leq \left(\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star)\right)^{\frac{1}{\alpha_{\mathcal{M}}}} + \left(\mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon}\right)^{\frac{2}{\alpha_{\mathsf{n}}}}.$$

**Proof of Lemma F.5.** We will have $M^\star \in \mathcal{M}^\ell$ as soon as $M^\star \in \mathcal{M}'$ and $M^\star \in \mathcal{M}''$ for

$$\mathcal{M}' \leftarrow \left\{M \in \mathcal{M} \;:\; \mathsf{n}^M_\varepsilon + \frac{1}{\Delta^M_{\min}} + \frac{4\mathsf{g}^M}{\Delta^M_{\min}} + \frac{2\mathsf{n}^M_\varepsilon}{\mathsf{g}^M} + \frac{4}{\Delta^M_{\min}\varepsilon} \leq \sqrt{\mathsf{n}^{2^{\ell-1}}}\right\}, \quad \mathcal{M}'' \leftarrow \mathcal{M}_{\Delta^\ell,\frac{1}{\Delta^\ell}}$$

where

$$\Delta^\ell = \arg\min_{\Delta \geq 0} \Delta \quad \text{s.t.} \quad \overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}_{\Delta,\frac{1}{\Delta}}) \leq s^{\alpha_{\mathcal{M}}}.$$

Note that if this occurs for some $\ell$, then the first exploration round $s$ such that $M^\star \in \mathcal{M}^\ell$ is $s = 2^{\ell-1}$. From the definition $\mathsf{n}^{2^{\ell-1}} = 2^{\alpha_{\mathsf{n}}(\ell-1)}$, we have $M^\star \in \mathcal{M}'$ as soon as

$$\begin{aligned}\sqrt{2^{\alpha_{\mathsf{n}}(\ell-1)}} &\geq \mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon} \\ \iff 2^{\ell-1} &\geq \left(\mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon}\right)^{2/\alpha_{\mathsf{n}}}.\end{aligned}$$

To have $M^\star \in \mathcal{M}''$, we need $\Delta^\ell \leq \Delta^\star_{\min}$ and $\mathsf{n}^\star_\varepsilon \leq \frac{1}{\Delta^\ell}$, which will be the case once

$$\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star) \leq 2^{(\ell-1)\cdot\alpha_{\mathcal{M}}} \iff \left(\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star)\right)^{\frac{1}{\alpha_{\mathcal{M}}}} \leq 2^{\ell-1}.$$

The result then follows by combining these bounds. □

**Lemma F.6.** *Let $s_T$ denote the total number of exploration rounds. For $\delta \leq 1/2$, running Algorithm 6, we can almost surely bound bound*

$$\begin{aligned}s_T \leq \frac{4}{\delta\mathsf{g}^\star}\Bigg(&\sum_{s=s^\star}^{s_T} 2s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\ &+ \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}}\cdot 2\mathsf{g}^\star \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\ &+ \sum_{s=s^\star}^{s_T} s^{3\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star\big(\tfrac{8}{\delta\mathsf{g}^\star}\big)^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q-\alpha_{\mathsf{n}}}}\Bigg) + s^\star.\end{aligned}$$

*In addition, the regret during exploration rounds is bounded as*

$$\begin{aligned}\mathbb{E}\Bigg[\sum_{s=s^\star}^{s_T}\Delta^\star(\pi^s)\Bigg] \leq \mathbb{E}\Bigg[&\sum_{s=s^\star}^{s_T} 2s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\ &+ \sum_{s=s^\star}^{s_T}(1+\delta)\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\} \\ &+ \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\Bigg] + 4\mathsf{g}^\star\big(\tfrac{8}{\delta\mathsf{g}^\star}\big)^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}.\end{aligned}$$

**Proof of Lemma F.6.** We first prove the bound on the regret during the exploration rounds, then use this result to prove the bound on $s_T$.

Recall that by definition, for exploration rounds, we have $p^s \leftarrow q^s\lambda^s + (1-q^s)\omega^s$. We consider three cases to bound the instantaneous expected regret, $\Delta^\star(p^s)$, for each $s \geq s^\star$.

**Case 1:** $M^\star \in \mathcal{M}^\ell \backslash \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}^s)$**.** Denote such rounds as $\mathcal{S}^1_{\mathrm{exp}}$. This case follows identically to Case 1 in Lemma F.3 with

$$\gamma^s = \frac{1+\mathsf{g}^\star\delta}{1+q^s} \cdot \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]}. \tag{56}$$

We therefore omit the proof and conclude that

$$\begin{aligned}\Delta^\star(p^s) &\le -\mathsf{g}^\star\delta + \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]] \\ &\le \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)].\end{aligned}$$

As regret is always lower-bounded by 0, we have $\Delta^\star(p^s) \ge 0$, so for rounds $s \in \mathcal{S}^1_{\mathrm{exp}}$, we can also write

$$\mathsf{g}^\star\delta \le \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]. \tag{57}$$

**Case 2:** $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}^s)$**,** $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^2_{\mathrm{exp}}$, and write

$$\Delta^\star(p^s) = [\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]] + (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]$$

for an arbitrary model $M \in \mathcal{M}^\star_{\mathrm{alt}}$. In this case, since $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}^s)$, we have that $\lambda^s \in \Lambda(M^\star; \varepsilon)$. This implies that

$$\Delta^\star(\lambda^s) \le (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s \quad \text{and} \quad \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \ge (1-\varepsilon)/\mathsf{n}^\star_s$$

for some $\mathsf{n}^\star_s \le \mathsf{n}^s$.

Since $M \in \mathcal{M}^\star_{\mathrm{alt}}$, it follows that $\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \ge (1-\varepsilon^s)/\mathsf{n}^\star_s$. Thus,

$$\begin{aligned}\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] &\le q^s\left[\Delta^\star(\lambda^s) - (1+\delta)\mathsf{g}^\star\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]\right] + 1 - q^s \\ &\le q^s\left[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s\right] + 1 - q^s \\ &= q^s\left[2\varepsilon - \delta(1-\varepsilon)\right] \cdot \frac{\mathsf{g}^\star}{\mathsf{n}^\star_s} + 1 - q^s \\ &\overset{(a)}{=} -\frac{(1-s^{-\alpha_q})\delta}{2} \cdot \frac{\mathsf{g}^\star}{\mathsf{n}^\star_s} + s^{-\alpha_q} \\ &\overset{(b)}{\le} \begin{cases} s^{-\alpha_q} & s < (\frac{8\mathsf{n}^\star_s}{\delta\mathsf{g}^\star})^{1/\alpha_q} \\ -\frac{\delta\mathsf{g}^\star}{4\mathsf{n}^\star_s} & s \ge (\frac{8\mathsf{n}^\star_s}{\delta\mathsf{g}^\star})^{1/\alpha_q} \end{cases} \\ &\overset{(c)}{\le} \begin{cases} s^{-\alpha_q} & s < (\frac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}} \\ -\frac{1}{4}\delta\mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}} & s \ge (\frac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}} \end{cases} \\ &\overset{(d)}{\le} 2\mathsf{g}^\star\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}}\end{aligned}$$

where $(a)$ follows from our choice of $\varepsilon = \frac{\delta/2}{2+\delta}$ and $q^s = 1 - s^{-\alpha_q}$, $(b)$ follows from some algebra, $(c)$ uses that $\mathsf{n}^\star_s \le \mathsf{n}^s$ and $\mathsf{n}^s = s^{\alpha_{\mathsf{n}}}$, and $(d)$ follows since we will always have $s^{-\alpha_q} \le 2\mathsf{g}^\star - \frac{1}{4}\delta\mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}}$. Thus:

$$\Delta^\star(p^s) \le (1+\delta)\mathsf{g}^\star\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}}$$

$$\leq (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}.$$

As this holds for all $M \in \mathcal{M}^\star_{\mathrm{alt}}$, we have

$$\Delta^\star(p^s) \leq (1+\delta)\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}.$$

Since $\Delta^\star(p^s) \geq 0$, this also implies that for $s \in \mathcal{S}^2_{\mathrm{exp}}$:

$$\begin{aligned}\frac{1}{4}\delta \mathsf{g}^\star &\leq s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star s^{\alpha_{\mathsf{n}}} \mathbb{I}\{s < (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\} \\ &\leq s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{\alpha_{\mathsf{n}}}{\alpha_q - \alpha_{\mathsf{n}}}} \cdot \mathbb{I}\{s < (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}.\end{aligned} \tag{58}$$

**Case 3:** $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}^s)$**,** $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^3_{\mathrm{exp}}$, and write

$$\begin{aligned}\Delta^\star(p^s) = \Big[\Delta^\star(p^s) - 2(1+\delta)\mathsf{g}^\star \sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]\Big] \\ + 2(1+\delta)\mathsf{g}^\star \sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]].\end{aligned}$$

Since $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s)$, we have that $\lambda^s \in \Lambda(M^\star; \varepsilon)$. This implies that for any $M \in \mathcal{M}^\star_{\mathrm{alt}}$:

$$\Delta^\star(\lambda^s) \leq (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s \quad \text{and} \quad \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \geq (1-\varepsilon)/\mathsf{n}^\star_s$$

for some $\mathsf{n}^\star_s \leq \mathsf{n}^s$. Assume that we are at epoch $\ell$. By construction we have that, for $M \in \mathcal{M}^\ell$, $\frac{1}{\Delta^M_{\min}} \leq \sqrt{\mathsf{n}^{2^\ell}} \iff \Delta^M_{\min} \geq \frac{1}{\sqrt{\mathsf{n}^{2^\ell}}}$. Since $\mathsf{n}^s$ is increasing in $s$, this implies that $\Delta^M_{\min} \geq \frac{1}{\sqrt{\mathsf{n}^s}}$. As $\xi^s$ is only supported on $\mathcal{M}^\ell$, since $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$, Lemma F.9 implies that $\mathbb{E}_{M \sim \xi^s}[\mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}] \geq \frac{1}{2\sqrt{\mathsf{n}^s}}$. Thus, we have

$$\begin{aligned}&2(1+\delta)\mathsf{g}^\star \sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] \\ &\geq 2(1+\delta)\mathsf{g}^\star \sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big) \cdot \mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}]] \\ &\geq 2(1+\delta)\mathsf{g}^\star \sqrt{\mathsf{n}^s} \cdot \frac{1-\varepsilon}{2\sqrt{\mathsf{n}^s}\mathsf{n}^\star_s} \\ &\geq \frac{(1+\delta)(1-\varepsilon)\mathsf{g}^\star}{\mathsf{n}^\star_s}.\end{aligned}$$

This implies that

$$\begin{aligned}&\Delta^\star(p^s) - 2(1+\delta)\mathsf{g}^\star \sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] \\ &\leq q^s[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s] + 1 - q^s \\ &\leq 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta \mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}},\end{aligned}$$

where the final inequality follows by the same argument as in Case 2. Thus,

$$\Delta^\star(p^s) \leq 2(1+\delta)\mathsf{g}^\star \sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta \mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}}$$

$$= s^{\alpha_{\mathsf{n}}/2} \cdot 2(1+\delta)\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}}$$
$$\leq s^{\alpha_{\mathsf{n}}/2} \cdot 2(1+\delta)\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}.$$

Just as in Case 2, using that $\Delta^\star(p^s) \geq 0$, this also implies that for $s \in \mathcal{S}^3_{\mathrm{exp}}$:

$$\frac{1}{4}\delta\mathsf{g}^\star \leq s^{3\alpha_{\mathsf{n}}/2} \cdot 2(1+\delta)\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{\alpha_{\mathsf{n}}}{\alpha_q - \alpha_{\mathsf{n}}}} \cdot \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}. \tag{59}$$

**Completing the Proof.** In total we have

$$\sum_{s=s^\star}^{s_T} \Delta^\star(p^s) \leq \sum_{s\in\mathcal{S}^1_{\mathrm{exp}}} \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)] + \sum_{s\in\mathcal{S}^2_{\mathrm{exp}}} (1+\delta)\mathsf{g}^\star \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]$$
$$+ \sum_{s\in\mathcal{S}^3_{\mathrm{exp}}} s^{\alpha_{\mathsf{n}}/2} \cdot 2(1+\delta)\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}.$$

By Lemma F.11, for $s \in \mathcal{S}^1_{\mathrm{exp}}$, we can bound $\gamma^s \leq (1+\mathsf{g}^\star\delta) \cdot s^{\alpha_q + \alpha_{\mathcal{M}}}$. This gives an upper bound on the above of

$$\leq \sum_{s=s^\star}^{s_T} (1+\mathsf{g}^\star) s^{\alpha_q+\alpha_{\mathcal{M}}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]$$
$$+ \sum_{s=s^\star}^{s_T} (1+\delta)\mathsf{g}^\star \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}$$
$$+ \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}/2} \cdot 4\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}},$$

which proves the regret bound.

We now bound the number of exploration rounds. Since for every $s \in \mathcal{S}^1_{\mathrm{exp}}$ Eq. (57) holds, for every $s \in \mathcal{S}^2_{\mathrm{exp}}$ Eq. (58) holds, and for every $s \in \mathcal{S}^3_{\mathrm{exp}}$ Eq. (59) holds, combining these inequalities gives

$$\frac{1}{4}\delta\mathsf{g}^\star |\mathcal{S}^1_{\mathrm{exp}} \cup \mathcal{S}^2_{\mathrm{exp}} \cup \mathcal{S}^3_{\mathrm{exp}}|$$
$$\leq \sum_{s\in\mathcal{S}^1_{\mathrm{exp}}} \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)] + \sum_{s\in\mathcal{S}^2_{\mathrm{exp}}} s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]$$
$$+ \sum_{s\in\mathcal{S}^3_{\mathrm{exp}}} s^{3\alpha_{\mathsf{n}}/2} \cdot 4\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q - \alpha_{\mathsf{n}}}}$$
$$\leq \sum_{s=s^\star}^{s_T} (1+\mathsf{g}^\star) s^{\alpha_q+\alpha_{\mathcal{M}}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]$$
$$+ \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}$$

$$+\sum_{s=s^\star}^{s_T} s^{3\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]+4\mathsf{g}^\star(\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q-\alpha_{\mathsf{n}}}}.$$

Using that $|\mathcal{S}^1_{\mathrm{exp}}\cup\mathcal{S}^2_{\mathrm{exp}}\cup\mathcal{S}^3_{\mathrm{exp}}|=s_T-s^\star$ and rearranging gives the bound on $s_T$.

□

### F.2.2. COMPLETING THE PROOF

**Theorem F.2** (Full version of Theorem A.2)**.** *Consider Algorithm 6, and suppose we set $\delta\le 1/2$, $D\big(\cdot\,\|\,\cdot\big)=D_{\mathsf{KL}}(\cdot\,\|\,\cdot)$, $\alpha_{\mathcal{M}}=1/2$, $\alpha_\zeta=1/8$, and $\alpha_q=1/4$, and instantiate $\mathbf{Alg}_D$ with Algorithm 4. Then if Assumptions 1.1 to 1.3 and A.1 to A.3 hold and $\mathsf{g}^\star>0$, the expected regret of is bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)]\le(1+\delta)\mathsf{g}^\star\log T+C_{\mathsf{aec}}\cdot\Big(\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star)\Big)^3\cdot\log^{3/2}\log T+C_{\mathsf{low}}\cdot\log^{6/7}T$$

*for $\varepsilon\leftarrow\frac{\delta}{4+2\delta}$,*

$$C_{\mathsf{aec}}:=\widetilde{O}\bigg(\frac{(V_{\mathcal{M}}+L_{\mathsf{KL}})V^3_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}})}{\delta\Delta^\star_{\min}}\bigg),$$

*and*

$$C_{\mathsf{low}}:=\widetilde{O}\bigg(\mathrm{poly}\bigg(V_{\mathcal{M}},L_{\mathsf{KL}},\mathsf{n}^\star_\varepsilon,d_{\mathsf{cov}},\log C_{\mathsf{cov}},\mathsf{g}^\star,\frac{1}{\Delta^\star_{\min}},\frac{1}{\delta},\log\log T\bigg)\bigg).$$

**Proof of Theorem F.2.** The bound on the regret incurred in the exploit phase follows identically to Theorem F.1, since the exploit test is the same. We turn to bounding the regret in the explore phase. First, since we can incur regret of at most 1 at every round, we bound

$$\mathbb{E}\bigg[\sum_{s=1}^{s_T}\Delta^{M^\star}(p^s)\bigg]\le\mathbb{E}\bigg[\sum_{s=s^\star}^{s_T}\Delta^{M^\star}(p^s)\bigg]+s^\star$$

By Lemma F.6, we can bound

$$\begin{aligned}\mathbb{E}\bigg[\sum_{s=s^\star}^{s_T}\Delta^\star(\pi^s)\bigg]&\le\mathbb{E}\bigg[\sum_{s=s^\star}^{s_T}(1+\mathsf{g}^\star)s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\\&\quad+\sum_{s=s^\star}^{s_T}(1+\delta)\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\\&\quad+\sum_{s=s^\star}^{s_T}s^{\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\bigg]+4\mathsf{g}^\star(\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\\&\le\mathbb{E}\bigg[\sum_{s=s^\star}^{s_T}(1+\mathsf{g}^\star)s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\\&\quad+\sum_{s=1}^{s_T}(1+\delta)\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\end{aligned}$$

$$+ \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}/2} \cdot 4\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]\Bigg] + 4\mathsf{g}^\star \big(\tfrac{8}{\delta \mathsf{g}^\star}\big)^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}.$$

Applying Lemma F.4 with $\alpha = 0$, we have

$$\mathbb{E}\Bigg[\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\Bigg] \le \log T + \mathbb{E}\Bigg[V_{\mathcal{M}} s_T^{1/2} \cdot \sqrt{1344 d_{\mathsf{cov}} \cdot \log(128 C_{\mathsf{cov}} s_T)}$$
$$+ (V_{\mathcal{M}} + L_{\mathsf{KL}})\Bigg(4 s_T^{1/2} + \sum_{s=1}^{s_T} s^{1/2} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\Bigg)\Bigg] + \log\log T + 7V_{\mathcal{M}}.$$

Note that for $s \ge s^\star$, our estimator is applied to a cover of a set containing $M^\star$. Furthermore, note that the estimation procedure used in Algorithm 6 is, other than the different choice of set to cover, identical to that used in Algorithm 7. It follows that the analysis of Algorithm 7 can be applied to the estimation procedure of Algorithm 6, with only the mild modification accounting for the difference in the size of the cover (as we are covering $\mathcal{M}^\ell$ instead of $\mathcal{M}$). However, as we can upper bound the size of the cover of $\mathcal{M}^\ell$ by the size of the cover of $\mathcal{M}$ via Lemma F.15, this change is inconsequential. Thus, by Lemma F.7, we can bound

$$\mathbb{E}\Bigg[\sum_{s=s^\star}^{s_T} 2 s^{\alpha_q + \alpha_{\mathcal{M}}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\Bigg] \le O\Big(\mathbb{E}\Big[V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} s_T) s_T^{\alpha_q+\alpha_{\mathcal{M}}} \sqrt{\log(s_T)}\Big]\Big)$$
$$\le O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \mathbb{E}[s_T]^{\alpha_q+\alpha_{\mathcal{M}}} \sqrt{\log(\mathbb{E}[s_T])}\Big), \tag{60}$$

where the second inequality uses Jensen's inequality and Lemma F.16 to pass the expectation through, which holds as long as $1/100 \le \alpha_q + \alpha_{\mathcal{M}} \le 3/4$. Similarly,

$$\mathbb{E}\Bigg[\sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}/2} \cdot 4\mathsf{g}^\star \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]\Bigg] \le O\Big(\mathsf{g}^\star V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \mathbb{E}[s_T]^{\alpha_{\mathsf{n}}/2} \sqrt{\log(\mathbb{E}[s_T])}\Big), \tag{61}$$

where we have again used Jensen's inequality and Lemma F.16 to pass the expectation through, which holds as long $1/100 \le \alpha_{\mathsf{n}}/2 \le 3/4$. For $s \le s^\star$, we do not have $M^\star \in \mathcal{M}^\ell$, and therefore the estimation guarantees are vacuous. In this regime, using that the KL divergence is always bounded by $2V_{\mathcal{M}}$ (Lemma F.13), we can simply upper bound the estimation error by $2V_{\mathcal{M}}$. Thus,

$$\mathbb{E}\Bigg[\sum_{s=1}^{s_T} s^{1/2} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\Bigg]$$
$$\le \mathbb{E}\Bigg[\sum_{s=s^\star}^{s_T} s^{1/2} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\Bigg] + 2V_{\mathcal{M}}(s^\star)^{3/2}$$
$$\le O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \mathbb{E}[s_T]^{1/2} \sqrt{\log(\mathbb{E}[s_T])}\Big) + 2V_{\mathcal{M}}(s^\star)^{3/2},$$

which gives, applying Lemma F.16 again:

$$\mathbb{E}\Bigg[\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\Bigg] \le \log T + O\Bigg(V_{\mathcal{M}} \mathbb{E}[s_T]^{1/2} \sqrt{d_{\mathsf{cov}} \cdot \log(C_{\mathsf{cov}} \mathbb{E}[s_T])}$$

$$+ (V_{\mathcal{M}} + L_{\mathsf{KL}})\Big(\mathbb{E}[s_T]^{1/2} + V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2}\sqrt{\log(\mathbb{E}[s_T])}\Big)\Big)$$
$$+ 2V_{\mathcal{M}}(s^\star)^{3/2} + \log\log T + 7V_{\mathcal{M}}. \tag{62}$$

Combining Eqs. (60) to (62), we have

$$\begin{aligned}\mathbb{E}\Big[\sum_{s=s^\star}^{s_T}\Delta^\star(\pi^s)\Big] \le (1+\delta)\mathsf{g}^\star\log T + O\Big(& V_{\mathcal{M}}\sqrt{d_{\mathsf{cov}}\mathbb{E}[s_T]\log(C_{\mathsf{cov}}\mathbb{E}[s_T])} \\ &+ (V_{\mathcal{M}} + L_{\mathsf{KL}})V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2}\sqrt{\log(\mathbb{E}[s_T])} \\ &+ V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\log\mathbb{E}[s_T]}((1+\mathsf{g}^\star)\mathbb{E}[s_T]^{\alpha_q+\alpha_{\mathcal{M}}} + \varepsilon\mathsf{g}^\star\mathbb{E}[s_T]^{\alpha_{\mathsf{n}}/2}) \\ &+ \mathsf{g}^\star(\tfrac{1}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}} + \log\log T + V_{\mathcal{M}}(s^\star)^{3/2}\Big).\end{aligned}$$

To control this, it remains to bound $s_T$. By Lemma F.6 we have the following almost sure bound:

$$\begin{aligned}s_T \le \frac{4}{\delta\mathsf{g}^\star}\Bigg( & \underbrace{\sum_{s=s^\star}^{s_T}(1+\mathsf{g}^\star)s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}_{(a)} \\ & + \underbrace{\sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}}\cdot 2\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}}_{(b)} \\ & + \underbrace{\sum_{s=1}^{s_T} s^{3\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star(\tfrac{4}{\delta\mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q-\alpha_{\mathsf{n}}}}}_{(c)}\Bigg) + s^\star.\end{aligned}$$

We bound the expectation of term $(a)$ as in Eq. (60). To bound term $(b)$, we apply Lemma F.4 with $\alpha = \alpha_{\mathsf{n}}$ to get

$$\begin{aligned}\mathbb{E}[(b)] \le \mathbb{E}[s_T^{\alpha_{\mathsf{n}}}]\log T + \mathbb{E}\Bigg[ & V_{\mathcal{M}} s_T^{1/2+\alpha_{\mathsf{n}}}\cdot\sqrt{1344 d_{\mathsf{cov}}\cdot\log(128 C_{\mathsf{cov}} s_T)} \\ & + (V_{\mathcal{M}} + L_{\mathsf{KL}})\Big(4\frac{s_T^{1/2+\alpha_{\mathsf{n}}/2}}{1-\alpha_{\mathsf{n}}} + \sum_{s=1}^{s_T} s^{1/2+\alpha_{\mathsf{n}}/2}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\Big)\Bigg] \\ & + \mathbb{E}[s_T^{\alpha_{\mathsf{n}}}]\log\log T + 7V_{\mathcal{M}}.\end{aligned}$$

Again applying Lemma F.7 and Lemma F.16, we have

$$\mathbb{E}[(c)] \le O\Big(\mathsf{g}^\star V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{3\alpha_{\mathsf{n}}/2}\sqrt{\log(\mathbb{E}[s_T])}\Big)$$

and

$$\mathbb{E}\Bigg[\sum_{s=1}^{s_T} s^{1/2+\alpha_{\mathsf{n}}/2}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\Bigg] \le O\Big(V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}/2}\sqrt{\log(\mathbb{E}[s_T])}\Big)$$

$$+ 2V_{\mathcal{M}}(s^\star)^{3/2+\alpha_{\mathsf{n}}/2},$$

as long as $1/100 \le \alpha_{\mathsf{n}} \le 1/4$. We therefore have (using Lemma F.16 to pass the expectation through):

$$\begin{aligned}\mathbb{E}[s_T] \le \frac{1}{\delta \mathsf{g}^\star} \cdot O\Bigg( &\mathbb{E}[s_T]^{\alpha_{\mathsf{n}}} \log T + V_{\mathcal{M}} \mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}} \sqrt{d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T])} + (V_{\mathcal{M}} + L_{\mathsf{KL}}) \mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}/2} \\ &+ (V_{\mathcal{M}} + L_{\mathsf{KL}}) V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}/2} \sqrt{\log(\mathbb{E}[s_T])} \\ &+ V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) (1+\mathsf{g}^\star) \mathbb{E}[s_T]^{\alpha_q + \alpha_{\mathcal{M}}} \sqrt{\log(\mathbb{E}[s_T])} \\ &+ \mathsf{g}^\star V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}} \mathbb{E}[s_T]) \mathbb{E}[s_T]^{3\alpha_{\mathsf{n}}/2} \sqrt{\log(\mathbb{E}[s_T])} + \mathsf{g}^\star (\tfrac{1}{\delta \mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q - \alpha_{\mathsf{n}}}} + V_{\mathcal{M}} (s^\star)^{3/2+\alpha_{\mathsf{n}}/2} \Bigg).\end{aligned}$$

We now set $\alpha_{\mathcal{M}} = 1/2$, $\alpha_{\mathsf{n}} = 1/8$, and $\alpha_q = 1/4$, and note that all of the preceding parameter restrictions are satisfied for these choices. Furthermore, this parameter choice implies that—using Lemma F.17 to handle log factors—we have

$$\mathbb{E}[s_T] \le \widetilde{O}\left( \frac{1}{(\delta \mathsf{g}^\star)^{8/7}} \log^{8/7} T + \mathrm{poly}\left( V_{\mathcal{M}}, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, L_{\mathsf{KL}}, \mathsf{g}^\star, \frac{1}{\delta}, \frac{1}{\mathsf{g}^\star} \right) + \frac{V_{\mathcal{M}} (s^\star)^{3/2+\alpha_{\mathsf{n}}/2}}{\delta \mathsf{g}^\star} \right).$$

Plugging this into the regret bound given above, we have

$$\begin{aligned}\mathbb{E}\left[ \sum_{s=s^\star}^{s_T} \Delta^\star(\pi^s) \right] \le (1+\delta)\mathsf{g}^\star \log T + \widetilde{O}\Bigg( &\mathrm{poly}\left( V_{\mathcal{M}}, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, L_{\mathsf{KL}}, \mathsf{g}^\star, \frac{1}{\delta}, \frac{1}{\mathsf{g}^\star}, \log\log T \right) \cdot \log^{6/7} T \\ &+ \frac{(1+1/\mathsf{g}^\star)(V_{\mathcal{M}} + L_{\mathsf{KL}}) V_{\mathcal{M}}^2 d_{\mathsf{cov}} \log(C_{\mathsf{cov}}) + V_{\mathcal{M}}}{\delta} \cdot (s^\star)^{3/2} \cdot \log^{3/2} \log T \Bigg).\end{aligned}$$

Finally, by Lemma F.5, we can bound $s^\star$ as

$$s^\star \le \left( \overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star) \right)^{\frac{1}{\alpha_{\mathcal{M}}}} + \left( \mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon} \right)^{\frac{2}{\alpha_{\mathsf{n}}}},$$

and, by Lemma E.3 and Lemma F.13, we can lower bound $\mathsf{g}^\star \ge \Delta^\star_{\min}/2V_{\mathcal{M}}$. Plugging this in gives the final bound. □

### F.3. Estimation Guarantees

In this section, we analyze Algorithm 7, which is a variant of the Tempered Aggregation algorithm (Algorithm 4) designed for infinite classes. This algorithm is used within Algorithm 5 in order to prove Theorem F.1.

Algorithm 7 simply applies Algorithm 4 to a sequence of covers for the class $\mathcal{M}$ on a doubling epoch schedule. In particular, at every epoch $\ell$, Algorithm 7 restarts the Tempered Aggregation algorithm (Algorithm 4), clearing the Tempered Aggregation instance from the previous epoch from memory. We denote the $\ell$th instantiation of Tempered Aggregation as $\mathsf{TemperedAggregation}^\ell$.

**Algorithm 7** Estimation with Adaptive Covering

**input:** Class $\mathcal{M}$.
$\ell \leftarrow 1$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
$\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$.
Initialize $\mathsf{TemperedAggregation}^\ell$ as an instance of Algorithm 4 with $\mathcal{M}^\ell_{\mathrm{cov}}$.
**for** $s = 1, 2, 3, \ldots$ **do**
  Receive $(\pi^s, r^s, o^s)$, $\mathfrak{D}^\ell \leftarrow \mathfrak{D}^\ell \cup \{(\pi^s, r^s, o^s)\}$.
  $\xi^s \leftarrow \mathsf{TemperedAggregation}^\ell(\mathfrak{D}^\ell)$, $\widehat{M}^s \leftarrow \mathbb{E}_{M\sim\xi^s}[M]$
  **if** $s \geq 2^\ell$ **then**.
    $\ell \leftarrow \ell + 1$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
    $\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$.
    Initialize $\mathsf{TemperedAggregation}^\ell$ as an instance of Algorithm 4 with $\mathcal{M}^\ell_{\mathrm{cov}}$.

**Lemma F.7.** *Let $\tau$ denote some stopping time with respect to the filtration $(\mathscr{F}^t)_{t=1}^T$ such that $\tau \leq T$ almost surely, and let $\alpha \in (0,1)$. When running Algorithm 7 under Assumption A.3, we have*

$$\mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right] \leq \mathbb{E}\left[(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1) \cdot \frac{2^{2\alpha}}{2^\alpha - 1}\tau^\alpha\right] + 4.$$

*In addition, if Assumption A.2 also holds, then*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]]\right] \\
&\quad \leq \mathbb{E}\left[(2 + 6V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1) \cdot \frac{2^{2\alpha}}{2^\alpha - 1}\tau^\alpha \sqrt{\log(2\tau)}\right] \\
&\qquad + \mathbb{E}[32(1 + V_{\mathcal{M}})\log(\tau)] + 8
\end{aligned} \tag{63}$$

*and*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M(\pi) \,\|\, M^\star(\pi))]]\right] \\
&\quad \leq \mathbb{E}\left[(2 + 6V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1) \cdot \frac{2^{2\alpha}}{2^\alpha - 1}\tau^\alpha \sqrt{\log(2\tau)}\right] \\
&\qquad + \mathbb{E}[32(1 + V_{\mathcal{M}})\log(\tau)] + 8.
\end{aligned} \tag{64}$$

**Proof of Lemma F.7.** Let $\mathcal{S}^k$ denote the set of $s$ values for which $\ell = k$ and note that $\mathcal{S}^k = \{2^{k-1}+1, 2^{k-1}+2, \ldots, 2^k\}$. We can bound

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right] \\
&\quad \leq \sum_{k=1}^{\lceil \log_2 T\rceil} \mathbb{E}\left[\sum_{s\in\mathcal{S}^k} s^\alpha \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s \leq \tau\}\right]
\end{aligned} \tag{65}$$

$$\leq \sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\alpha k} \cdot \mathbb{E}\left[\sum_{s \in \mathcal{S}^k} \mathbb{E}_{M \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s \leq \tau\}\right] \tag{66}$$

since we have that $\ell \leq \lceil \log_2 T \rceil$ by construction, and since $s \leq 2^k$ for $s \in \mathcal{S}^k$. Let $A_k$ denote the event

$$A_k := \left\{\forall s \in \mathcal{S}^k \ : \ \sum_{i=2^{k-1}}^{s} \mathbb{E}_{M \sim \xi^i}[\mathbb{E}_{\pi \sim p^i}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \leq 2 \log \frac{2^k \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho_k, \mu_k)}{\delta_k} + 2^k \rho_k\right\}.$$

By Proposition E.2 and a union bound, $\mathbb{P}[A_k] \geq 1 - \delta_k - 2^{2k}\mu_k$. Since the Hellinger distance is always bounded by 2 and $|\mathcal{S}^k| \leq 2^k$, we can upper bound

$$\text{Eq. (66)} \leq \sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\alpha k} \left( \sum_{s \in \mathcal{S}^k} \mathbb{E}\left[\mathbb{E}_{M \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s \leq \tau, A_k\}\right] + 2 \cdot 2^k \mathbb{E}[\mathbb{I}\{A_k^c\}] \right).$$

Choosing $\delta_k = 2^{-3k}$ and since $\mu_k = 2^{-5k}$, we have

$$\sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\alpha k} \cdot 2^{k+1} \mathbb{E}[\mathbb{I}\{A_k^c\}] \leq \sum_{k=1}^{\lceil \log_2 T \rceil} 2^{2k+1} \cdot (\delta_k + 2^{2k}\mu_k) = \sum_{k=1}^{\lceil \log_2 T \rceil} 2^{2k+1} \cdot 2 \cdot 2^{-3k} \leq 4.$$

Note that for $s \in \mathcal{S}^k$, if $s \leq \tau$, this implies that $2^{k-1} \leq \tau$. On the event $A_k$, for $\rho_k = 2^{-k}$, when $\alpha > 0$ we can bound

$$\begin{aligned}
&\sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\alpha k} \sum_{s \in \mathcal{S}^k} \mathbb{E}\left[\mathbb{E}_{M \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s \leq \tau, A_k\}\right] \\
&\quad \leq \mathbb{E}\left[\sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\alpha k} \left(2 \log \frac{2^k \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, 2^{-k}, 2^{-5k})}{2^{-3k}} + 1\right) \cdot \mathbb{I}\{2^{k-1} \leq \tau\}\right] \\
&\quad \leq \mathbb{E}\left[\left(2 \log \frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tau^{-1}/2, \tau^{-5}/32)}{\tau^{-4}/16} + 1\right) \cdot \max_k \left(\frac{2^\alpha}{2^\alpha - 1} 2^{\alpha k} \cdot \mathbb{I}\{2^{k-1} \leq \tau\}\right)\right] \\
&\quad \leq \mathbb{E}\left[\left(2 \log \frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tau^{-1}/2, \tau^{-5}/32)}{\tau^{-4}/16} + 1\right) \cdot \frac{2^{2\alpha}}{2^\alpha - 1} \tau^\alpha\right]
\end{aligned} \tag{67}$$

where the final two inequalities follow since $2^k \leq 2\tau$. Under Assumption A.3 we have

$$\log \frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tau^{-1}/2, \tau^{-5}/32)}{\tau^{-4}/16} \leq 10 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}} \tau),$$

which gives the first result.

**Bound on KL Estimation Error.** By Lemma F.14, for any $x > 0$ we have

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) \leq (2 + 2V_{\mathcal{M}} + x) \cdot D^2_{\mathsf{H}}\big(M(\pi), \widetilde{M}(\pi)\big) + 32(1 + V^2_{\mathcal{M}}/x + V^3_{\mathcal{M}}/x^2) \cdot \exp(-x^2/8V^2_{\mathcal{M}}).$$

In particular choosing $x = V_{\mathcal{M}}\sqrt{8\log s^2}$, we have

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) \le (2 + 2V_{\mathcal{M}} + V_{\mathcal{M}}\sqrt{8\log s}) \cdot D^2_{\mathsf{H}}\big(M(\pi), \widetilde{M}(\pi)\big) + 32(1+V_{\mathcal{M}})/s.$$

Repeating this for each step $s$, we can therefore bound

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} s^{\alpha}\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]]\right] \\
&\qquad \le (2 + 6V_{\mathcal{M}}) \cdot \mathbb{E}\left[\sum_{s=1}^{\tau} s^{\alpha}\sqrt{\log s}\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right] \\
&\qquad\qquad + 32(1+V_{\mathcal{M}}) \cdot \mathbb{E}[\log\tau].
\end{aligned}$$

The result in (63) then follows from a calculation nearly identical to our above bound on Hellinger estimation error. Applying Lemma F.14 in a similar fashion with the arguments flipped gives (64). □

In the following, we extend Lemma F.7 to the case when $\alpha = 0$.

**Lemma F.8.** *Let $\tau$ denote some stopping time with respect to the filtration $(\mathscr{F}^t)_{t=1}^T$ such that $\tau \le T$ almost surely. When running Algorithm 7, under Assumption A.3, we have*

$$\mathbb{E}\left[\sum_{s=1}^{\tau}\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right] \le \mathbb{E}[(20d_{\mathsf{cov}} \cdot \log(64C_{\mathsf{cov}}\tau) + 1) \cdot (2\log\tau + 1)] + 4.$$

*In addition, if Assumption A.2 also holds,*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau}\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]]\right] \\
&\qquad \le \mathbb{E}\Big[(2 + 5V_{\mathcal{M}})(20d_{\mathsf{cov}} \cdot \log(64C_{\mathsf{cov}}\tau) + 1)(2\log\tau + 1) \cdot \sqrt{\log(2\tau)}\Big] \\
&\qquad\qquad + \mathbb{E}[32(1+V_{\mathcal{M}})\log(\tau)] + 8
\end{aligned}$$

*and*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau}\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M(\pi) \,\|\, M^\star(\pi))]]\right] \\
&\qquad \le \mathbb{E}\Big[(2 + 5V_{\mathcal{M}})(20d_{\mathsf{cov}} \cdot \log(64C_{\mathsf{cov}}\tau) + 1)(2\log\tau + 1) \cdot \sqrt{\log(2\tau)}\Big] \\
&\qquad\qquad + \mathbb{E}[32(1+V_{\mathcal{M}})\log(\tau)] + 8.
\end{aligned}$$

**Proof of Lemma F.8.** This follows identically to Lemma F.7 but replacing Eq. (67) with

$$\begin{aligned}
&\sum_{k=1}^{\lceil\log_2 T\rceil} 2^{\alpha k}\sum_{s\in\mathcal{S}^k}\mathbb{E}\big[\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s \le \tau, A_k\}\big] \\
&\qquad \le \mathbb{E}\left[\sum_{k=1}^{\lceil\log_2 T\rceil}\left(2\log\frac{2^k\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, 2^{-k}, 2^{-5k})}{2^{-3k}} + 1\right) \cdot \mathbb{I}\{2^{k-1} \le \tau\}\right]
\end{aligned}$$

$$\leq \mathbb{E}\left[\left(2\log\frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\tau^{-1}/2,\tau^{-5}/32)}{\tau^{-4}/16}+1\right)\cdot(2\log\tau+1)\right].$$

The bound on the KL estimation error also follows from the same reasoning as in Lemma F.7. □

### F.4. Supporting Lemmas

**Lemma F.9.** *Consider running either Algorithm 5 or Algorithm 6. Assume that $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$ and that $\min_{M\in\mathcal{M}\,:\,\xi^s(M)>0}\Delta^M_{\min}\geq\Delta$. Then $\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\in\mathcal{M}^\star_{\mathrm{alt}}\}]\geq\frac{1}{2}\Delta$.*

**Proof of Lemma F.9.** Recall that $\widehat{M}^s=\mathbb{E}_{M\sim\xi^s}[M]$, so $\pi\in\boldsymbol{\pi}_{\widehat{M}^s}$ implies that $\pi\in\arg\max_{\pi'\in\Pi}\mathbb{E}_{M\sim\xi^s}[f^M(\pi')]$. If $\pi_\star\notin\boldsymbol{\pi}_{\widehat{M}^s}$, then there exists some $\widetilde{\pi}$ such that $\mathbb{E}_{M\sim\xi^s}[f^M(\widetilde{\pi})]>\mathbb{E}_{M\sim\xi^s}[f^M(\pi_\star)]$. Since $f^M(\pi)\in[0,1]$, this implies that

$$0<\mathbb{E}_{M\sim\xi^s}[f^M(\widetilde{\pi})-f^M(\pi_\star)]\leq\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\in\mathcal{M}^\star_{\mathrm{alt}}\}]-\mathbb{E}_{M\sim\xi^s}[(f^M(\pi_\star)-f^M(\widetilde{\pi}))\cdot\mathbb{I}\{M\notin\mathcal{M}^\star_{\mathrm{alt}}\}].$$

For $M\notin\mathcal{M}^\star_{\mathrm{alt}}$, we have $f^M(\pi_\star)-f^M(\widetilde{\pi})\geq\Delta^M_{\min}\geq\Delta$. Thus, the above implies

$$\begin{aligned}&0<\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\in\mathcal{M}^\star_{\mathrm{alt}}\}]-\Delta\cdot\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\notin\mathcal{M}^\star_{\mathrm{alt}}\}]\\ \iff&\Delta\cdot(1-\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\in\mathcal{M}^\star_{\mathrm{alt}}\}])<\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\in\mathcal{M}^\star_{\mathrm{alt}}\}].\end{aligned}$$

Rearranging gives $\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\in\mathcal{M}^\star_{\mathrm{alt}}\}]\geq\frac{\Delta}{1+\Delta}\geq\frac{1}{2}\Delta$. □

**Lemma F.10.** *When running Algorithm 5, on rounds $s$ for which $M^\star\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})$, we have*

$$\gamma^s\leq\frac{(1+\delta)(4\mathsf{n}_{\max}+2\delta\underline{\mathsf{g}}^{\mathcal{M}})}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}),$$

*for $\gamma^s$ as defined in Eq. (50).*

**Proof of Lemma F.10.** Recall that $\omega^s$ and $\lambda^s$ are chosen to minimize

$$\sup_{M\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})}\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]}.$$

Since $M^\star\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})$, we can therefore bound

$$\begin{aligned}\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}&\leq\inf_{\omega\in\triangle_\Pi}\sup_{M\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})}\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]}\\&=\inf_{\lambda,\omega\in\triangle_\Pi}\sup_{M\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}_{\max})}\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]}.\end{aligned}$$

Recall that we set

$$\mathsf{n}_{\max}=\left(\frac{1}{\Delta_{\min}\varepsilon}+\frac{2V_{\mathcal{M}}\mathsf{n}^{\mathcal{M}}_\varepsilon}{\Delta_{\min}}\right)\cdot\max_{M\in\mathcal{M}}\frac{32\mathsf{g}^M}{\Delta_{\min}}.$$

By Lemma F.13, under Assumption A.2, we can bound $D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \le 2V_{\mathcal{M}}$ for all $M, M' \in \mathcal{M}$ and $\pi \in \Pi$. It follows from Lemma E.3 that

$$\frac{2V_{\mathcal{M}}}{\Delta_{\min}} \ge \frac{1}{\min_{M \in \mathcal{M}: \mathsf{g}^M > 0} \mathsf{g}^M},$$

so

$$\mathsf{n}_{\max} \ge \left( \frac{1}{\Delta_{\min} \varepsilon} + \frac{\mathsf{n}^{\mathcal{M}}_{\varepsilon}}{\min_{M \in \mathcal{M}: \mathsf{g}^M > 0} \mathsf{g}^M} \right) \cdot \max_{M \in \mathcal{M}} \frac{32 \mathsf{g}^M}{\Delta_{\min}}.$$

Given this, straightforward calculation shows that $\mathsf{n}_{\max}$ meets the condition required by Lemma E.4, so Lemma E.4 implies

$$\begin{aligned}
\inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda; \mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega}[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]]} &\le \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon/2}(\lambda)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega}[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]]} \\
&= \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}; \xi^s) \\
&\le \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}).
\end{aligned}$$

By our choice of $q$ we have $\frac{1}{1-q} = \frac{4\mathsf{n}_{\max} + 2\delta \underline{\mathsf{g}}^{\mathcal{M}}}{\delta \underline{\mathsf{g}}^{\mathcal{M}}}$. We can then bound

$$\gamma^s = \frac{1+\delta}{1-q} \cdot \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]} \le \frac{(1+\delta)(4\mathsf{n}_{\max} + 2\delta \underline{\mathsf{g}}^{\mathcal{M}})}{\delta \underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}).$$

□

**Lemma F.11.** *Consider running Algorithm 6. Then on rounds $s$ for which $M^\star \in \mathcal{M}^\ell \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda^s; \mathsf{n}^s)$, we have*

$$\gamma^s \le (1 + \mathsf{g}^\star \delta) \cdot s^{\alpha_q + \alpha_{\mathcal{M}}}$$

*for $\gamma^s$ as defined in Eq. (56), and $\alpha_q, \alpha_{\mathcal{M}}$ parameters of Algorithm 6.*

**Proof of Lemma F.11.** Assume we are at epoch $\ell$. Recall that $\omega^s$ and $\lambda^s$ minimize

$$\sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda^s; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega}[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]]}.$$

Since $M^\star \in \mathcal{M}^\ell \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda^s; \mathsf{n}^s)$, we have can therefore bound

$$\begin{aligned}
\frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]} &\le \inf_{\omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda^s; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega}[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]]} \\
&= \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega}[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]]}.
\end{aligned}$$

By construction, for every $M \in \mathcal{M}^\ell$, we have

$$\mathsf{n}_\varepsilon^M + \frac{1}{\Delta_{\min}^M} + \frac{4\mathsf{g}^M}{\Delta_{\min}^M} + \frac{2\mathsf{n}_\varepsilon^M}{\mathsf{g}^M} + \frac{4}{\Delta_{\min}^M \varepsilon} \le \sqrt{\mathsf{n}^s}.$$

This implies that

$$\zeta := \min_{M \in \mathcal{M}^\ell} \min\left\{ \frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}_\varepsilon^M}, \frac{\Delta_{\min}^M \varepsilon}{4} \right\} \ge \frac{1}{\sqrt{\mathsf{n}^s}}$$

and

$$\max_{M \in \mathcal{M}^\ell} \max\left\{ \mathsf{n}_\varepsilon^M, \frac{4\mathsf{g}^M}{\Delta_{\min}^M} \right\} \le \sqrt{\mathsf{n}^s},$$

which together imply that

$$\max_{M \in \mathcal{M}^\ell} \max\left\{ \mathsf{n}_\varepsilon^M, \frac{4\mathsf{g}^M}{\Delta_{\min}^M}, \frac{2\mathsf{g}^M}{\zeta \Delta_{\min}^M} \right\} \le \mathsf{n}^s$$

By Lemma E.4 and since $\mathcal{M}^\ell$ is constructed such that $\inf_{M \in \mathcal{M}^\ell} \Delta_{\min}^M > 0$, we can therefore bound

$$\begin{aligned}\inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D(\widehat{M}(\pi) \,\|\, M(\pi))]]} &\le \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}_{\varepsilon/2}^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D(\widehat{M}(\pi) \,\|\, M(\pi))]]} \\ &\le \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}}; \xi^s)\end{aligned}$$

where the last inequality holds by the definition of $\mathcal{M}^\ell$ and $\Delta^\ell$. Note that by construction, $\xi^s$ is only supported on $\mathcal{M}^\ell$, so we can bound $\overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}}; \xi^s) \le \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}})$. By construction, we also have $\overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}}) \le 2^{\ell \alpha_{\mathcal{M}}} \le s^{\alpha_{\mathcal{M}}}$. Lastly, by our choice for $q^s$ we have $\frac{1}{1-q^s} = s^{\alpha_q}$. We can then bound

$$\gamma^s = \frac{1 + \mathsf{g}^\star \delta}{1 - q^s} \cdot \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega^s}[D(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]} \le (1 + \delta) \cdot s^{\alpha_q + \alpha_{\mathcal{M}}}.$$

□

### F.4.1. LIKELIHOOD RATIOS

**Lemma F.12.** *Consider running either Algorithm 5 or Algorithm 6. Under Assumption A.2, with probability at least $1 - \delta$, we can bound, for any $M \in \mathcal{M}$, $s$, and $\beta_i > 0$,*

$$\begin{aligned}\sum_{i=1}^s \mathbb{E}_{\widehat{M} \sim \xi^i}[\mathbb{E}_{\pi \sim p^i}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]] \le \sum_{i=1}^s \mathbb{E}_{\widehat{M} \sim \xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^i}(r^i, o^i)}{\mathbb{P}^{M, \pi^i}(r^i, o^i)}\right] + V_{\mathcal{M}} \sqrt{56 s \log 1/\delta} \\ + V_{\mathcal{M}} \cdot \left( \sum_{i=1}^s 1/\beta_i + \sum_{i=1}^s \beta_i \mathbb{E}_{\widehat{M} \sim \xi^i}[\mathbb{E}_{\pi \sim p^i}[D(\widehat{M}(\pi) \,\|\, M^\star(\pi))]] \right).\end{aligned}$$

**Proof of Lemma F.12.** By Theorem 2 of Shamir (2011), under Assumption A.2 we have that with probability at least $1-\delta$,

$$\sum_{i=1}^{s} \mathbb{E}_{(r,o)\sim M^\star}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] \mid \mathcal{H}^{i-1}\right] \le \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] + V_{\mathcal{M}}\sqrt{56s\log 1/\delta}. \tag{68}$$

Note that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \mid \mathcal{H}^{i-1}\right] &= \mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right]\right] \\ &= \mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right]\right]\end{aligned}$$

By Lemma B.4 of Foster et al. (2022b), we can bound, for any $\widehat{M}\in\mathcal{M}$,

$$\begin{aligned}&\left|\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] - \mathbb{E}_{(r,o)\sim\widehat{M}(\pi)}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right| \\ &\quad\le \sqrt{\frac{1}{2}\left(\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log^2\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] + \mathbb{E}_{(r,o)\sim\widehat{M}(\pi)}\left[\log^2\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right)\cdot D^2_{\mathsf{H}}\big(M^\star(\pi),\widehat{M}(\pi)\big)} \\ &\quad\le \sqrt{2V^2_{\mathcal{M}}\cdot D^2_{\mathsf{H}}\big(M^\star(\pi),\widehat{M}(\pi)\big)}\end{aligned}$$

where the second inequality follows under the subgaussian assumption, Assumption A.2. It follows that for any $\widehat{M}\in\mathcal{M}$,

$$\begin{aligned}&\mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right] \\ &\ge \mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{(r,o)\sim\widehat{M}(\pi)}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right] - \sqrt{2V^2_{\mathcal{M}}\cdot\mathbb{E}_{\pi\sim p^i}[D^2_{\mathsf{H}}\big(M^\star(\pi),\widehat{M}(\pi)\big)]} \\ &\overset{(a)}{=} \mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)] - \sqrt{2V^2_{\mathcal{M}}\cdot\mathbb{E}_{\pi\sim p^i}[D^2_{\mathsf{H}}\big(M^\star(\pi),\widehat{M}(\pi)\big)]} \\ &\overset{(b)}{\ge} \mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)] - V^2_{\mathcal{M}}\cdot(\beta\mathbb{E}_{\pi\sim p^i}[D^2_{\mathsf{H}}\big(M^\star(\pi),\widehat{M}(\pi)\big)] + 1/\beta)\end{aligned}$$

where $(a)$ follows by the definition of KL divergence, and $(b)$ from AM-GM for any $\beta>0$. Since this bound holds uniformly for all $\widehat{M}\in\mathcal{M}$, this implies that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \mid \mathcal{H}^{i-1}\right] &\ge \mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]] \\ &\quad - V^2_{\mathcal{M}}\cdot(\beta\mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D^2_{\mathsf{H}}\big(M^\star(\pi),\widehat{M}(\pi)\big)]] + 1/\beta).\end{aligned}$$

Combining this with Eq. (68), and using Assumption D.2 proves the result. □

**Lemma F.13.** *Under Assumption A.2, we have, for any $M, M' \in \mathcal{M}$,*

$$D_{\mathsf{KL}}\big(M'(\pi) \,\|\, M(\pi)\big) \le 2V_{\mathcal{M}}, \quad \forall \pi \in \Pi.$$

**Proof of Lemma F.13.** We have

$$D_{\mathsf{KL}}\big(M'(\pi) \,\|\, M(\pi)\big) = \mathbb{E}_{(r,o)\sim M'(\pi)}\left[\log \frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] \le \mathbb{E}_{(r,o)\sim M'(\pi)}\left[\left|\log \frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right|\right] \le 2V_{\mathcal{M}}$$

where the last inequality holds under Assumption A.2. □

**Lemma F.14.** *Under Assumption A.2, for any $x > 0$ and $M, \widetilde{M} \in \mathcal{M}$, we have*

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) \le (2 + 2V_{\mathcal{M}} + x) \cdot D^2_{\mathsf{H}}\Big(M(\pi), \widetilde{M}(\pi)\Big) + 32(1 + V^2_{\mathcal{M}}/x + V^3_{\mathcal{M}}/x^2) \cdot \exp(-x^2/8V^2_{\mathcal{M}}).$$

**Proof of Lemma F.14.** Fix $\pi$. Define

$$\mathcal{E} := \left\{\left|\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} - D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big)\right| \le x\right\}$$

and for $j \in \mathbb{N}$,

$$\mathcal{E}_j := \left\{e^{j-1} \cdot x < \left|\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} - D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big)\right| \le e^j \cdot x\right\}.$$

Note that $\mathcal{E}, (\mathcal{E}_j)_{j=1}^{\infty}$ form a partition of the probability space. Furthermore, since $D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) = \mathbb{E}_{o\sim M(\pi)}[\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}]$, under Assumption A.2 we have that $\mathbb{P}^{M,\pi}(\mathcal{E}_j) \le 2\exp(-x^2 e^{2(j-1)}/V^2_{\mathcal{M}})$ and $\mathbb{P}^{M,\pi}(\mathcal{E}^c) \le 2\exp(-x^2/V^2_{\mathcal{M}})$. Now,

$$\begin{aligned} D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) &= \int \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) \\ &= \int_{\mathcal{E}} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) + \sum_{j=1}^{\infty} \int_{\mathcal{E}_j} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o). \end{aligned}$$

Using that $D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) \le 2V_{\mathcal{M}}$ by Lemma F.13, we can bound

$$\begin{aligned} \sum_{j=1}^{\infty} \int_{\mathcal{E}_j} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) &\le \sum_{j=1}^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot \int_{\mathcal{E}_j} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) \\ &= \sum_{j=1}^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot \mathbb{P}^{M,\pi}(\mathcal{E}_j) \\ &\le \sum_{j=1}^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot 2\exp(-x^2 e^{2(j-1)}/V^2_{\mathcal{M}}) \\ &\le \int_0^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot 2\exp(-x^2 e^{2(j-1)}/V^2_{\mathcal{M}}) \mathrm{d}j \end{aligned}$$

$$\leq 4(x+V_{\mathcal{M}})\int_0^\infty e^j \exp(-e^j \cdot x^2 e^{-2}/V_{\mathcal{M}}^2)\mathrm{d}j$$
$$= 4(x+V_{\mathcal{M}})V_{\mathcal{M}}^2 e^2 \exp(-x^2e^{-2}/V_{\mathcal{M}}^2)/x^2$$
$$\leq 32(V_{\mathcal{M}}^2/x + V_{\mathcal{M}}^3/x^2)\cdot\exp(-x^2/8V_{\mathcal{M}}^2).$$

We turn now to the first term. Note that we can write

$$\int_{\mathcal{E}} \log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}\mathrm{d}\mathbb{P}^{M,\pi}(r,o) = \int_{\mathcal{E},\mathbb{P}^{\widetilde{M},\pi}(r,o)>\mathbb{P}^{M,\pi}(r,o)}\left(\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} + \frac{\mathbb{P}^{\widetilde{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} - 1\right)\mathrm{d}\mathbb{P}^{M,\pi}(r,o) - \mathbb{P}^{\widetilde{M},\pi}(\mathcal{E}^c)$$
$$+ \int_{\mathcal{E},\mathbb{P}^{\widetilde{M},\pi}(r,o)\leq\mathbb{P}^{M,\pi}(r,o)}\left(\frac{\mathbb{P}^{M}(o\mid\pi)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} + 1 - \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}\right)\mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,o) + \mathbb{P}^{M,\pi}(\mathcal{E}^c)$$

Following the proof of Lemma 4 of Yang and Barron (1998) and using that $\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \leq D_{\mathsf{KL}}\big(M(\pi)\,\|\,\widetilde{M}(\pi)\big) + x \leq 2V_{\mathcal{M}} + x$ on $\mathcal{E}$, we can bound this as

$$\leq (2+2V_{\mathcal{M}}+x)\int_{\mathcal{E}}(\sqrt{\mathrm{d}\mathbb{P}^{M,\pi}(r,o)} - \sqrt{\mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,o)})^2 + \mathbb{P}^{M,\pi}(\mathcal{E}^c)$$
$$\leq (2+2V_{\mathcal{M}}+x)\int(\sqrt{\mathrm{d}\mathbb{P}^{M,\pi}(r,o)} - \sqrt{\mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,o)})^2 + \mathbb{P}^{M,\pi}(\mathcal{E}^c)$$
$$\leq (2+2V_{\mathcal{M}}+x)\cdot D^2_{\mathsf{H}}\big(M(\pi),\widetilde{M}(\pi)\big) + 2\exp(-x^2/V_{\mathcal{M}}^2).$$

□

### F.4.2. COVERING NUMBERS

**Lemma F.15.** *For any subset $\mathcal{M}' \subseteq \mathcal{M}$, there exists some $(\rho,\mu)$-cover $\mathcal{M}'_{\mathrm{cov}} \subseteq \mathcal{M}'$ for $\mathcal{M}'$ such that $|\mathcal{M}'_{\mathrm{cov}}| \leq \mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho/2,\mu)$.*

**Proof of Lemma F.15.** Let $\mathcal{M}_{\mathrm{cov}}$ denote a $(\rho/2,\mu)$-cover of $\mathcal{M}$ with event $\mathcal{E}$. Throughout the proof we use the shorthand $(r,o,\pi)\in\mathcal{E}$ to denote that there exists $M\in\mathcal{M}$ such that $\mathbb{P}^{M,\pi}(r,o\mid\mathcal{E})>0$. By definition, it follows that for any $M\in\mathcal{M}$, there exists $M'\in\mathcal{M}_{\mathrm{cov}}$ such that

$$\sup_{r,o,\pi\,:\,(r,o,\pi)\in\mathcal{E}}\left|\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)}\right| \leq \rho/2. \tag{69}$$

Let $\mathcal{M}'_{\mathrm{cov}} = \varnothing$ and consider running the following procedure for every $M'\in\mathcal{M}_{\mathrm{cov}}$:

1. Choose a single $M\in\mathcal{M}'$ such that $\sup_{r,o,\pi:(r,o,\pi)\in\mathcal{E}}\left|\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)}\right| \leq \rho/2$ (if such an $M$ exists).
2. If there exists an $M\in\mathcal{M}'$ in step 1, set $\mathcal{M}'_{\mathrm{cov}} \leftarrow \mathcal{M}'_{\mathrm{cov}}\cup\{M\}$. Otherwise $\mathcal{M}'_{\mathrm{cov}}$ remains unchanged.

By construction $\mathcal{M}'_{\mathrm{cov}}\subseteq\mathcal{M}'$, and $|\mathcal{M}'_{\mathrm{cov}}|\leq|\mathcal{M}_{\mathrm{cov}}|$. We claim that $\mathcal{M}'_{\mathrm{cov}}$ is a $(\rho,\mu)$-cover of $\mathcal{M}'$. To see why, take some $M\in\mathcal{M}'$. Let $M'\in\mathcal{M}_{\mathrm{cov}}$ denote the point realizing Eq. (69) for $M$. Let

$M''$ denote the point chosen in the above procedure for $M'$. Note that there must exist some $M''$ chosen for this $M'$ since Eq. (69) holds for $M$, so in particular $M \in \mathcal{M}'$ satisfies the condition of step 1 in the above procedure. Then,

$$\begin{aligned}\sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left|\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M''}(o|\pi)}\right| &= \sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left|\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)} + \log \frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M''}(o|\pi)}\right| \\ &\le \sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left|\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)}\right| + \sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left|\log \frac{\mathbb{P}^{M''}(o|\pi)}{\mathbb{P}^{M',\pi}(r,o)}\right| \\ &\le \rho/2 + \rho/2 = \rho\end{aligned}$$

where the last inequality follows by our choice of $M'$ and the definition of $M''$. Thus, it follows that $\mathcal{M}'_{\text{cov}}$ is a $(\rho,\mu)$-cover of $\mathcal{M}'$. □

### F.4.3. FURTHER LEMMAS

**Lemma F.16.** *For $a > 0$, $\alpha \le [0, 3/4]$, and $\beta > 0$, the function $x^\alpha \log^\beta(ax)$ is concave in $x$ for $x \ge \frac{1}{a}\exp\left(\frac{4\beta}{\alpha}\right)$ when $\beta \le 1$, and for $x \ge \max\left\{\frac{1}{a}\exp\left(\sqrt{\frac{8(\beta-1)\beta}{\alpha}}\right), \frac{1}{a}\exp\left(\frac{4\beta}{\alpha}\right)\right\}$ when $\beta > 1$.*

**Proof of Lemma F.16.** By some calculation, we have

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}^2 x}\left(x^\alpha \log^\beta(ax)\right) = (-1+\beta)\beta x^{-2+\alpha}\log^{\beta-2}(ax) + (-\beta+2\alpha\beta)x^{-2+\alpha}\log^{\beta-1}(ax) \\ + (-1+\alpha)\alpha x^{-2+\alpha}\log^\beta(ax).\end{aligned}$$

If we restrict to $x \ge 1/a$, then to show that the function is concave it then suffices to show that

$$(-1+\beta)\beta\log^{-2}(ax) + (-\beta+2\alpha\beta)\log^{-1}(ax) + (-1+\alpha)\alpha \le 0$$

which, since $\alpha \le 3/4$, is implied by

$$(-1+\beta)\beta\log^{-2}(ax) + \frac{1}{2}\beta\log^{-1}(ax) \le \frac{1}{4}\alpha$$

which is further implied by

$$(-1+\beta)\beta\log^{-2}(ax) \le \frac{1}{8}\alpha \quad \text{and} \quad \frac{1}{2}\beta\log^{-1}(ax) \le \frac{1}{8}\alpha.$$

The former condition is met for $x > 1/a$ for all $\beta \in (0,1]$. For $\beta > 1$, it is met as long as

$$\frac{8(\beta-1)\beta}{\alpha} \le \log^2(ax) \iff x \ge \frac{1}{a}\exp\left(\sqrt{\frac{8(\beta-1)\beta}{\alpha}}\right).$$

The latter condition is met for

$$x \ge \frac{1}{a}\exp\left(\frac{4\beta}{\alpha}\right).$$

□

**Lemma F.17.** *For all $B, C, n \geq 1$, if $x \leq C \log^n(Bx)$, then $x \leq C(2n)^n \log^n(2nBC)$.*

**Proof of Lemma F.17.** This is a direct consequence of Lemma A.4 of Wagenmaker et al. (2022a). □

## Appendix G. Proofs for Examples

In this section, we provide proofs for the examples given in Appendix A. We begin in Appendix G.1 by introducing a condition which implies $\mathsf{n}_\varepsilon^{M^\star}$ is bounded, and is easy to verify for many classes of interest. Next, in Appendix G.2, we consider a variety of structured bandit settings, and in Appendix G.3 extend this to contextual bandits with finitely many arms. In Appendix G.4, we provide proofs for the informative arm example of Section 1.3. Finally, in Appendix G.5, we consider tabular MDPs.

### G.1. Preliminaries: Regular Models

To bound the quantity $\mathsf{n}_\varepsilon^\star = \mathsf{n}_\varepsilon^{M^\star}$ for the examples we consider, it will be helpful to introduce the following notion of a *regular model*.

**Definition G.1** (Regular Model)**.** *We say instance $M \in \mathcal{M}$ is a* regular model *if there exists some constant $L_{\mathcal{M}}^M > 0$ such that, for any $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ with $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$, there exists $M'' \in \mathcal{M}^{\mathrm{alt}}(M)$ such that $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M''(\pi_M)) = 0$ and, for all $\pi \in \Pi$,*

$$|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big)| \leq \sqrt{L_{\mathcal{M}}^M D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M))} + L_{\mathcal{M}}^M D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big). \tag{70}$$

Our definition of a regular model is a direct generalization of existing notions of class regularity found in the literature (Degenne et al., 2020b). As we will see, for a variety of standard bandit classes (including multi-armed bandits, linear bandits, and Lipschitz bandits), as well as tabular MDPs, one can show that$M^\star$ is a regular model with $L_{\mathcal{M}}^\star = L_{\mathcal{M}}^{M^\star}$ bounded by a polynomial function of problem parameters. Intuitively, $M^\star$ will be a regular model if, for any instance $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ for which it is sufficient to pull $\pi_\star$ in order to distinguish $M^\star$ and $M'$ (thereby ruling out $M'$ while incurring no regret), then there exists some other instance $M'' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ which is "close" to $M'$ in a certain sense, and which cannot be distinguished from $M^\star$ by simply pulling $\pi_\star$. As the following result shows, the quantity $\mathsf{n}_\varepsilon^\star$ can be bounded whenever $M^\star$ is a regular model.

**Proposition G.1.** *If $M$ is a regular model with $\Delta_{\min}^M > 0$, we can bound*

$$\mathsf{n}_\varepsilon^M \leq \frac{2\mathsf{g}^M}{\Delta_{\min}^M} \cdot \left(1 + L_{\mathcal{M}}^M + \frac{2\mathsf{g}^M}{\varepsilon \Delta_{\min}^M} \cdot L_{\mathcal{M}}^M\right).$$

Given Proposition G.1, for many of the examples in this section, rather than bounding $\mathsf{n}_\varepsilon^\star$ directly, we first show that $M^\star$ is a regular model with $L_{\mathcal{M}}^\star$ well-bounded, and then use Proposition G.1 to obtain a bound on $\mathsf{n}_\varepsilon^\star$.

**Proof of Proposition G.1.** To prove this result, it suffices to show that, for every normalized allocation $\lambda \in \Lambda(M, \varepsilon)$ with normalization factor $\mathsf{n}$, there exists some allocation $\eta \in \mathbb{R}_+^\Pi$ such that 1)

$\eta(\pi) = \mathsf{n}\lambda(\pi)$ for $\pi \neq \pi_M$, and $\eta(\pi_M) \leq \bar{\mathsf{n}}$ for some well-bounded $\bar{\mathsf{n}}$, and 2) $\Delta^M(\eta) \leq (1+2\varepsilon)\mathsf{g}^M$ and $I^M(\eta) \geq 1 - 2\varepsilon$.

Fix some $\lambda \in \Lambda(M, \varepsilon)$ with normalization factor $\mathsf{n} > 0$. Note that if $M$ is a regular model, then $I^M(\mathbb{I}_{\pi_M}) = 0$. Since $\lambda \in \Lambda(M, \varepsilon)$, this implies that $\lambda(\pi_M) < 1$. Let $\lambda'$ denote the allocation $\lambda'(\pi_M) = 1 - \zeta$ and $\lambda'(\pi) = \frac{\zeta}{1-\lambda(\pi_M)}\lambda(\pi)$ for $\pi \neq \pi_M$, for some $\zeta$ to be chosen.

We have

$$\Delta^M(\lambda') = \frac{\zeta}{1-\lambda(\pi_M)}\Delta^M(\lambda) \leq \frac{\zeta}{1-\lambda(\pi_M)}(1+\varepsilon)\mathsf{g}^M/\mathsf{n}. \tag{71}$$

Take $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ such that $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$ and let $M'' \in \mathcal{M}^{\mathrm{alt}}(M)$ denote the instance guaranteed to exist under Definition G.1. We then have that, for any $\pi$,

$$\begin{aligned} D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) &\geq D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) - \sqrt{L^M_{\mathcal{M}} D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M))} - L^M_{\mathcal{M}} D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \\ &\geq D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) - (1+\alpha) L^M_{\mathcal{M}} D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) - \frac{1}{\alpha} \end{aligned}$$

where the last inequality follows for any $\alpha > 0$ by AM-GM. Then

$$\begin{aligned} &\sum_\pi \lambda'(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \\ &= \sum_{\pi \neq \pi_M} \frac{\zeta}{1-\lambda(\pi_M)}\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) + (1-\zeta) D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \\ &\geq \sum_{\pi \neq \pi_M} \frac{\zeta\lambda(\pi)}{1-\lambda(\pi_M)} \Big( D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) - (1+\alpha) L^M_{\mathcal{M}} D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) - \frac{1}{\alpha} \Big) \\ &\quad + (1-\zeta) D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \\ &= \sum_{\pi \neq \pi_M} \frac{\zeta}{1-\lambda(\pi_M)}\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) + \Big( (1-\zeta) - (1+\alpha) L^M_{\mathcal{M}} \zeta \Big) D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) - \frac{\zeta}{\alpha}. \end{aligned}$$

We have $M'' \in \mathcal{M}^{\mathrm{alt}}(M)$, so by definition

$$\sum_\pi \lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) \geq (1-\varepsilon)/\mathsf{n}.$$

However, $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M''(\pi_M)) = 0$ by assumption, so it follows that

$$\sum_{\pi \neq \pi_M} \frac{\zeta}{1-\lambda(\pi_M)}\lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big) \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}}.$$

By assumption, $\Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n}$, and we can also lower bound $\Delta^M(\lambda) \geq (1-\lambda(\pi_M))\Delta^M_{\min}$. Rearranging these implies that

$$(1-\lambda(\pi_M))\mathsf{n} \leq (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}.$$

Set $\alpha = \frac{(1+\varepsilon)\mathsf{g}^M}{\varepsilon\Delta^M_{\min}}$, then we have

$$\frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}} - \frac{\zeta}{\alpha} = \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}} - \frac{\zeta\varepsilon}{(1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}}$$

$$\geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}} - \frac{\zeta\varepsilon}{(1-\lambda(\pi_M))\mathsf{n}}$$
$$= \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-2\varepsilon}{\mathsf{n}}.$$

Furthermore, with this choice of $\alpha$, if

$$\zeta \leq \frac{1}{1+(1+\alpha)L^M_{\mathcal{M}}} = \frac{1}{1+(1+(1+\varepsilon)\mathsf{g}^M/\varepsilon\Delta^M_{\min})L^M_{\mathcal{M}}},$$

we have

$$\Big((1-\zeta) - (1+\alpha)L^M_{\mathcal{M}}\zeta\Big) D_{\mathsf{KL}}\big(M(\pi_M) \,\|\, M'(\pi_M)\big) \geq 0.$$

We therefore have that, for any $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ with $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$, that

$$\sum_\pi \lambda'(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-2\varepsilon}{\mathsf{n}}.$$

Now consider $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ with $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) = 0$. In this case we have

$$\sum_\pi \lambda'(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) = \frac{\zeta}{1-\lambda(\pi_M)} \cdot \sum_\pi \lambda(\pi) D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}}$$

where the inequality follows since $\lambda \in \Lambda(M, \varepsilon)$ with normalization factor $\mathsf{n}$ by assumption. Together these bounds then imply that:

$$I^M(\lambda') \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-2\varepsilon}{\mathsf{n}}.$$

Combining this with our bound on $\Delta^M(\lambda')$ in (71) implies that $\lambda' \in \Lambda(M; 2\varepsilon)$ with parameter $\mathsf{n}' = \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}$.

To conclude, define the allocation $\eta := \mathsf{n}'\lambda'$. Then for $\pi \neq \pi_M$:

$$\eta(\pi) = \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}\lambda'(\pi) = \mathsf{n}\lambda(\pi)$$

and

$$\eta(\pi_M) = \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}\lambda'(\pi_M) \leq \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}$$

It follows then that $\Delta^M(\eta) = \mathsf{n}\Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M$, and $I^M(\eta) \geq \mathsf{n}' I^M(\lambda') \geq 1-2\varepsilon$. Therefore, $\eta$ satisfies the desired condition. Since $\lambda \in \Lambda(M, \varepsilon)$, we have

$$\Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies (1-\lambda(\pi_M))\mathsf{n} \leq (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min},$$

and thus

$$\eta(\pi_M) \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\Delta^M_{\min} \cdot \zeta} \leq \frac{2\mathsf{g}^M(1+(1+2\mathsf{g}^M/\varepsilon\Delta^M_{\min})L^\star_{\mathcal{M}})}{\Delta^M_{\min}}.$$

which proves the result. □

### G.2. Structured Bandits with Gaussian Noise

In this section, we consider the problem of structured bandits with Gaussian noise, in which $\mathcal{O} = \{\varnothing\}$, and the mean reward functions belong to a given function class $\mathcal{F}$. Concretely, we consider the model class

$$\mathcal{M} = \big\{M(\pi) = \mathcal{N}(f(\pi), \sigma^2) \ : \ f \in \mathcal{F}\big\}.$$

We set

$$D\big(M(\pi) \,\|\, \overline{M}(\pi)\big) \leftarrow D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big) = \frac{1}{2\sigma^2}(f^M(\pi) - f^{\overline{M}}(\pi))^2 \tag{72}$$

for $D$ the divergence used by $\mathsf{AE}^2_\star$. In general in the following examples we take $\sigma = 1$ for simplicity.

We begin by verifying that the basic regularity conditions required by our results are satisfied for generic classes $\mathcal{F}$, then provide bounds on the AEC for specific classes of interest.

**Lemma G.1.** *For bandits with Gaussian noise:*

1. *For $D \leftarrow D_{\mathsf{KL}}$, Assumptions A.2, D.1 and D.2 hold with parameters*

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = \frac{2\sqrt{2}}{\sigma}.$$

2. *We can bound $\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho, \mu)$ by the covering number of $\mathcal{M}$ in the distance $d(M, M') := \sup_{\pi\in\Pi} |f^M(\pi) - f^{M'}(\pi)|$ at tolerance $\frac{\sigma^2\cdot\rho}{2+\sqrt{2\sigma^2\log(2/\mu)}}$. Furthermore, it suffices to take $\mathcal{E} := \{|r| \le 1 + \sqrt{2\sigma^2\log(2/\mu)}\}$.*

**Proof of Lemma G.1.** In this setting, we have that for any $M, M' \in \mathcal{M}$ and any $\pi \in \Pi$,

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) = \frac{1}{2\sigma^2}(f^M(\pi) - f^{M'}(\pi))^2.$$

For $\overline{M} \in \mathcal{M}$, we therefore have

$$\begin{aligned}\big|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big)\big| &= \frac{1}{2\sigma^2}\Big|(f^M(\pi) - f^{M'}(\pi))^2 - (f^{\overline{M}}(\pi) - f^{M'}(\pi))^2\Big| \\ &\le \frac{2}{\sigma^2}|f^M(\pi) - f^{\overline{M}}(\pi)| \\ &= \frac{2\sqrt{2}}{\sigma}\sqrt{D\big(M(\pi) \,\|\, \overline{M}(\pi)\big)},\end{aligned}$$

where the inequality follows from the Mean Value Theorem and the assumption that $f^{\overline{M}}(\pi) \in [0, 1]$ for all $\pi \in \Pi$. This verifies that Assumption D.1 holds with $L_{\mathsf{KL}} = \frac{2\sqrt{2}}{\sigma}$.

To show that Assumption A.2 is met, we note that for all $M, \overline{M}, \overline{M}' \in \mathcal{M}$,

$$\begin{aligned}\log\frac{\mathbb{P}^{\overline{M},\pi}(r, o)}{\mathbb{P}^{M,\pi}(r, o)} &= \frac{1}{2\sigma^2}(r - f^M(\pi))^2 - \frac{1}{2\sigma^2}(r - f^{\overline{M}}(\pi))^2 \\ &= \frac{1}{2\sigma^2}\Big[f^M(\pi)^2 + f^{\overline{M}}(\pi)^2 - 2r(f^M(\pi) - f^{\overline{M}}(\pi))\Big]\end{aligned}$$

$$= \frac{1}{2\sigma^2}\left[f^M(\pi)^2 + f^{\overline{M}}(\pi)^2 - 2f^{\overline{M}'}(\pi)(f^M(\pi) - f^{\overline{M}}(\pi)) + 2(f^{\overline{M}'}(\pi) - r)(f^M(\pi) - f^{\overline{M}}(\pi))\right]$$
$$= \mathbb{E}_{(r,o)\sim\overline{M}'(\pi)}\left[\log\frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] + \frac{1}{\sigma^2}(f^{\overline{M}'}(\pi) - r)(f^M(\pi) - f^{\overline{M}}(\pi)).$$

It follows that

$$\log\frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} - \mathbb{E}_{(r,o)\sim\overline{M}'(\pi)}\left[\log\frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]$$

is sub-gaussian with parameter $\mathbb{E}_{r\sim\overline{M}'(\pi)}[(\frac{1}{\sigma^2}(f^{\overline{M}'}(\pi)-r)(f^M(\pi)-f^{\overline{M}}(\pi)))^2] \le \frac{4}{\sigma^2}$, which verifies Assumption A.2 with $V_{\mathcal{M}}^2 = \frac{8}{\sigma^2}$.

Finally, we bound the covering number. Let $\mathcal{E} := \{|r| \le 1 + \sqrt{2\sigma^2\log(2/\mu)}\}$. Elementary manipulations show that $\mathbb{P}^{\overline{M},\pi}(\mathcal{E}^c) \le \mu$ for any $\overline{M} \in \mathcal{M}$ and $\pi$. Using the same calculation as above, we have

$$\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)} = \frac{1}{2\sigma^2}(r - f^M(\pi))^2 - \frac{1}{2\sigma^2}(r - f^{M'}(\pi))^2 \le \frac{1+|r|}{\sigma^2}\cdot|f^M(\pi) - f^{M'}(\pi)|$$

where the inequality follows from the Mean Value Theorem. We therefore have that for any $M, M' \in \mathcal{M}$,

$$\sup_{r,o,\pi\,:\,|r|\le 1+\sqrt{2\sigma^2\log(2/\mu)}}\left|\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)}\right| \le \frac{2+\sqrt{2\sigma^2\log(2/\mu)}}{\sigma^2}\cdot\sup_{\pi}|f^M(\pi) - f^{M'}(\pi)|.$$

It follows that if we can form a $\frac{\sigma^2\cdot\rho}{2+\sqrt{2\sigma^2\log(2/\mu)}}$-cover of $\mathcal{M}$ in the distance $d(M,M') = \sup_{\pi\in\Pi}|f^M(\pi) - f^{M'}(\pi)|$, this will serve as an $(\rho,\mu)$ cover of $\mathcal{M}$. □

### G.2.1. DISCRETE STRUCTURED BANDITS

As a first example of bandits with Gaussian noise, we present an additional class that satisfies the uniformly regular assumption.

**Example G.1** (Discrete Structured Bandits)**.** Fix $\Delta_{\min} > 0$, and consider a discrete reward space $\mathcal{R} \subseteq [0,1]$ satisfying $\min_{r,r'\in\mathcal{R}}|r - r'| \ge \Delta_{\min}$. Consider any function class $\mathcal{F} \subseteq (\Pi \to \mathcal{R})$ defined such that each $f \in \mathcal{F}$ has a unique optimal decision. Let our model class be defined as

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(f(\pi), 1) \mid f \in \mathcal{F}\}.$$

I t is straightforward to show that Assumption A.1 and Assumption A.2 are met with $L_{\mathsf{KL}}, V_{\mathcal{M}} \le 4$, and that Assumption A.3 is satisfied with $d_{\mathsf{cov}}$ scaling with the log-covering number of $\mathcal{F}$ in the distance $d(f,f') = \sup_{\pi\in\Pi}|f(\pi) - f'(\pi)|$, and $C_{\mathsf{cov}} = O(1)$. Furthermore, Assumption A.4 is satisfied by construction of $\mathcal{R}$ and $\mathcal{F}$, and we can bound $\mathrm{n}_\varepsilon^{\mathcal{M}} \le \frac{2}{\Delta_{\min}^2}$.[8] We thus have the following corollary to Theorem A.1.

8. It is not difficult to see that, given the construction of $\mathcal{R}$, once the optimal arm has been played $\frac{2}{\Delta_{\min}^2}$ times, no additional information can be extracted from playing it.

**Corollary G.1.** *In the discrete structured bandits setting considered above, if* $\mathsf{g}^\star > 0$*, the regret of* AE$^2$ *is bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot \mathrm{poly}\left(\max_{M\in\mathcal{M}} \mathsf{g}^M, d_{\mathsf{cov}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta_{\min}}, \log\log T\right) \cdot \log^{1/2}(T)$$

As we show in Example A.3, the AEC in this setting can be bounded in terms of the eluder dimension of $\mathcal{F}$. ◁

**Proof for Example G.1.** We provide calculations for the discrete structured bandits setting of Example G.1. First, note that Assumptions A.1 to A.3 all hold by Lemma G.1. To bound $\mathsf{n}^{\mathcal{M}}$, consider $M \in \mathcal{M}$ and $M' \in \mathcal{M}^{\mathrm{alt}}(M)$. Note that either $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) = 0$, in which case there is no advantage to playing $\pi_M$, or, due to the discretization of the means, $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq \frac{1}{2}\Delta_{\min}^2$. Thus for any allocation $\eta \in \mathbb{R}_+^{\Pi}$, as long as $\eta(\pi_M) \geq \frac{2}{\Delta_{\min}^2}$, we have $\eta(\pi_M) D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq 1$. It follows that there is no advantage to choosing $\eta(\pi_M)$ larger than $\frac{2}{\Delta_{\min}^2}$, so we can bound $\mathsf{n}^{\mathcal{M}} \leq \frac{2}{\Delta_{\min}^2}$. □

### G.2.2. MULTI-ARMED BANDITS (EXAMPLE A.2)

In this section we prove the result in Example A.2. First, note that for any $M \in \mathcal{M}$, we have $\mathsf{g}^M \leq A/\Delta_{\min}^M$. Assumptions A.2, D.1 and D.2 are met due to Lemma G.1, and with constants $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$, $d_{\mathsf{cov}} = O(A)$, and $C_{\mathsf{cov}} = O(1)$. By Lemma G.2—stated and proven below—$M^\star$ is a regular model with $L_{\mathcal{M}}^\star = \sqrt{2}$ as long as $f^{M^\star}(\pi_\star) < 1$. It remains to bound the AEC.

**Proof of Proposition A.2.** It is immediate to see that $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \leq O(\frac{A}{\varepsilon})$ by choosing the exploration distribution to be uniform over $A$. By Lemma E.6, we then

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq c_1 \cdot \frac{A^3}{\varepsilon^2 \Delta_\star^6}$$

for a universal constant $c_1$. By Proposition G.1, we have

$$\mathsf{n}_{\varepsilon/36}^\star \leq c_2 \cdot \frac{\mathsf{g}^\star}{\Delta_{\min}^\star} \cdot \left(1 + \frac{\mathsf{g}^\star}{\varepsilon \Delta_{\min}^\star}\right) \leq c_2 \cdot \frac{A^2}{\varepsilon(\Delta_{\min}^\star)^4},$$

for a universal constant $c_2$, so we can lower bound $\Delta_\star \geq c_3 \varepsilon (\Delta_{\min}^\star)^4 / A^2$, giving

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq c_4 \cdot \frac{A^{15}}{\varepsilon^8 (\Delta_{\min}^\star)^{24}}.$$

□

**Lemma G.2.** *In the multi-armed bandit setting of Example A.2, any model* $M^\star \in \mathcal{M}$ *is a regular model with* $L_{\mathcal{M}}^\star = \sqrt{3}$ *as long as* $f^{M^\star}(\pi_\star) < 1$.

**Proof of Lemma G.2.** Let $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ and assume that $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) > 0$.

**Case 1:** $f^{M'}(\pi_{M'}) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) \le 1$**.** Let $M''$ denote the Gaussian bandit instance given by $f^{M''}(\pi) = f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)$; by our assumption, $M'' \in \mathcal{M}$, and $M'' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$. Furthermore, $f^{M''}(\pi_\star) = f^{M^\star}(\pi_\star)$ which implies $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$ as desired. Finally, for any $\pi$, we have

$$f^{M'}(\pi) - f^{M''}(\pi) = f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star).$$

This implies that for all $\pi$, since all models in $\mathcal{M}$ have unit Gaussian rewards, using the expression for the KL divergence given in Eq. (72):

$$|D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M''(\pi)\big)| \le |f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)| = \sqrt{2D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}$$

which implies the condition of Definition G.1 is met with $L^\star_{\mathcal{M}} = \sqrt{2}$.

**Case 2:** $f^{M'}(\pi_{M'}) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) > 1$**.** For this case the model $M''$ constructed in Case 1 will not be in $\mathcal{M}$. Assume first that $f^{M^\star}(\pi_\star) \ge f^{M'}(\pi_{M'})$ and in this case define $M''$ to be the instance

$$f^{M''}(\pi) = \begin{cases} \min\{f^{M^\star}(\pi_\star), f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} & \pi \neq \pi_{M'} \\ f^{M^\star}(\pi_\star) + \delta & \pi = \pi_{M'} \end{cases}$$

or some $\delta > 0$ such that $f^{M^\star}(\pi_\star) + \delta < 1$ (note that such a $\delta$ exists since we have assumed $f^{M^\star}(\pi_\star) < 1$). Note that we now have $M'' \in \mathcal{M}$, and $f^{M''}(\pi) < f^{M''}(\pi_{M'})$ for all $\pi \neq \pi_{M'}$, so $M'' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$. Furthermore, we have $f^{M''}(\pi_\star) = f^{M^\star}(\pi_\star)$, so $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$. For $\pi \neq \pi_{M'}$, if $\min\{f^{M^\star}(\pi_\star), f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M^\star}(\pi_\star)$, this implies that

$$f^{M^\star}(\pi_\star) \le f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) \implies f^{M^\star}(\pi_\star) - f^{M'}(\pi) \le f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star).$$

Since we have assumed $f^{M^\star}(\pi_\star) \ge f^{M'}(\pi_{M'})$, this implies that

$$|f^{M''}(\pi) - f^{M'}(\pi)| = |f^{M^\star}(\pi_\star) - f^{M'}(\pi)| \le |f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)|$$

So by the expression given for the KL divergence in Eq. (72), we have

$$|D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M''(\pi)\big)| \le \sqrt{3D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}. \tag{73}$$

For $\pi \neq \pi_{M'}$ with $\min\{f^{M^\star}(\pi_\star), f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)$, the bound on $|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi))|$ follows identically to Case 1. For $\pi = \pi_{M'}$, since we have assumed that $f^{M^\star}(\pi_\star) \ge f^{M'}(\pi_{M'})$ we have

$$|f^{M''}(\pi_{M'}) - f^{M'}(\pi_{M'})| = f^{M^\star}(\pi_\star) - f^{M'}(\pi_{M'}) + \delta \le f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) + \delta.$$

For small enough $\delta$, this implies that Eq. (73) is satisfied for $\pi = \pi_{M'}$ as well.

Consider now the case where $f^{M^\star}(\pi_\star) < f^{M'}(\pi_{M'})$. In this case define $M''$ by

$$f^{M''}(\pi) = \begin{cases} \min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} & \pi \neq \pi_{M'} \\ f^{M'}(\pi_{M'}) & \pi = \pi_{M'} \end{cases}$$

for $\delta > 0$ small enough that $f^{M'}(\pi_{M'}) - \delta > f^{M^\star}(\pi_\star)$. Note that $M''$ and $M'' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ by construction, and that $f^{M''}(\pi_\star) = f^{M^\star}(\pi_\star)$ by our choice of $\delta$, so $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$. For $\pi \neq \pi_{M'}$, if $\min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M'}(\pi_{M'}) - \delta$, then we have

$$\begin{aligned}
|f^{M''}(\pi) - f^{M'}(\pi)| &= |f^{M'}(\pi_{M'}) - \delta - f^{M'}(\pi)| \\
&\leq f^{M'}(\pi_{M'}) - f^{M'}(\pi) + \delta \\
&\leq |f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)| + \delta
\end{aligned}$$

where for the final inequality we have used that $\min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) = f^{M'}(\pi_{M'}) - \delta$. It follows that Eq. (73) is satisfied for this $\pi$ for sufficiently small $\delta$. If we instead have $\min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)$, then the bound on $|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi))|$ follows identically to Case 1.

For $\pi = \pi_{M'}$, we have $|f^{M''}(\pi_{M'}) - f^{M'}(\pi_{M'})| = 0$. This proves the result.

□

### G.2.3. STRUCTURED BANDITS WITH BOUNDED ELUDER DIMENSION (EXAMPLE A.3)

In this section, we give generic bounds on the uniform exploration coefficient and Allocation-Estimation Coefficient for structured bandit classes with bounded eluder dimension (cf. Definition A.5). These result are used by subsequent examples, including linear bandits.

**Lemma G.3.** *Let $\mathcal{M}\{M(\pi) = \mathcal{N}(f(\pi), 1) \mid f \in \mathcal{F}\}$. Then for all $\varepsilon > 0$, we have*

$$C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \leq \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\varepsilon}/2)}{\varepsilon}.$$

**Proof of Lemma G.3.** Let $\xi \in \triangle(\mathcal{M})$. Recall the expression for KL divergence between Gaussians of unit variance:

$$\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_p[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] = \frac{1}{2}\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_p[(f^M(\pi) - f^{\overline{M}}(\pi))^2]].$$

Abbreviate $d_{\mathrm{E}} := d_{\mathrm{E}}(\mathcal{F}, \sqrt{\varepsilon}/2)$ and let $\{\pi_1, \ldots, \pi_{d_{\mathrm{E}}}\}$ denote a maximal sequence of $\varepsilon$-independent points. By the definition of the eluder dimension, for any $\pi \in \Pi$ and any $M, \overline{M} \in \mathcal{M}$, we have:

$$\sqrt{\sum_{i=1}^{d_{\mathrm{E}}}(f^M(\pi_i) - f^{\overline{M}}(\pi_i))^2} \leq \sqrt{\varepsilon/2} \implies |f^M(\pi) - f^{\overline{M}}(\pi)| \leq \sqrt{\varepsilon/2}.$$

Now, set $p$ to be the uniform distribution over $\{\pi_1, \ldots, \pi_{d_{\mathrm{E}}}\}$. Assume that $M, M' \in \mathcal{M}$ are such that

$$\max_{M''\in\{M,M'\}} \mathbb{E}_{\overline{M}\sim\xi}\, \mathbb{E}_p[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)] = \frac{1}{2} \max_{M''\in\{M,M'\}} \mathbb{E}_{\overline{M}\sim\xi}\, \mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \leq \varepsilon/(16 d_{\mathrm{E}}),$$

Markov's inequality implies that for each $M'' \in \{M, M'\}$, with probability at least $3/4$ over the draw of $\overline{M} \sim \xi$,

$$\mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \leq \varepsilon/(2 d_{\mathrm{E}}).$$

Taking a union bound, we conclude that with probability at least $1/2$ over the draw of $\overline{M} \sim \xi$,

$$\max_{M'' \in \{M, M'\}} \mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \le \varepsilon/(2d_{\mathrm{E}}). \tag{74}$$

Going forward, let $\overline{M} \in \mathcal{M}$ be any model such that Eq. (74) holds; we have just proven that such a model exists. It follows from the maximality of $\pi_1, \dots, \pi_{d_{\mathrm{E}}}$ and the definition of $p$ that for all $\widetilde{\pi} \in \Pi$, and $M'' \in \mathcal{M}$,

$$\mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \le \varepsilon/(2d_{\mathrm{E}}) \implies (f^{M''}(\widetilde{\pi}) - f^{\overline{M}}(\widetilde{\pi}))^2 \le \varepsilon/2.$$

In particular, since this holds for both $M \in \{M', M''\}$, and since Eq. (74) holds, we have that for all $\pi$,

$$\begin{aligned} D_{\mathsf{KL}}\big(M'(\pi) \,\|\, M''(\pi)\big) &= \frac{1}{2}(f^{M'}(\pi) - f^{M''}(\pi))^2 \\ &\le (f^{M'}(\pi) - f^{\overline{M}}(\pi))^2 + (f^{\overline{M}}(\pi) - f^{M''}(\pi))^2 \\ &\le \varepsilon. \end{aligned}$$

As this is the condition required by Definition A.4, it suffices to take $C^{\xi}_{\mathrm{exp}}(\varepsilon) = 16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\varepsilon}/2)/\varepsilon$. Since this bound holds uniformly for all choices of $\xi$, the result follows. □

**Proof of Proposition A.3.** The bound

$$C_{\mathrm{exp}}(\mathcal{M}^\star, \delta) \le \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)}{\delta}.$$

follows from Lemma G.3, since $d_{\mathrm{E}}(\mathcal{F}', \delta) \le d_{\mathrm{E}}(\mathcal{F}, \delta)$ for all $\mathcal{F}' \subseteq \mathcal{F}$.

By Lemma E.6 we can bound $\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star)$:

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \le C_{\mathrm{exp}}(\mathcal{M}^\star, \delta) \quad \text{for} \quad \delta = \min_{M \in \mathcal{M}^\star} \min\left\{ \min\left\{ \frac{1}{81 L_{\mathsf{KL}}}, \frac{\Delta^M_{\mathrm{min}}}{34 V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M/\Delta^M_{\mathrm{min}} + \mathsf{n}^M_{\varepsilon/36}}, \frac{\Delta^M_{\mathrm{min}}}{3} \right\}^2.$$

Lemma Lemma E.2 implies that for all $M \in \mathcal{M}^\star$,

$$\mathsf{g}^M \le C_{\mathrm{exp}}(\mathcal{M}^\star, \tfrac{1}{4}(\Delta^M_{\mathrm{min}})^2) \le \frac{64 d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta^M_{\mathrm{min}})}{(\Delta^M_{\mathrm{min}})^2} \le \frac{64 d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta_\star)}{\Delta_\star^2}$$

where we have used that the eluder dimension increases as its scale $\varepsilon$ decreases; by Lemma G.1, it suffices to take $L_{\mathsf{KL}} = V_{\mathcal{M}} = 2$. A sufficient value for $\delta$ is therefore

$$\delta = c \cdot \frac{\varepsilon^2 \Delta_\star^8}{d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta_\star)^2}$$

The result follows.

□

### G.2.4. LINEAR BANDITS (EXAMPLE A.4)

**Proof of Proposition A.4.** The result follows directly from Proposition A.3, since it is known that linear bandits have eluder dimension which scales as $d_{\mathrm{E}}(\mathcal{F}, \varepsilon) = O(d \cdot \log 1/\varepsilon)$ (Russo and Van Roy, 2013). □

In what follows, we prove Proposition A.5, providing sufficient conditions under which it is possible to bound the regularity constant $L^\star_{\mathcal{M}}$ (and hence $\mathsf{n}^\star_\varepsilon$) for linear bandits.

We begin with an geometric assumption on $\Theta$, $\mathcal{X}$, and $\theta^\star$, which we will show ensures that $M^\star(\pi) = \mathcal{N}(\langle x_\pi, \theta^\star \rangle, 1)$ is a regular model. To state our condition, we denote, for any vectors $x$ and $y$, we define $x_y$ and $x_{\bar{y}}$ to be unique vectors satisfying $x = x_y + x_{\bar{y}}$ for $x_y \parallel y$ and $x_{\bar{y}} \perp y$.

**Assumption G.1** (Regular Linear Bandits)**.** *The sets $\Theta, \mathcal{X}$ and model parameter $\theta^\star$ satisfy:*

1. *$\Theta$ is a convex polytope.*
2. *For all $\theta \in \Theta$, we have that there exists some $\delta_\theta > 0$ such that $\{\theta' \in \mathbb{R}^d \ : \ \|\theta' - \theta\|_2 \le \delta_\theta\} \subseteq \Theta$.*
3. *Letting $x^\star \in \mathcal{X}$ denote the optimal action for $\theta^\star$, we have*

$$\left\{\theta \in \mathbb{R}^d \ : \ \|\theta - \theta^\star\|_2 \le \max_{x \in \mathcal{X}, x \neq x^\star} \Delta^\star(x)/\|x_{\bar{x}^\star}\|_2\right\} \subseteq \Theta.$$

The first two points above are quite mild. The primary restriction of Assumption G.1 is Point 3, which requires that $\theta^\star$ is located sufficiently far within the interior of $\Theta$. Using Assumption G.1 we can state the full version of Proposition A.5.

**Proposition G.2** (Full Version of Proposition A.5)**.** *If $\Theta, \mathcal{X}$, and $\theta^\star$ satisfy Assumption G.1, then $\mathsf{n}^\star_\varepsilon$ is bounded by a polynomial function of $d, 1/\Delta^\star_{\min}, 1/\varepsilon, \mathsf{g}^\star$, and a geometry-dependent term scaling with the structure of $\mathcal{X}$ and $\Theta$.*

**Remark G.1** (Comparison to Existing Work)**.** *We remark that Assumption G.1 is similar to the conditions required by existing works which achieve instance-optimality in linear bandits with polynomial lower-order terms (Tirinzoni et al., 2020; Kirschner et al., 2021). Though neither of these works explicitly states such a condition, closer inspection of their analysis reveals it is indeed required. In particular, the proof of Lemma 1 of Tirinzoni et al. (2020) relies on a result from Degenne et al. (2020b) which shows that a condition analogous to Definition G.1 is met for linear bandits. However, the proof given in Degenne et al. (2020b) appears to only hold when $\Theta$ is unbounded, or a condition such as Assumption G.1 holds. As Tirinzoni et al. (2020) assumes that $\Theta$ is bounded, their results therefore only appear to hold if a condition similar to Assumption G.1 also holds. Similarly, in the proof of Lemma 10 of Kirschner et al. (2021), it is assumed that for every arm $x \neq x_{\pi_\star}$, there exists some instance in the alternate set with optimal arm $x$. To satisfy this condition, it appears that an assumption similar to Assumption G.1 is required.*

*Thus, while not stated explicitly in the existing literature, it therefore seems that all existing results which obtain reasonable lower-order terms require an assumption similar to Assumption G.1. Removing this assumption (or showing it is necessary) is an interesting direction for future work.*

**Proof of Proposition G.2.** Under Assumption G.1, this follows directly from Lemma G.4 and Proposition G.1. □

**Lemma G.4.** *Under Assumption G.1, the linear bandit model $M^\star$ is regular for some $L^\star_{\mathcal{M}} < \infty$ whose value depends on the geometry of $\Theta$ and $\mathcal{X}$.*

**Proof of Lemma G.4.** Fix some $\theta^\star \in \Theta$ and let $x^\star$ denote its optimal arm. Let $\Theta^{\mathrm{alt}}(\theta^\star) \subseteq \Theta$ denote parameters with optimal arm $x \neq x^\star$. Assume there exists some $\theta \in \Theta^{\mathrm{alt}}(\theta^\star)$ such that $\langle \theta - \theta^\star, x^\star \rangle \neq 0$ (if this is not the case, $M^\star$ immediately satisfies Definition G.1 and we are done). Let $\Theta_x = \{\theta \in \Theta \ : \ \langle x, \theta \rangle \geq \langle x^\star, \theta \rangle\}$, $\Theta^\star = \{\theta \in \mathbb{R}^d \ : \ \langle \theta - \theta^\star, x^\star \rangle = 0\}$, and $\Theta^\star_x = \Theta_x \cap \Theta^\star$. We first show that, under Assumption G.1, $\Theta^\star_x \neq \varnothing$ for all $x \in \mathcal{X}$, $x \neq x^\star$. We then use this to show that $M^\star(\pi) = \mathcal{N}(\langle x_\pi, \theta^\star \rangle, 1)$ is a regular model.

**Part 1: $\Theta^\star_x \neq \varnothing$.** Fix some $x \in \mathcal{X}$ with $x \neq x^\star$. Consider $\theta = \theta^\star + a x_{\bar{x}^\star}$ for some $a \in \mathbb{R}$ to be chosen. By construction we have $\langle \theta, x^\star \rangle = \langle \theta^\star, x^\star \rangle$, which implies that $\theta \in \Theta^\star$ for all $a \in \mathbb{R}$. We wish to choose $a$ large enough that $\langle \theta, x \rangle \geq \langle \theta, x^\star \rangle$. Note that

$$\langle \theta, x \rangle = \langle \theta^\star, x \rangle + a \langle x_{\bar{x}^\star}, x_{\bar{x}^\star} + x_{x^\star} \rangle = \langle \theta^\star, x \rangle + a \|x_{\bar{x}^\star}\|_2^2$$

and $\langle \theta, x^\star \rangle = \langle \theta^\star, x^\star \rangle$. Thus, to satisfy $\langle \theta, x \rangle \geq \langle \theta, x^\star \rangle$, we need

$$a \|x_{\bar{x}^\star}\|_2^2 \geq \langle \theta^\star, x^\star - x \rangle \iff a \geq \Delta^\star(x) / \|x_{\bar{x}^\star}\|_2^2.$$

Let $a = \Delta^\star(x)/\|x_{\bar{x}^\star}\|_2^2$, then it follows that $\langle \theta, x \rangle \geq \langle \theta, x^\star \rangle$. Furthermore, we can bound

$$\|\theta - \theta^\star\|_2 \leq a \|x_{\bar{x}^\star}\|_2 = \Delta^\star(x) / \|x_{\bar{x}^\star}\|_2.$$

Under Assumption G.1, it follows that $\theta \in \Theta$.

**Part 2: $M^\star$ is a Regular Model.** Let $\overline{\Theta}_x = \{\theta - \theta^\star \ : \ \theta \in \Theta_x\}$. Note that, since $\Theta$ is a convex polytope, and $\Theta_x$ simply adds a linear inequality constraint, $\Theta_x$ is also convex. Let $\overline{\Theta}^\star_x = \{\phi \in \overline{\Theta}_x \ : \ \langle \phi, x^\star \rangle = 0\}$. From Part 1, we have $\overline{\Theta}^\star_x \neq \varnothing$. Lemma 23 of Kirschner et al. (2021) then gives that there exists a geometry-dependent constant $C(\Theta, \mathcal{X})$ such that, for all $\phi \in \overline{\Theta}_x$:

$$\min_{\phi' \in \overline{\Theta}^\star_x} \|\phi - \phi'\|_2 \leq C(\Theta, \mathcal{X}) \cdot |\langle \phi, x^\star \rangle|.$$

This implies that for all $\theta \in \Theta_x$, we have:

$$\min_{\theta' \in \Theta^\star_x} \|\theta - \theta'\|_2 \leq C(\Theta, \mathcal{X}) \cdot |\langle \theta - \theta^\star, x^\star \rangle|.$$

Now consider some $\theta \in \Theta^{\mathrm{alt}}(\theta^\star)$, and assume that $\langle \theta - \theta^\star, x^\star \rangle \neq 0$ (by assumption such a $\theta$ exists). Assume that $\theta$ has optimal arm $x$, which implies that $\theta \in \Theta_x$. By what we have just shown, we know that there exists some $\theta' \in \Theta$ with $\langle \theta', x \rangle > \langle \theta', x^\star \rangle$ so that $\theta' \in \Theta^{\mathrm{alt}}(\theta^\star)$, $\langle \theta' - \theta^\star, x^\star \rangle = 0$, and

$$\|\theta - \theta'\|_2 \leq C(\Theta, \mathcal{X}) \cdot |\langle \theta - \theta^\star, x^\star \rangle|.$$

Note that, for any $x' \in \mathcal{X}$, we have

$$|D_{\mathsf{KL}}\big(\theta^\star(x') \,\|\, \theta(x')\big) - D_{\mathsf{KL}}\big(\theta^\star(x') \,\|\, \theta'(x')\big)| = \frac{1}{2} |\langle \theta^\star - \theta, x' \rangle^2 - \langle \theta^\star - \theta', x' \rangle^2|$$

$$\leq 2 \max_{\theta'' \in \Theta} |\langle \theta'', x' \rangle| \cdot |\langle \theta - \theta', x' \rangle|$$
$$\leq 2 \max_{\theta'' \in \Theta} \|\theta''\|_2 \|x'\|_2^2 \cdot \|\theta - \theta'\|_2.$$

Furthermore, note that

$$\sqrt{D_{\mathsf{KL}}(\theta^\star(x^\star) \,\|\, \theta(x^\star))} = \frac{1}{\sqrt{2}} |\langle \theta^\star - \theta, x^\star \rangle|,$$

so

$$|D_{\mathsf{KL}}\big(\theta^\star(x') \,\|\, \theta(x')\big) - D_{\mathsf{KL}}\big(\theta^\star(x') \,\|\, \theta'(x')\big)|$$
$$\leq \left( 2\sqrt{2} C(\Theta, \mathcal{X}) \max_{\theta'' \in \Theta, x'' \in \mathcal{X}} \|\theta''\|_2 \|x''\|_2^2 \right) \cdot \sqrt{D_{\mathsf{KL}}(\theta^\star(x^\star) \,\|\, \theta(x^\star))}.$$

As $\theta \in \Theta^\star(\theta^\star)$ was arbitrary, we have therefore shown that $M^\star$ is a regular model with

$$L^\star_{\mathcal{M}} = \left( 2\sqrt{2} \cdot C(\Theta, \mathcal{X}) \cdot \max_{\theta'' \in \Theta, x'' \in \mathcal{X}} \|\theta''\|_2 \|x''\|_2^2 \right)^2.$$

□

### G.2.5. GENERALIZED LINEAR MODELS (EXAMPLE A.5)

**Proof Sketch for Example A.5.** The bound on the AEC follows as in Proposition A.4, using that the eluder dimension for generalized linear models is bounded as $O(d \cdot (\frac{g_{\max}}{g_{\min}})^2 \cdot \log \frac{g_{\max}}{\varepsilon})$ (Russo and Van Roy, 2013). For the other regularity assumptions, note that by the Mean Value Theorem, we have

$$|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(M(\pi) \,\|\, M''(\pi)\big)| = \frac{1}{2} |(g(\langle \theta, x \rangle) - g(\langle \theta', x \rangle))^2 - (g(\langle \theta, x \rangle) - g(\langle \theta'', x \rangle))^2|$$
$$\leq 2 g_{\max} |\langle \theta' - \theta'', x \rangle|$$

and

$$\sqrt{D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))} = \frac{1}{\sqrt{2}} |g(\langle \theta, x \rangle) - g(\langle \theta', x \rangle)| \geq \frac{g_{\min}}{\sqrt{2}} |\langle \theta - \theta', x \rangle|.$$

In light of these inequalities, bounds on all relevant regularity parameters for generalized linear bandits follow from similar reasoning to the proofs for linear bandits. In particular, the conclusion of Lemma G.4 holds for generalized linear bandits under Assumption G.1, with $L^\star_{\mathcal{M}}$ as in Lemma G.4, but scaled by $(\frac{g_{\max}}{g_{\min}})^2$).

□

### G.3. Contextual Bandits with Finitely Many Actions (Example A.6)

In this setting we take $D\big(M(\pi) \,\|\, \overline{M}(\pi)\big) \leftarrow D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)$ for $D$ the divergence employed by $\mathsf{AE}^2_\star$. Note that we have

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big) = \frac{1}{2}\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^M(x,a) - f^{\overline{M}}(x,a))^2]].$$

**Lemma G.5.** *In the contextual bandits setting of Example A.6:*

1. *Assumptions A.2, D.1 and D.2 hold with parameters*

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = 2\sqrt{2}$$

*and* $D(\cdot \,\|\, \cdot) = D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$.

2. *We can bound* $\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho, \mu)$ *by the covering number of* $\mathcal{M}$ *in the distance* $d(M, M') = \sup_{x\in\mathcal{X}, a\in\mathcal{A}} |f^M(x,a) - f^{M'}(x,a)|$ *at tolerance* $\frac{\sigma^2\cdot\rho}{2+\sqrt{2\log(2/\mu)}}$. *Furthermore, it suffices to take* $\mathcal{E} := \{|r| \le 1 + \sqrt{2\log(2/\mu)}\}$.

**Proof of Lemma G.5.** Using the expression for the KL divergence given above, for any $M, M', \overline{M} \in \mathcal{M}$ and $\pi \in \Pi$, we have

$$\begin{aligned}
&\big|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big)\big| \\
&\quad= \frac{1}{2}\left|\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^M(x,a) - f^{M'}(x,a))^2]] - \mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^{\overline{M}}(x,a) - f^{M'}(x,a))^2]]\right| \\
&\quad\le \frac{1}{2}\mathbb{E}_{x\sim p_{\mathcal{X}}}\left[\mathbb{E}_{a\sim\pi(x)}\left[\left|(f^M(x,a) - f^{M'}(x,a))^2 - (f^{\overline{M}}(x,a) - f^{M'}(x,a))^2\right|\right]\right] \\
&\quad\le 2\mathbb{E}_{x\sim p_{\mathcal{X}}}\left[\mathbb{E}_{a\sim\pi(x)}\left[\left|f^M(x,a) - f^{\overline{M}}(x,a)\right|\right]\right] \\
&\quad\le 2\sqrt{\mathbb{E}_{x\sim p_{\mathcal{X}}}\left[\mathbb{E}_{a\sim\pi(x)}\left[(f^M(x,a) - f^{\overline{M}}(x,a))^2\right]\right]} \\
&\quad= 2\sqrt{2}\sqrt{D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)}.
\end{aligned}$$

This verifies Assumption D.1 holds with $L_{\mathsf{KL}} = 2\sqrt{2}$. Assumption D.2 is immediate.

To show that Assumption A.2 is met, we note that

$$\log\frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} = \log\frac{\mathbb{P}^{\overline{M},\pi}(r \mid o)\mathbb{P}^{\overline{M},\pi}(o)}{\mathbb{P}^{M,\pi}(r \mid o)\mathbb{P}^{M,\pi}(o)} = \log\frac{\mathbb{P}^{\overline{M},\pi}(r \mid o)}{\mathbb{P}^{M,\pi}(r \mid o)},$$

where the second equality holds because the context distribution is identical for all models. As the reward likelihoods conditioned on the context are Gaussian, a calculation similar to Lemma G.1 shows that Assumption A.2 with $V_{\mathcal{M}} = 2\sqrt{2}$. The covering number bound also follows from the same reasoning as Lemma G.1. □

**Lemma G.6.** *For the contextual bandit setting described above, we can bound*

$$C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \le \frac{4A}{\varepsilon}.$$

**Proof of Lemma G.6.** Fix some $\xi \in \triangle_{\mathcal{M}}$ and let $\pi_{\mathrm{exp}}$ be uniform over $\mathcal{A}$ for each context. Then, for any $p' \in \triangle_{\Pi}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p'}[D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big)] \\
&= \frac{1}{2}\mathbb{E}_{\pi\sim p'}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^{M}(x,a) - f^{M'}(x,a))^2]]] \\
&\le \mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p'}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^{M}(x,a) - f^{\overline{M}}(x,a))^2 + (f^{\overline{M}}(x,a) - f^{M'}(x,a))^2]]]] \\
&\le \sum_{a\in\mathcal{A}} \mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[(f^{M}(x,a) - f^{\overline{M}}(x,a))^2 + (f^{\overline{M}}(x,a) - f^{M'}(x,a))^2]] \\
&= A\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim p_{\mathrm{exp}}(x)}[(f^{M}(x,a) - f^{\overline{M}}(x,a))^2 + (f^{\overline{M}}(x,a) - f^{M'}(x,a))^2]]] \\
&\le 2A\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi_{\mathrm{exp}}) \,\|\, M(\pi_{\mathrm{exp}})\big) + D_{\mathsf{KL}}\big(\overline{M}(\pi_{\mathrm{exp}}) \,\|\, M'(\pi_{\mathrm{exp}})\big)].
\end{aligned}$$

It follows that if

$$\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi_{\mathrm{exp}}) \,\|\, M(\pi_{\mathrm{exp}})\big)] \le \frac{\varepsilon}{4A} \quad \text{and} \quad \mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi_{\mathrm{exp}}) \,\|\, M'(\pi_{\mathrm{exp}})\big)] \le \frac{\varepsilon}{4A},$$

then we can bound $\mathbb{E}_{\pi\sim p'}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \le \varepsilon$. Thus, choosing $p_{\mathrm{exp}} \in \triangle(\Pi)$ to place probability mass 1 on $\pi_{\mathrm{exp}}$, a sufficient bound on $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon)$ is $4A/\varepsilon$. □

**Proof of Proposition A.7.** The bound on $C_{\mathrm{exp}}(\mathcal{M}^\star, \varepsilon)$ follows from Lemma G.6. Hence, by Lemma E.6 we can bound $\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star)$ as:

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \le \frac{4A}{\delta} \quad \text{for} \quad \delta = \min_{M\in\mathcal{M}^\star} \min\left\{ \min\left\{ \frac{1}{81L_{\mathsf{KL}}}, \frac{\Delta_{\mathrm{min}}^M}{34V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M/\Delta_{\mathrm{min}}^M + \mathsf{n}_{\varepsilon/36}^M}, \frac{\Delta_{\mathrm{min}}^M}{3} \right\}^2.$$

By Lemma E.2, we have that for all $M \in \mathcal{M}^\star$,

$$\mathsf{g}^M \le C_{\mathrm{exp}}(\mathcal{M}^\star, \frac{1}{4}(\Delta_{\mathrm{min}}^M)^2) \le \frac{16A}{\Delta_\star^2}.$$

By Lemma G.5, we can take $L_{\mathsf{KL}} = V_{\mathcal{M}} = 2\sqrt{2}$. A sufficient choice for $\delta$ is therefore

$$\delta = c \cdot \frac{\varepsilon^2\Delta_\star^8}{A^2}.$$

The result follows. □

### G.4. Informative Arms (Example A.1)

In this section, we provide calculations for the bandits with informative arms setting in Example A.1. We first show that Assumptions A.1 to A.3 are satisfied.

- Lemma G.1 If $\pi \in [A]$, then the response is simply Gaussian, so by Lemma G.1, the condition of Assumption A.1 is met with $L_{\mathsf{KL}} = 2$. If $\pi = \pi_i^\circ$, then by the Mean Value Theorem we have

$$\big|D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big)\big|$$

$$
\begin{aligned}
&= \left| \sum_{a \in [A]} \mathbb{P}^{M,\pi}(a) \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} - \sum_{a \in [A]} \mathbb{P}^{\overline{M},\pi}(a) \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \\
&\leq \left(1 + \max_{a \in [A]} \max \left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \right\}\right) \cdot \sum_{a \in [A]} |\mathbb{P}^{M,\pi}(a) - \mathbb{P}^{\overline{M},\pi}(a)| \\
&= \left(1 + \max_{a \in [A]} \max \left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \right\}\right) \cdot D_{\mathsf{TV}}\left(\mathbb{P}^{M,\pi}, \mathbb{P}^{\overline{M},\pi}\right) \\
&\leq \left(1 + \max_{a \in [A]} \max \left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \right\}\right) \cdot \sqrt{\frac{1}{2} D_{\mathsf{KL}}(\mathbb{P}^{M,\pi} \,\|\, \mathbb{P}^{\overline{M},\pi})}.
\end{aligned}
$$

Using the bound on the log-likelihood ratio given above, this verifies that Assumption A.1 holds with $L_{\mathsf{KL}} = \max\{2, 1 + \log \frac{A}{1-\beta}\}$.

- If $\pi \in [A]$, then the since the response is Gaussian, by Lemma G.1, the condition of Assumption A.2 is met with $V_{\mathcal{M}} = 2$. If $\pi = \pi_i^\circ$, then for $M \in \mathcal{M}$, either the observation is distributed as $1/A$, so $\mathbb{P}^{M,\pi}(r, o) = 1/A$ for all $o \in [A]$, or $i$ is the informative arm for instance $M$, in which case $\mathbb{P}^{M,\pi}(r, o) = (1-\beta)/A$ for $o \neq \pi_M$, and $\mathbb{P}^{M,\pi}(r, o) = \beta + (1-\beta)/A$ for $o = \pi_M$ (note that we can disregard $o = \perp$ since it occurs with probability 0 if an informative arm is pulled). The log-likehood ratio is then at most

$$
\log \frac{\beta + (1-\beta)/A}{(1-\beta)/A} \leq \log \frac{A}{1-\beta}.
$$

Thus, Assumption A.2 is satisfied with $V_{\mathcal{M}} = \max\{2, \log \frac{A}{1-\beta}\}$.

- Using Lemma G.1, it is easy to see that Assumption A.3 is met with $d_{\mathsf{cov}} = O(A)$ and $C_{\mathsf{cov}} = O(1)$.

To bound the parameter $\mathsf{n}^{\mathcal{M}}$, consider $M \in \mathcal{M}$ and $M' \in \mathcal{M}^{\mathrm{alt}}(M)$. Note that either $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) = 0$, in which case there is no advantage to playing $\pi_M$, or, due to the discretization of the means, $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq \frac{1}{2}\Delta_{\min}^2$. Thus, for any allocation $\eta \in \mathbb{R}_+^\Pi$, as long as $\eta(\pi_M) \geq \frac{2}{\Delta_{\min}^2}$, we have $\eta(\pi_M) D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq 1$. It follows that there is no advantage to choosing $\eta(\pi_M)$ larger than $\frac{2}{\Delta_{\min}^2}$, so we can bound $\mathsf{n}^{\mathcal{M}} \leq \frac{2}{\Delta_{\min}^2}$.

### G.4.1. BOUNDING THE ALLOCATION-ESTIMATION COEFFICIENT

We begin with some basic observations. First, since we restrict $\mathcal{M}$ to only contain instances with a single optimal decision, if $f^M(\pi_M) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$ for some $M \in \mathcal{M}$, this implies that $f^M(\pi) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$ for all $\pi \neq \pi_M$. Fix some $M \in \mathcal{M}$ satisfying $f^M(\pi_M) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$. It follows that, for every $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, it must be the case that $f^{M'}(\pi_M) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$. Therefore, since $M$ and $M'$ have different reward means at $\pi_M$, and since this holds for all $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, $M$ can be distinguished from every $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ by playing $\pi_M$. In this case, then, $\mathsf{g}^M = 0$, so any $\varepsilon$-optimal Graves-Lai allocation must put all its mass on $\pi_M$, implying $\Lambda(M, \varepsilon) = \{\mathbb{I}_{\pi_M}\}$. Denote such instances $M$ with $f^M(\pi_M) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$ as $\overline{\mathcal{M}}$. Note that for $M$ with $f^M(\pi_M) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$, we have $\mathbb{I}_{\pi_M^\circ} \in \Lambda(M, \varepsilon)$.

We proceed to bound the value of the  Fix $\overline{M} \in \mathrm{co}(\mathcal{M})$. For a first case, assume that that $f^{\overline{M}}(\pi_{\overline{M}}) \leq \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min} - \frac{1}{2}\Delta_{\min}$ and let $k = \arg\max_{i \in [N]} \mathbb{P}^{\overline{M}, \pi_i^\circ}(o = \pi_{\overline{M}})$ denote the index of the most informative arm for $\overline{M}$. Let $\lambda = \mathbb{I}_{\pi_k^\circ}$, and note that $\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ contains only instances in $\mathcal{M}$ that have informative arm $k$. Let $\mathcal{M}' = \{M \in \mathcal{M} \,:\, \pi_M^\circ \neq \pi_k^\circ\} \cup \overline{\mathcal{M}}$. Then $\mathcal{M} \backslash \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda) \subseteq \mathcal{M}'$. Let $\omega = \frac{1}{2}\mathrm{Unif}(\{\pi_i^\circ\}_{i \in [N]}) + \frac{1}{2}\mathrm{Unif}([A])$. Then,

$$\begin{aligned}
\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) &\leq \sup_{M \in \mathcal{M}'} \frac{1}{\mathbb{E}_{\pi \sim \omega}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]} \\
&\leq \sup_{M \in \mathcal{M}, \pi_M^\circ \neq \pi_k^\circ} \frac{2N}{\sum_{i=1}^N D_{\mathsf{KL}}\big(\overline{M}(\pi_i^\circ) \,\|\, M(\pi_i^\circ)\big)} + \sup_{M \in \overline{\mathcal{M}}} \frac{2A}{\sum_{\pi \in [A]} D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)}
\end{aligned}$$

If $\pi_M^\circ \neq \pi_k^\circ$, this implies that $o \sim M(\pi_i^\circ)$ is uniform on $[A]$ for $\pi_i^\circ \neq \pi_M^\circ$, and $o \sim M(\pi_i^\circ)$ is distributed as $\beta \mathbb{I}_{\pi_M} + (1-\beta)\mathrm{Unif}([A])$ for $\pi_i^\circ = \pi_M^\circ$. Note that since $k = \arg\max_{i \in [N]} \mathbb{P}^{\overline{M}, \pi_i^\circ}(o = \pi_{\overline{M}})$ and $\overline{M} \in \mathrm{co}(\mathcal{M})$, we can have at most $\mathbb{P}^{\overline{M}, \pi_i^\circ}(o) \leq 1/A + \beta/2$ for all $o$ if $i \neq k$, since if this were not the case, then $i$ must be $k$. It follows from Pinsker's inequality that for $M$ with $\pi_M^\circ \neq \pi_k^\circ$:

$$\begin{aligned}
D_{\mathsf{KL}}\big(\overline{M}(\pi_M^\circ) \,\|\, M(\pi_M^\circ)\big) &\geq 2 D_{\mathsf{TV}}\big(M(\pi_M^\circ), \overline{M}(\pi_M^\circ)\big)^2 \\
&\geq 2|\mathbb{P}^{M, \pi_M^\circ}(o = \pi_M) - \mathbb{P}^{\overline{M}, \pi_M^\circ}(o = \pi_M)|^2 \\
&= 2|\beta - 1/A - \beta/2|^2 \\
&\geq 2|\beta/4|^2
\end{aligned}$$

where the last inequality uses our assumption that $\beta \geq 4/A$. For $M \in \overline{\mathcal{M}}$, since $f^{\overline{M}}(\pi_{\overline{M}}) \leq \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min} - \frac{1}{2}\Delta_{\min}$, we have that

$$D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) \geq \frac{1}{8}\Delta_{\min}^2.$$

Thus, we can bound

$$\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \frac{64N}{\beta^2} + \frac{16A}{\Delta_{\min}^2}.$$

Now, consider the second case where $\overline{M}$ has $f^{\overline{M}}(\pi_{\overline{M}}) > \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min} - \frac{1}{2}\Delta_{\min}$. Note that in this case we must have $|\boldsymbol{\pi}_{\overline{M}}| = 1$. Set $\lambda = \mathbb{I}_{\pi_{\overline{M}}}$. Then we have that $\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ contains every instance except the single instance with $f^M(\pi_{\overline{M}}) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$. Let $\omega = \mathbb{I}_{\pi_{\overline{M}}}$. Note that for any instance with $f^M(\pi_{\overline{M}}) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$, i.e. every instance in $\mathcal{M} \backslash \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$, we have

$$D_{\mathsf{KL}}\big(\overline{M}(\pi_{\overline{M}}) \,\|\, M(\pi_{\overline{M}})\big) \geq \frac{1}{8}\Delta_{\min}^2.$$

It follows that with such an $\overline{M}$, we can bound

$$\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \frac{8}{\Delta_{\min}^2}.$$

### G.5. Tabular Reinforcement Learning (Appendix A.7)

In this section, we prove all of the claims in Appendix A.7 concerning tabular reinforcement learning.

Throughout this section we let $M_{sh}(a)$ denote the joint distribution of the next state and reward if we play action $a$ in state $s$ at step $h$ on MDP $M \in \mathcal{M}$. We also define

$$w_h^{M,\pi}(s,a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a)$$

as the state-action visitation probabilities on MDP $M$ under policy $\pi$ (and define $w_h^{M,\pi}(s)$ analogously). We let $r_h^M(s,a) = \mathbb{E}_{r\sim R_h^M(s,a)}[r]$ denote the mean reward on MDP $M$ at $(s,a,h)$, and let $r_h(s_h,a_h)$ the denote the realized (random reward) at step $h$. We let $r := (r_1(s_1,a_1),\dots,r_H(s_H,a_H))$ denote the vector of all random rewards in a given episode. $\tau = (s_1,\dots,s_H)$ denotes a trajectory of states, and $\tau_h = s_h$ the $h$th state in the trajectory. We denote the $Q$-value function on $M$ for policy $\pi$ by

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'}) \mid s_h = s, a_h = a\right]$$

and the value function by $V_h^{M,\pi}(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_h^{M,\pi}(s,a)]$. We denote the value of a policy by $V_1^{M,\pi} := V_1^{M,\pi}(s_1)$. For any function $V : \mathcal{S} \to \mathbb{R}$ we denote

$$\mathbb{P}_h^M[V](s,a) = \mathbb{E}_{s'\sim P_h^M(\cdot|s,a)}[V(s')].$$

For all results in this section concerning general divergences, we take $D(\cdot \,\|\, \cdot) \leftarrow D_{\mathsf{KL}}(\cdot\,\|\,\cdot)$.

**Proof of Proposition A.8.** To bound the AEC, we first move from KL divergence to Hellinger distance. Since we always have $D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big) \ge D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)$, we upper bound

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \le \sup_{\xi\in\triangle_{\mathcal{M}}} \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}^\star\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_\omega[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)]]}.$$

We then apply Lemma E.6 to bound this by $C^{\mathsf{D}}_{\mathrm{exp}}(\mathcal{M}^\star,\delta)$, with $D(\cdot\,\|\,\cdot) \leftarrow D^2_{\mathsf{H}}(\cdot,\cdot)$. The bound on $C^{\mathsf{D}}_{\mathrm{exp}}(\mathcal{M}^\star,\varepsilon)$ then follows directly from Lemma G.10, and gives

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \le \frac{SAH^2\cdot\log^2 H}{\delta^2} \quad \text{for} \quad \delta = \min_{M\in\mathcal{M}^\star} \min\left\{ \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\Delta^M_{\min}}{34V_{\mathcal{M}}}\right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M/\Delta^M_{\min} + \mathsf{n}^M_{\varepsilon/36}}, \frac{\Delta^M_{\min}}{3}\right\}^2.$$

By Lemma G.8, we have that Assumptions A.2 and D.1 hold with

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = 4H + \max_{\overline{M},\overline{M}'\in\mathcal{M}} \max_{\pi\in\Pi} \max_{\tau\in\mathcal{T}} \left|\log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau)}{\mathbb{P}^{\overline{M},\pi}(\tau)}\right|.$$

As we assume every transition has probability at least $P_{\min}$, we can lower bound $\mathbb{P}^{\overline{M},\pi}(\tau) \ge P_{\min}^H$, so it suffices to take

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = H(4 + \log 1/P_{\min}).$$

By Lemma E.2, for $M \in \mathcal{M}^\star$ we can bound

$$\mathrm{g}^M \leq C_{\exp}(\mathcal{M}^\star, \tfrac{1}{4}(\Delta^M_{\min})^2) \leq c \cdot \frac{SAH^2 \cdot \log^2 H}{\Delta_\star^4}.$$

A sufficient choice of $\delta$ is therefore

$$\delta = c \cdot \frac{\varepsilon^2 \Delta_\star^{12}}{S^2A^2H^6 \cdot \log^4 H \cdot \log^2 1/P_{\min}}.$$

□

### G.5.1. TABULAR MDPS ARE REGULAR CLASSES

**Lemma G.7.** *If $M^\star$ is such that $r_h^{M^\star}(s,a) \in [0, 1/H^2)$ for all $(s,a,h)$, then in the setting of $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$, $M^\star$ is a regular model with constant*

$$L^\star_{\mathcal{M}} = \frac{96}{\Delta^\star_{\min}}$$

**Proof of Lemma G.7.** Take some MDP $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ such that $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) > 0$. Let $M''$ be such that

$$D_{\mathsf{KL}}\big(M^\star_{sh}(\pi_\star(s,h)) \,\|\, M''_{sh}(\pi_\star(s,h))\big) = 0, \quad \forall s,h$$

and

$$D_{\mathsf{KL}}\big(M'_{sh}(a) \,\|\, M''_{sh}(a)\big) = 0, \quad \forall s,h,a \neq \pi_\star(s,h),$$

so that $M''$ is the MDP which is identical to $M^\star$ on optimal actions, and identical to $M'$ on suboptimal actions (recall that optimal actions for $M^\star$ are unique). By construction, we have that $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$. Furthermore, it is not difficult to see that $M'' \in \mathcal{M}$. In particular, to verify that $P_h^{M''}(s' \mid s,a) \geq P_{\min}$ for each $(s,a,h,s')$, we note that since $M^\star, M' \in \mathcal{M}$, for every $(s,a,h,s')$, we have $P_h^{M^\star}(s' \mid s,a) \geq P_{\min}$ and $P_h^{M'}(s' \mid s,a) \geq P_{\min}$. By construction, we have that $P_h^{M''}(\cdot \mid s,a)$ is identical to either $P_h^{M^\star}(\cdot \mid s,a)$ or $P_h^{M'}(\cdot \mid s,a)$ for each $(s,a,h)$, so it follows that $P_h^{M''}(s' \mid s,a) \geq P_{\min}$. The remaining conditions for inclusion in $\mathcal{M}$ are immediate. We consider two cases.

**Case 1:** $M'' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$**.** For $\pi \in \Pi$, by Lemma G.15 we have

$$D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M'(\pi)\big) = \sum_{s,a,h} w_h^{M^\star,\pi}(s,a) D_{\mathsf{KL}}\big(M^\star_{sh}(a) \,\|\, M'_{sh}(a)\big),$$

and

$$\begin{aligned} D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M''(\pi)\big) &= \sum_{s,a,h} w_h^{M^\star,\pi}(s,a) D_{\mathsf{KL}}\big(M^\star_{sh}(a) \,\|\, M''_{sh}(a)\big) \\ &= \sum_{s,a,h} w_h^{M^\star,\pi}(s,a) D_{\mathsf{KL}}\big(M^\star_{sh}(a) \,\|\, M'_{sh}(a)\big) \cdot \mathbb{I}\{a \neq \pi_\star(s,h)\} \end{aligned}$$

so that

$$\begin{aligned}
&|D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M''(\pi)\big)| \\
&= \sum_{s,h} w_h^{M^\star,\pi}(s,\pi_\star(s,h)) D_{\mathsf{KL}}\big(M^\star_{sh}(\pi_\star(s,h)) \,\|\, M'_{sh}(\pi_\star(s,h))\big) \\
&\le \sup_{s,h} \frac{w_h^{M^\star,\pi}(s,\pi_\star(s,h))}{w_h^{M^\star,\pi_\star}(s,\pi_\star(s,h))} \cdot D_{\mathsf{KL}}\big(M^\star(\pi_\star) \,\|\, M'(\pi_\star)\big) \\
&\le \sup_{s,h} \frac{1}{w_h^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}\big(M^\star(\pi_\star) \,\|\, M'(\pi_\star)\big) \\
&\le \frac{1}{\Delta^\star_{\min}} \cdot D_{\mathsf{KL}}\big(M^\star(\pi_\star) \,\|\, M'(\pi_\star)\big)
\end{aligned}$$

where the last inequality follows from Lemma G.13. Thus, in this case, $M^\star$ is a regular model with

$$L^\star_{\mathcal{M}} = \frac{1}{\Delta^\star_{\min}}.$$

**Case 2:** $M'' \notin \mathcal{M}^{\mathrm{alt}}(M^\star)$**.** Let $(\widetilde{s},\widetilde{a},\widetilde{h})$ be such that $Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\widetilde{a}) > Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$, and note that such a tuple is guaranteed to exist by Lemma G.11 since $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$. Let $\widetilde{M}''$ denote an MDP that is identical to $M''$ everywhere except for at $(\widetilde{s},\widetilde{h},\widetilde{a})$, where we set $r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})$ so that

$$Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\widetilde{a}) = Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) + \delta \tag{75}$$

for some $\delta > 0$ to be chosen. This will ensure $\widetilde{a}$ is the optimal action in $(\widetilde{s},\widetilde{h})$, so $\pi_{\widetilde{M}''} \neq \pi_\star$. By construction we have that $M^\star$ and $\widetilde{M}''$ behave identically on $\pi_\star$, which implies that $Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) = Q_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$. Furthermore, by assumption we have $r_h^{M^\star}(s,a) < 1/H^2$ for all $(s,a,h)$, which implies $Q_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) < 1/H$. As $Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\widetilde{a}) = r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) + \mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}) \ge r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})$, it follows that for small enough $\delta$, we can ensure Eq. (75) is met with $r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) < 1/H$, so that $\widetilde{M}'' \in \mathcal{M}$.

If we can show that, for all $\pi$, $|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widetilde{M}''(\pi)\big)|$ is bounded by some function of $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))$, we are then done. We proceed to show this. First, note that, similar to Case 1:

$$\begin{aligned}
&|D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widetilde{M}''(\pi)\big)| \\
&\qquad = \Big| \sum_{s,h} w_h^{M^\star,\pi}(s,\pi_\star(s,h)) D_{\mathsf{KL}}\big(M^\star_{sh}(\pi_\star(s,h)) \,\|\, M'_{sh}(\pi_\star(s,h))\big) \Big| \\
&\qquad\qquad + w_{\widetilde{h}}^{M^\star,\pi}(\widetilde{s},\widetilde{a}) \Big| D_{\mathsf{KL}}\big(M^\star_{\widetilde{s}\widetilde{h}}(\widetilde{a}) \,\|\, M'_{\widetilde{s}\widetilde{h}}(\widetilde{a})\big) - D_{\mathsf{KL}}\big(M^\star_{\widetilde{s}\widetilde{h}}(\widetilde{a}) \,\|\, \widetilde{M}''_{\widetilde{s}\widetilde{h}}(\widetilde{a})\big) \Big| \\
&\qquad \le \sup_{s,h} \frac{1}{w_h^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}\big(M^\star(\pi_\star) \,\|\, M'(\pi_\star)\big) \\
&\qquad\qquad + \frac{1}{2} w_{\widetilde{h}}^{M^\star,\pi}(\widetilde{s},\widetilde{a}) \Big| (r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}))^2 - (r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}))^2 \Big|
\end{aligned} \tag{76}$$

where the inequality follows by what we showed in Case 1, and since $M'$ and $\widetilde{M}''$ have identical transitions at $(\widetilde{s}, \widetilde{a}, \widetilde{h})$, so the contribution to the KL divergence from the transitions cancels, leaving only the KL divergence between unit-variance Gaussians. By the Mean Value Theorem and since rewards are in $[0, 1/H]$, we have

$$\frac{1}{2}\Big|(r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}))^2 - (r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}))^2\Big| \le \frac{1}{H}|r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})|.$$

Thus, it suffices to bound $|r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})|$.

By assumption $Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\widetilde{a}) > Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$. We can then ensure

$$Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\widetilde{a}) - Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) \le Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\widetilde{a}) - Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$$

for $\delta$ sufficiently small. This is equivalent to, abbreviating $\widetilde{a}_{M^\star} := \pi_\star(\widetilde{s},\widetilde{h})$:

$$\begin{aligned}
&r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) + \mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) \\
&\quad \le r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}) + \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}).
\end{aligned}$$

By construction we have that $V_h^{\widetilde{M}'',\pi_\star}(s) = V_h^{M^\star,\pi_\star}(s)$ for all $(s,h)$, $r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}_{M^\star}) = r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})$, and $\mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) = \mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})$, since $\widetilde{M}''$ behaves identically to $M$ on actions taken by $\pi_\star$. Furthermore, we have $\mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}) = \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a})$ since $\widetilde{M}''$ behaves identically to $M'$ on actions not taken by $\pi_\star$, other than the reward at $(\widetilde{s},\widetilde{a},\widetilde{h})$. Using these simplifications and rearranging, we get

$$\begin{aligned}
|r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a})| &\le |r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})| + |\mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| \\
&\quad + |\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a})| \\
&\le |r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})| + |\mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| \\
&\quad + |\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| \\
&\quad + |\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a})|.
\end{aligned}$$

Since rewards are unit-variance Gaussian, we have

$$|r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})| \le \sqrt{2D_{\mathsf{KL}}\big(M^\star_{\widetilde{h},\widetilde{s}}(\widetilde{a}_{M^\star}) \,\|\, M'_{\widetilde{h},\widetilde{s}}(\widetilde{a}_{M^\star})\big)} \le \sqrt{\frac{2}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})} D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.$$

Since $V_{\widetilde{h}+1}^{M^\star,\pi_\star} \in [0,1]$, we have

$$\begin{aligned}
|\mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| &\le \sum_{s'} |P_{\widetilde{h}}^{M}(s' \mid \widetilde{s},\widetilde{a}_{M^\star}) - P_{\widetilde{h}}^{M'}(s' \mid \widetilde{s},\widetilde{a}_{M^\star})| \\
&\le 2D_{\mathsf{TV}}\Big(P_{\widetilde{h}}^{M}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star}), P_{\widetilde{h}}^{M'}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star})\Big) \\
&\le \sqrt{2D_{\mathsf{KL}}\big(P_{\widetilde{h}}^{M}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star}) \,\|\, P_{\widetilde{h}}^{M'}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star})\big)}
\end{aligned}$$

$$\leq \sqrt{\frac{2}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})} D_{\mathsf{KL}}\big(P_{\widetilde{h}}^{M}(\cdot \mid \widetilde{s}, \widetilde{a}_{M^\star}) \,\|\, P_{\widetilde{h}}^{M'}(\cdot \mid \widetilde{s}, \widetilde{a}_{M^\star})\big)}$$
$$\leq \sqrt{\frac{2}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})} D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.$$

By Lemma G.12 we have

$$|\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s}, \widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s}, \widetilde{a}_{M^\star})| \leq \sum_{s'} P_{\widetilde{h}}^{M'}(s' \mid \widetilde{s}, \widetilde{a}_{M^\star}) |V_{\widetilde{h}+1}^{M^\star,\pi_\star}(s') - V_{\widetilde{h}+1}^{M',\pi_\star}(s')|$$
$$\leq \sum_{s'} P_{\widetilde{h}}^{M'}(s' \mid \widetilde{s}, \widetilde{a}_{M^\star}) \cdot \sqrt{\frac{8H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s')} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}$$
$$\leq \sup_s \sqrt{\frac{8H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}$$

and similarly

$$|\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s}, \widetilde{a}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s}, \widetilde{a})| \leq \sup_s \sqrt{\frac{8H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.$$

Altogether then:

$$|r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s}, \widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s}, \widetilde{a})| \leq \left(\sqrt{\frac{8}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})}} + \sup_s \sqrt{\frac{32H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s)}}\right) \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}$$
$$\leq \sup_{s,h} \sqrt{\frac{96H}{w_h^{M^\star,\pi_\star}(s)}} \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.$$

Combining this with Eq. (76), we have

$$|D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widetilde{M}''(\pi)\big)|$$
$$\leq \sup_{s,h} \frac{1}{w_h^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}\big(M^\star(\pi_\star) \,\|\, M'(\pi_\star)\big) + \sup_{s,h} \sqrt{\frac{96}{H w_h^{M^\star,\pi_\star}(s)}} \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}$$
$$\leq \frac{1}{\Delta_{\min}^\star} \cdot D_{\mathsf{KL}}\big(M^\star(\pi_\star) \,\|\, M'(\pi_\star)\big) + \sqrt{\frac{96}{H\Delta_{\min}^\star}} \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}$$

where the second inequality uses Lemma G.13. Thus, in this case $M^\star$ is a regular model with

$$L_{\mathcal{M}}^\star = \frac{96}{\Delta_{\min}^\star}.$$

□

### G.5.2. TABULAR MDPS SATISFY BASIC ASSUMPTIONS

**Lemma G.8.** *Tabular MDPs with unit-variance Gaussian rewards satisfy Assumptions A.2, D.1 and D.2 with*

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = 8H + \max_{\overline{M}, \overline{M}' \in \mathcal{M}} \max_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}', \pi}(\tau)}{\mathbb{P}^{\overline{M}, \pi}(\tau)} \right|,$$

*and $D(\cdot \,\|\, \cdot) = D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$, where $\mathcal{T} := \mathcal{S}^H$ and $\mathbb{P}^{\overline{M}, \pi}(\tau)$ denotes the probability of observing state sequence $\tau \in \mathcal{T}$ on $\overline{M}$ when playing policy $\pi$.*

**Proof of Lemma G.8.** We verify each assumption separately.

**Verifying Assumption D.1.** Fix some $M, M', \overline{M} \in \mathcal{M}$. Our goal is to bound

$$\left| D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big) \right|.$$

Let $\widetilde{M}$ denote the MDP with transitions identical to $M$ but rewards identical to $\overline{M}$. Then

$$\begin{aligned}\left| D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big) \right| \le{}& \left| D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) \right| \\ &+ \left| D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big) \right|\end{aligned}$$

We bound these terms separately. First, by Lemma G.15 we have

$$\begin{aligned}&D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) \\ &= \sum_{s,a,h} w_h^{M,\pi}(s,a) D_{\mathsf{KL}}\big(M_{sh}(\pi(s,h)) \,\|\, M'_{sh}(\pi(s,h))\big) \\ &= \sum_{s,a,h} w_h^{M,\pi}(s,a) \left[ \frac{1}{2}(r_h^M(s,a) - r_h^{M'}(s,a))^2 + D_{\mathsf{KL}}\Big(P_h^M(\cdot \mid s, \pi(s,h)) \,\|\, P_h^{M'}(\cdot \mid s, \pi(s,h))\Big) \right]\end{aligned}$$

and, given our definition of $\widetilde{M}$,

$$\begin{aligned}&D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) \\ &= \sum_{s,a,h} w_h^{M,\pi}(s,a) \left[ \frac{1}{2}(r_h^{\overline{M}}(s,a) - r_h^{M'}(s,a))^2 + D_{\mathsf{KL}}\Big(P_h^M(\cdot \mid s, \pi(s,h)) \,\|\, P_h^{M'}(\cdot \mid s, \pi(s,h))\Big) \right].\end{aligned}$$

Thus,

$$\begin{aligned}&\left| D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) \right| \\ &\quad = \left| \frac{1}{2} \sum_{s,a,h} w_h^{M,\pi}(s,a) \Big[ (r_h^M(s,a) - r_h^{M'}(s,a))^2 - (r_h^{\overline{M}}(s,a) - r_h^{M'}(s,a))^2 \Big] \right| \\ &\quad \overset{(a)}{\le} \sum_{s,a,h} w_h^{M,\pi}(s,a) |r_h^M(s,a) - r_h^{\overline{M}}(s,a)|\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{s,a,h} w_h^{M,\pi}(s,a)\sqrt{2D_{\mathsf{KL}}\big(M_{sh}(a) \,\|\, \overline{M}_{sh}(a)\big)} \\
&\leq \sqrt{2H \cdot \sum_{s,a,h} w_h^{M,\pi}(s,a) D_{\mathsf{KL}}\big(M_{sh}(a) \,\|\, \overline{M}_{sh}(a)\big)} \\
&= \sqrt{2H \cdot D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)},
\end{aligned}
$$

where $(a)$ holds by the Mean Value Theorem and the assumption that reward means are in $[0,1]$, and the final equality holds by Lemma G.15.

We turn now to bounding the second term. Let $\mathcal{T} = \mathcal{S}^H$ denote the space of all possible state trajectories. Let $\mathbb{P}^{M,\pi}(\tau = \cdot)$ denote the probability of observing $\tau \in \mathcal{T}$ when playing policy $\pi$ on $M$. We then have

$$
\begin{aligned}
D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) &= \int \log \frac{\mathbb{P}^{\widetilde{M},\pi}(r,\tau)}{\mathbb{P}^{M',\pi}(r,\tau)} \mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,\tau) \\
&= \int\int \log \frac{\mathbb{P}^{\overline{M},\pi}(r \mid \tau)\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(r \mid \tau)\mathbb{P}^{M',\pi}(\tau)} \mathrm{d}\mathbb{P}^{\overline{M},\pi}(r \mid \tau)\mathrm{d}\mathbb{P}^{M,\pi}(\tau) \\
&= \int \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \mathrm{d}\mathbb{P}^{M,\pi}(\tau) + \int \left( \int \log \frac{\mathbb{P}^{\overline{M},\pi}(r \mid \tau)}{\mathbb{P}^{M',\pi}(r \mid \tau)} \mathrm{d}\mathbb{P}^{\overline{M},\pi}(r \mid \tau) \right) \mathrm{d}\mathbb{P}^{M,\pi}(\tau) \\
&= \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} + \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) D_{\mathsf{KL}}\Big( \mathbb{P}^{\overline{M},\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau) \Big).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\Big| D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big) \Big| \\
&\leq \left| \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} - \sum_{\tau \in \mathcal{T}} \mathbb{P}^{\overline{M},\pi}(\tau) \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \\
&\quad + \sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| D_{\mathsf{KL}}\Big( \mathbb{P}^{M,\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau) \Big).
\end{aligned}
$$

Note that, since rewards at each state are independent,

$$
D_{\mathsf{KL}}\Big( \mathbb{P}^{M,\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau) \Big) = \sum_{h=1}^{H} D_{\mathsf{KL}}\Big( \mathbb{P}^{M,\pi}(r_h \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r_h \mid \tau) \Big) \leq H,
$$

since rewards have means are in $[0,1]$ and are unit Gaussian. This implies

$$
\begin{aligned}
\sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| D_{\mathsf{KL}}\Big( \mathbb{P}^{M,\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau) \Big) &\leq H \sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| \\
&= H D_{\mathsf{TV}}\big(M(\pi), \overline{M}(\pi)\big) \\
&\leq H\sqrt{\frac{1}{2} D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)}.
\end{aligned}
$$

Note that $\frac{\mathrm{d}}{\mathrm{d}x} x \log \frac{x}{y} = 1 + \log \frac{x}{y}$, so by the Mean Value Theorem we have

$$\begin{aligned}
&\left| \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} - \mathbb{P}^{\overline{M},\pi}(\tau) \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \\
&\quad \le \left(1 + \max\left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \right\}\right) \cdot |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| \\
&\quad \le \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)|.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\left| \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} - \sum_{\tau \in \mathcal{T}} \mathbb{P}^{\overline{M},\pi}(\tau) \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \\
&\quad \le \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot \sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| \\
&\quad = \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot D_{\mathsf{TV}}\big(M(\pi), \overline{M}(\pi)\big) \\
&\quad \le \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot \sqrt{\frac{1}{2} D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)}.
\end{aligned}$$

This verifies Assumption D.1 with

$$L_{\mathsf{KL}} = 1 + \sqrt{2H} + H + \max_{M' \in \mathcal{M}, \overline{M}' \in \mathcal{M}} \max_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right|.$$

**Verifying Assumption D.2.** That $D^2_{\mathsf{H}}\big(\overline{M}(\pi), \overline{M}'(\pi)\big) \le D\big(\overline{M}(\pi) \,\|\, \overline{M}'(\pi)\big)$ is immediate, since KL always upper bounds Hellinger squared.

**Verifying Assumption A.2.** We have

$$\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} = \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} + \log \frac{\mathbb{P}^{\overline{M},\pi}(r \mid \tau)}{\mathbb{P}^{M,\pi}(r \mid \tau)} = \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} + \sum_{h=1}^{H} \log \frac{\mathbb{P}^{\overline{M},\pi}(r_h \mid \tau)}{\mathbb{P}^{M,\pi}(r_h \mid \tau)}.$$

Using the same calculation as in Lemma G.1, we have that $\log \frac{\mathbb{P}^{\overline{M},\pi}(r_h|\tau)}{\mathbb{P}^{M,\pi}(r_h|\tau)}$ is 8-subgaussian, since rewards are unit-variance Gaussian. As $\log \frac{\mathbb{P}^{\overline{M},\pi}(r_h|\tau)}{\mathbb{P}^{M,\pi}(r_h|\tau)}$ and $\log \frac{\mathbb{P}^{\overline{M},\pi}(r_{h'}|\tau)}{\mathbb{P}^{M,\pi}(r_{h'}|\tau)}$ are independent for $h \neq h'$, it follows that $\log \frac{\mathbb{P}^{\overline{M},\pi}(r|\tau)}{\mathbb{P}^{M,\pi}(r|\tau)}$ is $8H$-subgaussian.

Furthermore, bounding

$$\log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} \le \sup_{\overline{M}, M \in \mathcal{M}} \sup_{\pi \in \Pi} \sup_{\tau \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} \right| =: V_{\mathcal{T}},$$

we have that $\log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)}$ is $V_{\mathcal{T}}^2$-subgaussian. Since the sum of subgaussian random variables is subgaussian, it follows that $\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} = \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} + \log \frac{\mathbb{P}^{\overline{M},\pi}(r|\tau)}{\mathbb{P}^{M,\pi}(r|\tau)}$ is $(V_{\mathcal{T}}^2 + 8H)$-subgaussian, which verifies Assumption A.2.

□

**Lemma G.9.** *Let* $P_{\min} := \inf_{M\in\mathcal{M}} \inf_{h,s',s,a} P_h^M(s' \mid s,a)$ *and assume* $\mathcal{M}$ *is such that* $P_{\min} > 0$. *We can construct a* $(\rho,\mu)$*-cover of* $\mathcal{M}$ *with respect to* $\mathcal{E} := \{|r_h| \le 1 + \sqrt{2\log(2H/\mu)}, \forall h \in [H]\}$, *with*

$$\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho,\mu) \le \frac{1}{\min\{\frac{\rho P_{\min}}{4H}, 2P_{\min}\}^{S^2AH}} \cdot \frac{(2H(2+\sqrt{\log(H/\mu)}))^{SAH}}{\rho^{SAH}}.$$

**Proof of Lemma G.9.** Throughout this proof, we use $M = \{(P_h^M)_{h=1}^H, (r_h^M)_{h=1}^H\} \in \mathcal{M}$ to denote the MDP in $\mathcal{M}$ with $R_h^M(s,a) = \mathcal{N}(r_h^M(s,a),1)$; for brevity, we $\mathcal{S}$, $\mathcal{A}$, and $s_1$ to be fixed and the dependence on them.

Observe that for any models $M, M' \in \mathcal{M}$, we have

$$\begin{aligned}\left|\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)}\right| &\le \left|\log\frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)}\right| + \left|\log\frac{\mathbb{P}^{M,\pi}(r \mid \tau)}{\mathbb{P}^{M',\pi}(r \mid \tau)}\right| \\ &= \sum_{h=1}^H \left|\log\frac{P_h^M(\tau_{h+1} \mid \tau_h, \pi(\tau_h,h))}{P_h^{M'}(\tau_{h+1} \mid \tau_h, \pi(\tau_h,h))}\right| + \sum_{h=1}^H \left|\log\frac{\mathbb{P}^{M,\pi}(r_h \mid \tau)}{\mathbb{P}^{M',\pi}(r_h \mid \tau)}\right|.\end{aligned}$$

Let $\mathcal{I}_\varepsilon = \{\varepsilon, 2\varepsilon, \dots, \lfloor 1/\varepsilon\rfloor\varepsilon\}$, so that $|\mathcal{I}_\varepsilon| \le 1/\varepsilon$. Let $\mathcal{P}_\varepsilon$ denote an $\varepsilon$ cover of $\triangle_{\mathcal{S}}$ in the $\ell_\infty$-norm, so that for any $P \in \triangle_{\mathcal{S}}$, there exists some $P' \in \mathcal{P}_\varepsilon$ such that $\sup_{s\in\mathcal{S}} |P_s - P'_s| \le \varepsilon$. It suffices to choose $\mathcal{P}_\varepsilon = \mathcal{I}_\varepsilon^S \cap \triangle_{\mathcal{S}}$, so we can bound $|\mathcal{P}_\varepsilon| \le 1/\varepsilon^S$. Let

$$\mathcal{M}_{\mathrm{cov}} = \left\{M = \{(P_h^M)_{h=1}^H, (r_h^M)_{h=1}^H\} \;:\; P_h^M(\cdot \mid s,a) \in \mathcal{P}_{\varepsilon_1}, r_h^M(s,a) \in \mathcal{I}_{\varepsilon_2}, \quad \forall s,a,h\right\}$$

for parameters $\varepsilon_1, \varepsilon_2 > 0$ to be chosen. Then

$$\mathcal{M}_{\mathrm{cov}} = (|\mathcal{P}_{\varepsilon_1}||\mathcal{I}_{\varepsilon_2}|)^{SAH} \le \frac{1}{\varepsilon_1^{S^2AH}} \cdot \frac{1}{\varepsilon_2^{SAH}}.$$

We will show that $\mathcal{M}_{\mathrm{cov}}$ is a $(\rho,\mu)$-cover of $\mathcal{M}$ for appropriately chosen $\mathcal{E}$ and $\varepsilon_1, \varepsilon_2 > 0$.

Let $\mathcal{E} := \{|r_h| \le 1 + \sqrt{2\log(2H/\mu)}, \forall h \in [H]\}$. As we assume rewards are unit-variance Gaussian and have means in $[0,1]$, it is straightforward to see that $\mathbb{P}[\mathcal{E}^c] \le \mu$. Fix $M$ and let $M' \in \mathcal{M}_{\mathrm{cov}}$ denote the instance such that

$$|r_h^M(s,a) - r_h^{M'}(s,a)| \le \varepsilon_2 \quad \text{and} \quad \sup_{s'\in\mathcal{S}} |P_h^M(s' \mid s,a) - P_h^{M'}(s' \mid s,a)| \le \varepsilon_1, \quad \forall s,a,h.$$

Note that such an instance is guaranteed to exist by definition of $\mathcal{M}_{\mathrm{cov}}$.

By a similar argument as in Lemma G.1, we can bound, on $\mathcal{E}$,

$$\begin{aligned}\sum_{h=1}^H \left|\log\frac{\mathbb{P}^{M,\pi}(r_h \mid \tau)}{\mathbb{P}^{M',\pi}(r_h \mid \tau)}\right| &\le \sum_{h=1}^H (1+|r_h|) \cdot \sup_{s,a} |r_h^M(s,a) - r_h^{M'}(s,a)| \\ &\le H(2+\sqrt{2\log(2H/\mu)}) \cdot \sup_{s,a,h} |r_h^M(s,a) - r_h^{M'}(s,a)| \\ &\le H(2+\sqrt{2\log(2H/\mu)}) \cdot \varepsilon_2.\end{aligned}$$

We also have

$$\begin{aligned}\sum_{h=1}^{H}\left|\log\frac{P_h^M(\tau_{h+1}\mid\tau_h,\pi(\tau_h,h))}{P_h^{M'}(\tau_{h+1}\mid\tau_h,\pi(\tau_h,h))}\right| &\le H\cdot\sup_{h,s',s,a}\left|\log\frac{P_h^M(s'\mid s,a)}{P_h^{M'}(s'\mid s,a)}\right| \\ &\le H\cdot\sup_{|x|\le\varepsilon_1}\sup_{h,s',s,a}\left|\log\frac{P_h^M(s'\mid s,a)}{P_h^M(s'\mid s,a)-x}\right| \\ &\le H\cdot\sup_{h,s',s,a}\frac{\varepsilon_1}{P_h^M(s'\mid s,a)-\varepsilon_1}\end{aligned}$$

where the last inequality holds as long as $\inf_{h,s',s,a}P_h^M(s'\mid s,a)-\varepsilon_1>0$. Denoting $P_{\min}:=\inf_{M\in\mathcal{M}}\inf_{h,s',s,a}P_h^M(s'\mid s,a)$, for $\mathcal{M}_{\mathrm{cov}}$ to be a $(\rho,\mu)$-cover, it therefore suffices that

$$H(2+\sqrt{2\log(2H/\mu)})\cdot\varepsilon_2\le\rho/2,\quad \frac{2H\varepsilon_1}{P_{\min}}\le\rho/2,\quad\text{and}\quad P_{\min}\ge\varepsilon_1/2$$

so it suffices to take

$$\varepsilon_1=\min\{\frac{\rho P_{\min}}{4H},2P_{\min}\}\quad\text{and}\quad\varepsilon_2=\frac{\rho}{2H(2+\sqrt{\log(H/\mu)})}.$$

The result now follows from our bound on $|\mathcal{M}_{\mathrm{cov}}|$. □

### G.5.3. TABULAR MDPS HAVE BOUNDED UNIFORM EXPLORATION COEFFICIENT

**Lemma G.10.** *For the tabular MDP class $\mathcal{M}$ in (27), we can bound, for all $\varepsilon>0$,*

$$C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M},\varepsilon)\le\frac{320000SAH^2\cdot\log^2H}{\varepsilon^2}.$$

*for $D(\cdot\,\|\,\cdot)\leftarrow D_{\mathsf{H}}^2(\cdot,\cdot)$.*

**Proof of Lemma G.10.** Let $\xi\in\Delta(\mathcal{M})$ be given. Define

$$p_{\mathrm{exp}}=\arg\min_{p\in\triangle_\Pi}\max_{q\in\triangle_\Pi}\sum_{s,a,h}\mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right].$$

We first show that, for any $M\in\mathcal{M}$ and any $\pi$,

$$\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{H}}^2(\overline{M}(\pi),M(\pi))]\le\sqrt{SAH^2\cdot\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p_{\mathrm{exp}}}[D_{\mathsf{H}}^2(\overline{M}(\pi),M(\pi))]]}.$$

Consider any policy $\pi$. We can bound

$$\begin{aligned}&\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{H}}^2(\overline{M}(\pi),M(\pi))]\\ &\overset{(a)}{\le}100\log(H)\cdot\sum_{s,a,h}\mathbb{E}_{\overline{M}\sim\xi}\left[w_h^{\overline{M},\pi}(s,a)D_{\mathsf{H}}^2(\overline{M}_{sh}(a),M_{sh}(a))\right]\end{aligned}$$

$$\overset{(b)}{\leq} 100\log(H)\cdot\sqrt{\sum_{s,a,h}\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)^2}{\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}[w_h^{\overline{M},\pi'}(s,a)]}\right]}\cdot\sqrt{\sum_{s,a,h}\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[w_h^{\overline{M},\pi'}(s,a)D_{\mathsf{H}}^4\big(\overline{M}_{sh}(a),M_{sh}(a)\big)\right]\right]}$$

$$\overset{(c)}{\leq} 200\log(H)\cdot\sqrt{\sum_{s,a,h}\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)^2}{\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}[w_h^{\overline{M},\pi'}(s,a)]}\right]}\cdot\sqrt{\sum_{s,a,h}\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[w_h^{\overline{M},\pi'}(s,a)D_{\mathsf{H}}^2\big(\overline{M}_{sh}(a),M_{sh}(a)\big)\right]\right]}$$

where $(a)$ follows from Lemma A.13 of Foster et al. (2021), $(b)$ follows from Cauchy-Schwarz and Jensen's inequality, and $(c)$ follows because the Hellinger distance is always bounded by 2. Now note that, by definition of $p_{\mathrm{exp}}$, we have

$$\sum_{s,a,h}\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)^2}{\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}[w_h^{\overline{M},\pi'}(s,a)]}\right]\leq\min_{p\in\triangle_\Pi}\max_{q\in\triangle_\Pi}\sum_{s,a,h}\mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right]$$

and by the minimax theorem, we can bound

$$\begin{aligned}\min_{p\in\triangle_\Pi}\max_{q\in\triangle_\Pi}\sum_{s,a,h}\mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right]&=\max_{q\in\triangle_\Pi}\min_{p\in\triangle_\Pi}\sum_{s,a,h}\mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right]\\&\leq\max_{q\in\triangle_\Pi}\sum_{s,a,h}\mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim q}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right]\\&\leq\max_{q\in\triangle_\Pi}\sum_{s,a,h}\mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)}{\mathbb{E}_{\pi'\sim q}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right]\\&=SAH.\end{aligned}$$

By Lemma A.9 of Foster et al. (2021), since $\mu^{\overline{M}}(s,a,h):=\frac{1}{H}w_h^{\overline{M},\pi'}(s,a)$ forms a valid distribution on $\mathcal{S}\times\mathcal{A}\times[H]$, we can upper bound

$$\sum_{s,a,h}w_h^{\overline{M},\pi'}(s,a)D_{\mathsf{H}}^2\big(\overline{M}_{sh}(a),M_{sh}(a)\big)\leq HD_{\mathsf{H}}^2\big(\overline{M}(\pi'),M(\pi')\big).$$

Altogether then, we have shown that for all $\pi\in\Pi$,

$$\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{H}}^2\big(\overline{M}(\pi),M(\pi)\big)]\leq 200\log(H)\cdot\sqrt{SAH^2\cdot\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p_{\mathrm{exp}}}[D_{\mathsf{H}}^2\big(\overline{M}(\pi),M(\pi)\big)]]}$$

as desired. Since the Hellinger distance is a metric and satisfies the triangle inequality, this in particular implies that, for any $M,M'$,

$$D_{\mathsf{H}}^2\big(M(\pi),M'(\pi)\big)\leq 2\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{H}}^2\big(\overline{M}(\pi),M(\pi)\big)]+2\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{H}}^2\big(\overline{M}(\pi),M'(\pi)\big)]$$

$$\leq 400 \log H \cdot \sqrt{SAH^2 \cdot \mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)]]}$$
$$+ 400 \log H \sqrt{SAH^2 \cdot \mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M'(\pi)\big)]]}.$$

Thus, if

$$\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M''(\pi)\big)]] \leq \frac{\varepsilon^2}{320000SAH^2 \cdot \log^2 H}$$

for both $M'' \in \{M, M'\}$, then $D^2_{\mathsf{H}}(M(\pi), M'(\pi)) \leq \varepsilon$. It follows that a sufficient choice for $C^{\mathsf{D}}_{\mathrm{exp}}$ is $320000SAH^2 \log^2 H/\varepsilon^2$. □

### G.5.4. SUPPORTING LEMMAS

**Lemma G.11.** *If $M$ has a unique optimal policy $\pi_M$ and $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, then there exists some $(\widetilde{s}, \widetilde{a}, \widetilde{h})$ such that*

$$Q^{M',\pi_M}_{\widetilde{h}}(\widetilde{s}, \widetilde{a}) > V^{M',\pi_M}_{\widetilde{h}}(\widetilde{s}).$$

**Proof of Lemma G.11.** Assume that this is not the case, i.e. that for all $(s, a, h)$,

$$Q^{M',\pi_M}_h(s, a) \leq V^{M',\pi_M}_h(s) = Q^{M',\pi_M}_h(s, \pi_M(s, h)).$$

Our goal is to show that in this case $\pi_{M'} = \pi_M$, which contradicts the fact that $M' \in \mathcal{M}^{\mathrm{alt}}(M)$. We proceed by induction.

**Base Case.** Let $h = H$ and assume that for all $(s, a)$,

$$Q^{M',\pi_M}_H(s, a) \leq Q^{M',\pi_M}_H(s, \pi_M(s, h)).$$

This contradicts the assumption that $\pi_M$ is unique.

**Inductive Case.** Assume that $\pi_{M'}(s, h') = \pi_M(s, h')$ for all $s$ and $h' > h$. This then implies that $V^{M',\pi_M}_{h+1}(s) = V^{M',\pi_{M'}}_{h+1}(s)$ for all $s$. It follows that for all $a$

$$Q^{M',\pi_M}_h(s, a) = r^{M'}_h(s, a) + \mathbb{P}^{M'}_h[V^{M',\pi_M}_{h+1}](s, a) = r^{M'}_h(s, a) + \mathbb{P}^{M'}_h[V^{M',\pi_{M'}}_{h+1}](s, a) = Q^{M',\pi_{M'}}_h(s, a)$$

so in particular $Q^{M',\pi_M}_h(s, \pi_M(s, h)) = Q^{M',\pi_{M'}}_h(s, \pi_M(s, h))$. Since we have assumed that for all $(s, a)$

$$Q^{M',\pi_M}_h(s, a) \leq Q^{M',\pi_M}_h(s, \pi_M(s, h))$$

we have

$$Q^{M',\pi_{M'}}_h(s, \pi_{M'}(s, h)) \leq Q^{M',\pi_{M'}}_h(s, \pi_M(s, h)).$$

However, since each $M \in \mathcal{M}$ has a unique optimal action at each state, this is a contradiction unless $\pi_{M'}(s, h) = \pi_M(s, h)$, which proves the inductive hypothesis. The result follows. □

**Lemma G.12.** *For MDPs $M, M'$ with unit variance Gaussian rewards, we have*

$$V_h^{M',\pi}(s) - V_h^{M,\pi}(s) \le \sqrt{\frac{8H}{w_h^{M,\pi}(s)} \cdot D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.$$

**Proof of Lemma G.12.** In the Gaussian reward setting, we have

$$r_{h'}^{M'}(s',a') - r_{h'}^{M}(s',a') \le \sqrt{(r_{h'}^{M'}(s',a') - r_{h'}^{M}(s',a'))^2} \le \sqrt{2D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)}.$$

Furthermore, since $V_{h'+1}^{M',\pi}(s') \in [0,1]$, we have

$$\begin{aligned}
\mathbb{P}_{h'}^{M'}[V_{h'+1}^{M',\pi}](s',a') - \mathbb{P}_{h'}^{M}[V_{h'+1}^{M',\pi}](s',a') &\le \sum_{s''} |P_{h'}^{M'}(s'' \mid s',a') - P_{h'}^{M}(s'' \mid s',a')| \\
&= 2D_{\mathsf{TV}}\Big(P_{h'}^{M'}(\cdot \mid s',a'), P_{h'}^{M}(\cdot \mid s',a')\Big) \\
&\le \sqrt{2D_{\mathsf{KL}}\big(P_{h'}^{M'}(\cdot \mid s',a') \,\|\, P_{h'}^{M}(\cdot \mid s',a')\big)} \\
&\le \sqrt{2D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)}.
\end{aligned}$$

By Lemma G.14, Jensen's inequality, and Lemma G.15, it follows that

$$\begin{aligned}
V_h^{M',\pi}(s) - V_h^{M,\pi}(s) &\le \sum_{h'=h}^{H} \sum_{s',a'} w_{h'}^{M,\pi}(s',a' \mid s_h = s) \cdot 2\sqrt{2D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)} \\
&\le 2\sqrt{2H \sum_{h'=h}^{H} \sum_{s',a'} w_{h'}^{M,\pi}(s',a' \mid s_h = s) D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)} \\
&\le 2\sqrt{\frac{2H}{w_h^{M,\pi}(s)} \cdot \sum_{h'=h}^{H} \sum_{s',a'} w_{h'}^{M,\pi}(s',a') D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)} \\
&\le 2\sqrt{\frac{2H}{w_h^{M,\pi}(s)} D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}
\end{aligned}$$

where we have used that, for $h < h'$,

$$w_{h'}^{M,\pi}(s',a') = \sum_{s''} w_{h'}^{M,\pi}(s',a' \mid s_h = s'') w_{\pi}^{M}(s'',h) \ge w_{h'}^{M,\pi}(s',a' \mid s_h = s) w_h^{M,\pi}(s).$$

□

**Lemma G.13.** *For any $M \in \mathcal{M}$ for which $\pi_M$ is unique, we have*

$$\Delta_{\min}^{M} \le \min_{s,h} w_h^{M,\pi_M}(s).$$

**Proof of Lemma G.13.** Let $\widetilde{\pi}$ denote the policy that differs from policy $\pi_M$ only at the state $\widetilde{s}$ and layer $\widetilde{h}$ given by $(\widetilde{s}, \widetilde{h}) = \arg\min_{s,h} w_h^{M,\pi_M}(s)$. Note that this implies, in particular, that $w_{\widetilde{h}}^{M,\pi_M}(\widetilde{s}) = w_{\widetilde{h}}^{M,\widetilde{\pi}}(\widetilde{s})$ since $\widetilde{\pi}$ and $\pi_M$ take identical actions up to step $\widetilde{h}$. By the Performance-Difference Lemma (Kakade, 2003), we have

$$\begin{aligned}
V_1^{M,\pi_M} - V_1^{M,\widetilde{\pi}} &= \sum_{h=1}^{H} \sum_{s,a} w_h^{M,\widetilde{\pi}}(s,a)(V_h^{M,\pi_M}(s) - Q_h^{M,\pi_M}(s,a)) \\
&\overset{(a)}{=} w_{\widetilde{h}}^{M,\widetilde{\pi}}(\widetilde{s}, \widetilde{\pi}(\widetilde{s}, \widetilde{h}))(V_{\widetilde{h}}^{M,\pi_M}(\widetilde{s}) - Q_{\widetilde{h}}^{M,\pi_M}(\widetilde{s}, \widetilde{\pi}(\widetilde{s}, \widetilde{h}))) \\
&= w_{\widetilde{h}}^{M,\pi_M}(\widetilde{s})(V_{\widetilde{h}}^{M,\pi_M}(\widetilde{s}) - Q_{\widetilde{h}}^{M,\pi_M}(\widetilde{s}, \widetilde{\pi}(\widetilde{s}, \widetilde{h}))) \\
&\leq w_{\widetilde{h}}^{M,\pi_M}(\widetilde{s})
\end{aligned}$$

where $(a)$ holds since $V_h^{M,\pi_M}(s) = Q_h^{M,\pi_M}(s,a)$ for all $(s,a,h)$ with $w^{M,\widetilde{\pi}}(s,a,h) > 0$ other than at $(\widetilde{s}, \widetilde{h})$. By assumption, the optimal policy is unique, so $V_1^{M,\pi_M} - V_1^{M,\widetilde{\pi}} > 0$, and thus

$$\Delta_{\min}^M = \min_{\pi \in \Pi \,:\, V_1^{M,\pi_M} - V_1^{M,\pi} > 0} V_1^{M,\pi_M} - V_1^{M,\pi} \leq V_1^{M,\pi_M} - V_1^{M,\widetilde{\pi}} \leq w_{\widetilde{h}}^{M,\pi_M}(\widetilde{s}) = \min_{s,h} w_h^{M,\pi_M}(s).$$

□

**Lemma G.14** (Lemma E.15 of Dann et al. (2017))**.** *For MDPs $M, M'$ and policy $\pi$, we have*

$$V_h^{M',\pi}(s) - V_h^{M,\pi}(s) = \sum_{h'=h}^{H} \sum_{s',a'} w_{h'}^{M,\pi}(s', a' \mid s_h = s) \cdot \Big[ r_{h'}^{M'}(s',a') - r_{h'}^{M}(s',a') + \mathbb{P}_{h'}^{M'}[V_{h'+1}^{M',\pi}](s',a') - \mathbb{P}_{h'}^{M}[V_{h'+1}^{M',\pi}](s',a') \Big].$$

**Lemma G.15.** *For MDPs $M, M'$ and policy $\pi$, we have*

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, M'(\pi)\big) = \sum_{h=1}^{H} \sum_{s,a} w_h^{M,\pi}(s,a) D_{\mathsf{KL}}\big(M_{hs}(a) \,\|\, M'_{hs}(a)\big).$$

**Proof of Lemma G.15.** This is a standard calculation; see e.g. (Simchowitz and Jamieson, 2019; Tirinzoni et al., 2021). □

## Appendix H. Proofs and Additional Results from Appendix B

### H.1. Technical Lemmas

Throughout this section, when the class $\mathcal{M}$ is clear from context, we define

$$\Lambda(M; \varepsilon, \bar{\mathsf{n}}) := \{\lambda \in \triangle_\Pi \;:\; \exists \mathsf{n} \leq \bar{\mathsf{n}} \text{ s.t. } \Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n}, I^M(\lambda; \mathcal{M}) \geq (1-\varepsilon)/\mathsf{n}\}. \quad (77)$$

**Lemma H.1** (Derandomization). *Let $\bar{\mathsf{n}} > 0$ be given. For any $p \in \triangle(\triangle(\Pi))$, defining $\bar{\lambda}_p = \mathbb{E}_{\lambda\sim p}[\lambda] \in \Delta(\Pi)$, we have that for all $M \in \mathcal{M}$,*

$$\mathbb{I}\{\bar{\lambda}_p \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}})\} \le \delta^{-1} \cdot \mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})],$$

*where $\delta := \frac{\varepsilon}{2} \cdot \min\left\{1, \frac{\mathsf{g}^M}{\bar{\mathsf{n}}}\right\}$.*

**Proof of Lemma H.1.** Let $M \in \mathcal{M}$ be fixed and abbreviate $I^M(\lambda) = I^M(\lambda;\mathcal{M})$. Fix $p \in \triangle(\triangle(\Pi))$. For any $\lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})$, let $\mathsf{n}_\lambda > 0$ denote the least $\mathsf{n} > 0$ such that

$$\Delta^M(\lambda) \le (1+\varepsilon/2)\mathsf{g}^M/\mathsf{n}, \quad \text{and} \quad I^M(\lambda;\mathcal{M}) \ge (1-\varepsilon/2)/\mathsf{n}.$$

Define

$$\mathsf{n} = \left(\mathbb{E}_{\lambda\sim p}\left[\frac{1}{\mathsf{n}_\lambda} \mid \lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})\right]\right)^{-1},$$

and note that by Jensen's inequality,

$$\mathsf{n} \le \mathbb{E}_{\lambda\sim p}[\mathsf{n}_\lambda \mid \lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \le \bar{\mathsf{n}}.$$

We first observe that since $\Delta^M \in [0,1]$,

$$\begin{aligned}
\Delta^M(\bar{\lambda}_p) &\le \mathbb{E}_{\lambda\sim p}[\Delta^M(\lambda) \mid \lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] + \mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \\
&\le (1+\varepsilon/2)\mathsf{g}^M \cdot \mathbb{E}_{\lambda\sim p}\left[\frac{1}{\mathsf{n}_\lambda} \mid \lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})\right] + \mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \\
&= (1+\varepsilon/2)\frac{\mathsf{g}^M}{\mathsf{n}} + \mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})].
\end{aligned}$$

Next, note that the map $\lambda \mapsto I^M(\lambda)$ is concave and non-negative (it is an infimum over non-negative linear functions), so we have

$$\begin{aligned}
I^M(\bar{\lambda}_p) &\ge \mathbb{E}_{\lambda\sim p}[I^M(\lambda)] \\
&\ge \mathbb{E}_{\lambda\sim p}[I^M(\lambda) \mid \lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \cdot \mathbb{P}_{\lambda\sim p}[\lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \\
&\ge (1-\varepsilon/2)\,\mathbb{E}_{\lambda\sim p}\left[\frac{1}{\mathsf{n}_\lambda} \mid \lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})\right] \cdot \mathbb{P}_{\lambda\sim p}[\lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \\
&= (1-\varepsilon/2)\frac{1}{\mathsf{n}} \cdot \mathbb{P}_{\lambda\sim p}[\lambda \in \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \\
&= (1-\varepsilon/2)\frac{1}{\mathsf{n}} \cdot (1 - \mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})]).
\end{aligned}$$

It follows that as long as

$$\mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] \le \delta := \frac{\varepsilon}{2} \cdot \min\left\{1, \frac{\mathsf{g}^M}{\bar{\mathsf{n}}}\right\} \le \frac{\varepsilon}{2} \cdot \min\left\{1, \frac{\mathsf{g}^M}{\mathsf{n}}\right\},$$

we have

$$\Delta^M(\bar{\lambda}_p) \le (1+\varepsilon)\mathsf{g}^M/\mathsf{n}, \quad \text{and} \quad I^M(\bar{\lambda}_p;\mathcal{M}) \ge (1-\varepsilon)/\mathsf{n},$$

so that $\bar{\lambda}_p \in \Lambda(M;\varepsilon,\mathsf{n}) \subseteq \Lambda(M;\varepsilon,\bar{\mathsf{n}})$. We conclude that

$$\mathbb{I}\{\bar{\lambda}_p \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}})\} \le \mathbb{I}\{\mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})] > \delta\} \le \delta^{-1} \cdot \mathbb{P}_{\lambda\sim p}[\lambda \notin \Lambda(M;\varepsilon/2,\bar{\mathsf{n}})].$$

□

**Lemma H.2.** *Let $M \in \mathcal{M}$ be given, and suppose Assumption A.4 holds. Fix $T \in \mathbb{N}$ and consider an algorithm $\mathbb{A}$ such that for all $M' \in \mathcal{M}^{\text{alt}}(M) \cup \{M\}$,*

$$\mathbb{E}^{M', \mathbb{A}}[\mathbf{Reg}(T)] \leq R^{M'} \cdot \log(T).$$

*for some bound $R^M \geq 2$. Define $\eta^M \in \mathbb{R}_+^{\Pi}$ via*

$$\eta^M(\pi) = \mathbb{E}^{M, \mathbb{A}}\left[\frac{T(\pi)}{\log(T)}\right].$$

*Then if*

$$\log(T) \geq \frac{6}{\varepsilon} \log\left(\sup_{M' \in \mathcal{M}^{\text{alt}}(M) \cup \{M\}} \frac{R^{M'}}{\Delta_{\min}^{M'}} \cdot \log(T)\right),$$

*we must have*

$$I^M(\eta^M; \mathcal{M}) \geq (1 - \varepsilon).$$

**Proof of Lemma H.2.** Throughout this proof we will use that $\pi_M$ is uniquely defined for all $M \in \mathcal{M}$ by Assumption A.4. Note that

$$\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq R^M \log T \implies \sum_{\pi \neq \pi_M} \mathbb{E}^{M, \mathbb{A}}[T(\pi)] \leq \frac{R^M \log T}{\Delta_{\min}^M}.$$

Fix some $M' \in \mathcal{M}^{\text{alt}}(M)$. Then $\pi_M \neq \pi_{M'}$ (recall that under Assumption A.4, each $M \in \mathcal{M}$ has a unique optimal), so

$$\mathbb{E}^{M, \mathbb{A}}[T(\pi_{M'})] \leq \frac{R^M \log T}{\Delta_{\min}^M}, \quad \mathbb{E}^{M', \mathbb{A}}[T(\pi_{M'})] \geq T - \frac{R^{M'} \log T}{\Delta_{\min}^{M'}}.$$

By Lemma H.1 of Simchowitz and Jamieson (2019), we have that for any $\mathcal{H}^T$-measurable variable $Z \in [0, 1]$, that

$$\sum_{\pi} \mathbb{E}^{M, \mathbb{A}}[T(\pi)] D_{\mathsf{KL}}(M(\pi), M'(\pi)) \geq d(\mathbb{E}^{M, \mathbb{A}}[Z], \mathbb{E}^{M', \mathbb{A}}[Z])$$

for $d(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y}$. Choosing $Z = T(\pi_{M'})/T$, and using that

$$d(x, y) \geq (1 - x) \log \frac{1}{1 - y} - \log 2,$$

we have

$$\begin{aligned}
\sum_{\pi} \mathbb{E}^{M, \mathbb{A}}[T(\pi)] D_{\mathsf{KL}}(M(\pi), M'(\pi)) &\geq \left(1 - \frac{R^M \log T}{\Delta_{\min}^M T}\right) \log \frac{T}{T - (T - \frac{R^{M'} \log T}{\Delta_{\min}^{M'}})} - \log 2 \\
&= \left(1 - \frac{R^M \log T}{\Delta_{\min}^M T}\right) \left(\log T - \log \frac{R^{M'} \log T}{\Delta_{\min}^{M'}}\right) - \log 2.
\end{aligned}$$

Now, if

$$\log(T) \geq \frac{2}{\varepsilon} \log\left( \frac{\sup_{M' \in \mathcal{M}^{\mathrm{alt}}(M) \cup \{M\}} R^{M'} \log T}{\Delta_{\min}^{M'}} \vee 2 \right),$$

we have $\log(2) \leq \varepsilon \log(T)$,

$$\log T - \log\left( \frac{R^{M'} \log T}{\Delta_{\min}^{M'}} \right) \geq (1-\varepsilon) \log(T),$$

and $1 - \frac{R^M \log T}{\Delta_{\min}^M T} \geq (1-\varepsilon)$, so we can lower bound

$$\sum_{\pi} \mathbb{E}^{M, \mathbb{A}}[T(\pi)] D_{\mathsf{KL}}(M(\pi), M'(\pi)) \geq \left((1-\varepsilon)^2 - \varepsilon\right) \log(T) \geq (1-3\varepsilon) \log(T).$$

As this is true for every $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, the result follows. □

**Lemma B.1.** *Let $\varepsilon \in (0,2)$, and suppose that Assumption A.4 holds. Fix $T \in \mathbb{N}$ and consider an algorithm $\mathbb{A}$ such that for all $M \in \mathcal{M}$,*

$$\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq (1+\varepsilon) \mathsf{g}^M \cdot \log(T).$$

*For each $M \in \mathcal{M}$, define $\eta^M \in \mathbb{R}_+^{\Pi}$ via $\eta^M(\pi) = \mathbb{E}^{M, \mathbb{A}}\left[ \frac{T(\pi)}{\log(T)} \right]$, where $T(\pi)$ denotes the number of pulls of decision $\pi$, and define $\lambda^M = \eta^M / \|\eta^M\|_1$. Then if*

$$\log(T) \geq \frac{6}{\varepsilon} \log\left( \sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \right),$$

*we have that for all $M \in \mathcal{M}$,*

$$\lambda^M \in \Lambda(M; \varepsilon). \tag{35}$$

**Proof of Lemma B.1.** Immediate consequence of Lemma H.2. □

## H.2. Proof of Theorem B.1

**Theorem H.1** (Full Statement of Theorem B.1)**.** *Let $\varepsilon > 0$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Let $\{\overline{\mathsf{n}}^M\}_{M \in \mathcal{M}_0}$ be a collection of non-negative scalars indexed by $\mathcal{M}_0$, and set $\delta := \frac{\varepsilon}{2} \cdot \min\left\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\overline{\mathsf{n}}^M}\right\}$. Unless*

$$T > \frac{\delta}{8} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0, \overline{M}),$$

*any algorithm must have, for some $M \in \mathcal{M}_0$:*

$$\mathbb{P}^{M, \mathbb{A}}\left[ \widehat{\lambda} \notin \Lambda(M; \varepsilon, \overline{\mathsf{n}}^M) \right] \geq \frac{\delta}{6}.$$

**Proof of Theorem H.1.** Fix $\varepsilon > 0$, and let an algorithm $\mathbb{A}$ be given. For any $\overline{M} \in \mathcal{M}^+$, define $q^{\overline{M}} = \mathbb{P}^{\overline{M}, \mathbb{A}}(\widehat{\lambda} = \cdot) \in \triangle(\triangle(\Pi))$, and let $\omega^{\overline{M}} := \mathbb{E}^{\overline{M}, \mathbb{A}}\left[\frac{1}{T}\sum_{t=1}^T p^t\right] \in \triangle(\Pi)$.

Fix $\alpha > 0$ and $\overline{M} \in \mathcal{M}^+$ be fixed. Define

$$M = \arg\max_{M \in \mathcal{M}_0}\left\{\mathbb{P}_{\lambda \sim q^{\overline{M}}}[\lambda \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi \sim \omega^{\overline{M}}}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\};$$

we assume that such an $M \in \mathcal{M}_0$ does exist, as otherwise the claim we will prove is trivial. It is immediate from this definition that we have

$$\begin{aligned}
&\mathbb{P}^{\overline{M}, \mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\right] \\
&= \mathbb{P}_{\lambda \sim q^{\overline{M}}}[\lambda \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)] \\
&= \sup_{M \in \mathcal{M}_0}\left\{\mathbb{P}_{\lambda \sim q^{\overline{M}}}[\lambda \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi \sim \omega^{\overline{M}}}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\} \\
&\ge \inf_{q \in \triangle(\triangle(\Pi)), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0}\left\{\mathbb{P}_{\lambda \sim q}[\lambda \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\} \\
&=: \mathsf{opt},
\end{aligned}$$

with the convention that this value is zero if the set $\left\{M \in \mathcal{M}_0 \mid \mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\}$ is empty. In addition, we have

$$\mathbb{E}_{\pi \sim \omega^{\overline{M}}}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2. \tag{78}$$

Now, define $\delta := \frac{\varepsilon}{2} \cdot \min\left\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\right\}$, and let $\overline{\lambda}_q := \mathbb{E}_{\lambda \sim q}[\lambda]$. By Lemma H.1, we have

$$\begin{aligned}
\mathsf{opt} &= \inf_{q \in \triangle(\triangle(\Pi)), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0}\left\{\mathbb{P}_{\lambda \sim q}[\lambda \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\} \\
&\ge \delta \cdot \inf_{q \in \triangle(\triangle(\Pi)), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0}\left\{\mathbb{I}\left\{\overline{\lambda}_q \notin \Lambda(M; 2\varepsilon, \bar{\mathsf{n}}^M)\right\} \mid \mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\} \\
&\ge \delta \cdot \inf_{\lambda \in \triangle(\Pi), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0}\left\{\mathbb{I}\{\lambda \notin \Lambda(M; 2\varepsilon, \bar{\mathsf{n}}^M)\} \mid \mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\} \\
&\ge \delta \cdot \inf_{\lambda \in \triangle(\Pi), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0}\left\{\mathbb{I}\{\lambda \notin \Lambda(M; 2\varepsilon)\} \mid \mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right] \le \alpha^2\right\} \\
&= \delta \cdot \mathbb{I}\left\{\alpha^2 \ge \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M})\right)^{-1}\right\}.
\end{aligned} \tag{79}$$

We conclude that

$$\mathbb{P}^{\overline{M}, \mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\right] \ge \delta \cdot \mathbb{I}\left\{\alpha^2 \ge \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M})\right)^{-1}\right\}. \tag{80}$$

To proceed, using Lemma A.11 of Foster et al. (2021), we have

$$\begin{aligned}
\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\right] &\ge \frac{1}{3}\mathbb{P}^{\overline{M}, \mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\right] - \frac{4}{3}D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}) \\
&\ge \frac{\delta}{3}\mathbb{I}\left\{\alpha^2 \ge \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M})\right)^{-1}\right\} - \frac{4}{3}D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}).
\end{aligned}$$

Using (78) gives

$$D_{\mathsf{KL}}(\mathbb{P}^{\overline{M},\mathbb{A}} \,\|\, \mathbb{P}^{M,\mathbb{A}}) = \mathbb{E}^{M,\mathbb{A}}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t} D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\right] = T\cdot\mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2 T,$$

so we have

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \ge \frac{\delta}{3}\mathbb{I}\Big\{\alpha^2 \ge \big(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0,\overline{M})\big)^{-1}\Big\} - \frac{4}{3}\alpha^2 T.$$

We set $\alpha^2 = \frac{\delta}{8T}$, so that

$$\begin{aligned}\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] &\ge \frac{\delta}{6}\cdot\mathbb{I}\Big\{\alpha^2 \ge \big(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0,\overline{M})\big)^{-1}\Big\}\\ &= \frac{\delta}{6}\cdot\mathbb{I}\Big\{T \le \frac{\delta}{8}\cdot\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0,\overline{M})\Big\}.\end{aligned}$$

By taking the supremum over all possible choices for $\overline{M} \in \mathcal{M}^+$, we conclude that unless

$$T > \frac{\delta}{8}\cdot\sup_{\overline{M}\in\mathcal{M}^+}\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0,\overline{M}),$$

the algorithm must have $\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \ge \frac{\delta}{6}$. □

### H.3. Proof of Theorem B.2

Recall that for $M \in \mathcal{M}^+$ we define

$$\boldsymbol{\pi}_M = \arg\max_{\pi\in\Pi} f^M(\pi)$$

as the set $\boldsymbol{\pi}_M \subseteq \Pi$ of all optimal decisions $\pi$ with $f^M(\pi) = \max_{\pi'\in\Pi} f^M(\pi')$. Unless otherwise stated, the results in this subsection do not make use of Assumption A.4. For $\overline{M} \in \mathcal{M}^+$ and $\mathcal{M}_0 \subseteq \mathcal{M}$, we define

$$\mathcal{M}_0^{\mathrm{opt}}(\overline{M}) = \big\{M \in \mathcal{M}_0 \mid \boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}},\ \ D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) = 0\ \forall \pi \in \boldsymbol{\pi}_{\overline{M}}\big\}.$$

For a subset $\Pi' \subseteq \Pi$, let

$$N_{\neg\Pi'} = \big|\big\{t \in [T] \mid \pi^t \notin \Pi'\big\}\big|.$$

Note that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, since $\boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}$, we have

$$N_{\neg\boldsymbol{\pi}_{\overline{M}}} \le N_{\neg\boldsymbol{\pi}_M}.$$

**Theorem B.2** (Main lower bound—strong variant)**.** *Let $\varepsilon > 0$, $\mathsf{n}_{\max} > 0$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given, and define $\delta = \frac{\varepsilon}{2}\cdot\min\{1, \inf_{M\in\mathcal{M}_0}\mathsf{g}^M/\mathsf{n}_{\max}\}$. Unless*

$$\sup_{M\in\mathcal{M}_0}\frac{\mathsf{g}^M}{\Delta^M_{\min}}\cdot\log(T) \ge \Omega(\delta^2)\cdot\sup_{\overline{M}\in\mathcal{M}^+}\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M}),$$

*there is no algorithm that simultaneously ensures that*

1. $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq 2 \cdot \mathsf{g}^M \log(T), \quad \forall M \in \mathcal{M}_0.$
2. $\mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\mathsf{n}_{\max})\right] \leq \frac{\delta}{12}, \quad \forall M \in \mathcal{M}_0.$

**Proof of Theorem B.2.** For each $M \in \mathcal{M}$, if $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$, then $\mathbb{E}^{M,\mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}] \leq 2\frac{\mathsf{g}^M}{\Delta^M_{\min}} \log(T)$. The result now follows by appealing to Theorem H.2 with $\bar{\mathsf{n}}^M = \mathsf{n}_{\max}$ and $R = 2\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta^M_{\min}} \log(T)$. □

**Theorem H.2.** *Let $T \in \mathbb{N}$, $\varepsilon > 0$, $R \geq 1$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Let $\{\bar{\mathsf{n}}^M\}_{M \in \mathcal{M}_0}$ be a collection of non-negative scalars indexed by $\mathcal{M}_0$. Define $\delta = \frac{\varepsilon}{2} \cdot \min\left\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\right\}$. Unless*

$$R \geq \frac{\delta^2}{192} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

*there is no algorithm that simultaneously ensures that*

1. $\mathbb{E}^{M,\mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}] \leq R, \quad \forall M \in \mathcal{M}_0.$
2. $\mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] \leq \frac{\delta}{12}, \quad \forall M \in \mathcal{M}_0.$

**Proof of Theorem H.2.** Let $\varepsilon > 0$ be fixed. To prove the result, it suffices to lower bound the constrained minimax value

$$\mathfrak{M} := \sup_{\overline{M} \in \mathcal{M}^+} \inf_{\mathbb{A}} \left\{ \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] \mid \mathbb{E}^{M',\mathbb{A}}[N_{\neg \boldsymbol{\pi}_{M'}}] \leq R \ \forall M' \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M}) \right\}. \tag{81}$$

We begin by appealing to the following technical lemma.

**Lemma H.3.** *Let $\overline{M} \in \mathcal{M}^+$ and $T \in \mathbb{N}$ be given. Consider any algorithm $\mathbb{A}$ with the property that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$,*

$$\mathbb{E}^{M,\mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}] \leq R$$

*for some $R \geq 1$. For any $\beta \in (0,1)$, there exists a modified algorithm $\mathbb{A}'$ with the following properties:*

- $\mathbb{P}^{M,\mathbb{A}'}\left[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > \lceil \frac{R}{\beta} \rceil\right] = 0$ *for all models $M \in \mathcal{M}^+$.*
- *For all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$,*

$$\mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] \geq \mathbb{P}^{M,\mathbb{A}'}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] - \beta.$$

By Lemma H.3, for any $\beta \in (0,1)$, we have

$$\mathfrak{M} \geq \sup_{\overline{M} \in \mathcal{M}^+} \inf_{\mathbb{A}} \left\{ \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] \mid \mathbb{P}^{M',\mathbb{A}}\left[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > \left\lceil \frac{R}{\beta} \right\rceil\right] = 0 \ \forall M' \in \mathcal{M}^+ \right\} - \beta.$$

Now, consider an arbitrary choice for $\overline{M}$ above. We lower bound the minimax value using another technical lemma.

**Lemma H.4.** *Let $T \in \mathbb{N}$ and $\varepsilon > 0$ be given. Let $\{\bar{\mathsf{n}}^M\}_{M \in \mathcal{M}}$ be a collection of non-negative scalars indexed by $\mathcal{M}$. Consider any algorithm $\mathbb{A}$ with the property that*

$$\mathbb{P}^{\overline{M}, \mathbb{A}}\big[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > R\big] = 0$$

*for some $R \geq 1$. For any $\overline{M} \in \mathcal{M}^+$, if we set $\delta := \frac{\varepsilon}{2} \cdot \min\Big\{1, \inf_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\Big\}$, then unless*

$$R > \frac{\delta}{8} \cdot \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

*the algorithm must have*

$$\sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M, \mathbb{A}}\Big[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\Big] \geq \frac{\delta}{6}.$$

Bounding $\left\lceil \frac{R}{\beta} \right\rceil \leq \frac{2R}{\beta}$, it follows from Lemma H.4 that unless

$$\frac{2R}{\beta} > \frac{\delta}{8} \cdot \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

where $\delta := \frac{\varepsilon}{2} \cdot \min\Big\{1, \inf_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\Big\}$, we have

$$\mathfrak{M} \geq \inf_{\mathbb{A}} \left\{ \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M, \mathbb{A}}\Big[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\Big] \mid \mathbb{P}^{M', \mathbb{A}}\left[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > \left\lceil \frac{R}{\beta} \right\rceil\right] = 0 \ \forall M' \in \mathcal{M}^+ \right\} - \beta \geq \frac{\delta}{6} - \beta.$$

To conclude, we set $\beta = \frac{\delta}{12}$ and maximize over $\overline{M} \in \mathcal{M}^+$. □

**Proof of Lemma H.3.** Fix $\beta \in (0, 1)$ and let $C := \lceil \frac{R}{\beta} \rceil$. Fix $\mathbb{A} = (p, q)$ and consider the algorithm $\mathbb{A}' = (p', q')$ defined implicitly as follows. For $t = 1, \dots, T$:

- Sample $\pi^t \sim p^t(\cdot \mid \mathcal{H}^{t-1})$.
- If $|\{i \leq t \mid \pi^i \notin \boldsymbol{\pi}_{\overline{M}}\}| = C$, break and play an arbitrary decision $\pi \in \boldsymbol{\pi}_{\overline{M}}$ until round $T$.

Return $\widehat{\lambda} \sim q(\cdot \mid \mathcal{H}^T)$.

It is immediate from this construction that $\mathbb{A}' = (p', q')$ has $N_{\neg \boldsymbol{\pi}_{\overline{M}}} \leq C$ almost surely under all possible models $M \in \mathcal{M}^+$. We now focus on bounding the performance. Let $T_0$ be the greatest value of $t$ for which $|\{i \leq t \mid \pi^i \notin \boldsymbol{\pi}_{\overline{M}}\}| \leq C$. First, observe that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, since the algorithms behave identically in law whenever $T_0 = T$,

$$\begin{aligned}
\mathbb{P}^{M, \mathbb{A}'}\Big[\widehat{\lambda} \in \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\Big] &\geq \mathbb{P}^{M, \mathbb{A}'}\Big[\widehat{\lambda} \in \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M) \wedge T_0 = T\Big] \\
&= \mathbb{P}^{M, \mathbb{A}}\Big[\widehat{\lambda} \in \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M) \wedge T_0 = T\Big] \\
&= \mathbb{P}^{M, \mathbb{A}}\Big[\widehat{\lambda} \in \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M) \wedge N_{\neg \boldsymbol{\pi}_{\overline{M}}} \leq C\Big].
\end{aligned}$$

By the union bound, we have

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \in \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M) \wedge N_{\neg\boldsymbol{\pi}_{\overline{M}}} \leq C\Big] \geq \mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \in \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] - \mathbb{P}^{M,\mathbb{A}}\big[N_{\neg\boldsymbol{\pi}_{\overline{M}}} > C\big].$$

Finally, we observe that by Markov's inequality we have

$$\mathbb{P}^{M,\mathbb{A}}\big[N_{\neg\boldsymbol{\pi}_{\overline{M}}} > C\big] \leq \frac{\mathbb{E}^{M,\mathbb{A}}\big[N_{\neg\boldsymbol{\pi}_{\overline{M}}}\big]}{C} \leq \frac{\mathbb{E}^{M,\mathbb{A}}[N_{\neg\boldsymbol{\pi}_M}]}{C} \leq \beta,$$

where we have used that $N_{\neg\boldsymbol{\pi}_{\overline{M}}} \leq N_{\neg\boldsymbol{\pi}_M}$, since $\boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}$. Rearranging, we obtain

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \geq \mathbb{P}^{M,\mathbb{A}'}\Big[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] - \beta.$$

□

**Proof of Lemma H.4.** Fix $\varepsilon > 0$, and let an algorithm $\mathbb{A}$ be given. For any $\overline{M} \in \mathcal{M}^+$, define $q^{\overline{M}} = \mathbb{P}^{\overline{M},\mathbb{A}}(\widehat{\lambda} = \cdot) \in \triangle(\triangle(\Pi))$, and let $\omega^{\overline{M}} := \mathbb{E}^{\overline{M},\mathbb{A}}\Big[\frac{1}{N_{\neg\boldsymbol{\pi}_{\overline{M}}}} \sum_{t:\pi^t \notin \boldsymbol{\pi}_{\overline{M}}} p^t\Big] \in \triangle(\Pi)$, with the convention that the value inside the expectation is zero whenever $N_{\neg\boldsymbol{\pi}_{\overline{M}}} = 0$.[9]

Fix $\alpha > 0$ and $\overline{M} \in \mathcal{M}^+$ be fixed. Define

$$M = \operatorname*{arg\,max}_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{\mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \leq \alpha^2\Big\};$$

we assume that such an $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$ does exist, as otherwise the claim we will prove is trivial. It is immediate from this definition that we have

$$\begin{aligned}
&\mathbb{P}^{\overline{M},\mathbb{A}}\Big[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \\
&= \mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)] \\
&= \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{\mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \leq \alpha^2\Big\} \\
&\geq \inf_{q\in\triangle(\triangle(\Pi)),\omega\in\triangle(\Pi)} \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \big\{\mathbb{P}_{\lambda\sim q}[\lambda \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \leq \alpha^2\big\} =: \mathsf{opt},
\end{aligned}$$

with the convention that this value is zero if the set $\Big\{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M}) \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \leq \alpha^2\Big\}$ is empty. In addition, we have

$$\mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \leq \alpha^2. \tag{82}$$

Now, define $\delta := \frac{\varepsilon}{2} \cdot \min\Big\{1, \inf_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\Big\}$, and let $\bar{\lambda}_q := \mathbb{E}_{\lambda\sim q}[\lambda]$. By Lemma H.1, we have

$$\mathsf{opt} = \inf_{q\in\triangle(\triangle(\Pi)),\omega\in\triangle(\Pi)} \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \big\{\mathbb{P}_{\lambda\sim q}[\lambda \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \leq \alpha^2\big\}$$

9. If $N_{\neg\boldsymbol{\pi}_{\overline{M}}} = 0$ almost surely under $\overline{M}$, we can take $R = 0$, in which case the statement of the lemma is vacuous.

$$
\begin{aligned}
&\geq \delta \cdot \inf_{q\in\triangle(\triangle(\Pi)),\omega\in\triangle(\Pi)} \sup_{M\in\mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \left\{\mathbb{I}\left\{\overline{\lambda}_q \notin \Lambda(M;2\varepsilon,\bar{\mathsf{n}}^M)\right\} \mid \mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right] \leq \alpha^2\right\} \\
&\geq \delta \cdot \inf_{\lambda\in\triangle(\Pi),\omega\in\triangle(\Pi)} \sup_{M\in\mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \left\{\mathbb{I}\left\{\lambda \notin \Lambda(M;2\varepsilon,\bar{\mathsf{n}}^M)\right\} \mid \mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right] \leq \alpha^2\right\} \\
&\geq \delta \cdot \inf_{\lambda\in\triangle(\Pi),\omega\in\triangle(\Pi)} \sup_{M\in\mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \left\{\mathbb{I}\left\{\lambda \notin \Lambda(M;2\varepsilon)\right\} \mid \mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right] \leq \alpha^2\right\} \\
&= \delta \cdot \mathbb{I}\left\{\alpha^2 \geq \left(\mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\right)^{-1}\right\}. \qquad (83)
\end{aligned}
$$

Hence, we have

$$
\mathbb{P}^{\overline{M},\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] \geq \delta \cdot \mathbb{I}\left\{\alpha^2 \geq \left(\mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\right)^{-1}\right\}. \qquad (84)
$$

To proceed, using Lemma A.11 of Foster et al. (2021), we have

$$
\begin{aligned}
\mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] &\geq \frac{1}{3}\mathbb{P}^{\overline{M},\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] - \frac{4}{3}D_{\mathsf{KL}}(\mathbb{P}^{\overline{M},\mathbb{A}}\,\|\,\mathbb{P}^{M,\mathbb{A}}) \\
&\geq \frac{\delta}{3}\mathbb{I}\left\{\alpha^2 \geq \left(\mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\right)^{-1}\right\} - \frac{4}{3}D_{\mathsf{KL}}(\mathbb{P}^{\overline{M},\mathbb{A}}\,\|\,\mathbb{P}^{M,\mathbb{A}}).
\end{aligned}
$$

Now, recall that from the definition, we have that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$,

$$
\begin{aligned}
D_{\mathsf{KL}}(\mathbb{P}^{\overline{M},\mathbb{A}}\,\|\,\mathbb{P}^{M,\mathbb{A}}) &= \mathbb{E}^{M,\mathbb{A}}\left[\sum_{t:\pi^t\notin\boldsymbol{\pi}_{\overline{M}}} \mathbb{E}_{\pi\sim p^t} D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right] \\
&= \mathbb{E}^{M,\mathbb{A}}\left[\frac{N_{\neg\boldsymbol{\pi}_{\overline{M}}}}{N_{\neg\boldsymbol{\pi}_{\overline{M}}}} \sum_{t:\pi^t\notin\boldsymbol{\pi}_{\overline{M}}} \mathbb{E}_{\pi\sim p^t} D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right] \\
&\leq R\cdot\,\mathbb{E}^{M,\mathbb{A}}\left[\frac{1}{N_{\neg\boldsymbol{\pi}_{\overline{M}}}} \sum_{t:\pi^t\notin\boldsymbol{\pi}_{\overline{M}}} \mathbb{E}_{\pi\sim p^t} D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right] \\
&= R\cdot\mathbb{E}_{\pi\sim\omega^{\overline{M}}}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right] \leq \alpha^2 R,
\end{aligned}
$$

where the first inequality uses that $\mathbb{P}^{\overline{M},\mathbb{A}}\left[N_{\neg\boldsymbol{\pi}_{\overline{M}}} > R\right] = 0$, and the second inequality uses (82).

With this, we have

$$
\mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] \geq \frac{\delta}{3}\mathbb{I}\left\{\alpha^2 \geq \left(\mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\right)^{-1}\right\} - \frac{4}{3}\alpha^2 R.
$$

We set $\alpha^2 = \frac{\delta}{8R}$, so that

$$
\begin{aligned}
\mathbb{P}^{M,\mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M;\varepsilon,\bar{\mathsf{n}}^M)\right] &\geq \frac{\delta}{6}\cdot\mathbb{I}\left\{\alpha^2 \geq \left(\mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\right)^{-1}\right\} \\
&= \frac{\delta}{6}\cdot\mathbb{I}\left\{R \leq \frac{\delta}{8}\cdot\mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\right\}.
\end{aligned}
$$

We conclude that unless

$$R > \frac{\delta}{8} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

the algorithm must have $\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Lambda(M; \varepsilon, \bar{\mathsf{n}}^M)\right] \geq \frac{\delta}{6}$.

□

### H.4. Proofs for Lower Bound Examples

**Proof of Example B.2.** Let $\Delta \in (0,1)$ and $A \geq 2$ be given and set

$$\mathcal{M} = \left\{ M(\pi) = \mathcal{N}(f^M(\pi), 1/2) \mid f^M \in [0,1]^A \right\}$$

and

$$\mathcal{M}_0 = \{M \in \mathcal{M} \ : \ \Delta^M_{\min} \geq \Delta/2\}$$

Define $\overline{M} \in \mathcal{M}$ via $f^{\overline{M}}(\pi) = \Delta \mathbb{I}\{\pi = A\}$. Fix $\varepsilon \in (0, 1/2)$ and define a subclass

$$\mathcal{M}' = \{\overline{M}\} \cup \{M_i\}_{i \in [A-1]}$$

via

$$f^{M_i}(\pi) = \Delta \mathbb{I}\{\pi = A\} + \varepsilon \cdot \Delta \mathbb{I}\{\pi = i\}.$$

Since $\varepsilon \leq 1/2$, we have $\mathcal{M}' \subseteq \mathcal{M}^{\mathrm{opt}}(\overline{M})$ and $\mathcal{M}' \subseteq \mathcal{M}_0$. In addition, we have

$$D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M_i(\pi)\big) = (f^{\overline{M}}(\pi) - f^{M_i}(\pi))^2.$$

Let $\mathcal{M}'' \subseteq \mathcal{M}$ denote the set of instances such that, for $M' \in \mathcal{M}''$, $D_{\mathsf{KL}}(M_i(A) \,\|\, M'(A)) = 0$, and $M' \in \mathcal{M}^{\mathrm{alt}}(M_i)$, for all $i \in [A-1]$. Then,

$$\begin{aligned} I^{M_i}(\lambda; \mathcal{M}) &= \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M_i)} \mathbb{E}_{\pi \sim \lambda}[D_{\mathsf{KL}}\big(M_i(\pi) \,\|\, M'(\pi)\big)] \\ &\leq \inf_{M' \in \mathcal{M}''} \mathbb{E}_{\pi \sim \lambda}[D_{\mathsf{KL}}\big(M_i(\pi) \,\|\, M'(\pi)\big)] \\ &= \min_{j \in [A-1]} \big\{ \lambda_i \cdot (1-\varepsilon)^2 \Delta^2 \mathbb{I}\{j = i\} + \lambda_j \cdot \Delta^2 \mathbb{I}\{j \neq i\} \big\}. \end{aligned} \tag{85}$$

We also have that

$$\mathsf{g}^{M_i} = \mathsf{g} := \frac{(A-2)}{\Delta} + \frac{1}{(1-\varepsilon)\Delta},$$

where we have used again that $\varepsilon \leq 1/2$.

Fix any pair $\lambda, \omega \in \triangle_{\Pi}$ and consider the value

$$\sup_{M \in \mathcal{M}' \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\right]} \right\}.$$

Pick $\delta \le \varepsilon/32$, and let $\mathcal{J} \subseteq [A-1]$ be the set of models $i$ for which $\lambda \in \Lambda(M_i;\delta)$. For each such model, by definition, there exists $\mathsf{n}_i > 0$ such that

$$\Delta^{M_i}(\lambda) \le (1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i}, \quad \text{and} \quad I^{M_i}(\lambda;\mathcal{M}) \ge (1-\delta)\frac{1}{\mathsf{n}_i}.$$

Define $\bar{\mathsf{n}} = \max_{i\in\mathcal{J}} \mathsf{n}_i$ and $\underline{\mathsf{n}} = \min_{i\in\mathcal{J}} \mathsf{n}_i$. Using Eq. (85), it is immediate that for all $j \in [A-1]$, $\lambda_j \ge (1-\delta)\frac{\Delta^{-2}}{\underline{\mathsf{n}}}$, and that for all $i \in \mathcal{J}$,

$$\lambda_i \ge (1-\delta)\frac{(1-\varepsilon)^{-2}\Delta^{-2}}{\mathsf{n}_i}.$$

In particular, this implies that for all $i \in \mathcal{J}$,

$$\begin{aligned}(1-\delta)\frac{(A-2)\Delta^{-1}}{\underline{\mathsf{n}}} + (1-\delta)\frac{(1-\varepsilon)^{-1}\Delta^{-1}}{\mathsf{n}_i} &\le \Delta^{M_i}(\lambda)\\ &\le (1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i}\\ &= (1+\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i} + (1+\delta)\frac{(1-\varepsilon)^{-1}\Delta^{-1}}{\mathsf{n}_i},\end{aligned}$$

or by rearranging,

$$\begin{aligned}(1-\delta)\frac{(A-2)\Delta^{-1}}{\underline{\mathsf{n}}} &\le (1+\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i} + 2\delta\frac{(1-\varepsilon)^{-1}\Delta^{-1}}{\mathsf{n}_i},\\ &\le (1+\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i} + 4\delta\frac{\Delta^{-1}}{\mathsf{n}_i},\\ &\le (1+2\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i}\end{aligned}$$

as long as $A \ge 6$. Since this holds uniformly for all $i \in \mathcal{J}$, rearranging once more gives

$$\bar{\mathsf{n}} \le \frac{(1+2\delta)}{1-\delta}\underline{\mathsf{n}} \le (1+2\delta)^2\underline{\mathsf{n}} \le (1+8\delta)\underline{\mathsf{n}},$$

where we have used that $\delta \le 1/2$.

Now, observe that for all $i \in \mathcal{J}$, we have

$$\begin{aligned}\Delta^{M_i}(\lambda) &\ge (1-\delta)(A-|\mathcal{J}|-1)\frac{1}{\Delta\underline{\mathsf{n}}} + (1-\delta)|\mathcal{J}|\frac{1}{(1-\varepsilon)^2\Delta\bar{\mathsf{n}}} + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},\\ &\ge (1-\delta)(A-|\mathcal{J}|-1)\frac{1}{\Delta\underline{\mathsf{n}}} + \frac{(1-\delta)}{1+8\delta}|\mathcal{J}|\frac{1}{(1-\varepsilon)^2\Delta\underline{\mathsf{n}}} + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},\\ &\ge \frac{(1-\delta)}{\Delta\underline{\mathsf{n}}}\left((A-|\mathcal{J}|-1) + |\mathcal{J}|\frac{1-8\delta}{(1-\varepsilon)^2}\right) + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},\\ &\ge \frac{(1-\delta)}{\Delta\underline{\mathsf{n}}}\left((A-|\mathcal{J}|-1) + |\mathcal{J}|\frac{1}{(1-\varepsilon)}\right) + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},\end{aligned}$$

where we have used that $\frac{1}{1+x} \ge 1-x$, and that $\delta \le \varepsilon/8$. Suppose that $|\mathcal{J}| \ge \frac{A}{2}$. Then we have

$$(A-|\mathcal{J}|-1) + |\mathcal{J}|\frac{1}{(1-\varepsilon)} \ge \frac{A}{2}\left(1+\frac{1}{1-\varepsilon}\right) - 1 \ge (1+\varepsilon/2)A - 1$$

$$\geq (1+\varepsilon/2)(A-1),$$

so that

$$\Delta^{M_i}(\lambda) \geq \frac{(1-\delta)(1+\varepsilon/2)}{\Delta \underline{\mathsf{n}}}(A-1) + (1-\delta)\frac{1}{(1-\varepsilon)\Delta \mathsf{n}_i}.$$

Noting that $\underline{\mathsf{n}} \leq \mathsf{n}_i$ and $\delta \leq \varepsilon/8$, we further have

$$\Delta^{M_i}(\lambda) \geq (1+\varepsilon/4)\frac{A-1}{\Delta}\frac{1}{\mathsf{n}_i} + (1-\delta)\frac{1}{(1-\varepsilon)\Delta \mathsf{n}_i}.$$

Observe that the right-hand side above is greater than $(1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i}$ if and only if

$$(1+\varepsilon/4)(A-1) > (1+\delta)(A-1) + \frac{2\delta}{1-\varepsilon},$$

which is satisfied if $\delta \leq \varepsilon/32$. In this case, we have

$$\Delta^{M_i}(\lambda) > (1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i},$$

which contradicts the assumption that $i \in \mathcal{J}$. It follows that we must have $|\mathcal{J}| < A/2$.

Now, to conclude, select $i = \arg\min_{i \in [A-1]\setminus\mathcal{J}} \omega_i$, and consider the value

$$\sup_{M \in \mathcal{M}' \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\} \geq \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M_i(\pi)\right)\right]} = \frac{1}{\omega_i \cdot \varepsilon^2\Delta^2},$$

where the first inequality follows because $\lambda \notin \Lambda(M_i;\delta)$ by definition, and the equality follows from the construction of $\overline{M}$ and $M_i$. Since $\sum_{i\in[A-1]\setminus\mathcal{J}} \omega_i \leq 1$ and $|[A-1]\setminus\mathcal{J}| \geq \frac{A}{2}$, we must have $\omega_i \leq \frac{2}{A}$, so that

$$\frac{1}{\omega_i \cdot \varepsilon^2\Delta^2} \geq \frac{A}{2\varepsilon^2\Delta^2}$$

as desired. To complete the proof, note that this holds uniformly for all choices for $\lambda$ and $\omega$, and that

$$\begin{aligned}
\mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0, \overline{M}) &= \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}_0\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\} \\
&\geq \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}'\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\}.
\end{aligned}$$

To obtain the parameter setting in the theorem statement, we rescale $\Delta \leftarrow 2\Delta$ and $\varepsilon \leftarrow 32\varepsilon$. □

**Proof of Example B.3.** We reduce the lower bound to that of multi-armed bandits via a standard tree construction (Osband and Van Roy, 2016; Domingues et al., 2021); as the argument is standard, we only sketch the approach. Assume without loss of generality that $H$ is a multiple of 2. Set $H = \log_2(S/2)$. Consider a sub-class $\mathcal{M}' \subseteq \mathcal{M}$ defined as follows. All models $M \in \mathcal{M}'$ have

identical, deterministic dynamics given by a binary tree. Each layer $h$ has $2^{h-1}$ states, so that layer $H$ has $S/4$ states, and the total number of states is $S-1$. The agent begins from a root state $s_1$ deterministically. For $h \le H-1$, there are two available actions, left and right. Choosing left leads to the left successor for the current layer, and right leads to the right successor for the current layer. There are no rewards for layer $h \le H-1$. For layer $H$, there are $A$ available actions, and rewards are arbitrary, subject to the constraint that the mean lies in $[0,1]$ and the noise follows $\mathcal{N}(0,1/2)$.

It is clear that the class $\mathcal{M}'$ is equivalent to the class of multi-armed bandit instances with $SA/4$ actions. As a consequence, the lower bound follows from Example B.2.

□

**Proof of Example B.1.** Let $\Delta \in (0,1/6)$, $\beta \in (0,1)$, and $A, N \ge 2$ be given. Consider the reference model $\overline{M} \in \mathcal{M}^+$ defined as follows:

- For each bandit arm $k \in [A]$, we have $f^{\overline{M}}(k) = \frac{1}{2} + \Delta\mathbb{I}\{k = A\}$ and $r \sim \mathcal{N}(f^{\overline{M}}(k), 1)$. There are no observations, i.e. $o = \perp$ almost surely.
- For each revealing arm $\pi^\circ_k$, we receive zero reward almost surely (so $f^{\overline{M}}(\pi^\circ_k) = 0$) and $o \sim \mathrm{Unif}([A])$.

We define a subclass

$$\mathcal{M}' = \{M_j\}_{j\in[N]} \subset \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$$

as follows

- For each bandit arm $k \in [A]$, we have $f^{M_j}(k) = \frac{1}{2} + \Delta\mathbb{I}\{k = A\}$ and $r \sim \mathcal{N}(f^{M_j}(k), 1)$. There are no observations, i.e. $o = \perp$ almost surely.
- For each revealing arm $\pi^\circ_k$, we receive zero reward almost surely (so $f^{M_j}(\pi^\circ_k) = 0$). We have

$$o \sim \begin{cases} \mathrm{Unif}([A]), & k \neq j, \\ \beta\mathbb{I}_i + (1-\beta)\mathrm{Unif}([A]), & k = j. \end{cases}$$

Note that $\mathcal{M}' \subseteq \mathcal{M}_0$. For all $j \in [N]$, a direct calculation gives

$$D_{\mathsf{KL}}\big(\overline{M}(\pi^\circ_j) \,\|\, M_j(\pi^\circ_j)\big) = \frac{A-1}{A}\log\left(\frac{1}{1-\beta}\right) + \frac{1}{A}\log\left(\frac{1}{1+\beta(A-1)}\right) =: \alpha$$

and

$$D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M_j(\pi)\big) = \alpha \cdot \mathbb{I}\{\pi = \pi^\circ_j\}. \tag{86}$$

In addition, it is straightforward to see that $\alpha \le 2\beta$ whenever $\beta \le 1/2$. Next we calculate that for any $j \in [N]$ and $M \in \mathcal{M}$ with $\pi^\circ_M = \pi^\circ_j$, and $\pi_M \neq A$,

$$D_{\mathsf{KL}}\big(M_j(\pi^\circ_j) \,\|\, M(\pi^\circ_j)\big) = \beta\log\left(1 + \frac{\beta A}{1-\beta}\right) =: \gamma,$$

which has $\gamma \leq O(\beta)$ whenever $\beta \leq 1/A$ and $\gamma \geq \beta \log(1+\beta A)$. Lastly, we have that for all $i \in [A]$, all $M, M' \in \mathcal{M}$ have

$$D_{\mathsf{KL}}\big(M(i) \,\|\, M'(i)\big) = \frac{1}{2}(f^{M}(i) - f^{M'}(i))^2.$$

Let $\mathcal{M}''$ denote the set of instances such that for $M' \in \mathcal{M}''$, $\pi_{M'} \neq A$, and $f^{M'}(A) = \frac{1}{2} + \Delta$, so that $D_{\mathsf{KL}}(M_j(A) \,\|\, M'(A)) = 0$ and $\mathcal{M}'' \subseteq \mathcal{M}^{\mathrm{alt}}(M_j)$ for all $j \in [N]$. Using the above calculations and the definition of $I^{M_j}(\lambda; \mathcal{M})$, we can then compute, for all $j \in [N]$,

$$I^{M_j}(\lambda; \mathcal{M}) \leq \inf_{M' \in \mathcal{M}''} \mathbb{E}_\lambda[D_{\mathsf{KL}}\big(M_j(\pi) \,\|\, M'(\pi)\big)] = \frac{\Delta^2}{2} \cdot \min_{k \in [A-1]} \lambda_k + \gamma \cdot \lambda_{\pi_j^\circ} \tag{87}$$

and

$$\mathsf{g}^{M_j} = \mathsf{g} := \min\left\{2\frac{A-1}{\Delta}, \left(\frac{1}{2} + \Delta\right)\frac{1}{\gamma}\right\} = \left(\frac{1}{2} + \Delta\right)\frac{1}{\gamma}$$

whenever $\gamma \geq \Delta/2(A-1)$.

Fix any pair $\lambda, \omega \in \triangle_\Pi$ and consider the value

$$\sup_{M \in \mathcal{M}' \setminus \mathcal{M}^{\mathrm{gl}}_{1/2}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi \sim \omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big]} \right\}.$$

Let $\mathcal{J} \subseteq [N]$ be the set of models $j$ for which $\lambda \in \Lambda(M_j; 1/2)$. For each such model, by definition, there exists $\mathsf{n}_j > 0$ such that

$$\Delta^{M_j}(\lambda) \leq (1 + 1/2)\frac{\mathsf{g}}{\mathsf{n}_j} \leq \frac{1}{\gamma \mathsf{n}_j}, \quad \text{and} \quad I^{M_j}(\lambda; \mathcal{M}) \geq \frac{1}{2\mathsf{n}_j}.$$

Define $\bar{\mathsf{n}} = \max_{j \in \mathcal{J}} \mathsf{n}_j$ and $\underline{\mathsf{n}} = \min_{j \in \mathcal{J}} \mathsf{n}_j$. Let us begin with some basic observations.

- Since $\Delta^{M_j} = \Delta^{\overline{M}}$ for all $j$, we have $\Delta^{M_j}(\lambda) \leq \frac{1}{\gamma \mathsf{n}_{j'}}$ for all $j, j' \in \mathcal{J}$, and hence

$$\Delta^{M_j}(\lambda) \leq \frac{1}{\gamma \bar{\mathsf{n}}}. \tag{88}$$

- Any $j \in \mathcal{J}$ must have

$$\lambda_{\pi_j^\circ} \gamma \geq \frac{1}{4\mathsf{n}_j} \geq \frac{1}{4\bar{\mathsf{n}}}. \tag{89}$$

  To see this, observe that if it were not the case, we would need $\min_{i \in [A-1]} \lambda_i \frac{\Delta^2}{2} \geq \frac{1}{4\mathsf{n}_j}$ to satisfy the constraint that $I^{M_j}(\lambda; \mathcal{M}) \geq \frac{1}{2\mathsf{n}_j}$ (by (87)). But if this were to occur, we would have

$$\frac{A-1}{2\Delta \mathsf{n}_j} \leq \Delta^{M_j}(\lambda) \leq \frac{1}{\gamma \mathsf{n}_j},$$

  which would contradict the assumption that $\gamma \geq 2\Delta/(A-1)$.

Combing the inequalities (88) and (89), it follows that any $j \in \mathcal{J}$ must have

$$\frac{|\mathcal{J}|}{8\gamma \mathsf{n}_j} \le \frac{1}{2}\sum_{k\in\mathcal{J}} \lambda_{\pi_k^\circ} \le \Delta^{M_j}(\lambda) \le \frac{1}{\gamma \mathsf{n}_j},$$

which implies that $|\mathcal{J}| \le 8$. Hence, as long as $N \ge 16$, we have $|[N] \setminus \mathcal{J}| \ge N/2$.

To conclude, select $k = \arg\min_{j\in[N]\setminus\mathcal{J}} \omega_{\pi_j^\circ}$, and consider the value

$$\sup_{M\in\mathcal{M}'\setminus\mathcal{M}^{\mathrm{gl}}_{1/2}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M(\pi)\right)\right]} \right\} \ge \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi)\,\|\,M_k(\pi)\right)\right]} = \frac{1}{\omega_{\pi_k^\circ}\cdot\alpha},$$

where the first inequality follows because $\lambda \notin \Lambda(M_k; 1/2)$ by definition, and the equality follows from (86). Since $\sum_{j\in[N]\setminus\mathcal{J}} \omega_{\pi_j^\circ} \le 1$ and $|[N]\setminus\mathcal{J}| \ge \frac{N}{2}$, we must have $\omega_{\pi_k^\circ} \le \frac{2}{N}$, so that

$$\frac{1}{\omega_{\pi_k^\circ}\cdot\alpha} \ge \frac{N}{2\alpha}.$$

as desired. Since this holds uniformly for all choices for $\lambda$ and $\omega$, the proof is completed. □

### H.5. Lower Bound on Regret for Algorithms with Well-Behaved Tails

In this section, we present an additional result, Theorem H.3, which shows that for algorithms for which the tail behavior is "well-behaved" in a certain sense, the Allocation-Estimation Coefficient directly leads to lower bounds on the least possible value of $T$ for which any algorithm can achieve (approximate) instance-optimality.

**Theorem H.3.** *Let the time horizon $T \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Suppose that there exists an algorithm $\mathbb{A}$ with the property that for all $M \in \mathcal{M}_0$,*

1. $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^M \log(T)$.
2. *For all $\pi \in \Pi$, if $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \ne 0$, then $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \ge 1$.*
3. $\sqrt{\mathbb{E}^{M,\mathbb{A}}[(\mathbf{Reg}(T))^2]} \le 2\mathsf{g}^M \log(T)$.

*In addition, assume that 1) $\mathsf{g}^M \ge 1$ for all $M \in \mathcal{M}_0$, 2) Assumption A.4 holds, 3) Assumption A.2 holds with parameter $V_{\mathcal{M}} \ge 1$, and 4) that*

$$\log(T) \ge \frac{12}{\varepsilon}\log\left(\sup_{M\in\mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}}\cdot\log(T)\right).$$

*Then if we define $\delta = \varepsilon\cdot\min\{1, \inf_{M\in\mathcal{M}_0} \frac{\mathsf{g}^M}{3\mathsf{g}^M/\Delta^M_{\min} + \mathsf{n}^{\mathcal{M}}_\varepsilon}\}$, it must be the case that*

$$\log^3(T) \ge \frac{\delta^2}{C}\cdot \sup_{\overline{M}\in\mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{4\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

*for $C \le O\left((\sup_{M\in\mathcal{M}} \frac{\mathsf{g}^M}{\Delta^M_{\min}})^4 \cdot \frac{V^2_{\mathcal{M}}\log(\delta^{-1})}{\varepsilon^2}\right)$.*

We prove Theorem H.3 by combining Theorem B.2 with an another technical result, Proposition H.1 (stated and proven in the sequel), which shows that for algorithms that satisfy the assumptions of Theorem H.3, it is possible to use the empirical decision frequencies to compute an allocation that is approximately optimal with high probability.

**Proof of Theorem H.3.** Define $\bar{\mathsf{n}}^M = 3\frac{\mathsf{g}^M}{\Delta^M_{\min}} + \mathsf{n}^{\mathcal{M}}_{\varepsilon}$, and set

$$\delta := \varepsilon \cdot \min\left\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\right\}.$$

Assume that

$$\log(T) \geq \frac{12}{\varepsilon} \log\left(\sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T)\right).$$

Let $\mathbb{A}$ be the algorithm in the statement of the proposition, and let $\mathbb{A}'$ be the modified algorithm created through Proposition H.1 with parameter $\delta$. By assumption, we have that $\sqrt{\mathbb{E}^{M,\mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}]} \leq R := 2\sup_{M \in \mathcal{M}} \frac{\mathsf{g}^M}{\Delta^M_{\min}} \log(T)$. We define $n = c \cdot \frac{\log(24\delta^{-1})}{\varepsilon^2} \cdot \frac{R^3 V^2_{\mathcal{M}}}{\log(T)}$ for a sufficiently large numerical constant $c$ and $T' = T \cdot n$. Proposition H.1 implies that for time $T'$, the algorithm $\mathbb{A}'$ satisfies

$$\sqrt{\mathbb{E}^{M,\mathbb{A}'}[N_{\neg \boldsymbol{\pi}_M}]} \leq R' := R \cdot n$$

and

$$\mathbb{P}^{M,\mathbb{A}'}\left[\widehat{\lambda} \in \Lambda(M; 2\varepsilon, \bar{\mathsf{n}}^M)\right] \geq 1 - \frac{\delta}{24}.$$

On the other hand, since the precondition of Theorem H.2 is now satisfied with parameter $R'$, we have that unless

$$R' \geq \frac{\delta^2}{192} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{4\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}), \tag{90}$$

the algorithm must have

$$\mathbb{P}^{M,\mathbb{A}'}\left[\widehat{\lambda} \in \Lambda(M; 2\varepsilon, \bar{\mathsf{n}}^M)\right] \leq 1 - \frac{\delta}{12}.$$

As $\frac{\delta}{12} > \frac{\delta}{24}$, this is a contradiction unless (90) holds.

□

**Proposition H.1.** *Let the time horizon $T \in \mathbb{N}$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Let $\mathbb{A}$ be an algorithm with the property that for all $M \in \mathcal{M}_0$,*

1. $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^M \log(T)$ *for some* $\varepsilon \in (0,1)$.
2. *For all* $\pi \in \Pi$, *if* $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \neq 0$, *then* $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \geq 1$.
3. $\sqrt{\mathbb{E}^{M,\mathbb{A}}\left[N^2_{\neg \boldsymbol{\pi}_M}\right]} \leq R$ *for some* $R \geq 2$.

*In addition, assume that 1)* $\mathsf{g}^M \geq 1$ *for all* $M \in \mathcal{M}_0$*, 2) Assumption A.4 holds, 3) Assumption A.2 holds with parameter* $V_{\mathcal{M}} \geq 1$*, and 4) that*

$$\log(T) \geq \frac{12}{\varepsilon} \log\left( \sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T) \right).$$

*Then for any* $\delta \in (0, e^{-1})$*, if we define* $n = c \cdot \frac{\log(\delta^{-1})}{\varepsilon^2} \cdot \frac{R^3 V^2_{\mathcal{M}}}{\log(T)}$ *for a sufficiently large numerical constant* $c$*, there exists an algorithm* $\mathbb{A}'$ *that, using* $T' := T \cdot n$ *rounds, returns a normalized allocation* $\widehat{\lambda} \in \triangle_\Pi$ *such that*

$$\mathbb{P}^{M, \mathbb{A}'}\left[ \widehat{\lambda} \in \Lambda(M; 2\varepsilon, \bar{\mathsf{n}}^M) \right] \geq 1 - \delta,$$

*for* $\bar{\mathsf{n}}^M \leq 3\frac{\mathsf{g}^M}{\Delta^M_{\min}} + \mathsf{n}^{\mathcal{M}}_\varepsilon$ *and that* $\sqrt{\mathbb{E}^{M, \mathbb{A}'}\left[N^2_{\neg \boldsymbol{\pi}_M}\right]} \leq R \cdot n$ *and*

$$\mathbb{E}^{M, \mathbb{A}'}\left[ \mathbf{Reg}(T') \right] \leq (1 + \varepsilon)\mathsf{g}^M \log(T) \cdot n$$

*for all* $M \in \mathcal{M}_0$*.*

**Proof of Proposition H.1.** We first state a technical lemma regarding robust mean estimation.

**Lemma H.5.** *Let* $X \in \mathbb{R}^d$ *be a random variable with* $\mu := \mathbb{E}[X]$*. Assume that* $\|\mu\|_0 \leq s$*, where* $s$ *is a known upper bound. For any* $\delta \leq e^{-1}$*, there exists an estimator* $\widehat{\mu}_n$ *that, given* $n$ *independent samples from* $X$*, ensures that with probability at least* $1 - \delta$*,*

$$\|\widehat{\mu}_n - \mu\|_1 \leq 24\sqrt{\frac{2s \cdot \mathbb{E}\|X - \mu\|_2^2 \cdot \log(\delta^{-1})}{n}}.$$

*In addition,* $\|\widehat{\mu}_n\|_0 \leq s$ *with probability* 1*.*

Throughout, we will use that since Assumption A.4 holds, $\pi_M$ is unique for all $M \in \mathcal{M}$. Fix $M \in \mathcal{M}_0$. Let $\widehat{\eta} \in \mathbb{R}^\Pi_+$ denote the vector of empirical decision frequencies when $\mathbb{A}$ is run with horizon $T$, i.e. $\widehat{\eta}(\pi) = T(\pi)$. Let

$$\eta^M = \mathbb{E}^{M, \mathbb{A}}[\widehat{\eta}].$$

For parameters $n \in \mathbb{N}$ and $\delta \leq e^{-1}$ we define $\mathbb{A}'$ as follows:

- Run $\mathbb{A}$ a total of $n$ times independently (so that $T' = T \cdot n$), and let $\widehat{\eta}^1, \dots, \widehat{\eta}^n$ be the empirical decision frequencies.
- Apply the algorithm from Lemma H.5 to $\widehat{\eta}^1, \dots, \widehat{\eta}^n$ with parameters $\delta$ and $s = 2R$, and let $\check{\eta} \in \mathbb{R}^\Pi_+$ be the resulting vector (note that we can take $\check{\eta}$ to have non-negative entries without loss of generality).
- Set $\widehat{\pi} = \arg\max_{\pi \in \Pi} \check{\eta}$, and set $\widetilde{\eta}(\pi) = \check{\eta}(\pi)\mathbb{I}\{\pi \neq \widehat{\pi}\}$ and $\widetilde{\eta}(\widehat{\pi}) = \mathsf{n}^{\mathcal{M}}_\varepsilon \cdot \log T$ (note that $\mathsf{n}^{\mathcal{M}}_\varepsilon$ is a class-dependent quantity, and so is known to the learner).
- Set $\widehat{\lambda} = \widetilde{\eta}/\|\widetilde{\eta}\|_1$.

It is immediate that this construction satisfies $\sqrt{\mathbb{E}^{M,\mathbb{A}'}\left[N_{\neg\pi_M}^2\right]} \leq R \cdot n$ and

$$\mathbb{E}^{M,\mathbb{A}'}\left[\mathbf{Reg}(T')\right] \leq (1+\varepsilon)\mathsf{g}^M \log(T) \cdot n,$$

so it remains to show that $\widehat{\lambda}$ is a near-optimal allocation with high probability when $n$ is chosen appropriately.

We start by applying Lemma H.5. To do so, we carry out some prerequisite calculations. First, by assumption, we have $\eta^M(\pi) \geq 1$ if $\eta^M \neq 0$. Using this, along with Assumption A.4, we have

$$\|\eta^M\|_0 \leq 1 + \sum_{\pi \neq \pi_M} \eta^M(\pi) \leq 1 + \mathbb{E}^{M,\mathbb{A}}[N_{\neg\pi_M}] \leq 1 + R \leq 2R.$$

Second,

$$\begin{aligned}\mathbb{E}^{M,\mathbb{A}}\|\widehat{\eta} - \eta^M\|_2^2 &\leq \mathbb{E}^{M,\mathbb{A}}\left[\left(\widehat{\eta}(\pi_M) - \eta^M(\pi_M)\right)^2\right] + \sum_{\pi \neq \pi_M} \mathbb{E}^{M,\mathbb{A}}\left[\widehat{\eta}(\pi)^2\right] \\ &\leq \mathbb{E}^{M,\mathbb{A}}\left[\left(\widehat{\eta}(\pi_M) - \eta^M(\pi_M)\right)^2\right] + \mathbb{E}^{M,\mathbb{A}}\left[N_{\neg\pi_M}^2\right].\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}^{M,\mathbb{A}}\left[\left(\widehat{\eta}(\pi_M) - \eta^M(\pi_M)\right)^2\right] &= \mathbb{E}^{M,\mathbb{A}}\left[\left(\sum_{\pi \neq \pi_M} \widehat{\eta}(\pi) - \sum_{\pi \neq \pi_M} \eta^M(\pi)\right)^2\right] \\ &\leq \mathbb{E}^{M,\mathbb{A}}\left[\left(\sum_{\pi \neq \pi_M} \widehat{\eta}(\pi)\right)^2\right] \\ &= \mathbb{E}^{M,\mathbb{A}}\left[N_{\neg\pi_M}^2\right].\end{aligned}$$

so that

$$\mathbb{E}^{M,\mathbb{A}}\|\widehat{\eta} - \eta^M\|_2^2 \leq 2\,\mathbb{E}^{M,\mathbb{A}}\left[N_{\neg\pi_M}^2\right] \leq 2R^2.$$

As a result, Lemma H.5 implies that with probability $1-\delta$,

$$\|\check{\eta} - \eta^M\|_1 \leq \sqrt{C_1 \frac{\log(\delta^{-1})}{n}} =: \varepsilon_{\mathrm{stat}} \tag{91}$$

where $C_1 = O(R^3)$.

Next, we appeal to Lemma H.2, which implies that as long as

$$\log(T) \geq \frac{12}{\varepsilon} \log\left(\sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T)\right), \tag{92}$$

we have

$$\Delta^M(\eta^M) \leq (1+\varepsilon/2)\mathsf{g}^M \log(T), \quad \text{and} \quad I^M(\eta^M; \mathcal{M}) \geq (1-\varepsilon/2)\log(T).$$

Applying (91) and using that $\Delta^M \in [0,1]$ and $D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \le 2V_{\mathcal{M}}$ by Lemma F.13, we have

$$\Delta^M(\check{\eta}) \le (1+\varepsilon/2)\mathsf{g}^M \log(T) + \varepsilon_{\mathrm{stat}}, \quad \text{and} \quad I^M(\check{\eta}; \mathcal{M}) \ge (1-\varepsilon/2)\log(T) - \varepsilon_{\mathrm{stat}} \cdot 2V_{\mathcal{M}}.$$

Thus, as long as

$$\varepsilon_{\mathrm{stat}} \le \frac{1}{2}\varepsilon \cdot \min\{\mathsf{g}^M \log(T), (2V_{\mathcal{M}})^{-1}\log(T)\},$$

we have

$$\Delta^M(\check{\eta}) \le (1+\varepsilon)\mathsf{g}^M \log(T), \quad \text{and} \quad I^M(\check{\eta}; \mathcal{M}) \ge (1-\varepsilon)\log(T). \tag{93}$$

Next, we claim that $\widehat{\pi} = \pi_M$. To see this, note that since $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \le 2\mathsf{g}^M \log(T)$, we have

$$\eta^M(\pi_M) \ge T - 2\frac{\mathsf{g}^M}{\Delta^M_{\min}}\frac{\log(T)}{T} \ge \frac{3}{4}T$$

as long as $T > 8\frac{\mathsf{g}^M}{\Delta^M_{\min}}\log(T)$, which is implied by the condition (92). Hence, as long as $\varepsilon_{\mathrm{stat}} \le \frac{T}{4}$, we have

$$\check{\eta}(\pi_M) > \frac{T}{2},$$

which implies that $\widehat{\pi} = \pi_M$. By definition of $\mathsf{n}^{\mathcal{M}}_\varepsilon$, and since $\check{\eta}$ satisfies Eq. (93) above, we then have that setting $\check{\eta}(\widehat{\pi}) = \mathsf{n}^{\mathcal{M}}_\varepsilon \log(T)$ does not affect the regret, and only decreases the information gain by a factor of $\varepsilon\log(T)$. It follows that

$$\Delta^M(\widetilde{\eta}) \le (1+\varepsilon)\mathsf{g}^M \log(T), \quad \text{and} \quad I^M(\widetilde{\eta}; \mathcal{M}) \ge (1-2\varepsilon)\log(T).$$

We conclude that $\widehat{\lambda} \in \Lambda(M; 2\varepsilon, \bar{\mathsf{n}})$ for

$$\bar{\mathsf{n}} = \|\widetilde{\eta}\|_1 / \log(T).$$

To wrap up, we compute that

$$\|\widetilde{\eta}\|_1 \le \sum_{\pi \neq \pi_M} \eta^M(\pi) + \varepsilon_{\mathrm{stat}} + \mathsf{n}^{\mathcal{M}}_\varepsilon \log(T) \le 2\frac{\mathsf{g}^M}{\Delta^M_{\min}}\log(T) + \varepsilon_{\mathrm{stat}} + \mathsf{n}^{\mathcal{M}}_\varepsilon \log(T) \le 3\frac{\mathsf{g}^M}{\Delta^M_{\min}}\log(T) + \mathsf{n}^{\mathcal{M}}_\varepsilon \log(T)$$

whenever $\varepsilon_{\mathrm{stat}} \le \mathsf{g}^M \log(T)$.

□

**Proof of Lemma H.5.** From Proposition 1 of Lugosi and Mendelson (2019), we have that for any $\delta \le e^{-1}$, there exists an estimator $\widetilde{\mu}_n$ that, given $n$ independent samples of $X$, ensures that with probability at least $1-\delta$,

$$\|\widetilde{\mu}_n - \mu\|_2 \le 12\sqrt{\frac{\mathbb{E}\|X-\mu\|_2^2 \cdot \log(\delta^{-1})}{n}}.$$

Given $\widetilde{\mu}_n$, we define $\widehat{\mu}_n = \arg\min_{u \in \mathbb{R}^d : \|u\|_0 \le s} \|u - \widetilde{\mu}_n\|_2$. Since $\|\mu\|_0 \le s$, we have $\|\widetilde{\mu}_n - \widehat{\mu}_n\|_2 = \min_{u:\|u\|_0 \le s}\|\widetilde{\mu}_n - u\|_2 \le \|\widetilde{\mu}_n - \mu\|_2$. It follows that

$$\|\widehat{\mu}_n - \mu\|_2 \le \|\widetilde{\mu}_n - \widehat{\mu}_n\|_2 + \|\widetilde{\mu}_n - \mu\| \le 2\|\widetilde{\mu}_n - \mu\|_2.$$

Finally, we note that since $\widehat{\mu}_n$ and $\mu$ are both $s$-sparse, $\|\widehat{\mu}_n - \mu\|_1 \le \sqrt{2s}\|\widehat{\mu}_n - \mu\|_2$. □