On the Complexity of Adversarial Decision Making

Dylan J. Foster dylanfoster@microsoft.com
Alexander Rakhlin rakhlin@mit.edu
Ayush Sekhari sekhari@mit.edu
Karthik Sridharan ks999@cornell.edu

Abstract

A central problem in online learning and decision making—from bandits to reinforcement learning—is to understand what modeling assumptions lead to sample-efficient learning guarantees. We consider a general *adversarial decision making* framework that encompasses (structured) bandit problems with adversarial rewards and reinforcement learning problems with adversarial dynamics. Our main result is to show—via new upper and lower bounds—that the Decision-Estimation Coefficient, a complexity measure introduced by Foster et al. [17] in the stochastic counterpart to our setting, is necessary and sufficient to obtain low regret for adversarial decision making. However, compared to the stochastic setting, one must apply the Decision-Estimation Coefficient to the *convex hull* of the class of models (or, hypotheses) under consideration. This establishes that the price of accommodating adversarial rewards or dynamics is governed by the behavior of the model class under convexification, and recovers a number of existing results—both positive and negative. En route to obtaining these guarantees, we provide new structural results that connect the Decision-Estimation Coefficient to variants of other well-known complexity measures, including the Information Ratio of Russo and Van Roy [47] and the Exploration-by-Optimization objective of Lattimore and György [32].

1 Introduction

To reliably deploy data-driven decision making methods in real-world systems where safety is critical, such methods should satisfy two desiderata: (i) provable robustness in the face of dynamic or even adversarial environments, and (ii) ability to effectively take advantage of problem structure as modeled by the practitioner. In high-dimensional problems, this entails efficiently generalizing across states and actions while delicately exploring new decisions.

For decision making in static, stochastic environments, recent years have seen extensive investigation into optimal sample complexity and algorithm design principles, and the foundations are beginning to take shape. With an emphasis on reinforcement learning, a burgeoning body of research identifies specific modeling assumptions under which sample-efficient interactive decision making is possible [12, 54, 20, 39, 5, 27, 14, 36, 13, 56], as well as general structural conditions that aim to unify these assumptions [45, 19, 51, 53, 15, 21, 17]. For dynamic or adversarial settings, however, comparatively little is known outside of (i) positive results for special cases such as adversarial bandit problems [4, 3, 18, 10, 1, 7, 26, 16, 8, 29], and (ii) a handful of negative results suggesting that online reinforcement learning in agnostic or adversarial settings can actually be statistically intractable [48, 37]. These developments raise the following questions: (a) what are the underlying phenomena that govern the statistical complexity of decision making in adversarial settings? (b) what are the corresponding algorithmic design principles that attain optimal statistical complexity?

Contributions. We consider an adversarial variant of the *Decision Making with Structured Observations* (DMSO) framework introduced in Foster et al. [17], where a learner or decision-maker interacts with a sequence of *models* (reward distributions in the case of bandits, or MDPs in the case of reinforcement learning) chosen by an adaptive adversary, and aims to minimize regret against the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

best decision in hindsight. Models are assumed to belong to a known *model class*, which reflects the learner's prior knowledge about the problem. The main question we investigate is: *How does the structure of the model class determine the minimax regret for adversarial decision making?* We show:

1. For *any* model class, one can obtain high-probability regret bounds that scale with a *convexified* version of the *Decision-Estimation Coefficient* (DEC), a complexity measure introduced by Foster et al. [17].
2. For any algorithm with "reasonable" tail behavior, the optimal regret for adversarial decision making is lower bounded by (a suitably localized version of) the convexified DEC.

In the process of obtaining these results, we draw new connections to several existing complexity measures.

1.1 Problem Setting

We adopt an adversarial variant of the DMSO framework of Foster et al. [17] consisting of T rounds, where at each round $t = 1, \ldots, T$:

1. The learner selects a *decision* $\pi^{(t)} \in \Pi$, where Π is the *decision space*.
2. Nature selects a *model* $M^{(t)} \in \mathcal{M}$, where $\mathcal{M}$ is a *model class*.
3. The learner receives a reward $r^{(t)} \in \mathcal{R} \subseteq \mathbb{R}$ and observation $o^{(t)} \in \mathcal{O}$ sampled via $(r^{(t)}, o^{(t)}) \sim M^{(t)}(\pi^{(t)})$, where $\mathcal{O}$ is the *observation space*. We abbreviate $z^{(t)} := (r^{(t)}, o^{(t)})$ and $\mathcal{Z} := \mathcal{R} \times \mathcal{O}$.

Here, each model $M = M(\cdot, \cdot \mid \cdot) \in \mathcal{M}$ is a conditional distribution $M : \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ that maps the learner's decision to a distribution over rewards and observations. This setting subsumes (adversarial) bandit problems, where models correspond to reward functions (or distributions), as well as adversarial reinforcement learning, where models correspond to Markov decision processes (MDPs). In both cases, the model class $\mathcal{M}$ encodes prior knowledge about the decision making problem, such as structure of rewards or dynamics (e.g., linearity or convexity). The model class might be parameterized by linear models, neural networks, or other rich function approximators depending on the problem domain.

For a model $M \in \mathcal{M}$, $\mathbb{E}^{M,\pi}[\cdot]$ denotes expectation under the process $(r, o) \sim M(\pi)$. We define $f^M(\pi) := \mathbb{E}^{M,\pi}[r]$ as the mean reward function and $\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$ as the decision with greatest reward for M. We let $\mathcal{F}_{\mathcal{M}} = \{f^M \mid M \in \mathcal{M}\}$ denote the induced class of reward functions. We measure performance via *regret* to the best fixed decision in hindsight:[1]

$$\mathbf{Reg}_{\mathsf{DM}} := \sup_{\pi^\star \in \Pi} \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \left[f^{M^{(t)}}(\pi^\star) - f^{M^{(t)}}(\pi^{(t)}) \right]. \tag{1}$$

This formulation—in which models are selected by a potentially adaptive adversary—generalizes Foster et al. [17], who considered a *stochastic* setting where $M^{(t)} = M^\star$ is fixed across all rounds. Examples include:

- **Adversarial bandits.** With no observations ($\mathcal{O} = \{\varnothing\}$), the adversarial DMSO framework is equivalent to the *adversarial bandit* problem with structured rewards. In this context, $\pi^{(t)}$ is typically referred to as an *action* or *arm* and Π is referred to as the *action space*. The most basic example here is the adversarial finite-armed bandit problem with A actions [4, 3, 18], where $\Pi = \{1, \ldots, A\}$ and $\mathcal{F}_{\mathcal{M}} = \mathbb{R}^A$. Other well-studied examples include adversarial linear bandits [10, 1, 7], bandit convex optimization [26, 16, 8, 29], and nonparametric bandits [26, 6, 38].[2]
- **Reinforcement learning.** The adversarial DMSO framework encompasses finite-horizon, episodic online reinforcement learning, with each round t corresponding to a single episode:

[1] The results in this paper immediately extend to the regret $\sup_{\pi^\star \in \Pi} \sum_{t=1}^{T} r^{(t)}(\pi^\star) - r^{(t)}(\pi^{(t)})$ through standard tail bounds.

[2] Typically, these examples are formulated with deterministic rewards, which we encompass by restricting models in $\mathcal{M}$ to be deterministic. Our formulation is more general and allows for, e.g., semi-stochastic adversaries.

$\pi^{(t)}$ is a *policy* (a mapping from state to actions) to play in the episode, $r^{(t)}$ is the cumulative reward in the episode, and the observation $o^{(t)}$ is the episode's trajectory (sequence of observed states, actions, and rewards). Online reinforcement learning in the stochastic setting where $M^{(t)} = M^\star$ is fixed has received extensive attention [19, 51, 20, 53, 15, 21, 17], but the adversarial setting we study has received less investigation. Examples include the adversarial MDP problem where an adversary chooses a sequence of tabular MDPs, which is known to be intractable [37], and the easier problem in which there is a fixed (known) MDP but rewards are adversarial [40, 57, 41, 22]. See Appendix D for more details.

We refer to Appendix B for additional measure-theoretic details and background, and to Foster et al. [17] for further examples and detailed discussion.[3]

Understanding statistical complexity (i.e., minimax regret) for the DMSO setting at this level of generality is a challenging problem. Even if one restricts only to bandit-type problems with no observations, any complexity measure must capture the role of structural assumptions such as convexity or smoothness in determining the optimal rates. To go beyond bandit problems and handle the general setting, one must accommodate problems with rich, structured feedback such as reinforcement learning, where observations (as well as subtle features of the noise distribution) can reveal information about the underlying model.

1.2 Overview of Results

For a model class $\mathcal{M}$, reference model $\overline{M} \in \mathcal{M}$, and scale parameter $\gamma > 0$, the Decision-Estimation Coefficient [17] is defined via

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big], \tag{2}$$

where we recall that for probability measures $\mathbb{P}$ and $\mathbb{Q}$ with a common dominating measure ν, (squared) Hellinger distance is given by

$$D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) = \int \left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}} \right)^2. \tag{3}$$

We define $\mathsf{dec}_\gamma(\mathcal{M}) = \sup_{\overline{M} \in \mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$, and let $\mathrm{co}(\mathcal{M})$ denote the convex hull of $\mathcal{M}$, which can be viewed as the set of all mixtures of models in $\mathcal{M}$. Our main results show that the *convexified Decision-Estimation Coefficient*,

$$\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})),$$

leads to upper and lower bounds on the optimal regret for adversarial decision making.

Theorem (informal). *For any model class $\mathcal{M}$, Algorithm 1 ensures that with high probability,*

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \tag{4}$$

where γ satisfies the balance $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \propto \frac{\gamma}{T} \log|\Pi|$. Moreover, for any algorithm with "reasonable" tail behavior (Section 2.2), regret must scale with a localized version of the same quantity.

As a consequence, there exists an algorithm for which $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \tilde{o}(T)$ if and only if $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \propto \gamma^{-\rho}$ for some $\rho > 0$.

For the stochastic version of our setting, Foster et al. [17] give upper and lower bounds that scale with $\mathsf{dec}_\gamma(\mathcal{M})$, without convexifying (under appropriate technical assumptions; cf. Section 2.3). Hence, our results show that in general, the gap in optimal regret for stochastic and adversarial decision making (or, "price of adversarial outcomes") is governed by the behavior of the DEC under convexification. For example, multi-armed bandits, linear bandits, and convex bandits correspond to convex model classes (where $\mathrm{co}(\mathcal{M}) = \mathcal{M}$), which gives a post-hoc explanation for why these problems are tractable in the adversarial setting. Finite state/action Markov decision processes do not correspond to a convex model class, and have $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))$ exponentially large compared to $\mathsf{dec}_\gamma(\mathcal{M})$; in this case, our results recover lower bounds of Liu et al. [37].

Beyond these results, we prove that the convexified Decision-Estimation Coefficient is equivalent to:

[3] We mention in passing that the upper bounds in this paper encompass the more general setting where rewards are not observed by the learner (i.e., $z^{(t)}$ does not contain the reward), thus subsuming the partial monitoring problem. Our lower bounds, however, require that rewards are observed. See Appendix A.

1. a "parameterized" variant of the generalized Information Ratio of Lattimore and György [32].
2. a novel high-probability variant of the *Exploration-by-Optimization* objective of Lattimore and Szepesvári [35], Lattimore and György [32].

Our techniques. On the lower bound side, we strengthen the approach from Foster et al. [17] with an improved change-of-measure argument (leading to improved results even in the stochastic setting), and combine this with the simple idea of constructing adversaries based on static mixture models. On the upper bound side, we extend the powerful Exploration-by-Optimization machinery of Lattimore and György [32] to the DMSO setting, and give a novel high-probability variant of the technique which leads to regret bounds for adaptive adversaries. We show that the performance of this method is controlled by a complexity measure whose value is equivalent to the convexified DEC, as well as parameterized variant of the Information Ratio (we present results in terms of the former to draw comparison to the stochastic setting).

Overall, our results heavily draw on the work of Foster et al. [17] and Lattimore and György [32], but we believe they play a valuable role in bridging these lines of research and formalizing connections.

Organization. Section 2 presents our main results, including upper and lower bounds on regret and a characterization of learnability. In Section 3, we provide new structural results connecting the DEC to Exploration-by-Optimization and the Information Ratio. We close with discussion of future directions (Section 4). Additional comparison to related work is deferred to Appendix A. The appendix also contains proofs and additional results, including examples (Appendix D) and further structural results (Appendix E).

2 Main Results

We now present our main results. First, using a new high-probability variant of the Exploration-by-Optimization technique [35, 32], we provide an upper bound on regret based on the (convexified) Decision-Estimation Coefficient (Section 2.1). Next, we present a lower bound that scales with a localized version of the same quantity (Section 2.2). Finally, we use these results to give a characterization for learnability (Section 2.3), and discuss the gap between stochastic and adversarial decision making.

To keep presentation as simple as possible, we make the following assumption.

Assumption 2.1. *The decision space* Π *has* $|\Pi| < \infty$*, and we have* $\mathcal{R} = [0, 1]$.

This assumption only serves to facilitate the use of the minimax theorem, and we expect that our results can be generalized (e.g., with covering numbers as in Section 3.4 of Foster et al. [17]).

2.1 Upper Bound

In this section we give regret bounds for adversarial decision making based on the (convexified) Decision-Estimation Coefficient. A-priori, it is not obvious why the DEC should bear any relevance to the adversarial setting we consider: The algorithms and regret bounds based on the DEC that Foster et al. [17] introduce for the stochastic setting heavily rely on the ability to estimate a static underlying model, yet in the adversarial setting, the learner may only interact with each model a single time. This renders any sort of global estimation (e.g., for dynamics of an MDP) impossible. In spite of this difficulty, we show that regret bounds can be achieved by building on the *Exploration-by-Optimization* technique of Lattimore and Szepesvári [35], Lattimore and György [32], which provides an elegant approach to estimating rewards that exploits the structure of the model class under consideration.

Exploration-by-Optimization—introduced by Lattimore and Szepesvári [35] and substantially expanded in Lattimore and György [32]—can be thought of as a generalization of the classical EXP3 algorithm [4] for finite-action bandits, which applies the exponential weights method for full-information online learning to a sequence of unbiased importance-weighted estimators for rewards. While EXP3 is near-optimal for bandits, it is unsuitable for general model classes because the reward estimators the algorithm uses do not exploit the structure of the decision space. Consequently, the regret scales linearly with $|\Pi|$ rather than with, e.g., dimension, as one might hope for problems like linear bandits. The idea behind Exploration-by-Optimization is to solve an optimization problem at each round to search for a (potentially biased) reward estimator and modified sampling distribution that better exploit the structure of the model class $\mathcal{M}$, leading to information sharing and improved regret. Lattimore and György [32] showed that for a general partial monitoring setting (cf. Appendix A),

Algorithm 1 High-Probability Exploration-by-Optimization (ExO^+)

1: **parameters**: Learning rate $\eta > 0$.

2: **for** $t = 1, 2, \cdots, T$ **do**

3: Define $q^{(t)} \in \Delta(\Pi)$ via exponential weights update:

$$q^{(t)}(\pi) = \frac{\exp\left(\eta \sum_{i=1}^{t-1} \widehat{f}^{(i)}(\pi)\right)}{\sum_{\pi' \in \Pi} \exp\left(\eta \sum_{i=1}^{t-1} \widehat{f}^{(i)}(\pi')\right)}. \tag{5}$$

4: Solve *high-probability exploration-by-optimization* objective: // See Eq. (8)

$$(p^{(t)}, g^{(t)}) \leftarrow \operatorname*{arg\,min}_{p \in \Delta(\Pi), g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^\star \in \Pi} \Gamma_{q^{(t)},\eta}(p, g\,; \pi^\star, M). \tag{6}$$

5: Sample decision $\pi^{(t)} \sim p^{(t)}$ and observe $z^{(t)} = (r^{(t)}, o^{(t)})$.

6: Form reward estimator:

$$\widehat{f}^{(t)}(\pi) = \frac{g^{(t)}(\pi; \pi^{(t)}, z^{(t)})}{p^{(t)}(\pi^{(t)})}. \tag{7}$$

the expected regret for this method—and for more general family of algorithms based on Bregman divergences—is bounded by a generalization of the Information Ratio of Russo and Van Roy [46, 47].

Our development builds on that of Lattimore and György [32], but we pursue *high-probability* guarantees rather than in-expectation guarantees. This allows us to provide regret bounds that hold for *adaptive adversaries*, rather than oblivious adversaries as considered in prior work.[4] Beyond this basic motivation, our interest in high-probability guarantees comes from the lower bound in the sequel (Section 2.2), which shows that the convexified Decision-Estimation Coefficient lower bounds regret for algorithms with "reasonable" tail behavior. To develop high-probability regret bounds and complement this lower bound, we use a novel variant of the Exploration-by-Optimization objective and a specialized analysis that goes beyond the Bregman divergence framework.

Our algorithm, ExO^+, is displayed in Algorithm 1. At each round t, the algorithm computes a *reference distribution* $q^{(t)} \in \Delta(\Pi)$ by applying the standard exponential weights update (with learning rate $\eta > 0$) to a sequence of reward estimators $\widehat{f}^{(1)}, \ldots, \widehat{f}^{(t-1)}$ from previous rounds (Line 3). For the main step (Line 4), the algorithm obtains a *sampling distribution* $p^{(t)} \in \Delta(\Pi)$ and an *estimation function* $g^{(t)} \in \mathcal{G} := (\Pi \times \Pi \times \mathcal{Z} \to \mathbb{R})$ by solving a minimax optimization problem based on a new objective we term *high-probability exploration-by-optmization*: Defining

$$\begin{aligned}\Gamma_{q,\eta}(p, g\,; \pi^\star, M) := \; & \mathbb{E}_{\pi \sim p}[f^M(\pi^\star) - f^M(\pi)] \\ & + \frac{1}{\eta} \cdot \mathbb{E}_{\pi \sim p, z \sim M(\pi)} \mathbb{E}_{\pi' \sim q}\left[\exp\left(\frac{\eta}{p(\pi)}(g(\pi'; \pi, z) - g(\pi^\star; \pi, z))\right) - 1\right],\end{aligned} \tag{8}$$

we solve

$$(p^{(t)}, g^{(t)}) \leftarrow \operatorname*{arg\,min}_{p \in \Delta(\Pi), g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^\star \in \Pi} \Gamma_{q^{(t)},\eta}(p, g\,; \pi^\star, M). \tag{9}$$

Finally (Lines 5 and 6), the algorithm samples $\pi^{(t)} \sim p^{(t)}$, observes $z^{(t)} = (r^{(t)}, o^{(t)})$, and then forms an importance-weighted reward estimator via $\widehat{f}^{(t)}(\pi) := g^{(t)}(\pi; \pi^{(t)}, z^{(t)}) / p^{(t)}(\pi^{(t)})$.

The interpretation of the high-probability Exploration-by-Optimization objective (8) is as follows: For a given round t, the model $M \in \mathcal{M}$ and decision $\pi^\star \in \Pi$ should be thought of as a proxy for the true model $M^{(t)}$ and optimal decision, respectively. By solving the minimax problem in (9), the min-player aims to—in the face of an unknown, worst-case model—find a sampling distribution that minimizes instantaneous regret, yet ensures good tail behavior for the importance-weighted estimator $g(\cdot; \pi, z)/p(\pi)$. Tail behavior is captured by the moment generating function-like term in (8), which penalizes the learner for over-estimating rewards under the reference distribution q or under-estimating rewards under $\pi^\star$.

[4] In general, in-expectation regret bounds do not imply high-probability bounds. For example, in adversarial bandits, the EXP3 algorithm can experience linear regret with constant probability [34].

We show that this approach leads to a bound on regret that scales with the convexified DEC.

Theorem 2.1 (Main upper bound)**.** *For any choice of $\eta > 0$, Algorithm 1 ensures that for all $\delta > 0$, with probability at least $1 - \delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \mathsf{dec}_{1/8\eta}(\mathrm{co}(\mathcal{M})) \cdot T + \frac{2}{\eta} \cdot \log(|\Pi|/\delta). \tag{10}$$

In particular, for any $\delta > 0$, with appropriate η, the algorithm ensures that with probability at least $1 - \delta$,

$$\mathbf{Reg}_{\mathsf{DM}} \leq O(1) \cdot \inf_{\gamma > 0} \{\mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot \log(|\Pi|/\delta)\}. \tag{11}$$

This regret bound holds for arbitrary, potentially adaptive adversaries. The result should be compared to the upper bound for the stochastic setting in Foster et al. [17] (e.g., Theorem 3.3), which takes a similar form, but scales with the weaker quantity $\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$.[5] See Appendix A for comparison to Lattimore and Szepesvári [35], Lattimore and György [32].

Equivalence of Exploration-by-Optimization and Decision-Estimation Coefficient. We now discuss a deeper connection between Exploration-by-Optimization and the DEC. Define the minimax value of the high-probability Exploration-by-Optimization objective via

$$\mathsf{exo}_{\eta}(\mathcal{M}, q) := \inf_{p \in \Delta(\Pi), g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^{\star} \in \Pi} \Gamma_{q,\eta}(p, g\,; \pi^{\star}, M), \tag{12}$$

and let $\mathsf{exo}_{\eta}(\mathcal{M}) := \sup_{q \in \Delta(\Pi)} \mathsf{exo}_{\eta}(\mathcal{M}, q)$. This quantity can be interpreted as a complexity measure for $\mathcal{M}$ whose, value reflects the difficulty of exploration. The following structural result (Corollary 3.1 in Section 3), which is critical to the proof of Theorem 2.1, shows that this complexity measure is equivalent to the convexified Decision-Estimation Coefficient:

$$\mathsf{dec}_{(4\eta)^{-1}}(\mathrm{co}(\mathcal{M})) \leq \mathsf{exo}_{\eta}(\mathcal{M}) \leq \mathsf{dec}_{(8\eta)^{-1}}(\mathrm{co}(\mathcal{M})), \quad \forall \eta > 0. \tag{13}$$

As we show, the regret of Algorithm 1 is controlled by the value of $\mathsf{exo}_{\eta}(\mathcal{M})$, and thus Theorem 2.1 follows. In the process of proving (13), we also establish equivalence of the Exploration-by-Optimization objective and a *parameterized* version of the Information Ratio, which is of independent interest (cf. Section 3). Both results build on, but go beyond the Bregman divergence-based framework in Lattimore and György [32], and exploit a somewhat obscure connection between Hellinger distance and the moment generating function (MGF) for the logarithmic loss. In particular, we use a technical lemma (proven in Appendix C), which shows that up to constants, the Hellinger distance between two probability distributions can be expressed as variational problem based on the associated MGFs.

Lemma 2.1. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\mathcal{X}, \mathscr{F})$. Then*

$$\frac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) \leq \sup_{g : \mathcal{X} \to \mathbb{R}} \left\{1 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right]\right\} \leq D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}). \tag{14}$$

The lower inequality in Lemma 2.1 is proven using a trick similar to one used by Zhang [55] to prove high-probability bounds for maximum likelihood estimation based on Hellinger distance. To prove the upper bound in (13), we apply the lower inequality in (14) with the test function g taking the role of the estimation function in the Exploration-by-Optimization objective.

Further remarks. The main focus of this work is statistical complexity (in particular, minimax regret), and the runtime and memory requirements of Algorithm 1, which are linear in $|\Pi|$, are not practical for large decision spaces. Improving the computational efficiency is an interesting question for future work. We mention in passing that Theorem 2.1 answers a question raised by Foster et al. [17] of obtaining in the frequentist setting a regret bound matching the Bayesian regret bound in their Theorem 3.6.

[5] If a proper estimation algorithm (i.e., an algorithm producing estimators that lie in $\mathcal{M}$) is available, Foster et al. [17] (Theorem 4.1) gives tighter bounds scaling with $\mathsf{dec}_{\gamma}(\mathcal{M})$.

2.2 Lower Bound

We now complement the regret bound in the prequel with a lower bound based on the convexified DEC. Our most general result shows that for any algorithm, either the expected regret or its (one-sided) second moment must scale with a localized version of the convexified DEC.

To state the result, we define the *localized model class* around a model $\overline{M}$ via

$$\mathcal{M}_\varepsilon(\overline{M}) = \{M \in \mathcal{M} : f^{\overline{M}}(\pi_{\overline{M}}) \geq f^M(\pi_M) - \varepsilon\},$$

and define $\mathsf{dec}_{\gamma,\varepsilon}(\mathcal{M}) := \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M})$ as the *localized Decision-Estimation Coefficient*. We let $(x)_+ := \max\{x, 0\}$ and define $V(\mathcal{M}) := \sup_{M,M'\in\mathcal{M}} \sup_{\pi\in\Pi} \sup_{A\in\mathscr{R}\otimes\mathscr{O}} \{\frac{M(A|\pi)}{M'(A|\pi)}\} \vee e$;[6] finiteness of $V(\mathcal{M})$ is not necessary, but removes a $\log(T)$ factor from Theorem 2.2.

Theorem 2.2 (Main lower bound)**.** *Let $C(T) := c\cdot\log(T \wedge V(\mathcal{M}))$ for a sufficiently large numerical constant $c > 0$. Set $\varepsilon_\gamma := \frac{\gamma}{4C(T)T}$. For any algorithm, there exists an oblivious adversary for which*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{E}(\mathbf{Reg}_{\mathsf{DM}})_+^2} \geq \Omega(1) \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T - O(T^{1/2}). \tag{15}$$

Theorem 2.2 implies that for any algorithm (such as Algorithm 1) with tail behavior beyond what is granted by control of the first moment, the regret in Theorem 2.1 cannot be substantially improved. In more detail, consider the notion of a *sub-Chebychev* algorithm.

Definition 2.1 (Sub-Chebychev Algorithm)**.** *A regret minimization algorithm is said to be sub-Chebychev with parameter R if for all $t > 0$,*

$$\mathbb{P}((\mathbf{Reg}_{\mathsf{DM}})_+ \geq t) \leq R^2/t^2. \tag{16}$$

For sub-Chebychev algorithms, both the mean and (root) second moment of regret are bounded by the parameter R (cf. Appendix F.4), which has the following consequence.

Corollary 2.1. Any regret minimization algorithm with sub-Chebychev parameter $R > 0$ must have

$$R \geq \widetilde{\Omega}(1) \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T - O(T^{1/2}). \tag{17}$$

To interpret this result, suppose for simplicity that $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))$ and $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M}))$ are continuous with respect to $\gamma > 0$, and that $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \gtrsim \gamma^{-1}$, which is satisfied for non-trivial classes.[7] In this case, it follows from Theorem 2.1 (cf. Proposition F.2 for details) that by setting $\delta = 1/T^2$, Algorithm 1 is sub-Chebychev with parameter

$$R = \widetilde{O}\Big(\inf_{\gamma>0}\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot \log(|\Pi|)\}\Big) = \widetilde{O}(\mathsf{dec}_{\gamma_{\mathrm{u}}}(\mathrm{co}(\mathcal{M})) \cdot T), \tag{18}$$

where γ_{u} satisfies the balance $\mathsf{dec}_{\gamma_{\mathrm{u}}}(\mathrm{co}(\mathcal{M})) \propto \frac{\gamma_{\mathrm{u}}}{T}\log|\Pi|$. On the other hand, the lower bound in (17) can be shown to scale with

$$R \geq \widetilde{\Omega}\Big(\mathsf{dec}_{\gamma_\ell,\varepsilon_{\gamma_\ell}}(\mathrm{co}(\mathcal{M})) \cdot T\Big), \tag{19}$$

where γ_ℓ satisfies the balance $\mathsf{dec}_{\gamma_\ell,\varepsilon_{\gamma_\ell}}(\mathrm{co}(\mathcal{M})) \propto \frac{\gamma_\ell}{T}$. We conclude that the upper bound from Theorem 2.1 cannot be improved beyond (i) localization and (ii) dependence on $\log|\Pi|$.

As an example, we show in Appendix D.3 that for the multi-armed bandit problem with $\Pi = \{1, \ldots, A\}$, the upper bound in (18) yields $R = \widetilde{O}(\sqrt{AT\log A})$, while the lower bound in (19) yields $R = \Omega(\sqrt{AT})$. See Appendix D for additional examples which further illustrate the scaling in the upper and lower bounds.

[6] Recall (Appendix B) that $M(\cdot, \cdot \mid \pi)$ is the conditional distribution given π.

[7] The dominant term $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T$ in (15) scales with $T^{1/2}$ for any class that is non-trivial in the sense that it embeds the two-armed bandit problem, so that the $-O(T^{1/2})$ term can be discarded.

The dependence on $\log|\Pi|$ cannot be removed from the upper bound or made to appear in the lower bound in general (cf. Section 3.5 of Foster et al. [17]). As shown in Foster et al. [17], localization is inconsequential for most model classes commonly studied in the literature. The same is true for the examples we consider here (Appendix D), where Theorem 2.2 leads to the correct rate up to small polynomial factors. However, improving the upper bound to achieve localization, which Foster et al. [17] show is possible in the stochastic setting, is an interesting future direction.

See Appendix A for further discussion and for comparison to a related lower bound in Lattimore [31].

Why convexity? At this point, a natural question is *why* the convex hull $\mathrm{co}(\mathcal{M})$ plays a fundamental role in the adversarial setting. For the lower bound, the intuition is simple: Given a model class $\mathcal{M}$, the adversary can pick any mixture distribution $\mu \in \Delta(\mathcal{M})$, then choose the sequence of models $M^{(1)}, \ldots, M^{(T)}$ by sampling $M^{(t)} \sim \mu$ independently at each round. This is equivalent to playing a static mixture model $M^\star = \mathbb{E}_{M\sim\mu}[M] \in \mathrm{co}(\mathcal{M})$, which is what allows us to prove a lower bound based on the DEC for the set $\mathrm{co}(\mathcal{M})$ of all such models. In view of the fact that the lower bound is obtained through this static (and stochastic) adversary, we believe the more surprising result here is that good behavior of the convexified DEC is also *sufficient* for low regret for fully adversarial decision making.

2.3 Learnability and Comparison to Stochastic Setting

Building on the upper and lower bounds in the prequel, we give a characterization for *learnability* (i.e., when non-trivial regret is possible) for adversarial decision making. This extends the learnability characterization for the stochastic setting in Foster et al. [17], and follows a long tradition in learning theory [52, 2, 49, 44, 11]. To state the result, we define the minimax regret for model class $\mathcal{M}$ as

$$\mathfrak{M}(\mathcal{M}, T) = \inf_{\boldsymbol{p}^{(1)},\ldots,\boldsymbol{p}^{(T)}} \sup_{\boldsymbol{M}^{(1)},\ldots,\boldsymbol{M}^{(T)}} \mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}],$$

where $\boldsymbol{p}^{(t)} : (\Pi \times \mathcal{Z})^{t-1} \to \Delta(\Pi)$ and $\boldsymbol{M}^{(t)} : (\Pi \times \mathcal{Z})^{t-1} \to \mathcal{M}$ are policies for the learner and adversary, respectively. Our characterization is as follows.

Theorem 2.3. *Suppose there exists $M_0 \in \mathcal{M}$ such that f^{M_0} is a constant function, and that $|\Pi| < \infty$.*

1. *If there exists $\rho > 0$ such that $\lim_{\gamma\to\infty} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot \gamma^\rho = 0$, then $\lim_{T\to\infty} \frac{\mathfrak{M}(\mathcal{M},T)}{T^p} = 0$ for $p < 1$.*
2. *If $\lim_{\gamma\to\infty} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot \gamma^\rho > 0$ for all $\rho > 0$, then $\lim_{T\to\infty} \frac{\mathfrak{M}(\mathcal{M},T)}{T^p} = \infty$ for all $p < 1$.*

The same conclusion holds when $\Pi = \Pi_T$ grows with T, but has $\log|\Pi_T| = O(T^q)$ for any $q < 1$.[8]

Theorem 2.3 shows that polynomial decay of the convexified DEC is necessary and sufficient for low regret. We emphasize that this result is complementary to Theorem 2.2, and does not require localization or any assumption on the tail behavior of the algorithm. This is a consequence of the coarse, asymptotic nature of the result, which allows us the use of rescaling arguments to remove these conditions.

Comparison to stochastic setting. Having shown that the convexified Decision-Estimation Coefficient leads to upper and lower bounds on the optimal regret for the adversarial DMSO setting, we now contrast with the stochastic setting. There, Foster et al. [17] obtain upper bounds on regret that have the same form as (11), but scale with the weaker quantity $\max_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$.[9] For classes that are not convex, but where "proper" estimators are available (including finite-state/action MDPs), the upper bounds in Foster et al. [17] can further be improved to scale with $\mathsf{dec}_\gamma(\mathcal{M})$. Hence, our results show that in general, the price of adversarial outcomes can be as large as $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))/\mathsf{dec}_\gamma(\mathcal{M})$. Examples (see Appendix D for details and more) include:

- For tabular (finite-state/action) MDPs with horizon H, S states, and A actions, Foster et al. [17] show that $\mathsf{dec}_\gamma(\mathcal{M}) \leq \mathrm{poly}(H, S, A)/\gamma$, and use this to obtain regret

[8] Allowing Π to grow with T is useful when considering infinite decision spaces, because it facilitates covering arguments.

[9] Theorem 3.1 of Foster et al. [17] attains $\mathbf{Reg}_{\mathsf{DM}} \lesssim \inf_{\gamma>0}\{\max_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) + \gamma \cdot \log|\mathcal{M}|\}$ with high probability.

$\sqrt{\mathrm{poly}(H,S,A)\cdot T}$. Tabular MDPs are *not* a convex class, and $\mathrm{co}(\mathcal{M})$ is equivalent to the class of so-called *latent MDPs*, which are known to be intractable [28, 37]. Indeed, we show (Appendix D) that $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \geq \Omega(A^{\min\{S,H\}})$, which implies an exponential lower bound on regret through Theorem 2.2. This highlights that in general, the gap between stochastic and adversarial outcomes can be quite large.

- For many common bandit problems, one has $\mathrm{co}(\mathcal{M}) = \mathcal{M}$, leading to polynomial bounds on regret in the adversarial setting. For example, the multi-armed bandit problem with A actions has $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \leq O(A/\gamma)$, leading to $\sqrt{AT\log A}$ regret (via Theorem 2.1), and the linear bandit problem in d dimensions has $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \leq O(d/\gamma)$, leading to regret $\sqrt{dT\log|\Pi|}$.

3 Connections Between Complexity Measures

The Decision-Estimation Coefficient bears a resemblance to the *generalized Information Ratio* introduced by Lattimore and György [32], Lattimore [30] which extends the original Information Ratio of Russo and Van Roy [46, 47]. In what follows, we establish deeper connections between these complexity measures. All of the results in this section are proven in Appendix E.

Let us call any function $D(\cdot \parallel \cdot) : \Delta(\Pi) \times \Delta(\Pi) \to \mathbb{R}^+$ a *divergence-like function*. We restate the generalized Information Ratio from Lattimore [31]. For a distribution $\mu \in \Delta(\mathcal{M} \times \Pi)$ and decision distribution $p \in \Delta(\Pi)$, let $\mathbb{P}$ be the law of the process $(M, \pi^\star) \sim \mu, \pi \sim p, z \sim M(\pi)$, and define $\mu_{\mathrm{pr}}(\pi') := \mathbb{P}(\pi^\star = \pi')$ and $\mu_{\mathrm{po}}(\pi'; \pi, z) := \mathbb{P}(\pi^\star = \pi' \mid (\pi, z))$. The distribution μ_{pr} should be thought of as the prior over $\pi^\star$, and μ_{po} should be thought of as the posterior over $\pi^\star$ after observing (z, π); note that the law μ_{po} does not depend on the distribution p. For parameter $\lambda > 1$, Lattimore [31] defines the generalized Information Ratio for a class $\mathcal{M}$ via[10]

$$\Psi_\lambda(\mathcal{M}) = \sup_{\mu \in \Delta(\mathcal{M}\times\Pi)} \inf_{p \in \Delta(\Pi)} \left\{ \frac{(\mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)])^\lambda}{\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})]} \right\}. \tag{20}$$

Here, we have slightly generalized the original definition in Lattimore [31] by incorporating models in $\mathcal{M}$ rather than placing an arbitrary prior over observations z directly. We also use a general divergence-like function, while Lattimore [31] uses KL divergence and Lattimore and György [32] use Bregman divergences.

To understand the connection to the Decision-Estimation Coefficient, it will be helpful to introduce another variant of the Information Ratio, which we call the *parameterized Information Ratio*.

Definition 3.1. *For a divergence-like function $D(\cdot \parallel \cdot) : \Delta(\Pi) \times \Delta(\Pi) \to \mathbb{R}_+$, the* parameterized Information Ratio *is given by*

$$\begin{aligned} &\mathsf{inf}^D_\gamma(\mathcal{M}) \\ &= \sup_{\mu \in \Delta(\mathcal{M}\times\Pi)} \inf_{p \in \Delta(\Pi)} \mathbb{E}_{\pi\sim p}\big[\mathbb{E}_{(M,\pi^\star)\sim\mu}[f^M(\pi^\star) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})]\big]. \end{aligned} \tag{21}$$

The parameterized Information Ratio is always bounded by the generalized Information Ratio in (20); in particular, we have $\mathsf{inf}^D_\gamma(\mathcal{M}) \leq (\Psi_\lambda(\mathcal{M})/\gamma)^{\frac{1}{\lambda-1}}\ \forall \gamma > 0$. All regret bounds based on the generalized Information Ratio that we are aware of [32, 31] implicitly bound regret by the parameterized Information Ratio, and then invoke the inequality above to move to the generalized Information Ratio. In general though, it does not appear that these notions are equivalent. Informally, this is because the notion in (20) is equivalent to requiring that a single distribution p certify a certain bound on the value in (21) for all values of the parameter γ simultaneously, while the parameterized Information Ratio allows the distribution p to vary as a function of $\gamma > 0$ (hence the name); see also Appendix E.

Letting $\mathsf{inf}^{\mathsf{H}}_\gamma(\mathcal{M})$ denote the parameterized Information Ratio with $D = D^2_{\mathsf{H}}(\cdot,\cdot)$, we show that this notion is equivalent to the convexified Decision-Estimation Coefficient.

Theorem 3.1. *For all $\gamma > 0$, $\mathsf{inf}^{\mathsf{H}}_\gamma(\mathcal{M}) \leq \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \leq \mathsf{inf}^{\mathsf{H}}_{\gamma/4}(\mathcal{M})$.*

[10] Lattimore and György [32] give a slightly different but essentially equivalent definition; cf. Appendix E.

This result is a special case of Theorem E.1 in Appendix E, which shows that a similar equivalence holds for a class of "well-behaved" f-divergences that includes KL divergence, but not necessarily for general Bregman divergences. The idea is to use Bayes' law to move from the Decision-Estimation Coefficient, which considers distance between *distributions over observations*, to the Information Ratio, which considers distance between *distributions over decisions*.

In light of this characterization, the results in this paper could have equivalently been presented in terms of the parameterized Information Ratio. We chose to present them in terms of the Decision-Estimation Coefficient in order to draw parallels to the stochastic setting, where guarantees that scale with $\mathsf{dec}_\gamma(\mathcal{M})$ (without convexification) are available. It is unclear whether the Information Ratio can accurately reflect the complexity for both stochastic and adversarial settings in the same fashion, because—unlike the DEC—it is invariant under convexification, as the following proposition shows.[11]

Proposition 3.1. *For any divergence-like function $D(\cdot \parallel \cdot)$, we have*

$$\mathsf{inf}^D_\gamma(\mathcal{M}) = \mathsf{inf}^D_\gamma(\mathrm{co}(\mathcal{M})), \quad \forall \gamma > 0.$$

For a final structural result, we show that up to constants, the parameterized Information Ratio is equivalent to the high-probability Exploration-by-Optimization objective.

Theorem 3.2. *For all $\eta > 0$,*

$$\mathsf{inf}^{\mathsf{H}}_{\eta^{-1}}(\mathcal{M}) \leq \mathsf{exo}_\eta(\mathcal{M}) \leq \mathsf{inf}^{\mathsf{H}}_{(8\eta)^{-1}}(\mathcal{M})$$

This result is proven through a direct argument (cf. Section 2.1), and the equivalence of the DEC and Exploration-by-Optimization in (13) is proven by combining with Theorem 3.1. Summarizing the equivalence:

Corollary 3.1. For all $\eta > 0$,

$$\mathsf{dec}_{(4\eta)^{-1}}(\mathrm{co}(\mathcal{M})) \leq \mathsf{inf}^{\mathsf{H}}_{\eta^{-1}}(\mathcal{M}) \leq \mathsf{exo}_\eta(\mathcal{M}) \leq \mathsf{inf}^{\mathsf{H}}_{(8\eta)^{-1}}(\mathcal{M}) \leq \mathsf{dec}_{(8\eta)^{-1}}(\mathrm{co}(\mathcal{M})).$$

Since this equivalence depends of the value of the parameter $\gamma > 0$ in the parameterized Information Ratio, it is not clear whether a similar equivalence can be established using the generalized Information Ratio in (20). We note in passing that one can use similar arguments to lower bound the Bregman divergence-based Exploration-by-Optimization objective in Lattimore and György [32] by the parameterized Information Ratio for the Bregman divergence of interest, complementing their upper bound.

4 Discussion

We have shown that the convexified Decision-Estimation Coefficient is necessary and sufficient to achieve low regret for adversarial interactive decision making, establishing that convexity governs the price of adversarial outcomes. Our results elucidate the relationship between the DEC, Exploration-by-Optimization, and the Information Ratio, and we hope they will find broader use.

This work adds to a growing body of research which shows that online reinforcement learning with agnostic or adversarial outcomes can be statistically intractable [48, 37]. A promising future direction is to extend our techniques to natural semi-adversarial models in which reinforcement learning is tractable. Another interesting direction is to address the issue of computational efficiency for large decision spaces.

Acknowledgements

We thank Zak Mhammedi for useful comments and feedback. AR acknowledges support from the ONR through awards N00014-20-1-2336 and N00014-20-1-2394, from the ARO through award W911NF-21-1-0328, and from the DOE through award DE-SC0022199. Part of this work was completed while DF was visiting the Simons Institute for the Theory of Computing. KS acknowledges support from NSF CAREER Award 1750575.

[11] The variants in Lattimore and György [32], Lattimore [31] are also invariant under convexification.

References

[1] J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proc. of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

[3] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[5] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[6] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

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

[8] S. Bubeck, Y. T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

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

[10] V. Dani, T. P. Hayes, and S. Kakade. The price of bandit information for online optimization. 2007.

[11] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the ERM principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 207–232, 2011.

[12] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[13] S. Dong, B. Van Roy, and Z. Zhou. Provably efficient reinforcement learning with aggregated states. *arXiv preprint arXiv:1912.06366*, 2019.

[14] S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudik, and J. Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

[15] S. S. Du, S. M. Kakade, J. D. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in RL. *International Conference on Machine Learning*, 2021.

[16] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

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

[18] E. Hazan and S. Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

[19] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pages 1704–1713, 2017.

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

[21] C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Neural Information Processing Systems*, 2021.

[22] T. Jin and H. Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. *Advances in neural information processing systems*, 33:16557–16566, 2020.

[23] J. Kirschner. *Information-Directed Sampling-Frequentist Analysis and Applications*. PhD thesis, ETH Zurich, 2021.

[24] J. Kirschner, T. Lattimore, and A. Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369. PMLR, 2020.

[25] J. Kirschner, T. Lattimore, C. Vernade, and C. Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

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

[27] A. Krishnamurthy, A. Agarwal, and J. Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

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

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

[30] T. Lattimore. Minimax regret for bandit convex optimisation of ridge functions. *arXiv preprint arXiv:2106.00444*, 2021.

[31] T. Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. *arXiv preprint arXiv:2202.10997*, 2022.

[32] T. Lattimore and A. György. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

[33] T. Lattimore and C. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pages 2111–2139. PMLR, 2019.

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

[35] T. Lattimore and C. Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pages 2488–2515. PMLR, 2020.

[36] L. Li. *A unifying framework for computational reinforcement learning theory*. Rutgers, The State University of New Jersey—New Brunswick, 2009.

[37] Q. Liu, Y. Wang, and C. Jin. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. *arXiv preprint arXiv:2203.06803*, 2022.

[38] S. Magureanu, R. Combes, and A. Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR, 2014.

[39] A. Modi, N. Jiang, A. Tewari, and S. Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[40] G. Neu, A. György, C. Szepesvári, et al. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pages 231–243. Citeseer, 2010.

[41] G. Neu, A. György, C. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, 2014.

[42] Y. Polyanskiy. Information theoretic methods in statistics and computer science. 2020. URL https://people.lids.mit.edu/yp/homepage/sdpi_course.html.

[43] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. 2014.

[44] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems 23*, pages 1984–1992, 2010.

[45] D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

[46] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[47] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[48] A. Sekhari, C. Dann, M. Mohri, Y. Mansour, and K. Sridharan. Agnostic reinforcement learning with low-rank MDPs and rich observations. *Advances in Neural Information Processing Systems*, 34, 2021.

[49] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability, and uniform convergence. *Journal of Machine Learning Research (JMLR)*, 2010.

[50] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958.

[51] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. 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.

[52] V. N. Vapnik. The nature of statistical learning theory. 1995.

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

[54] L. Yang and M. Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[55] T. Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

[56] D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[57] A. Zimin and G. Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents of Appendix

A Detailed Discussion of Related Work **16**

B Preliminaries **17**

C Technical Tools **18**

C.1 Tail Bounds 18

C.2 Minimax Theorem 18

C.3 Information Theory 18

C.4 Online Learning 20

D Examples **20**

D.1 Structured Bandits 20

D.2 Reinforcement Learning 22

D.3 Proofs for Examples 22

E Structural Results **27**

E.1 Background on Complexity Measures 27

E.2 Decision-Estimation Coefficient and Information Ratio 28

E.3 High-Probability Exploration-By-Optimization and Information Ratio 30

F Proofs for Main Results (Section 2) **33**

F.1 Proof of Theorem 2.1 33

F.2 Proof of Theorem 2.2 35

F.3 Proof of Theorem 2.3 39

F.4 Sub-Chebychev Algorithms 40

A Detailed Discussion of Related Work

Beyond Foster et al. [17], which was the starting point for this work, our results build on a long line of research on partial monitoring and the Information Ratio [46, 47, 33, 29, 30, 32, 31, 23, 24, 25]; most closely related are the works the works of Lattimore and György [32] and Lattimore [31]. Below we discuss and compare to these results in greater detail.

Comparison to partial monitoring setting. Lattimore and György [32], Lattimore [31] and other works in this sequence consider a general partial monitoring setting in which each outcome $z^{(t)}$ is directly chosen by an adversary, and needs not contain the reward signal. In terms of reward signal, our setting is more restrictive because we assume that $r^{(t)}$ is observed. In fact, our upper bounds encompass the more general setting where rewards are not observed by the learner (thus subsuming the partial monitoring problem) but our lower bounds that require that rewards are observed. In terms of data generation process, our setting is more general because we restrict to models in a known class in $\mathcal{M}$. This setup recovers the case where $z^{(t)}$ is fully adversarial because by considering the special case where $\mathcal{M}$ consists of point masses over $\mathcal{Z}$. However, our framework also allows for semi-stochastic adversaries, and for problems like (structured) adversarial MDPs. For example, if we constrain all models in $\mathcal{M}$ place ε probability mass on a particular outcome z, any adversary in our framework must place ε mass on this outcome as well.

Upper bounds. Our upper bounds build on the Exploration-by-Optimization technique, which was introduced in Lattimore and Szepesvári [35] and generalized significantly by Lattimore and György [32]. The latter result shows that for a general family of mirror descent-based Exploration-by-Optimization algorithms parameterized by Bregman divergences, the regret can be bounded by the generalized Information Ratio for the associated Bregman divergence (cf. Appendix E). This approach yields bounds on pseudoregret with a similar form to Theorem 2.1 (with $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))$ replaced by the generalized Information Ratio), but does appear to give regret bounds for adaptive adversaries, or give high-probability guarantees. Note that in general, in-expectation regret bounds do not imply high-probability regret bounds (for example, even for multi-armed bandits, the EXP3 algorithm can experience linear regret with constant probability [34]). Likewise, bounds on pseudoregret do not translate to bounds on (expected) true regret unless one restricts to oblivious adversaries. To develop high-probability regret bounds that complement our lower bounds, we depart from the Bregman divergence-based framework and exploit refined properties of Hellinger distance. We note that the work of Lattimore and Szepesvári [35] also proposes a high-probability Exploration-by-Optimization variant, but it is unclear whether this objective can be related to the information ratio or Decision-Estimation Coefficient for general models.[12]

Lower bounds. Our lower bounds refine the proof strategy from Foster et al. [17]. Our most important technical result is a lower bound for the stochastic setting (Theorem F.1), which improves upon Theorem 3.1 from Foster et al. [17] by using a more refined change of measure argument. In particular, Theorem 3.1 of Foster et al. [17] gives a lower bound based on the DEC that holds with low probability, and thus provides a meaningful converse only to algorithms with sub-Gaussian or sub-exponential tail behavior. Our lower bound gives a meaningful converse to any upper bound with sub-Chebychev tail behavior, which is a significantly weaker assumption. We note that while Theorem 3.2 of Foster et al. [17] provides lower bounds on expected regret without assumptions on tail behavior, this result requires a stronger notion of localization than we consider here. Of course, proving a lower bound on expected regret that matches our lower bound remains an interesting open problem.

Recent work of Lattimore [31], provides lower bounds on regret in a general partial monitoring setting based on a generalized Information Ratio (cf. Appendix E). This result is somewhat complementary to our lower bound (Theorem 2.2). On the positive side, Lattimore [31] give a lower bound on *expected regret* that is always tight in terms of dependence on T, while our result only leads to tight dependence on T if one restricts to sub-Chebychev algorithms. In addition, the lower bound applies to the general partial monitoring setting, while our lower bound requires that rewards are observed. However, the lower bound of Lattimore [31] is loose in $\mathrm{poly}(|\Pi|)$ factors, and thus does not lead to meaningful dependence on problem parameters such as dimension in the same fashion as our results.

[12] This objective is based on a Bernstein-type tail bound, which scales with the range of the estimation functions. We avoid explicit dependence on the range using a more specialized tail bound based on Lemma C.1.

An interesting question for future work is to investigate whether these techniques can be combined with our own to get the best of both worlds.

B Preliminaries

Basic notation. For a set $\mathcal{X}$, we let $\Delta(\mathcal{X})$ denote the set of all Radon probability measures over $\mathcal{X}$. We let $\mathrm{co}(\mathcal{X})$ denote the set of all finitely supported convex combinations of elements in $\mathcal{X}$. We use the shorthand $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$.

We adopt non-asymptotic big-oh notation: For functions $f, g : \mathcal{X} \to \mathbb{R}_+$, we write $f = O(g)$ (resp. $f = \Omega(g)$) if there exists a constant $C > 0$ such that $f(x) \le Cg(x)$ (resp. $f(x) \ge Cg(x)$) for all $x \in \mathcal{X}$. We write $f = \widetilde{O}(g)$ if $f = O(g \cdot \mathrm{polylog}(T))$, $f = \widetilde{\Omega}(g)$ if $f = \Omega(g/\mathrm{polylog}(T))$, and $f = \widetilde{\Theta}(g)$ if $f = \widetilde{O}(g)$ and $f = \widetilde{\Omega}(g)$. We write $f \propto g$ if $f = \widetilde{\Theta}(g)$.

Probability spaces. We formalize the probability spaces for the DMSO framework in the same fashion as Foster et al. [17], which we briefly summarize here. 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 up to time 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 $(\Omega^{(t)}, \mathscr{F}^{(t)})$.

Formally, a model $M = M(\cdot, \cdot \mid \cdot) \in \mathcal{M}$ is a probability kernel from $(\Pi, \mathscr{P})$ to $(\mathcal{R} \times \mathcal{O}, \mathscr{R} \otimes \mathscr{O})$; we use the convention $M(\pi) = M(\cdot, \cdot \mid \pi)$ throughout the paper.[13] An *algorithm* for horizon T is a sequence $p^{(1)}, \ldots, p^{(T)}$, where $p^{(t)}(\cdot \mid \cdot)$ is a probability kernel from $(\Omega^{(t-1)}, \mathscr{F}^{(t-1)})$ to $(\Pi, \mathscr{P})$.

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

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

Hellinger distance is defined as

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

and Kullback-Leibler divergence is defined as

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

For a convex function $f : (0, \infty) \to \mathbb{R}$, the associated f-divergence for measures $\mathbb{P}$ and $\mathbb{Q}$ with $\mathbb{P} \ll \mathbb{Q}$ is given by

$$D_f(\mathbb{P} \,\|\, \mathbb{Q}) := \mathbb{E}_{\mathbb{Q}} \left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}} \right) \right] \tag{22}$$

whenever $\mathbb{P} \ll \mathbb{Q}$. More generally, defining $p = \frac{d\mathbb{P}}{d\nu}$ and $q = \frac{d\mathbb{Q}}{d\nu}$ for a common dominating measure ν, we have

$$D_f(\mathbb{P} \,\|\, \mathbb{Q}) := \int_{q>0} q f\left(\frac{p}{q}\right) d\nu + \mathbb{P}(q = 0) \cdot f'(\infty), \tag{23}$$

where $f'(\infty) := \lim_{x \to 0^+} x f(1/x)$.

[13] For measurable spaces $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$ a probability kernel $P(\cdot \mid \cdot)$ from $(\mathcal{X}, \mathscr{X})$ to $(\mathcal{Y}, \mathscr{Y})$ has the property that (i) For all $x \in \mathcal{X}$, $P(\cdot \mid x)$ is a probability measure, (ii) for all $Y \in \mathscr{Y}$, $x \mapsto P(Y \mid x)$ is measurable.

C Technical Tools

C.1 Tail Bounds

Lemma C.1 (e.g., Lemma A.4 of Foster et al. [17])**.** *For any sequence of real-valued random variables $(X_t)_{t\le T}$ adapted to a filtration $(\mathscr{F}_t)_{t\le T}$, we have that with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \le \sum_{t=1}^{T} \log\big(\mathbb{E}\big[e^{X_t} \mid \mathscr{F}_{t-1}\big]\big) + \log(\delta^{-1}). \tag{24}$$

C.2 Minimax Theorem

Lemma C.2 (Sion's Minimax Theorem [50])**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be convex sets in linear topological spaces, and assume $\mathcal{X}$ is compact. Let $F : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ be such that (i) $F(x,\cdot)$ is concave and upper semicontinuous over $\mathcal{Y}$ for all $x\in\mathcal{X}$ and (ii) $F(\cdot,y)$ is convex and lower semicontinuous over $\mathcal{X}$ for all $y\in\mathcal{Y}$. Then*

$$\inf_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} F(x,y) = \sup_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} F(x,y). \tag{25}$$

C.3 Information Theory

C.3.1 Basic Results

Proposition C.1. *For any f-divergence $D_f(\cdot\,\|\,\cdot)$, one has that for any pair of random variables (X,Y) with joint law $\mathbb{P}_{X,Y}$,*

$$\mathbb{E}_{X\sim\mathbb{P}_X}\big[D_f\big(\mathbb{P}_{Y|X}\,\|\,\mathbb{P}_Y\big)\big] = \mathbb{E}_{Y\sim\mathbb{P}_Y}\big[D_f\big(\mathbb{P}_{X|Y}\,\|\,\mathbb{P}_X\big)\big].$$

Proof of Proposition C.1. Recalling that $D_f(\mathbb{P}\,\|\,\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\Big[f\Big(\frac{d\mathbb{P}}{d\mathbb{Q}}\Big)\Big]$ for $\mathbb{P}\ll\mathbb{Q}$, we have

$$\begin{aligned}
\mathbb{E}_{X\sim\mathbb{P}_X}\big[D_f\big(\mathbb{P}_{Y|X}\,\|\,\mathbb{P}_Y\big)\big] &= \mathbb{E}_{X\sim\mathbb{P}_X}\,\mathbb{E}_{Y\sim\mathbb{P}_Y}\Big[f\Big(\tfrac{d\mathbb{P}_{Y|X}}{d\mathbb{P}_Y}\Big)\Big]\\
&= \mathbb{E}_{X\sim\mathbb{P}_X}\,\mathbb{E}_{Y\sim\mathbb{P}_Y}\Big[f\Big(\tfrac{d\mathbb{P}_{X,Y}}{d(\mathbb{P}_X\otimes\mathbb{P}_Y)}\Big)\Big]\\
&= \mathbb{E}_{Y\sim\mathbb{P}_Y}\,\mathbb{E}_{X\sim\mathbb{P}_X}\Big[f\Big(\tfrac{d\mathbb{P}_{X|Y}}{d\mathbb{P}_X}\Big)\Big] = \mathbb{E}_{Y\sim\mathbb{P}_Y}\big[D_f\big(\mathbb{P}_{X|Y}\,\|\,\mathbb{P}_X\big)\big],
\end{aligned}$$

where we have used that $\mathbb{P}_{Y|X}\ll\mathbb{P}_Y$, $\mathbb{P}_{X|Y}\ll\mathbb{P}_X$, and $\mathbb{P}_{X,Y}\ll\mathbb{P}_X\otimes\mathbb{P}_Y$. □

C.3.2 Change of Measure

Lemma C.3 (Donsker-Varadhan (e.g., Polyanskiy and Wu [43]))**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\mathcal{X},\mathscr{F})$. Then*

$$D_{\mathsf{KL}}(\mathbb{P}\,\|\,\mathbb{Q}) = \sup_{h:\mathcal{X}\to\mathbb{R}}\{\mathbb{E}_{\mathbb{P}}[h(X)] - \log(\mathbb{E}_{\mathbb{Q}}[\exp(h(X))])\}. \tag{26}$$

Lemma C.4. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\mathcal{X},\mathscr{F})$. Then for all functions $h:\mathcal{X}\to\mathbb{R}$,*

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \le \sqrt{2^{-1}(\mathbb{E}_{\mathbb{P}}[h^2(X)] + \mathbb{E}_{\mathbb{Q}}[h^2(X)])\cdot D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q})}. \tag{27}$$

Proof of Lemma C.4. From Polyanskiy and Wu [43], it holds that for all functions $h:\mathcal{X}\to\mathbb{R}$, if $\mathbb{P}\ll\mathbb{Q}$,

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \le \sqrt{\mathbb{V}_{\mathbb{Q}}[h(X)]\cdot D_{\chi^2}(\mathbb{P}\,\|\,\mathbb{Q})} \le \sqrt{\mathbb{E}_{\mathbb{Q}}[h^2(X)]\cdot D_{\chi^2}(\mathbb{P}\,\|\,\mathbb{Q})}, \tag{28}$$

where $D_{\chi^2}(\mathbb{P}\,\|\,\mathbb{Q}) := \int\frac{(d\mathbb{P}-d\mathbb{Q})^2}{d\mathbb{Q}}$ and $\mathbb{V}_{\mathbb{Q}}$ denotes the variance under $\mathbb{Q}$. The result follows by using that $D_{\chi^2}\big(\mathbb{P}\,\|\,\frac{\mathbb{P}+\mathbb{Q}}{2}\big)\le D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q})$. □

Lemma 2.1. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\mathcal{X}, \mathscr{F})$. Then*

$$\frac{1}{2} D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) \leq \sup_{g:\mathcal{X}\to\mathbb{R}} \left\{1 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right]\right\} \leq D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}). \tag{14}$$

Proof of Lemma 2.1. We first show that Hellinger distance is lower bounded by the quantity in (14). Recall that Hellinger distance is the f-divergence associated with $f(x) = (1 - \sqrt{x})^2$ (cf. (23)). Let $f^\star(y) := \sup_{x\geq 0}\{xy - f(x)\}$ be the Fenchel dual of f, which has the form

$$f^\star(y) = \begin{cases} \frac{y}{1-y}, & y < 1, \\ \infty, & y \geq 1. \end{cases}$$

Using Theorem 7.14 of Polyanskiy [42], we express Hellinger distance as a variational problem based on the dual:

$$\begin{aligned} D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) &= \sup_{h:\mathcal{X}\to(-\infty,1)} \{\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[f^\star(h(X))]\} \\ &= \sup_{h:\mathcal{X}\to(-\infty,1)} \left\{\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}\left[\frac{h(X)}{1-h(X)}\right]\right\}. \end{aligned}$$

Reparameterizing via $h(X) = 1 - h'(X)$ for $h' : \mathcal{X} \to (0, \infty)$, this gives

$$D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) = \sup_{h:\mathcal{X}\to(0,\infty)} \left\{2 - \mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}\left[\frac{1}{h(X)}\right]\right\}.$$

To conclude, observe that for any test function $g : \mathcal{X} \to \mathbb{R}$, by setting $h(x) = e^{g(x)} \cdot \mathbb{E}_{\mathbb{Q}}[e^{-g}]$, we have

$$\begin{aligned} 2 - \mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}\left[\frac{1}{h(X)}\right] &= 2 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right] - \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right] / \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right] \\ &= 1 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right], \end{aligned}$$

so that

$$D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) \geq \sup_{g:\mathcal{X}\to\mathbb{R}} \left\{1 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right]\right\}.$$

We now prove the other direction of the inequality in (14). Let ν be a common dominating measure for $\mathbb{P}$ and $\mathbb{Q}$, and set $p = \frac{d\mathbb{P}}{d\nu}$ and $q = \frac{d\mathbb{Q}}{d\nu}$. We first consider the case where $p, q > 0$ everywhere. Set $g(x) = \frac{1}{2}\log(q(x)/p(x))$. Then we have $\mathbb{E}_{\mathbb{P}}\left[e^{g}\right] = \int \sqrt{pq}d\nu = 1 - \frac{1}{2}D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q})$, and likewise, $\mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right] = \int \sqrt{pq}d\nu = 1 - \frac{1}{2}D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q})$. As a result,

$$\sup_{g:\mathcal{X}\to\mathbb{R}} \left\{1 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right]\right\} \geq 1 - (1 - \tfrac{1}{2}D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}))^2 \geq \frac{1}{2}D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}),$$

where we have used that $D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) \in [0, 2]$. For the general case, we appeal to Lemma C.5 below and take $\varepsilon \to 0$. □

The following result is generalization of Lemma 2.1 which shows that up to small approximation error, the lower bound in (14) can be obtained using test functions with small magnitude.

Lemma C.5. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\mathcal{X}, \mathscr{F})$. Then for any $\alpha \geq 1$,*

$$\frac{1}{2} D_{\mathsf{H}}^2(\mathbb{P}, \mathbb{Q}) \leq \sup_{g\in\mathcal{G}_\alpha} \left\{1 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right]\right\} + 4e^{-\alpha}, \tag{29}$$

where $\mathcal{G}_\alpha := \{g : \mathcal{X} \to \mathbb{R} \mid \|g\|_\infty \leq \alpha\}$.

Proof of Lemma C.5. Fix $\alpha \geq 1$ and let $\varepsilon := e^{-2\alpha}$. Note that $\varepsilon \in (0, e^{-2})$. Given measures $\mathbb{P}$ and $\mathbb{Q}$, set $\mathbb{P}_\varepsilon = (1 - \varepsilon)\mathbb{P} + \varepsilon\mathbb{Q}$ and $\mathbb{Q}_\varepsilon = (1 - \varepsilon)\mathbb{Q} + \varepsilon\mathbb{P}$. Consider the test function $g = \frac{1}{2}\log(\frac{d\mathbb{Q}_\varepsilon}{d\mathbb{P}_\varepsilon})$, which has the following properties:

- $\|g\|_\infty \leq \frac{1}{2}\log\left(\frac{1-\varepsilon}{\varepsilon} + \frac{\varepsilon}{1-\varepsilon}\right) \leq \frac{1}{2}\log(\varepsilon^{-1})$, where we have used that $\varepsilon \leq 1/2$. This establishes that $g \in \mathcal{G}_\alpha$.

- $\mathbb{E}_{\mathbb{P}}\big[e^{g}\big] \leq (1-\varepsilon)^{-1/2} \int \sqrt{d\mathbb{P} d\mathbb{Q}_{\varepsilon}} = (1-\varepsilon)^{-1/2}\big(1 - \frac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}_{\varepsilon})\big)$.
- $\mathbb{E}_{\mathbb{Q}}\big[e^{-g}\big] \leq (1-\varepsilon)^{-1/2} \int \sqrt{d\mathbb{P}_{\varepsilon} d\mathbb{Q}} = (1-\varepsilon)^{-1/2}\big(1 - \frac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}_{\varepsilon}, \mathbb{Q})\big)$.

Using these bounds yields

$$\begin{aligned}\sup_{g:\mathcal{X}\to\mathbb{R}} \big\{1 - \mathbb{E}_{\mathbb{P}}\big[e^{g}\big] \cdot \mathbb{E}_{\mathbb{Q}}\big[e^{-g}\big]\big\} &\geq 1 - (1-\varepsilon)^{-1}(1 - \tfrac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}_{\varepsilon}, \mathbb{Q}))(1 - \tfrac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}_{\varepsilon})) \\ &\geq 1 - (1-\varepsilon)^{-1}(1 - \tfrac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}_{\varepsilon}, \mathbb{Q})) \\ &\geq (1-\varepsilon)^{-1} \cdot \frac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}_{\varepsilon}, \mathbb{Q}) - 2\varepsilon.\end{aligned}$$

Finally, we note that by the triangle inequality for Hellinger distance and convexity of squared Hellinger distance,

$$D_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) \leq D_{\mathsf{H}}(\mathbb{P}_{\varepsilon}, \mathbb{Q}) + D_{\mathsf{H}}(\mathbb{P}, \mathbb{P}_{\varepsilon}) \leq D_{\mathsf{H}}(\mathbb{P}_{\varepsilon}, \mathbb{Q}) + \varepsilon^{1/2} D_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}),$$

so that $D^2_{\mathsf{H}}(\mathbb{P}_{\varepsilon}, \mathbb{Q}) \geq (1-\varepsilon^{1/2})^2 D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q})$, and

$$\sup_{g:\mathcal{X}\to\mathbb{R}} \big\{1 - \mathbb{E}_{\mathbb{P}}\big[e^{g}\big] \cdot \mathbb{E}_{\mathbb{Q}}\big[e^{-g}\big]\big\} \geq \frac{(1-\varepsilon^{1/2})^2}{1-\varepsilon} \frac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) - 2\varepsilon \geq \frac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) - 4\varepsilon^{1/2},$$

where we have used that $\varepsilon \in (0,1)$ and $D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) \in [0, 2]$.

□

C.4 Online Learning

Lemma C.6 (e.g., Cesa-Bianchi and Lugosi [9])**.** *Let Π be a finite set. Consider the exponential weights method with learning rate $\eta > 0$ and initial point $q^{(1)} = \mathrm{unif}(\Pi)$, which has the update:*

$$q^{(t+1)}(\pi) = \frac{\exp(\eta \sum_{i \leq t} f^{(i)}(\pi))}{\sum_{\pi'} \exp(\eta \sum_{i \leq t} f^{(t)}(\pi'))},$$

for an arbitrary (potentially adaptively selected) sequence of reward vectors $f^{(1)}, \ldots, f^{(T)}$ in $\mathbb{R}^{\Pi}$. This strategy ensures that with probability 1,

$$\sum_{t=1}^{T} \langle q - q^{(t)}, f^{(t)} \rangle \leq \sum_{t=1}^{T} \langle q^{(t+1)} - q^{(t)}, f^{(t)} \rangle - \frac{1}{\eta} \sum_{t=1}^{T} D_{\mathsf{KL}}(q^{(t+1)} \,\|\, q^{(t)}) + \frac{D_{\mathsf{KL}}(q \,\|\, q^{(1)})}{\eta},$$

for all $q \in \Delta(\Pi)$.

D Examples

In this section, we instantiate our upper bounds for basic examples of interest.

D.1 Structured Bandits

We first consider adversarial (structured) bandit problems, which correspond to the special case of the adversarial DMSO setting in which there are no observations (i.e., $\mathcal{O} = \{\varnothing\}$). We consider three examples: finite-armed bandits, linear bandits, and convex bandits. For each example, we take $\mathcal{R} = [0, 1]$, fix a *reward function class* $\mathcal{F} \subseteq (\Pi \to [0, 1])$, and take $\mathcal{M}_{\mathcal{F}} = \{M \mid f^{M} \in \mathcal{F}\}$ to be the induced model class. The class $\mathcal{M}_{\mathcal{F}}$ should be thought of as the set of all reward distributions over $[0, 1]$ with mean rewards in $\mathcal{F}$.

Example D.1 (Finite-armed bandit)**.** In the finite-armed bandit problem, we take $\Pi = \{1, \ldots, A\}$ as the decision space for $A \in \mathbb{N}$, let $\mathcal{F} = [0, 1]^{A}$, and take $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$ as the induced model class. For this setting, whenever $A \geq 2$, it holds that

$$\mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \leq \frac{A}{\gamma} \quad \forall \gamma > 0, \quad \text{and} \quad \mathsf{dec}_{\gamma, \varepsilon_{\gamma}}(\mathrm{co}(\mathcal{M})) \geq 2^{-6} \cdot \frac{A}{\gamma} \quad \forall \gamma \geq \frac{A}{3}, \tag{30}$$

where $\varepsilon_{\gamma} = \frac{A}{12\gamma}$. ◁

This result follows from Foster et al. [17, Proposition 5.2 and 5.3], noting that $\mathrm{co}(\mathcal{M}) = \mathcal{M}$. Plugging (30) into Theorem 2.1 yields a $O(\sqrt{AT \log A})$ upper bound on regret, and plugging into Theorem 2.2 gives a $\widetilde{\Omega}(\sqrt{AT})$ lower bound for sub-Chebychev algorithms.[14]

Example D.2 (Linear bandit)**.** In the linear bandit problem, we have $\Pi \subseteq \mathbb{R}^d$ and

$$\mathcal{F} = \{f : \Pi \to [0,1] \mid f \text{ is linear}\},$$

and take $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$ as the induced model class. For this setting, it holds that[15]

$$\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \leq \frac{d}{4\gamma} \quad \forall \gamma > 0, \quad \text{and} \quad \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \geq \frac{d}{12\gamma} \quad \forall \gamma \geq \frac{2d}{3}, \tag{31}$$

where $\varepsilon_\gamma := \frac{d}{3\gamma}$. ◁

This result follows from Foster et al. [17, Proposition 6.1 and 6.2], again noting that $\mathrm{co}(\mathcal{M}) = \mathcal{M}$. Plugging (30) into Theorem 2.1 yields a $O(\sqrt{dT \log |\Pi|})$ upper bound on regret, and plugging into Theorem 2.2 gives a $\widetilde{\Omega}(\sqrt{dT})$ lower bound for sub-Chebychev algorithms.

Example D.3 (Convex bandit)**.** In the convex bandit problem, we have $\Pi \subseteq \mathbb{R}^d$ and

$$\mathcal{F} = \{f : \Pi \to [0,1] \mid f \text{ is convex}\},$$

and take $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$ as the induced model class. For this setting, it holds that for all $\gamma > 0$,

$$\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \leq O\left(\frac{d^4}{\gamma} \cdot \mathrm{polylog}(d, \mathrm{diam}(\Pi), \gamma)\right). \tag{32}$$

◁

This result follows from Foster et al. [17, Proposition 6.3] (which itself is a restatement of Lattimore and Szepesvári [34, Theorem 3]), and by noting once more that $\mathrm{co}(\mathcal{M}) = \mathcal{M}$.

Remark D.1. The adversarial bandit literature [4, 3, 18, 10, 1, 7, 26, 16, 8, 29, 26, 6] typically considers a slightly different formulation in which the adversary selects a deterministic reward function. This can be captured by restricting $\mathcal{M}$ to deterministic models. It is clear that the upper bounds on $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))$ in the examples above lead to upper bounds for this formulation. The lower bounds in Examples D.1 and D.2 easily extend as well.

The following example gives a bandit-type problem in which $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \gg \mathsf{dec}_\gamma(\mathcal{M})$, leading to a separation between the optimal regret for the stochastic setting and the adversarial setting.

Example D.4 (Noiseless bandits with gap)**.** Let $\Delta \in (0,1]$. Consider a deterministic multi-armed bandit problem with the decision space $\Pi = \{1, \ldots, A\}$, such that for any bandit instance M_f and $\pi \in \Pi$, the rewards are deterministically given by $f(\pi) \in [0,1]$ and f satisfies the gap assumption $\sup_{\pi,\pi'} |f(\pi) - f(\pi')| \leq \Delta$. Then, for any $\gamma \geq A$, there exists a model class $\mathcal{M}$ consisting of deterministic multi-armed bandit instances such that

$$\mathsf{dec}_\gamma(\mathcal{M}) \leq \Delta^2 \cdot \frac{A}{\gamma}$$

which leads to regret $\mathbf{Reg}_{\mathsf{DM}} = \widetilde{O}(\Delta\sqrt{AT})$ in the stochastic setting via Foster et al. [17]. However,

$$\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \geq 64^{-1} \cdot \Delta \cdot \frac{A}{\gamma},$$

which leads to a $\widetilde{\Omega}(\sqrt{\Delta \cdot AT})$ lower bound for the adversarial setting via Theorem 2.2. The lower bound on the regret in the adversarial setting is larger by a factor of at least $\sqrt{\Delta}$ than the regret upper bound in the stochastic case. ◁

The proof of Example D.4 is given in Appendix D.3.2.

[14] For this example and Example D.2, the lower bound on $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M}))$ in Foster et al. [17] is witnessed by a subfamily $\mathcal{M}' \subseteq \mathcal{M}$ with $V(\mathcal{M}') = O(1)$. As a result, we can take $C(T) = O(1)$ in Theorem 2.2.

[15] The upper bound here holds for all Π, while the lower bound holds for a specific choice for Π.

D.2 Reinforcement Learning

We now consider an example in reinforcement learning. We begin by recalling how to view the episodic reinforcement learning problem in the DMSO framework.

Model class. For episodic reinforcement learning, we fix a *horizon* H and let the model class $\mathcal{M}$ consist of a set of non-stationary Markov Decision Processes (MDP). Each model $M \in \mathcal{M}$ is specified by

$$M = \big\{\{\mathcal{S}_h\}_{h=1}^{H+1}, \mathcal{A}, \{P_h^M\}_{h=1}^{H}, \{R_h^M\}_{h=1}^{H}, d_1\big\},$$

where $\mathcal{S}_h$ is the state space for layer h, $\mathcal{A}$ is the action space, $P_h^M : \mathcal{S}_h \times \mathcal{A} \mapsto \Delta(\mathcal{S}_{h+1})$ is the probability transition kernel for layer h, $R_h^M : \mathcal{S}_h \times \mathcal{A} \mapsto \Delta([0,1])$ is the reward distribution for layer h and $d_1 \in \Delta(\mathcal{S}_1)$ is the initial state distribution. This formulation allows the reward distribution and transition kernel to vary across models in $\mathcal{M}$, but keeps the initial state distribution is fixed. We adopt the convention that $S_{H+1} = \{s_{H+1}\}$ where s_{H+1} is a deterministic terminal state.

For each episode, the learner selects a non-stationary policy $\pi = (\pi_1, \ldots, \pi_H)$, where $\pi_h : \mathcal{S}_h \mapsto \mathcal{A}$; we let Π_{NS} denote the set of all such policies. For a given MDP $M \in \mathcal{M}$, an episode proceeds by first sampling $s_1 \sim d_1$, then for $h = 1, \ldots, H$:

- $a_h = \pi_h(s_h)$.
- $r_h \sim R_h^M(s_h, a_h)$ and $s_{h+1} \sim P_h^M(\cdot \mid s_h, a_h)$.

The value of the policy π under M is given by $f^M(\pi) := \mathbb{E}^{M,\pi}\big[\sum_{h=1}^H r_h\big]$, where $\mathbb{E}^{M,\pi}[\cdot]$ denotes expectation under the process above.

Adversarial protocol. Within the adversarial DMSO framework, model classes above give rise to the following adversarial reinforcement learning protocol. At each time t, the learner plays selects a policy $\pi \in \Pi_{\mathrm{NS}}$ and the adversary chooses an MDP $M^{(t)} \in \mathcal{M}$. The policy $\pi^{(t)}$ is executed in $M^{(t)}$, resulting in a trajectory $\tau^{(t)} = (s_1^{(t)}, r_1^{(t)}, r_1^{(t)}), \ldots, (s_H^{(t)}, r_H^{(t)}, r_H^{(t)})$. The learner then observes feedback $(r^{(t)}, o^{(t)})$, where $r^{(t)} := \sum_{h=1}^H r_H^{(t)}$ is the cumulative reward for the episode, and $o^{(t)} = \tau^{(t)}$ is the trajectory.

With this setting in mind, we give our main example.

Example D.5 (Tabular MDP)**.** Let $\mathcal{M}$ be the class of finite-state/action (tabular) MDPs with horizon H, $S \geq 2$ states, $A \geq 2$ actions, and $\sum_{h=1}^H r_h \in [0,1]$. Then, for any $\gamma \geq A^{\min\{S-1,H\}}/6$,

$$\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \geq \frac{A^{\min\{S-1,H\}}}{24\gamma},$$

where $\varepsilon_\gamma := A^{\min\{S-1,H\}}/24\gamma$. ◁

Using this result with Theorem 2.2 leads to a lower bound on regret that scales with $\Omega(A^{\min S-1,H})$, which recovers existing intractability results for this setting [28, 37]. Note that we have $\mathsf{dec}_\gamma(\mathcal{M}) = \mathrm{poly}(S, A, H)/\gamma$ for this setting [17], so this is a case where there is a large separation between stochastic and adversarial decision making.

We briefly mention that the set $\mathrm{co}(\mathcal{M})$ can be interpreted as the set of *latent MDPs* [28], which is also known to be intractable. In the latent MDP problem, each model is a mixture of MDPs. At the beginning of each episode, the underlying MDP is drawn from the mixture (the identity is not observed), and the learner's policy is executed in this MDP for the duration of the episode.

D.3 Proofs for Examples

D.3.1 Preliminaries

Our lower bounds on the Decision-Estimation Coefficient involve a constructing hard sub-family of models. Recall the following definition from [17].

Definition D.1 ((α,β,δ)-family)**.** *A reference model $\overline{M} \in \mathcal{M}$ and collection $\{M_1, \ldots, M_N\}$ with $N \geq 2$ are said to be an (α,β,δ)-family if the following properties hold:*

- *Regret property. There exist functions* $u^M : \Pi \mapsto [0,1]$*, with* $\sum_{M\in\mathcal{M}} u^M(\pi) \le \frac{N}{2}$ *for all* π *such that*

$$f^M(\pi_M) - f^M(\pi) \ge \alpha \cdot (1 - u^M(\pi))$$

  for all $M \in \mathcal{M}$*.*

- *Information property. There exist functions* $v^M : \Pi \mapsto [0,1]$*, with* $\sum_{M\in\mathcal{M}} v^M(\pi) \le 1$ *for all* π*, such that*

$$D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big) \le \beta \cdot v^M(\pi) + \delta.$$

Any (α, β, δ)-family leads to a difficult decision making problem because a given decision can have low regret or large information gain on (roughly) one model in the family. This is formalized through the following lemma.

Lemma D.1 (Lemma 5.1, [17])**.** *Let* $\mathcal{M} = \{M_1, \ldots, M_N\}$ *be an* (α, β, δ)*-family with respect to* $\overline{M}$*. Then, for all* $\gamma \ge 0$*,*

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \ge \frac{\alpha}{2} - \gamma\left(\frac{\beta}{N} + \delta\right).$$

The following technical lemma bounds Hellinger distance for Bernoulli distributions.

Lemma D.2 (Lemma A.7, [17])**.** *For any* $\Delta \in (0, 1/2)$*,*

$$D^2_{\mathsf{H}}\Big(\mathrm{Ber}\Big(\frac{1}{2} + \Delta\Big), \mathrm{Ber}\Big(\frac{1}{2}\Big)\Big) \le 3\Delta^2.$$

D.3.2 Proof for Example D.4

In this section, we prove the separation between the stochastic and adversarial setting in the deterministic multi-armed bandit problem with gaps given in Example D.4.

For $i \in [A]$, define the vector $e_i \in [0,1]^A$ such that $e_i[j] = \Delta\mathbb{I}\{j = i\}$. Additionally, define $e_{\mathbf{0}}$ and $e_{\mathbf{1}}$ as the vector of all zeros and ones respectively. We consider a multi-armed bandit setting with the decision space $\Pi = [A]$. For a given reward function $f \in [0,1]^A$, let δ_f denote the noiseless multi-armed bandit model with $r = f(\pi)$ almost surely for action π. Define the model class

$$\mathcal{M} := \{\delta_{\mathbf{0}}, \delta_{\mathbf{1}}\} \cup \{e_i \mid i \in [A]\},$$

consisting of $A + 2$ models.

In the following, we first show the upper bound $\mathsf{dec}_\gamma(\mathcal{M}) \le \Delta A/\gamma$ in the stochastic setting, and then provide the lower bound $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \ge A/64\gamma$ in the adversarial setting .

Stochastic setting. Consider any $\overline{M} \in \mathcal{M}$. Define the ε-greedy strategy $\widehat{p}_{\overline{M}} = (1-\varepsilon)\pi_{\overline{M}} + \varepsilon \cdot \mathrm{Uniform}([A] \setminus \{\pi_{\overline{M}}\})$ where we set $\varepsilon = \Delta A/\gamma$, and note that

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \le \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \tag{33}$$

When $M = \overline{M}$, we have

$$\begin{aligned}\mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] &= \mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] \\ &\le \mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}[\Delta \cdot \mathbb{I}\{\pi \ne \pi_{\overline{M}}\}] \\ &= \Delta\varepsilon = \frac{\Delta^2 A}{\gamma}.\end{aligned} \tag{34}$$

When $M \ne \overline{M}$, note that

$$\mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \ge \mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}\big[\mathbb{I}\{M(\pi) \ne \overline{M}(\pi)\}\big] \ge \frac{\varepsilon}{A-1}$$

where the first inequality holds since both M and $\overline{M}$ have deterministic reward function, and the second inequality holds since any π for which $M(\pi) \ne \bar{M}(\pi)$ is chosen with probability at least

$\varepsilon/(A-1)$ under $\widehat{p}$. Using the fact that $\sup_{i,j\in[A]^2}|f^M(i)-f^M(j)| \le \Delta$ for all $M \in \mathcal{M}$ (by construction), we get that

$$\mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}\big[f^M(\pi_M)-f^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big] \le \mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}[f^M(\pi_M)-f^M(\pi)] - \frac{\gamma\varepsilon}{A-1}$$
$$\le \Delta - \frac{\Delta A}{A-1},$$

where the second line holds because $\varepsilon = \Delta A/\gamma$. Since $A/A-1 \ge 1$, we get that

$$\mathbb{E}_{\pi\sim\widehat{p}_{\overline{M}}}\big[f^M(\pi_M)-f^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big] < 0.$$

Using this fact with (34) in the relation (33) implies that

$$\mathsf{dec}_\gamma(\mathcal{M},\overline{M}) \le \frac{\Delta^2 A}{\gamma}.$$

Since the above holds for any $\overline{M}\in\mathcal{M}$, we immediately get that $\mathsf{dec}_\gamma(\mathcal{M}) \le \Delta^2 A/\gamma$.

Adversarial setting. Define $\overline{M}\in\mathrm{co}(\mathcal{M})$ as $\overline{M} = \frac{1}{2}\delta_{\mathbf{1}} + \frac{1}{2}\delta_{\mathbf{0}}$, and observe that for any $\pi\in[A]$, $\overline{M}(\pi) = \mathrm{Ber}(1/2)$. Next, for each $i\in[A]$, define $M^{(i)}\in\mathrm{co}(\mathcal{M})$ as $M^{(i)} = \frac{1}{2}\delta_{\mathbf{1}} + \frac{A}{8\gamma}\delta_{e_i} + (\frac{1}{2} - \frac{A}{8\gamma})\delta_{e_{\mathbf{0}}}$, and observe that for any $\pi\in[A]$

$$M^{(i)}(\pi) = \mathrm{Ber}(1/2)\mathbb{I}\{\pi\neq i\} + \mathrm{Ber}(1/2 + {}^{\Delta A}/_{8\gamma})\mathbb{I}\{\pi = i\}.$$

Consider the class $\mathcal{M}' = \big\{\overline{M}\big\}\cup\{M^{(i)} \mid i\in[A]\}$ consisting of $A+1$ models. Note that for any i and π,

$$f^{M^{(i)}}(\pi_{M^{(i)}}) - f^{M^{(i)}}(\pi_{M^{(i)}}) = \frac{\Delta A}{8\gamma}\mathbb{I}\{\pi\neq i\},$$

and $f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi) = 0$. Furthermore, for any i and π,

$$D^2_{\mathsf{H}}\big(M^{(i)}(\pi),\overline{M}(\pi)\big) = D^2_{\mathsf{H}}(\mathrm{Ber}(1/2),\mathrm{Ber}(1/2+{}^{\Delta A}/_{8\gamma}))\mathbb{I}\{\pi = i\} \le \frac{3\Delta^2A^2}{64\gamma^2}\mathbb{I}\{\pi = i\},$$

where the last inequality is due to Lemma D.2. Thus, the model class $\mathcal{M}'$ is a $(\Delta A/8\gamma, 3\Delta^2A^2/64\gamma^2, 0)$-family (D.1), and Lemma D.1 implies the lower bound

$$\mathsf{dec}_\gamma(\mathcal{M}') \ge \frac{\Delta A}{16\gamma} - \frac{3\Delta^2 A}{64\gamma} \ge \Delta\cdot\frac{A}{64\gamma},$$

where the last inequality holds since $\Delta \le 1$. Since $\mathcal{M}'\subset\mathrm{co}(\mathcal{M})$, the above implies that $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \ge \Delta A/64\gamma$.

D.3.3 Proof for Example D.5 (Tabular MDP)

In this section, we prove the lower bound in Example D.5. We first derive an intermediate result which gives a lower bound on the Decision-Estimation Coefficient when the model class $\mathcal{M}$ consists of *mixtures of K MDPs*; this is equivalent to the subset of $\mathrm{co}(\mathcal{M})$ where we restrict to support size K, as well as the so-called latent MDP setting [28].

Lemma D.3. *Let $K\ge 1$ be given. Let $\mathcal{M}$ be the class of* mixtures of K MDPs *with horizon H, $S\ge 2$ states, $A\ge 2$ actions, and $\sum_{h=1}^H r_h\in[0,1]$. Then there exists $\overline{M}\in\mathcal{M}$ such that for all $\gamma \ge A^{\min\{S-1,H,K\}}/6$,*

$$\mathsf{dec}_\gamma(\mathcal{M}_{\varepsilon_\gamma}(\overline{M}),\overline{M}) \ge \frac{A^{\min\{S-1,H,K\}}}{24\gamma},$$

where $\varepsilon_\gamma := \frac{A^{\min\{S-1,H,K\}}}{24\gamma}$.

The proof of this result proceeds by constructing a hard sub-family of models and appealing to Lemma D.1. Our construction is based of the lower bound for latent MDPs in Kwon et al. [28].

Proof of Lemma D.3. Let $\mathcal{S}$ and $\mathcal{A}$ be arbitrary sets with $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. Let $\Delta\in(0,1/2)$ be a parameter to be chosen later, and define $\overline{K} := \min\{S-1,K,H\}$. Partition the state space

$\mathcal{S}$ into sets $\mathcal{S}'$ and $\mathcal{S} \setminus \mathcal{S}'$ such that $|\mathcal{S}'| = \overline{K} + 1$, and label the states in $\mathcal{S}'$ as $\{s^{(1)}, \ldots, s^{(\overline{K}+1)}\}$. Additionally, define sets via $\mathcal{S}_h = \{s^{(h)}, s^{(\overline{K}+1)}\}$ for $h \leq \overline{K}$ and $\mathcal{S}_h = \{s^{(\overline{K}+1)}\} \cup (S \setminus S')$ for $\overline{K} < h \leq H + 1$. Recall that the decision space Π_{NS} is the set of all deterministic non-stationary policies $\pi = (\pi_1, \ldots, \pi_H)$ where $\pi_h : \mathcal{S}_h \mapsto \mathcal{A}$.

We construct a class $\mathcal{M}' \subseteq \mathcal{M}$ in which each model $M \in \mathcal{M}'$ is specified by

$$M = \big\{\{\mathcal{S}_h\}_{h=1}^{H+1}, \mathcal{A}, \{\mathbb{M}_k^M\}_{k=1}^{\overline{K}}, \{a_k^M\}_{k=1}^{K}\big\},$$

where for each $k \in [\overline{K}]$, $a_k^M \in \mathcal{A}$, and where $\mathbb{M}_k^M$ is a tabular MDP specified by

$$\mathbb{M}_k^M = \big\{\{\mathcal{S}_h\}_{h=1}^{H+1}, \mathcal{A}, \big\{P_{h,k}^M\big\}_{h=1}^{H}, \big\{R_{h,k}^M\big\}_{h=1}^{H}, \delta_{s^{(1)}}\big\}.$$

Here, $d_1 = \delta_{s^{(1)}}$, so that the initial state s_1 is $s^{(1)}$ deterministically. The transitions $P_{h,k}^M$ and rewards $R_{h,k}^M$ are constructed as follows.

- Construction of $\mathbb{M}_1^M$.
  - (i) For all $h \leq H$, the dynamics $P_{h,k}^M$ are deterministic. For an action a_h in the state s_h, the next state s_{h+1} is

$$s_{h+1} = \begin{cases} s^{(h+1)}, & \text{if } h \leq \overline{K},\, s_h = s^{(h)}, \text{ and } a_h = a_i^M, \\ s^{(\overline{K}+1)}, & \text{if } h \leq \overline{K},\, s_h = s^{(h)}, \text{ and } a_h \neq a_i^M, \\ s_h, & \text{otherwise.} \end{cases}$$

  - (ii) The reward distribution is given by

$$R_{h,k}^M(s_h, a_h) = \begin{cases} \mathrm{Ber}\big(\frac{1}{2} + \Delta\big), & \text{if} \quad h = \overline{K}, s_h = s^{(\overline{K})}, \text{ and } a_h = a_{\overline{K}}^M, \\ \mathrm{Ber}\big(\frac{1}{2}\big), & \text{if} \quad h = \overline{K}, s_h = s^{(\overline{K})}, \text{ and } a_h \neq a_{\overline{K}}^M, \\ 0, & \text{otherwise.} \end{cases}$$

- Construction of $\mathbb{M}_j^M$ for $2 \leq j \leq \overline{K}$.
  - (i) For each $h \leq H$, the dynamics $P_{h,k}^M$ are deterministic. For action a_h in state s_h, the next state s_{h+1} is

$$s_{h+1} = \begin{cases} s^{(h+1)} & \text{if} \quad s_h = s^{(h)} \text{ and } h < j \\ s^{(\overline{K}+1)} & \text{if} \quad s_h = s^{(h)}, h = j \text{ and } a_h = a_h^M \\ s^{(h+1)} & \text{if} \quad s_h = s^{(h)}, h = j \text{ and } a_h \neq a_h^M \\ s^{(h+1)} & \text{if} \quad s_h = s^{(h)}, h > j \text{ and } a_h = a_h^M \\ s^{(\overline{K}+1)} & \text{if} \quad s_h = s^{(h)}, h > j \text{ and } a_h \neq a_h^M \\ s^{(\overline{K}+1)} & \text{if} \quad h = \overline{K} - 1 \text{ or } h = \overline{K} \\ s_h & \text{otherwise} \end{cases}.$$

  - (ii) The reward distribution is given by

$$R_{h,k}^M(s_h, a_h) = \begin{cases} \mathrm{Ber}\big(\frac{1}{2}\big), & \text{if} \quad h = \overline{K}, \\ 0, & \text{otherwise.} \end{cases}$$

Each model $M \in \mathcal{M}'$ is a uniform mixture of $\overline{K}$ MDPs $\{\mathbb{M}_1^M, \ldots, \mathbb{M}_{\overline{K}}^M\}$ as described above, parameterized by the action sequence $a_{1:\overline{K}}^M$. The model class $\mathcal{M}'$ is defined as the set of all such mixture models (one for each sequence in $\mathcal{A}^{\overline{K}}$, so that $|\mathcal{M}'| = A^{\overline{K}}$.

At the start of each episode, an MDP $\mathbb{M}_z^M$ is chosen by sampling $z \sim \mathrm{Unif}([\overline{K}])$. The trajectory is then drawn by setting $s_1 = s^{(1)}$, and for $h = 1, \ldots, H$:

- $a_h = \pi_h(s_h)$.
- $r_h \sim R_{h,z}^M(s_h, a_h)$ and $s_{h+1} \sim P_{h,z}^M(\cdot \mid s_h, a_h)$.

Note that rewards can be non-zero only at layer $h = \overline{K}$. We receive a reward from $\mathrm{Ber}\big(\frac{1}{2} + \Delta\big)$ only when $z = 1$ and the first $\overline{K}$ actions match $a^M_{1:\overline{K}}$, i.e. $a_{1:\overline{K}} = a^M_{1:\overline{K}}$. For every other action sequence, the reward is sampled from $\mathrm{Ber}\big(\frac{1}{2}\big)$. Thus, for any policy π,

$$f^M(\pi) = \tfrac{1}{2} + \Delta \mathbb{I}\{\pi(s_{1:\overline{K}}) = a^M_{1:\overline{K}}\},$$

which implies that

$$f^M(\pi_M) - f^M(\pi) = \Delta(1 - \mathbb{I}\{\pi(s_{1:\overline{K}}) = a^M_{1:\overline{K}}\}). \tag{35}$$

Finally, we define the reference model $\overline{M}$. The model $\overline{M}$ is specified by $\big\{\{\mathcal{S}_h\}_{h=1}^{H+1}, \mathcal{A}, \mathbb{M}^{\overline{M}}\big\}$ where $\mathbb{M}^{\overline{M}}$ is a tabular MDP given by

$$\mathbb{M}^{\overline{M}} = \big\{\{\mathcal{S}_h\}_{h=1}^{H+1}, \mathcal{A}, P_h^{\overline{M}}, R_h^{\overline{M}}, \delta_{s^{(1)}}\big\}.$$

Here, the initial state s_1 is $s^{(1)}$ deterministically, and the transitions $P^{\overline{M}}_{h,k}$ and rewards $R^{\overline{M}}_{h,k}$ are as follows:

(i) Transitions are stochastic and independent of the chosen action. In particular, for each $h \le H$, the dynamics $P_h^{\overline{M}}$ are given by

$$P_h^{\overline{M}}(s_{h+1} \mid s_h, a_h) = \begin{cases} \frac{\overline{K}-h}{\overline{K}-h+1} & \text{if} \quad h \le \overline{K}, s_h = s^{(h)} \text{ and } s_{h+1} = s^{(h+1)} \\ \frac{1}{\overline{K}-h+1} & \text{if} \quad h \le \overline{K}, s_h = s^{(h)} \text{ and } s_{h+1} = s^{(\overline{K}+1)} \\ 1 & \text{if} \quad h \le \overline{K}, s_h \ne s^{(h)} \text{ and } s_h = s_{h+1} \\ 1 & \text{if} \quad h > \overline{K} \text{ and } s_h = s_{h+1} \\ 0 & \text{otherwise} \end{cases}.$$

(ii) The reward distribution is given by

$$R_h^{\overline{M}}(s_h, a_h) = \begin{cases} \mathrm{Ber}\big(\frac{1}{2}\big), & \text{if} \quad h = \overline{K}, \\ 0, & \text{otherwise.} \end{cases}.$$

Note that $\overline{M}$ can be thought of as a mixture of $\overline{K}$ identical tabular MDPs each given by $\mathbb{M}^{\overline{M}}$. Note that for any policy π, the rewards for any trajectory in $\overline{M}$ are sampled from $\mathrm{Ber}\big(\frac{1}{2}\big)$, and thus $f^{\overline{M}}(\pi) = \frac{1}{2}$ which implies that

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi) = 0. \tag{36}$$

We define $\mathcal{M}'' = \mathcal{M}' \cup \{\overline{M}\} \subseteq \mathcal{M}$, and note that for any policy π, the distribution over the trajectories is identical in all mixture models in $\mathcal{M}''$. However, as mentioned before, the rewards in $\overline{M}$ are sampled from $\mathrm{Ber}\big(\frac{1}{2}\big)$ and for any $M \in \mathcal{M}'$, the rewards in M are sampled from $\mathrm{Ber}\big(\frac{1}{2} + \frac{\Delta}{\overline{M}}\mathbb{I}\Big\{\pi(s_{1:\overline{K}}) = a^M_{1:\overline{K}}\Big\}\big)$. Thus, for any policy π and $M \in \mathcal{M}'$,

$$\begin{aligned} D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big) &= D^2_{\mathsf{H}}\Big(\mathrm{Ber}\Big(\tfrac{1}{2} + \tfrac{\Delta}{\overline{K}}\mathbb{I}\{\pi(s_{1:\overline{K}}) = a^M_{1:\overline{K}}\}\Big), \mathrm{Ber}\Big(\tfrac{1}{2}\Big)\Big) \\ &\le 3\frac{\Delta^2}{\overline{K}^2} \cdot \mathbb{I}\{\pi(s_{1:\overline{K}}) = a^M_{1:\overline{K}}\}, \end{aligned} \tag{37}$$

where the last line uses Lemma D.2.

The bounds in (35), (36) and (37) together imply that the model class $\mathcal{M}''$ is a $(\frac{\Delta}{\overline{K}}, 3\frac{\Delta^2}{\overline{K}^2}, 0)$-family in the sense of Definition D.1, where for each $\pi \in \Pi$ and$M \in \mathcal{M}''$ we take

$$u^M(\pi) := \mathbb{I}\{\pi(s_{1:\overline{K}}) = a^M_{1:\overline{K}}\} \qquad \text{and} \qquad v^M(\pi) := \mathbb{I}\{\pi(s_{1:\overline{K}}) = a^M_{1:\overline{K}}\},$$

with $u^{\overline{M}}(\pi) := 1$ and $v^{\overline{M}}(\pi) := 0$. As a result, Lemma D.1 implies that

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \ge \frac{\Delta}{2\overline{K}} - \frac{3\gamma\Delta^2}{\overline{K}^2 N},$$

for $N := A^{\overline{K}} + 1$. Setting $\Delta = \frac{\overline{K}N}{12\gamma}$ leads to the lower bound $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \geq \frac{N}{24\gamma}$. We conclude by noting that all $M \in \mathcal{M}''$ have $M \in \mathcal{M}_{\varepsilon_\gamma}(\overline{M})$ with $\varepsilon_\gamma = \frac{N}{24\gamma}$, and thus the lower bound on the DEC also applies to the class $\mathcal{M}_{\varepsilon_\gamma}(\overline{M})$. □

Proof for Example D.5. let $\mathcal{M}$ be the class of all tabular MDPs, and let $\mathcal{M}^{(K)}$ denote the set of all mixture models in which each $M \in \mathcal{M}^{(K)}$ is a mixture of K MDPs from $\mathcal{M}$. Additionally, define $\widetilde{\mathcal{M}} = \mathrm{co}(\mathcal{M})$, and note that $\mathcal{M}^{(K)} \subseteq \widetilde{\mathcal{M}}$ for all $K \geq 1$. For any $\varepsilon > 0$ and $\overline{M} \in \mathcal{M}^{(K)}$, we have that $\mathcal{M}^{(k)}_\varepsilon(\overline{M}) \subseteq \widetilde{\mathcal{M}}_\varepsilon(\overline{M})$, which implies that

$$\mathsf{dec}_\gamma(\widetilde{\mathcal{M}}_\varepsilon(\overline{M}), \overline{M}) \geq \mathsf{dec}_\gamma(\mathcal{M}^{(K)}_\varepsilon(\overline{M}), \overline{M}),$$

because $\mathsf{dec}_\gamma(\cdot, \overline{M})$ is a non-decreasing function with respect to inclusion. Using Lemma D.3, it holds that for any $K \geq 1$ and $\gamma \geq A^{\min\{S-1,H,K\}}/6$, with $\varepsilon_\gamma := A^{\min\{S-1,H,K\}}/24\gamma$,

$$\mathsf{dec}_\gamma(\widetilde{\mathcal{M}}_\varepsilon(\overline{M}), \overline{M}) \geq \mathsf{dec}_\gamma(\mathcal{M}^{(K)}_\varepsilon(\overline{M}), \overline{M}) \geq \frac{A^{\min\{S-1,H,K\}}}{24\gamma}.$$

Setting $K = S$ above gives the desired lower bound. □

E Structural Results

This section is organized as follows.

- In Appendix E.1, we recall the existing variants of Information Ratio and state some basic properties.
- In Appendix E.2, we prove equivalence of the Decision-Estimation Coefficient and the parameterized Information Ratio with Hellinger distance (Theorem 3.1), as well as a generalization of this result (Theorem E.1).
- In Appendix E.3, we prove equivalence of the parameterized Information Ratio with Hellinger distance and the high-probability exploration-by-optimization objective.

E.1 Background on Complexity Measures

Throughout this section, we refer to any non-negative function $D(\cdot \parallel \cdot)$ that is defined over $\Delta(\mathcal{X}) \times \Delta(\mathcal{X})$ for all measurable spaces $(\mathcal{X}, \mathscr{F})$ as a *divergence-like* function.

Generalized Information Ratio. Below we recall two notions of *generalized Information Ratio* introduced by Lattimore and György [32] and Lattimore [31], which extend the original definition of Russo and Van Roy [46, 47].

For a given prior $\mu \in \Delta(\mathcal{M} \times \Pi)$ and distribution $p \in \Delta(\Pi)$, let $\mathbb{P}$ denote the law of the process $(M, \pi^\star) \sim \mu, \pi \sim p, z \sim M(\pi)$, and define $\mu_{\mathrm{pr}}(\pi') := \mathbb{P}(\pi^\star = \pi')$ and $\mu_{\mathrm{po}}(\pi'; \pi, z) := \mathbb{P}(\pi^\star = \pi' \mid (\pi, z))$.

1. Lattimore and György [32] define a class $\mathcal{M}$ to have generalized Information Ratio (α, β, λ) (where $\alpha, \beta \geq 0$ and $\lambda > 1$) if for each prior $\mu \in \Delta(\mathcal{M} \times \Pi)$, there exists a distribution $p \in \Delta(\Pi)$ such that

$$\mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] \leq \alpha + \beta^{1-1/\lambda}\big(\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot; \pi, z) \parallel \mu_{\mathrm{pr}})]\big)^{1/\lambda}. \tag{38}$$

2. Lattimore [31] defines the generalized Information Ratio for a class $\mathcal{M}$ (for $\lambda > 1$) via

$$\Psi_\lambda(\mathcal{M}) = \sup_{\mu \in \Delta(\mathcal{M}\times\Pi)} \inf_{p \in \Delta(\Pi)} \left\{ \frac{(\mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)])^\lambda}{\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot; \pi, z) \parallel \mu_{\mathrm{pr}})]} \right\}. \tag{39}$$

As mentioned in Section 3, the formulations above slightly generalize the original versions in Lattimore and György [32], Lattimore [31] by incorporating models $M \in \mathcal{M}$ and considering general distances.

The following proposition shows that boundedness of the generalized Information Ratio implies boundedness of the parameterized Information Ratio (Definition 3.1).

Proposition E.1. *Fix $\alpha, \beta \geq 0$ and $\lambda > 1$. If a class $\mathcal{M}$ has generalized Information Ratio (α, β, λ) in the sense of (38), then*

$$\mathsf{inf}^D_\gamma(\mathcal{M}) \leq \alpha + \beta/\gamma^{\frac{1}{\lambda-1}} \quad \forall \gamma > 0.$$

Likewise, the generalized Information Ratio in (39) satisfies

$$\mathsf{inf}^D_\gamma(\mathcal{M}) \leq (\Psi_\lambda(\mathcal{M})/\gamma)^{\frac{1}{\lambda-1}} \quad \forall \gamma > 0.$$

Proof of Proposition E.1. Suppose $\mathcal{M}$ has generalized Information Ratio (α, β, λ). Then there exists $p \in \Delta(\Pi)$ such that for all $\mu \in \Delta(\mathcal{M} \times \Pi)$,

$$\begin{aligned}\mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] &\leq \alpha + \beta^{1-1/\lambda}\big(\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})]\big)^{1/\lambda} \\ &\leq \alpha + \frac{\beta}{\gamma^{\frac{1}{\lambda-1}}} + \gamma \cdot \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})],\end{aligned}$$

where we have applied Young's inequality, which gives that $xy \leq \frac{\lambda-1}{\lambda} x^{\frac{\lambda}{\lambda-1}} + \frac{1}{\lambda} y^\lambda$ for $x, y \geq 0$.

For the second result, we use that the definition of $\Psi_\lambda(\mathcal{M})$ implies generalized Information Ratio $(0, (\Psi_\lambda(\mathcal{M}))^{\frac{1}{\lambda-1}}, \lambda)$. □

This results show that an upper bound in terms of the parameterized Information Ratio in Definition 3.1 implies an upper bound in terms of either version of the generalized Information Ratio. It is also straightforward to see that generalized Information Ratio $(0, \beta, \lambda)$ in (38) implies that $\Psi_\lambda(\mathcal{M}) \leq \beta^{\lambda-1}$ and vice-versa. Note that $\alpha = 0$ is the most interesting regime, as the regret bounds in Lattimore and György [32] scale with $\alpha \cdot T$ when $\alpha > 0$.

Another important property of the parameterized Information Ratio (as well as the generalized Information Ratio) is that it is invariant under convexification.

Proposition 3.1. *For any divergence-like function $D(\cdot \parallel \cdot)$, we have*

$$\mathsf{inf}^D_\gamma(\mathcal{M}) = \mathsf{inf}^D_\gamma(\mathrm{co}(\mathcal{M})), \quad \forall \gamma > 0.$$

Proof of Proposition 3.1. Fix $\mu \in \Delta(\mathrm{co}(\mathcal{M}) \times \Pi)$. We can represent any $\overline{M} \in \mathrm{co}(\mathcal{M})$ as a mixture $\nu \in \Delta(\mathcal{M})$, so that $\overline{M} = \mathbb{E}_{M\sim\nu}[M]$. Let $\widetilde{\mu} \in \Delta(\Delta(\mathcal{M}) \times \Pi)$ be such that the process $(\nu, \pi^\star) \sim \widetilde{\mu}$, $\overline{M} = \mathbb{E}_{M\sim\nu}[M]$ has the same law as $(\overline{M}, \pi^\star) \sim \widetilde{\mu}$. Finally, let $\mu' \in \Delta(\mathcal{M} \times \Pi)$ be the law of $(M, \pi^\star)$ induced by sampling $(\nu, \pi^\star) \sim \widetilde{\mu}$ and $M \sim \nu$.

We observe that for any distribution $p \in \Delta(\Pi)$,

$$\begin{aligned}\mathbb{E}_{(\overline{M},\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}\big[f^{\overline{M}}(\pi^\star) - f^{\overline{M}}(\pi)\big] &= \mathbb{E}_{(\nu,\pi^\star)\sim\widetilde{\mu}}\, \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{M\sim\nu}[f^M(\pi^\star) - f^M(\pi)] \\ &= \mathbb{E}_{(M,\pi^\star)\sim\mu'}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)].\end{aligned}$$

Next, observe that $(\pi, \pi^\star, z)$ are identically distributed under the processes $\pi \sim p$, $(\overline{M}, \pi^\star) \sim \mu$, $z \sim \overline{M}(\pi)$ and $\pi \sim p$, $(M, \pi^\star) \sim \mu'$, $z \sim M(\pi)$. As a result, we have $\mu_{\mathrm{pr}} = \mu'_{\mathrm{pr}}$ and $\mu_{\mathrm{po}} = \mu'_{\mathrm{po}}$, so

$$\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})] = \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D\big(\mu'_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu'_{\mathrm{pr}}\big)\big].$$

This establishes that $\mathsf{inf}^D_\gamma(\mathrm{co}(\mathcal{M})) \leq \mathsf{inf}^D_\gamma(\mathcal{M})$; the other direction is trivial. □

E.2 Decision-Estimation Coefficient and Information Ratio

Theorem 3.1. *For all $\gamma > 0$, $\mathsf{inf}^{\mathsf{H}}_\gamma(\mathcal{M}) \leq \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \leq \mathsf{inf}^{\mathsf{H}}_{\gamma/4}(\mathcal{M})$.*

Theorem 3.1 is a special case of the following theorem, which concerns general divergence-like functions.

Theorem E.1. *Let $D(\cdot \parallel \cdot)$ be any divergence-like function defined over $\Delta(\Pi) \times \Delta(\Pi)$ and $\Delta(\mathcal{Z}) \times \Delta(\mathcal{Z})$, for which there exists constants $c_1, c_2 \geq 1$ such that:*

1. *For all $\mathbb{Q} \in \Delta(\mathcal{Z})$, $\mathbb{P} \mapsto D(\mathbb{P} \parallel \mathbb{Q})$ is convex.*

2. *For all pairs of random variables* $(\pi, z) \in \Pi \times \mathcal{Z}$,

$$\mathbb{E}_{\pi\sim\mathbb{P}_\pi}\big[D\big(\mathbb{P}_{z|\pi} \,\|\, \mathbb{P}_z\big)\big] \le c_1 \cdot \mathbb{E}_{z\sim\mathbb{P}_z}\big[D\big(\mathbb{P}_{\pi|z} \,\|\, \mathbb{P}_\pi\big)\big]$$

and

$$\mathbb{E}_{z\sim\mathbb{P}_z}\big[D\big(\mathbb{P}_{\pi|z} \,\|\, \mathbb{P}_\pi\big)\big] \le c_1 \cdot \mathbb{E}_{\pi\sim\mathbb{P}_\pi}\big[D\big(\mathbb{P}_{z|\pi} \,\|\, \mathbb{P}_z\big)\big].$$

3. *For all pairs of random variables* $(\pi, z) \in \Pi \times \mathcal{Z}$,

$$\mathbb{E}_{\pi\sim\mathbb{P}_\pi}\big[D\big(\mathbb{P}_{z|\pi} \,\|\, \mathbb{P}_z\big)\big] \le c_2 \cdot \inf_{\mathbb{Q}} \mathbb{E}_{\pi\sim\mathbb{P}_\pi}\big[D\big(\mathbb{P}_{z|\pi} \,\|\, \mathbb{Q}\big)\big].$$

4. *For all* $\varepsilon > 0$ *sufficiently small, and all* $\mathbb{Q} \in \Delta(\mathcal{Z})$, *there exists* $\mathbb{Q}' \in \Delta(\mathcal{Z})$ *such that* $D(\mathbb{P} \,\|\, \mathbb{Q}) \ge D(\mathbb{P} \,\|\, \mathbb{Q}') - \varepsilon$ *and* $\sup_{\mathbb{P}\in\Delta(\mathcal{Z})} D(\mathbb{P} \,\|\, \mathbb{Q}') < \infty$.

Then we have

$$\mathsf{inf}^D_{c_1\gamma}(\mathcal{M}) \le \mathsf{dec}^D_\gamma(\mathrm{co}(\mathcal{M})) \le \mathsf{inf}^D_{\gamma/c_1c_2}(\mathcal{M}). \tag{40}$$

All f-divergences satisfy Property 2 with $c_1 = 1$, but may not satisfy Property 3. On the other hand, Bregman divergences[16] satisfy Property 3 with $c_2 = 1$, but may not satisfy Property 2 (consider squared euclidean distance). KL-divergence, being both an f-divergence and a Bregman divergence, satisfies both properties with $c_1 = c_2 = 1$ (this fact has been used tacitly in many prior works). Squared Hellinger distance is an f-divergence but not a Bregman divergence, yet satisfies Property 3 with $c_2 = 4$ as a consequence of the triangle inequality.

Proof of Theorem E.1. We first bound the DEC by the Information Ratio, then proceed to bound the Information Ratio by the DEC.

Bounding the DEC by the Information Ratio. Fix any reference model $M' \in \mathcal{M}$, $\varepsilon > 0$ and let M'' be such that $D^2_{\mathsf{H}}(\cdot, M'(\pi)) \ge D^2_{\mathsf{H}}(\cdot, M''(\pi)) - \varepsilon$ and $D^2_{\mathsf{H}}(\cdot, M''(\pi)) < \infty$ (M'' can be obtained by applying property 4 uniformly for all π). Using the minimax theorem (Lemma C.2), we have

$$\begin{aligned}
\mathsf{dec}^D_\gamma(\mathcal{M}, M') &\le \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D(M(\pi) \,\|\, M''(\pi))] + \gamma\varepsilon \\
&= \inf_{p\in\Delta(\Pi)} \sup_{\nu\in\Delta(\mathcal{M})} \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{M\sim\nu}[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D(M(\pi) \,\|\, M''(\pi))] + \gamma\varepsilon \\
&= \sup_{\nu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{M\sim\nu}[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D(M(\pi) \,\|\, M''(\pi))] + \gamma\varepsilon.
\end{aligned}$$

Note that the application of the minimax theorem is admissible here, since $\Delta(\Pi)$ is compact (a consequence of finiteness of Π) and the objective value is bounded (a consequence of the choice of M'' and the fact that $f^M \in [0, 1]$).

Fix any $\nu \in \Delta(\mathcal{M})$, and define $\mu \in \Delta(\mathcal{M} \times \Pi)$ be the induced law of (M, π_M). Define $\overline{M}_{\pi'}(\pi) = \mathbb{E}_{M\sim\nu}[M(\pi) \mid \pi_M = \pi']$ and $\overline{M}(\pi) = \mathbb{E}_{M\sim\mu}[M(\pi)] = \mathbb{E}_{\pi^\star\sim\mu}[M_{\pi^\star}(\pi)]$. Then, for any distribution $p \in \Delta(\Pi)$,

$$\begin{aligned}
&\mathbb{E}_{M\sim\nu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D(M(\pi) \,\|\, M''(\pi))] \\
&\qquad = \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi) - \gamma \cdot D(M(\pi) \,\|\, M''(\pi))] \\
&\qquad \le \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}\big[f^M(\pi^\star) - f^M(\pi) - \gamma \cdot D\big(\overline{M}_{\pi^\star}(\pi) \,\|\, M''(\pi)\big)\big] \\
&\qquad \le \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}\left[f^M(\pi^\star) - f^M(\pi) - \frac{\gamma}{c_2} \cdot D\big(\overline{M}_{\pi^\star}(\pi) \,\|\, \overline{M}(\pi)\big)\right],
\end{aligned}$$

where the first inequality uses convexity of D in the first argument (Property 1), and the second inequality uses Property 3. To proceed, let $\mathbb{P}$ be the law of the process $\pi \sim p$, $(M, \pi^\star) \sim \mu$, $z \sim M(\pi)$, and define $\mu_{\mathrm{pr}}(\pi') := \mathbb{P}(\pi^\star = \pi')$ and $\mu_{\mathrm{po}}(\pi'; \pi, z) := \mathbb{P}(\pi^\star = \pi' \mid (\pi, z))$. Observe that $\overline{M}_{\pi^\star}(\pi) = \mathbb{P}_{z|\pi,\pi^\star}$ and $\overline{M}(\pi) = \mathbb{P}_{z|\pi}$. Hence, using Property 2, it holds that for all π,

$$\mathbb{E}_{\pi^\star\sim\nu}\big[D\big(\overline{M}_{\pi^\star}(\pi) \,\|\, \overline{M}(\pi)\big)\big] \ge \frac{1}{c_1}\, \mathbb{E}_{z|\pi}\big[D\big(\mathbb{P}_{\pi^\star|\pi,z} \,\|\, \mathbb{P}_{\pi^\star|\pi}\big)\big] = \frac{1}{c_1}\, \mathbb{E}_{z|\pi}\big[D\big(\mathbb{P}_{\pi^\star|\pi,z} \,\|\, \mathbb{P}_{\pi^\star}\big)\big],$$

[16] Recall that for a convex set $\mathcal{X}$ and regularizer $\mathcal{R} : \mathcal{X} \to \mathbb{R}$, $D_{\mathcal{R}}(x \,\|\, y) := \mathcal{R}(x) - \mathcal{R}(y) - \langle \nabla\mathcal{R}(y), x - y\rangle$ is the associated Bregman divergence.

where the last equality holds because π and $\pi^\star$ are independent (marginally). Since $D\big(\mathbb{P}_{\pi^\star|\pi,z} \parallel \mathbb{P}_{\pi^\star}\big) = D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})$, if we choose p to attain the minimum in (21) for μ we are guaranteed that

$$\begin{aligned}
&\mathbb{E}_{M\sim\nu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi) - \gamma\cdot D(M(\pi) \parallel M'(\pi))] \\
&\leq \mathbb{E}_{(M,\pi^\star)\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] - \frac{\gamma}{c_1c_2}\cdot\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})] + \gamma\varepsilon \\
&\leq \mathsf{inf}^D_{\gamma/c_1c_2}(\mathcal{M}) + \gamma\varepsilon.
\end{aligned}$$

Taking $\varepsilon \to 0$, we conclude that $\mathsf{dec}^D_\gamma(\mathcal{M}) \leq \mathsf{inf}^D_{\gamma/c_1c_2}(\mathcal{M})$. By Proposition 3.1, $\mathsf{inf}^D_\gamma(\mathcal{M}) = \mathsf{inf}^D_\gamma(\mathrm{co}(\mathcal{M}))$, so applying the result to $\mathrm{co}(\mathcal{M})$ yields

$$\mathsf{dec}^D_\gamma(\mathrm{co}(\mathcal{M})) \leq \mathsf{inf}^D_{\gamma/c_1c_2}(\mathcal{M}).$$

Bounding the Information Ratio by the DEC. We now consider the opposite direction. Fix a prior $\mu \in \Delta(\mathcal{M}\times\Pi)$ and consider the value for the parameterized Information Ratio:

$$\mathbb{E}_{(M,\pi^\star)\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] - \gamma\cdot\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})],$$

with μ_{pr} and μ_{po} defined as before. Define $\overline{M}_{\pi'}(\pi) := \mathbb{E}_\mu[M(\pi) \mid \pi^\star = \pi']$ and $\overline{M}(\pi) = \mathbb{E}_{M\sim\mu}[M(\pi)]$. Using that $\pi^\star$ and π are independent (marginally), along with Property 2, we have

$$\begin{aligned}
\mathbb{E}_{z|\pi}[D(\mu_{\mathrm{po}}(\cdot;\pi,z) \parallel \mu_{\mathrm{pr}})] &= \mathbb{E}_{z|\pi}\big[D\big(\mathbb{P}_{\pi^\star|\pi,z} \parallel \mathbb{P}_{\pi^\star}\big)\big] \\
&= \mathbb{E}_{z|\pi}\big[D\big(\mathbb{P}_{\pi^\star|\pi,z} \parallel \mathbb{P}_{\pi^\star|\pi}\big)\big] \geq \frac{1}{c_1}\,\mathbb{E}_{\pi^\star\sim\mu}\big[D\big(\overline{M}_{\pi^\star}(\pi) \parallel \overline{M}(\pi)\big)\big].
\end{aligned}$$

Next, observe that

$$\begin{aligned}
\mathbb{E}_{(M,\pi^\star)\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] &= \mathbb{E}_{\pi\sim p}\,\mathbb{E}_{\pi^\star\sim\mu}\,\mathbb{E}[f^M(\pi^\star) - f^M(\pi) \mid \pi^\star] \\
&= \mathbb{E}_{\pi\sim p}\,\mathbb{E}_{\pi^\star\sim\mu}\big[f^{\overline{M}_{\pi^\star}}(\pi^\star) - f^{\overline{M}_{\pi^\star}}(\pi)\big] \\
&\leq \mathbb{E}_{\pi^\star\sim\mu}\,\mathbb{E}_{\pi\sim p}\Big[\max_{\pi'} f^{\overline{M}_{\pi^\star}}(\pi') - f^{\overline{M}_{\pi^\star}}(\pi)\Big].
\end{aligned}$$

Recall that the definition of $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))$ implies the following: For any $\kappa \in \Delta(\mathcal{M})$ there exists a distribution $p \in \Delta(\Pi)$ such that for all $\nu \in \Delta(\mathcal{M})$, defining $\overline{M}_\kappa(\pi) := \mathbb{E}_{M\sim\kappa}[M(\pi)]$ and $\overline{M}_\nu(\pi) := \mathbb{E}_{M\sim\nu}[M(\pi)]$, we have

$$\mathbb{E}_{\pi\sim p}\Big[\max_{\pi'} f^{\overline{M}_\nu}(\pi') - f^{\overline{M}_\nu}(\pi) - \gamma\cdot D\big(\overline{M}_\nu(\pi) \parallel \overline{M}_\kappa(\pi)\big)\Big] \leq \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})). \tag{41}$$

By invoking (41) with $\overline{M}_\kappa = \overline{M}$ and $\overline{M}_\nu = \overline{M}_{\pi^\star}$, we are guaranteed that for every draw of $\pi^\star$,

$$\mathbb{E}_{\pi\sim p}\Big[\max_{\pi'} f^{\overline{M}_{\pi^\star}}(\pi') - f^{\overline{M}_{\pi^\star}}(\pi)\Big] \leq \frac{\gamma}{c_1}\cdot\mathbb{E}_{\pi\sim p}\big[D\big(\overline{M}_{\pi^\star}(\pi) \parallel \overline{M}(\pi)\big)\big] + \mathsf{dec}_{\gamma/c_1}(\mathrm{co}(\mathcal{M})).$$

Taking the expectation over $\pi^\star \sim \mu$, we conclude that

$$\mathsf{inf}^D_\gamma(\mathcal{M}) \leq \mathsf{dec}_{\gamma/c_1}(\mathrm{co}(\mathcal{M})).$$

□

E.3 High-Probability Exploration-By-Optimization and Information Ratio

Theorem 3.2. *For all $\eta > 0$,*

$$\mathsf{inf}^{\mathsf{H}}_{\eta^{-1}}(\mathcal{M}) \leq \mathsf{exo}_\eta(\mathcal{M}) \leq \mathsf{inf}^{\mathsf{H}}_{(8\eta)^{-1}}(\mathcal{M})$$

Proof of Theorem 3.2. We first prove the upper bound. We divide the proof into two parts as follows: **Upper bound: Minimax theorem.** We first use the minimax theorem to move to a Bayesian counterpart to the Exploration-by-Optimization objective. This requires some care to ensure

boundedness and compactness, but otherwise is conceptually straightforward. To begin, observe that we can write the Exploration-by-Optimization objective as

$$\begin{aligned}\mathsf{exo}_\eta(\mathcal{M}) &= \sup_{q\in\Delta(\Pi)} \inf_{p\in\Delta(\Pi), g\in\mathcal{G}} \sup_{M\in\mathcal{M},\pi^\star\in\Pi} [\Gamma_{q,\eta}(p,g\,;\pi^\star,M)] \\ &= \sup_{q\in\Delta(\Pi)} \inf_{p\in\Delta(\Pi), g\in\mathcal{G}} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)].\end{aligned}$$

Fix $\alpha \geq 1 \vee \eta^{-1}$ and $\varepsilon \in (0,1)$, and define

$$\mathcal{G}_\alpha = \{g\in\mathcal{G} \mid \|g\|_\infty \leq \alpha\}, \quad \text{and} \quad \mathcal{P}_\varepsilon = \big\{p\in\Delta(\Pi) \mid p(\pi)\geq \varepsilon|\Pi|^{-1} \ \ \forall\pi\big\}.$$

Then by restricting to these classes, we have[17]

$$\mathsf{exo}_\eta(\mathcal{M}) \leq \sup_{q\in\Delta(\Pi)} \inf_{p\in\mathcal{P}_\varepsilon, g\in\mathcal{G}_\alpha} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)].$$

We now verify that the conditions required to apply the minimax theorem are satisfied.

- The map $\mu \mapsto \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)]$ is linear. Furthermore, by Lemma E.1 (given below), the map $(p,g)\mapsto \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)]$ is jointly convex.
- Since we have restricted to $p\in\mathcal{P}_\varepsilon$ and $g\in\mathcal{G}_\alpha$, the value $\Gamma_{q,\eta}(p,g;\pi^\star,M)$ is uniformly bounded, as well as continuous with respect to p and g (so long as $\varepsilon>0$ and $\alpha<\infty$).
- The set $\Delta(\mathcal{M}\times\Pi)$ is convex. Since $|\Pi|<\infty$, the set $\mathcal{P}_\varepsilon\times\mathcal{G}_\alpha$ is convex and compact (for $\mathcal{P}_\varepsilon$ equipped with the usual topology and $\mathcal{G}_\alpha$ equipped with the product topology; see Lattimore and György [32] for details).

Hence, using Lemma C.2 we can bound by the value of the Bayesian game as follows:

$$\mathsf{exo}_\eta(\mathcal{M}) \leq \sup_{q\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\mathcal{P}_\varepsilon, g\in\mathcal{G}_\alpha} \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)]. \tag{42}$$

Upper bound: Moving to Hellinger distance. For any $q\in\Delta(\Pi)$, $\mu\in\Delta(\mathcal{M}\times\Pi)$, and $p\in\mathcal{P}_\varepsilon$, the value of the game in (42) is

$$\begin{aligned}&\mathbb{E}_{(M,\pi^\star)\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi^\star)-f^M(\pi)] \\ &\quad + \frac{1}{\eta}\inf_{g\in\mathcal{G}_\alpha}\mathbb{E}_{(M,\pi^\star)\sim\mu}\left[\mathbb{E}_{\pi\sim p, z\sim M(\pi)}\,\mathbb{E}_{\pi'\sim q}\exp\left(\frac{\eta}{p(\pi)}(g(\pi'\,;\pi,z)-g(\pi^\star\,;\pi,z))\right)-1\right].\end{aligned}$$

Let $\mathbb{P}$ be the law of the process $(M,\pi^\star)\sim\mu, \pi\sim p, z\sim M(\pi)$, and define $\mu_{\mathrm{pr}}(\pi') := \mathbb{P}(\pi^\star=\pi')$ and $\mu_{\mathrm{po}}(\pi';\pi,z) := \mathbb{P}(\pi^\star=\pi' \mid (\pi,z))$. Using Bayes' rule, we can rewrite the second term above as

$$\inf_{g\in\mathcal{G}_\alpha}\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\left[\mathbb{E}_{\pi'\sim q}\left[\exp\left(\eta\frac{g(\pi'\,;\pi,z)}{p(\pi)}\right)\right]\cdot\mathbb{E}_{\pi^\star\sim\mu_{\mathrm{po}}(\cdot\,;\pi,z)}\left[\exp\left(-\eta\frac{g(\pi^\star\,;\pi,z)}{p(\pi)}\right)\right]-1\right]$$

By reparameterizing via $g(\pi'\,;\pi,z)\leftarrow\frac{p(\pi)}{\eta}g(\pi'\,;\pi,z)$, the value is upper bounded by

$$\inf_{g\in\mathcal{G}_{\alpha\eta}}\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\big[\mathbb{E}_{\pi'\sim q}[\exp(g(\pi'\,;\pi,z))]\cdot\mathbb{E}_{\pi^\star\sim\mu_{\mathrm{po}}(\cdot\,;\pi,z)}[\exp(-g(\pi^\star\,;\pi,z))]-1\big].$$

Furthermore, by skolemizing, we can rewrite this as

$$V(p,q,\mu) := \mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\inf_{g:\Pi\to\mathbb{R},\|g\|_\infty\leq\alpha\eta}\big\{\mathbb{E}_{\pi'\sim q}[\exp(g(\pi'))]\cdot\mathbb{E}_{\pi^\star\sim\mu_{\mathrm{po}}(\cdot\,;\pi,z)}[\exp(-g(\pi^\star))]-1\big\}.$$

We now appeal to Lemma C.5, which grants that

$$V(p,q,\mu) \leq -\frac{1}{2}\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big]+4e^{-\alpha\eta}. \tag{43}$$

[17] Restricting to these sets allows us to enforce boundedness and continuity of the Exploration-by-Optimization objective, which is necessary to appeal to the minimax theorem. The parameters α and ε will not enter the final bound quantitatively.

Using (43), we can upper bound the value as

$$\mathsf{exo}_\eta(\mathcal{M}) \le \sup_{q\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\mathcal{P}_\varepsilon} \Bigg\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] - \frac{1}{2\eta}\, \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big] \Bigg\} + \frac{4}{\eta} e^{-\alpha\eta}.$$

In addition, since $f^M \in [0,1]$ and $D^2_{\mathsf{H}}(\cdot,\cdot) \in [0,2]$, we can further bound by

$$\sup_{q\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\Delta(\Pi)} \Bigg\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] - \frac{1}{2\eta}\, \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big] \Bigg\} + O(\eta^{-1}e^{-\alpha\eta} + \varepsilon\cdot(1+\eta^{-1})).$$

Since this expression only depends on α and ε through the additive approximation terms, taking the limit as $\alpha \to \infty$ and $\varepsilon \to 0$ yields

$$\mathsf{exo}_\eta(\mathcal{M}) \le \sup_{q\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\Delta(\Pi)} \Bigg\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] - \frac{1}{2\eta}\, \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big] \Bigg\}.$$

Finally, recall that since Hellinger distance satisfies the triangle inequality,

$$\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),\mu_{\mathrm{pr}})\big] \le 2\,\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big] + 2D^2_{\mathsf{H}}(\mu_{\mathrm{pr}},q).$$

Using that $\mu_{\mathrm{pr}}(\pi') = \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}[\mu_{\mathrm{po}}(\pi'\,;\pi,z)]$ and that squared Hellinger distance is convex, we have $D^2_{\mathsf{H}}(\mu_{\mathrm{pr}},q) \le \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big]$, and so

$$\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),\mu_{\mathrm{pr}})\big] \le 4\cdot\, \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big].$$

It follows that

$$\begin{aligned}\mathsf{exo}_\eta(\mathcal{M}) &\le \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\Delta(\Pi)} \Bigg\{ \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] - \frac{1}{8\eta}\, \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),\mu_{\mathrm{pr}})\big] \Bigg\} \\ &= \mathsf{inf}^{\mathsf{H}}_{1/8\eta}(\mathcal{M}).\end{aligned}$$

Lower bound. It is immediate (without invoking the minimax theorem) that

$$\begin{aligned}\mathsf{exo}_\eta(\mathcal{M}) &= \sup_{q\in\Delta(\Pi)} \inf_{p\in\Delta(\Pi),g\in\mathcal{G}} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)] \\ &\ge \sup_{q\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\Delta(\Pi),g\in\mathcal{G}} \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)].\end{aligned}$$

Performing the same sequence of calculations as in the upper bound, it holds that for any $q \in \Delta(\Pi)$, $\mu \in \Delta(\mathcal{M}\times\Pi)$, and $p \in \Delta(\Pi)$,

$$\begin{aligned}&\inf_{g\in\mathcal{G}} \mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(p,g\,;\pi^\star,M)] \\ &= \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] \\ &\quad + \eta^{-1} \inf_{g\in\mathcal{G}} \mathbb{E}_{(M,\pi^\star)\sim\mu}\Bigg[\mathbb{E}_{\pi\sim p, z\sim M(\pi)}\, \mathbb{E}_{\pi'\sim q} \exp\Bigg(\frac{\eta}{p(\pi)}(g(\pi'\,;\pi,z) - g(\pi^\star\,;\pi,z))\Bigg) - 1\Bigg] \\ &= \mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)] \\ &\qquad + \eta^{-1}\, \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|\pi} \inf_{g\in\mathcal{G}} \big\{\mathbb{E}_{\pi'\sim q}[\exp(g(\pi'))]\cdot \mathbb{E}_{\pi^\star\sim\mu_{\mathrm{po}}(\cdot\,;\pi,z)}[\exp(-g(\pi^\star))] - 1\big\},\end{aligned}$$

where we define μ_{pr} and μ_{po} as in the prequel. Using Lemma 2.1 yields

$$\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\inf_{g\in\mathcal{G}}\big\{\mathbb{E}_{\pi'\sim q}[\exp(g(\pi'))]\cdot\mathbb{E}_{\pi^\star\sim\mu_{\mathrm{po}}(\cdot\,;\pi,z)}[\exp(-g(\pi^\star))]-1\big\}$$
$$\geq\ -\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big].$$

We conclude that

$$\begin{aligned}\mathsf{exo}_\eta(\mathcal{M})\geq\sup_{q\in\Delta(\Pi)}\sup_{\mu\in\Delta(\mathcal{M}\times\Pi)}\inf_{p\in\Delta(\Pi)}&\Bigg\{\mathbb{E}_{(M,\pi^\star)\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi^\star)-f^M(\pi)]\\&\qquad-\frac{1}{\eta}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),q)\big]\Bigg\}\\\geq\sup_{\mu\in\Delta(\mathcal{M}\times\Pi)}\inf_{p\in\Delta(\Pi)}&\Bigg\{\mathbb{E}_{(M,\pi^\star)\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi^\star)-f^M(\pi)]\\&\qquad-\frac{1}{\eta}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{z|\pi}\big[D^2_{\mathsf{H}}(\mu_{\mathrm{po}}(\cdot\,;\pi,z),\mu_{\mathrm{pr}})\big]\Bigg\}\\=\mathsf{inf}^{\mathsf{H}}_{1/\eta}(\mathcal{M}).&\end{aligned}$$

□

Lemma E.1. *For any fixed $M\in\mathcal{M}$ and $\pi^\star\in\Pi$, the map $(p,g)\mapsto\Gamma_{q,\eta}(p,g\,;\pi^\star,M)$ is jointly convex with respect to $(p,g)\in\Delta(\Pi)\times\mathcal{G}$, where $\mathcal{G}:=(\Pi\times\Pi\times\mathcal{Z}\to\mathbb{R})$.*

Proof of Lemma E.1. Let $M\in\mathcal{M}$ and $\pi^\star\in\Pi$ be fixed. The map $p\mapsto\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)]$ is linear, so our main task is to show that the function

$$(p,g)\mapsto\sum_{\pi}p(\pi)\,\mathbb{E}_{z\sim M(\pi)}\left[\sum_{\pi'}q(\pi')\exp\left(\frac{\eta}{p(\pi)}(g(\pi'\,;\pi,z)-g(\pi^\star\,;\pi,z))\right)\right]$$

is jointly convex. We can rewrite this as

$$\sum_{\pi}q(\pi')\sum_{\pi}p(\pi)\,\mathbb{E}_{z\sim M(\pi)}\left[\exp\left(\frac{\eta}{p(\pi)}(g(\pi'\,;\pi,z)-g(\pi^\star\,;\pi,z))\right)\right].$$

Since convexity is preserved under summation with non-negative weights, it suffices to show that for any fixed (π,π'), the map

$$(p(\pi),g)\mapsto p(\pi)\,\mathbb{E}_{z\sim M(\pi)}\left[\exp\left(\frac{\eta}{p(\pi)}(g(\pi'\,;\pi,z)-g(\pi^\star\,;\pi,z))\right)\right]\tag{44}$$

is convex. Since the function $g\mapsto\mathbb{E}_{z\sim M(\pi)}[\exp(\eta(g(\pi'\,;\pi,z)-g(\pi^\star\,;\pi,z)))]$ is convex over $\mathcal{G}$, convexity for (44) follows from the following standard result. □

Proposition E.2 (Convexity of perspective transformation)**.** *Let $f:\mathbb{R}^d\to(-\infty,\infty)$ be a convex function. Then the function*

$$(x,t)\mapsto t\cdot f(x/t)$$

is convex over $\mathbb{R}^d\times\mathbb{R}_+$.

F Proofs for Main Results (Section 2)

F.1 Proof of Theorem 2.1

Theorem 2.1 (Main upper bound)**.** *For any choice of $\eta>0$, Algorithm 1 ensures that for all $\delta>0$, with probability at least $1-\delta$,*

$$\mathbf{Reg}_{\mathsf{DM}}\leq\mathsf{dec}_{1/8\eta}(\mathrm{co}(\mathcal{M}))\cdot T+\frac{2}{\eta}\cdot\log(|\Pi|/\delta).\tag{10}$$

In particular, for any $\delta>0$, with appropriate η, the algorithm ensures that with probability at least $1-\delta$,

$$\mathbf{Reg}_{\mathsf{DM}}\leq O(1)\cdot\inf_{\gamma>0}\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}))\cdot T+\gamma\cdot\log(|\Pi|/\delta)\}.\tag{11}$$

Proof of Theorem 2.1. Let us adopt convention $\langle p, f\rangle = \sum_\pi p(\pi)\cdot f(\pi)$ and let e_π denote the πth standard basis vector in $\mathbb{R}^\Pi$. For each $\pi^\star \in \Pi$, we write regret as

$$\mathbf{Reg}_{\mathsf{DM}}(\pi^\star) = \sum_{t=1}^T \mathbb{E}_{\pi\sim p^{(t)}}\left[f^{M^{(t)}}(\pi^\star) - f^{M^{(t)}}(\pi)\right] = \sum_{t=1}^T \left\langle e_{\pi^\star} - p^{(t)}, f^{M^{(t)}}\right\rangle.$$

Adding and subtracting $\sum_{t=1}^T \left\langle e_{\pi^\star} - q^{(t)}, \widehat{f}^{(t)}\right\rangle$, we rewrite this as

$$\sum_{t=1}^T \left\langle e_{\pi^\star} - p^{(t)}, f^{M^{(t)}}\right\rangle = \sum_{t=1}^T \left\langle e_{\pi^\star} - p^{(t)}, f^{M^{(t)}}\right\rangle + \sum_{t=1}^T \left\langle e_{\pi^\star} - q^{(t)}, \widehat{f}^{(t)}\right\rangle - \sum_{t=1}^T \left\langle e_{\pi^\star} - q^{(t)}, \widehat{f}^{(t)}\right\rangle. \tag{45}$$

The exponential weights update ensures (Lemma C.6) that with probability 1,

$$\begin{aligned}\sum_{t=1}^T \left\langle e_{\pi^\star} - q^{(t)}, \widehat{f}^{(t)}\right\rangle &\le \sum_{t=1}^T \left\langle q^{(t+1)} - q^{(t)}, \widehat{f}^{(t)}\right\rangle - \frac{1}{\eta}\sum_{t=1}^T D_{\mathsf{KL}}(q^{(t+1)} \,\|\, q^{(t)}) + \frac{D_{\mathsf{KL}}(e_{\pi^\star} \,\|\, q^{(1)})}{\eta} \\ &\le \sum_{t=1}^T \left\langle q^{(t+1)} - q^{(t)}, \widehat{f}^{(t)}\right\rangle - \frac{1}{\eta}\sum_{t=1}^T D_{\mathsf{KL}}(q^{(t+1)} \,\|\, q^{(t)}) + \frac{\log|\Pi|}{\eta}.\end{aligned}$$

In addition, using Lemma C.3, we have that for all t,

$$\left\langle q^{(t+1)}, \widehat{f}^{(t)}\right\rangle - \frac{1}{\eta} D_{\mathsf{KL}}(q^{(t+1)} \,\|\, q^{(t)}) \le \frac{1}{\eta}\log\left(\sum_\pi q^{(t)}(\pi)\exp\left(\eta\cdot \widehat{f}^{(t)}(\pi)\right)\right).$$

Hence, combining this with (45), we have

$$\mathbf{Reg}_{\mathsf{DM}}(\pi^\star) \le \sum_{t=1}^T \left\langle e_{\pi^\star} - p^{(t)}, f^{M^{(t)}}\right\rangle - \left\langle e_{\pi^\star}, \widehat{f}^{(t)}\right\rangle + \frac{1}{\eta}\sum_{t=1}^T \log\left(\sum_\pi q^{(t)}(\pi)\exp\left(\eta\cdot \widehat{f}^{(t)}(\pi)\right)\right) + \frac{\log|\Pi|}{\eta}.$$

Let $\mathscr{F}_t := \sigma(\pi^{(1)}, z^{(1)}, \ldots, \pi^{(t)}, z^{(t)})$ be a filtration, and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathscr{F}_t]$. For each $\pi \in \Pi$, define a sequence of random variables $\{X_t(\pi)\}_{t=1}^T$ via

$$X_t(\pi) = \frac{1}{\eta}\log\left(\sum_{\pi'} q^{(t)}(\pi')\exp\left(\eta\cdot \widehat{f}^{(t)}(\pi')\right)\right) - \left\langle e_{\pi^\star}, \widehat{f}^{(t)}\right\rangle.$$

Using Lemma C.1 and a union bound, we have that for any $\eta > 0$, with probability at least $1-\delta$, for all $\pi \in \Pi$

$$\sum_{t=1}^T X_t(\pi) \le \frac{1}{\eta}\sum_{t=1}^T \log(\mathbb{E}_{t-1}[\exp(\eta X_t(\pi))]) + \frac{\log(|\Pi|/\delta)}{\eta}.$$

Since this bounded holds uniformly for all π, we are guaranteed that with probability at least $1-\delta$, for all $\pi^\star \in \Pi$,

$$\mathbf{Reg}_{\mathsf{DM}}(\pi^\star) \le \sum_{t=1}^T \left\langle e_{\pi^\star} - p^{(t)}, f^{M^{(t)}}\right\rangle + \frac{1}{\eta}\sum_{t=1}^T \log(\mathbb{E}_{t-1}[\exp(\eta X_t(\pi^\star))]) + 2\frac{\log(|\Pi|/\delta)}{\eta}.$$

We compute that for any $\pi^\star \in \Pi$,

$$\begin{aligned}&\log(\mathbb{E}_{t-1}[\exp(\eta X_t(\pi^\star))]) \\ &= \log\left(\mathbb{E}_{\pi\sim p^{(t)}}\,\mathbb{E}_{z\sim M^{(t)}(\pi)}\,\mathbb{E}_{\pi'\sim q^{(t)}}\left[\exp\left(\frac{\eta}{p^{(t)}(\pi)}\cdot\left(g^{(t)}(\pi';\pi,z) - g^{(t)}(\pi^\star;\pi,z)\right)\right)\right]\right) \\ &\le \mathbb{E}_{\pi\sim p^{(t)}}\,\mathbb{E}_{z\sim M^{(t)}(\pi)}\,\mathbb{E}_{\pi'\sim q^{(t)}}\left[\exp\left(\frac{\eta}{p^{(t)}(\pi)}\cdot\left(g^{(t)}(\pi';\pi,z) - g^{(t)}(\pi^\star;\pi,z)\right)\right)\right] - 1,\end{aligned}$$

where we have used that $\log(x) \le x - 1$ for $x > 0$. Hence, with probability at least $1-\delta$, for all $\pi^\star \in \Pi$,

$$
\begin{aligned}
&\mathbf{Reg}_{\mathsf{DM}}(\pi^\star) \\
&\le \sum_{t=1}^{T} \big\langle e_{\pi^\star} - p^{(t)}, f^{M^{(t)}} \big\rangle + 2\frac{\log(|\Pi|/\delta)}{\eta} \\
&\quad + \frac{1}{\eta}\left(\mathbb{E}_{\pi\sim p^{(t)}} \mathbb{E}_{z\sim M^{(t)}(\pi)} \mathbb{E}_{\pi'\sim q^{(t)}} \left[\exp\left(\frac{\eta}{p^{(t)}(\pi)} \cdot \big(g^{(t)}(\pi'\,;\pi,z) - g^{(t)}(\pi^\star\,;\pi,z) \big) \right) \right] - 1 \right) \\
&= \sum_{t=1}^{T} \Gamma_{q^{(t)},\eta}(p^{(t)}, g^{(t)}\,; \pi^\star, M^{(t)}) + 2\frac{\log(|\Pi|/\delta)}{\eta} \\
&\le \mathsf{exo}_\eta(\mathcal{M}) \cdot T + 2\frac{\log(|\Pi|/\delta)}{\eta},
\end{aligned}
$$

where the last line uses that $(p^{(t)}, g^{(t)})$ are chosen to minimize the Exploration-By-Optimization objective. Finally, using Corollary 3.1, $\mathsf{exo}_\eta(\mathcal{M}) \le \mathsf{dec}_{(8\eta)^{-1}}(\mathrm{co}(\mathcal{M}))$.

□

F.2 Proof of Theorem 2.2

In this section we prove Theorem 2.2. Most of the work consists of proving an improved lower bound for the *stochastic* setting in which $M^{(t)} = M^\star$ is fixed across t (Theorem F.1). We then appeal to this stochastic lower bound with the class $\mathrm{co}(\mathcal{M})$. Since $\mathrm{co}(\mathcal{M})$ is equivalent to the set of mixtures of models in $\mathcal{M}$, this establishes existence of distribution $\mu \in \Delta(\mathcal{M})$ and mixture model $M_\mu = \mathbb{E}_{M\sim\mu}[M]$ for which regret in the stochastic setting must scale with $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M}))$. The proof concludes by arguing that this yields a lower bound for the adversarial setting when we sample $M^{(t)} \sim \mu$.

Throughout this section, we define the *one-sided variance* for a random variable Z as

$$\mathbb{V}_+[Z] := \mathbb{E}\big[(Z - \mathbb{E}[Z])_+^2\big].$$

Theorem 2.2 (Main lower bound)**.** *Let $C(T) := c \cdot \log(T \wedge V(\mathcal{M}))$ for a sufficiently large numerical constant $c > 0$. Set $\varepsilon_\gamma := \frac{\gamma}{4C(T)T}$. For any algorithm, there exists an oblivious adversary for which*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{E}(\mathbf{Reg}_{\mathsf{DM}})_+^2} \ge \Omega(1) \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T - O(T^{1/2}). \tag{15}$$

We also have the following slight variant of Theorem 2.2.

Theorem 2.2a. *Let $C(T) := c \cdot \log(T \wedge V(\mathcal{M}))$ for a sufficiently large numerical constant $c > 0$. Set $\varepsilon_\gamma := \frac{\gamma}{4C(T)T}$. For any algorithm, there exists an oblivious adversary for which $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \ge 0$ and*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \cdot T} \ge \Omega(1) \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T, \tag{46}$$

Proof of Theorem 2.2. We invoke Theorem F.1 with the model class $\mathrm{co}(\mathcal{M})$, which implies that there exists a distribution $\mu \in \Delta(\mathcal{M})$ for which

$$\mathbb{E}\big[\widetilde{\mathbf{Reg}}_{\mathsf{DM}}\big] + \sqrt{\mathbb{V}_+\big[\widetilde{\mathbf{Reg}}_{\mathsf{DM}}\big]} \ge L := 8^{-1} \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T,$$

where

$$\widetilde{\mathbf{Reg}}_{\mathsf{DM}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}} \mathbb{E}_{M\sim\mu}[f^M(\pi_\mu) - f^M(\pi^{(t)})],$$

and $\pi_\mu := \arg\max_{\pi\in\Pi} \mathbb{E}_{M\sim\mu}[f^M(\pi)]$, with the data generating process is (for each $t = 1, \ldots, T$):

- The learner samples $\pi^{(t)} \sim p^{(t)}$.
- Nature samples $z^{(t)} \sim \mathbb{E}_{M\sim\mu}[M(\pi^{(t)})]$.

Observe that this is equivalent in law to the following data-generating process, which constitutes an admissible adversary (with $M^{(t)} \in \mathcal{M}$):

- The learner samples $\pi^{(t)} \sim p^{(t)}$.
- Nature samples $M^{(t)} \sim \mu$ and $z^{(t)} \sim M^{(t)}(\pi^{(t)})$.

Likewise, we can equivalently write

$$\widetilde{\mathbf{Reg}}_{\mathsf{DM}} = \sum_{t=1}^{T} \mathbb{E}_{M^{(t)}\sim\mu}\, \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^{(t)}}(\pi_\mu) - f^{M^{(t)}}(\pi^{(t)})\right].$$

Hence, all that remains is to relate the quantity $\widetilde{\mathbf{Reg}}_{\mathsf{DM}}$ to the realized regret $\mathbf{Reg}_{\mathsf{DM}}$ for the sequence $M^{(1)}, \ldots M^{(T)}$, which entails removing the conditional expectation over $M^{(t)} \sim \mu$. To this end, we first observe that

$$\begin{aligned}
\mathbb{E}\big[\widetilde{\mathbf{Reg}}_{\mathsf{DM}}\big] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^{(t)}}(\pi_\mu) - f^{M^{(t)}}(\pi^{(t)})\right]\right] \\
&\leq \mathbb{E}\left[\max_{\pi^\star\in\Pi} \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^{(t)}}(\pi^\star) - f^{M^{(t)}}(\pi^{(t)})\right]\right] = \mathbb{E}\big[\mathbf{Reg}_{\mathsf{DM}}\big].
\end{aligned}$$

Next, note that since $\widetilde{\mathbf{Reg}}_{\mathsf{DM}}$ is non-negative, $\mathbb{V}_+\big[\widetilde{\mathbf{Reg}}_{\mathsf{DM}}\big] \leq \mathbb{E}\big[(\widetilde{\mathbf{Reg}}_{\mathsf{DM}})_+^2\big]$. Define

$$\widehat{\mathbf{Reg}}_{\mathsf{DM}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^{(t)}}(\pi_\mu) - f^{M^{(t)}}(\pi^{(t)})\right].$$

Then we have

$$\begin{aligned}
\mathbb{E}\big[(\widetilde{\mathbf{Reg}}_{\mathsf{DM}})_+^2\big] &\leq 2\,\mathbb{E}\big[(\widehat{\mathbf{Reg}}_{\mathsf{DM}})_+^2\big] + 2\,\mathbb{E}\big[(\widetilde{\mathbf{Reg}}_{\mathsf{DM}} - \widehat{\mathbf{Reg}}_{\mathsf{DM}})^2\big] \\
&\leq 2\,\mathbb{E}\big[(\mathbf{Reg}_{\mathsf{DM}})_+^2\big] + 2\,\mathbb{E}\big[(\widetilde{\mathbf{Reg}}_{\mathsf{DM}} - \widehat{\mathbf{Reg}}_{\mathsf{DM}})^2\big] \\
&\leq 2\,\mathbb{E}\big[(\mathbf{Reg}_{\mathsf{DM}})_+^2\big] + 2T,
\end{aligned}$$

where the first inequality uses that $\widehat{\mathbf{Reg}}_{\mathsf{DM}} \leq \mathbf{Reg}_{\mathsf{DM}}$ almost surely, and the second inequality uses (i) $f^M \in [0,1]$, and (ii) for any sequence of random variables $(Z_t)_{t=1}^T$ with $\mathbb{E}\big[Z_t \mid Z_1, \ldots, Z_{t-1}\big] = 0$, $\mathbb{E}\Big[\big(\sum_{t=1}^T Z_t\big)^2\Big] = \sum_{t=1}^T \mathbb{E}\big[Z_t^2\big]$. Putting everything together, we conclude that

$$\mathbb{E}\big[\mathbf{Reg}_{\mathsf{DM}}\big] + \sqrt{2\,\mathbb{E}\big[\big(\mathbf{Reg}_{\mathsf{DM}}\big)_+^2\big]} \geq L - \sqrt{2T}.$$

This proves Theorem 2.2. To prove Theorem 2.2a, we use that since $\widetilde{\mathbf{Reg}}_{\mathsf{DM}} \in [0, T]$,

$$\mathbb{V}_+\big[\widetilde{\mathbf{Reg}}_{\mathsf{DM}}\big] \leq T \cdot \mathbb{E}\big[\widetilde{\mathbf{Reg}}_{\mathsf{DM}}\big] \leq T \cdot \mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}].$$

□

The following result concerns the *stochastic setting* in Foster et al. [17]. Here, there is a (unknown) underlying model $M^\star \in \mathcal{M}$. For $t = 1, \ldots, T$, data is generated through the process:

- Learner samples $\pi^{(t)} \sim p^{(t)}$.
- Nature samples $z^{(t)} \sim M^\star(\pi^{(t)})$.

In addition, regret simplifies to

$$\mathbf{Reg}_{\mathsf{DM}} = \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] \tag{47}$$

For a fixed algorithm, let $\mathbb{P}^M$ denote the law of $\mathcal{H}^{(T)}$ when $M^\star = M$, and let $\mathbb{E}^M[\cdot]$ and $\mathbb{V}_+ \sup M[\cdot]$ denote the corresponding expectation non-negative variance. Recall that $V(\mathcal{M}) = \sup_{M,M'\in\mathcal{M}} \sup_{\pi\in\Pi} \sup_{A\in\mathscr{A}\otimes\mathscr{O}} \left\{\frac{M(A|\pi)}{M'(A|\pi)}\right\} \vee e$. Our main lower bound for the stochastic setting is as follows.

Theorem F.1. *Let $C(T) := 2^9 \log(T \wedge V(\mathcal{M}))$, and set $\varepsilon_\gamma = \frac{\gamma}{4C(T)T}$. For any algorithm, there exists a model in $\mathcal{M}$ for which*

$$\mathbb{E}^M[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{V}_+^M[\mathbf{Reg}_{\mathsf{DM}}]} \geq 8^{-1} \cdot \sup_{\gamma \geq 4\sqrt{C(T)T}} \sup_{\overline{M} \in \mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \cdot T.$$

The general structure of the lower bound follows that of Theorem 3.1 in Foster et al. [17], with the main difference being that we use a more refined change-of-measure argument to move from a "reference" model $\overline{M} \in \mathcal{M}$ to a worst-case alternative. Specifically, we replace Lemma A.11 in Foster et al. [17], which requires an almost sure bound on the random variables under consideration (in our case, regret), with Lemma C.4, which requires only boundedness of the second moment. Combining this with a self-bounding argument that takes advantage of the localized model class yields the result.

Proof of Theorem F.1. Throughout this proof we will use that $\mathbf{Reg}_{\mathsf{DM}}$ is non-negative in the stochastic setting, which can be seen by inspecting (47) (in the general adversarial setting, it is possible for $\mathbf{Reg}_{\mathsf{DM}}$ to be negative).

Let us introduce additional notation. For $M \in \mathcal{M}$, define $g^M(\pi) = f^M(\pi_M) - f^M(\pi)$, and for $p \in \Delta(\Pi)$, let $g^M(p) := \mathbb{E}_{\pi \sim p}[g^M(\pi)]$. Let $\widehat{p} := \frac{1}{T}\sum_{t=1}^T p^{(t)}$, and $p_M := \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^T p^{(t)}\right]$.

To begin, fix $\overline{M} \in \mathcal{M}$, $\gamma > 0$, and $\varepsilon > 0$, and set

$$M = \operatorname*{arg\,max}_{M \in \mathcal{M}_\varepsilon(\overline{M})} \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right].$$

Abbreviate $\mathsf{dec}_\gamma \equiv \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M})$. The definition of the DEC implies that

$$\begin{aligned}\mathsf{dec}_\gamma &\leq \mathbb{E}_{p_{\overline{M}}}[g^M(\pi)] - \gamma \cdot \mathbb{E}_{p_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \\ &= \mathbb{E}^{\overline{M}}[g^M(\widehat{p})] - \gamma \cdot \mathbb{E}_{p_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]. \end{aligned} \tag{48}$$

Change of measure. To proceed, we write

$$\begin{aligned}\mathbb{E}^{\overline{M}}[g^M(\widehat{p})] &= \mathbb{E}^{\overline{M}}\left[g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})]\right] + \mathbb{E}^{\overline{M}}\left[g^{\overline{M}}(\widehat{p})\right] + \mathbb{E}^M[g^M(\widehat{p})] \\ &\leq \mathbb{E}^{\overline{M}}\left[(g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})])_+\right] + \mathbb{E}^{\overline{M}}\left[g^{\overline{M}}(\widehat{p})\right] + \mathbb{E}^M[g^M(\widehat{p})]. \end{aligned} \tag{49}$$

We recall the following technical lemma.

Lemma C.4. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\mathcal{X}, \mathscr{F})$. Then for all functions $h : \mathcal{X} \to \mathbb{R}$,*

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2^{-1}(\mathbb{E}_{\mathbb{P}}[h^2(X)] + \mathbb{E}_{\mathbb{Q}}[h^2(X)]) \cdot D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q})}. \tag{27}$$

Defining $h(\widehat{p}) = (g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})])_+$, Lemma C.4 implies that

$$\begin{aligned}&\mathbb{E}^{\overline{M}}\left[(g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})])_+\right] \\ &\leq \mathbb{E}^M\left[(g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})])_+\right] + \sqrt{\left(\mathbb{E}^M[h(\widehat{p})^2] + \mathbb{E}^{\overline{M}}[h(\widehat{p})^2]\right) \cdot D^2_{\mathsf{H}}(\mathbb{P}^M, \mathbb{P}^{\overline{M}})} \\ &\leq \mathbb{E}^M[g^M(\widehat{p})] + \sqrt{\left(\mathbb{E}^M[h(\widehat{p})^2] + \mathbb{E}^{\overline{M}}[h(\widehat{p})^2]\right) \cdot D^2_{\mathsf{H}}(\mathbb{P}^M, \mathbb{P}^{\overline{M}})}, \end{aligned} \tag{50}$$

where we have used that $g^M, g^{\overline{M}} \geq 0$. We proceed to bound the second moment terms. For the first such term, we bound by the non-negative variance as follows:

$$\begin{aligned}\mathbb{E}^M\left[h(\widehat{p})^2\right] &= \mathbb{E}^M\left[(g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})])_+^2\right] \\ &\leq \mathbb{E}^M\left[(g^M(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})])_+^2\right] \\ &= \mathbb{V}_+^M[g^M(\widehat{p})]. \end{aligned} \tag{51}$$

where the first inequality uses that $g^{\overline{M}} \geq 0$. For the second variance term, we begin with the bound

$$\mathbb{E}^{\overline{M}}\left[h(\widehat{p})^2\right] = \mathbb{E}^{\overline{M}}\left[(g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^M[g^M(\widehat{p})])_+^2\right] \leq \mathbb{E}^{\overline{M}}\left[(g^M(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+^2\right].$$

We further bound this quantity by

$$\begin{aligned}
&\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+^2\big] \\
&\qquad = \mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+(f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) + f^{\overline{M}}(\widehat{p}) - f^{M}(\widehat{p}))_+\big] \\
&\qquad \le \mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+(f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}))_+\big] \\
&\qquad\qquad\qquad + \mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+(f^{\overline{M}}(\widehat{p}) - f^{M}(\widehat{p}))_+\big].
\end{aligned}$$

For the first term above, we have

$$\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+(f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}})_+\big] \le \varepsilon \cdot \mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+\big] \le \varepsilon \cdot \mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})],$$

where we have used the localization property and the fact that $g^{M}, g^{\overline{M}} \ge 0$. For the second term, using the AM-GM inequality yields

$$\begin{aligned}
&\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+(f^{\overline{M}}(\widehat{p}) - f^{M}(\widehat{p}))_+\big] \\
&\qquad \le \frac{1}{2}\,\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+^2\big] + \frac{1}{2}\,\mathbb{E}^{\overline{M}}\big[(f^{\overline{M}}(\widehat{p}) - f^{M}(\widehat{p}))_+^2\big] \\
&\qquad \le \frac{1}{2}\,\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+^2\big] + \frac{1}{2}\,\mathbb{E}_{\pi\sim p_{\overline{M}}}\big[(f^{M}(\pi) - f^{\overline{M}}(\pi))^2\big] \\
&\qquad \le \frac{1}{2}\,\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+^2\big] + \frac{1}{2}\,\mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big],
\end{aligned}$$

where the second line uses Jensen's inequality, and the last line uses that rewards are observed and bounded in $[0, 1]$. After combining these results and rearranging, we have

$$\mathbb{E}^{\overline{M}}\big[h(\widehat{p})^2\big] \le \mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}))_+^2\big] \le 2\varepsilon \cdot \mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big]. \quad (52)$$

To proceed, we recall Lemma A.13 from Foster et al. [17], which states that

$$D_{\mathsf{H}}^2\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) \le C(T) \cdot T \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big], \quad (53)$$

where $C(T) \le 2^8 \cdot \log(T \wedge V(\mathcal{M}))$. Combining this with the variance bounds above and (50), we have

$$\begin{aligned}
&\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^{M}[g^{M}(\widehat{p})])_+\big] \\
&\qquad \le \mathbb{E}^{M}[g^{M}(\widehat{p})] \\
&\qquad\qquad + \sqrt{\big(\mathbb{V}_+^{M}[g^{M}(\widehat{p})] + 2\varepsilon \cdot \mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big]\big) \cdot D_{\mathsf{H}}^2(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}})} \\
&\qquad \le \mathbb{E}^{M}[g^{M}(\widehat{p})] + \sqrt{2\mathbb{V}_+^{M}[g^{M}(\widehat{p})]} \\
&\qquad\qquad\qquad + \sqrt{\big(2\varepsilon \cdot \mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big]\big) \cdot D_{\mathsf{H}}^2(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}})} \\
&\qquad \le \mathbb{E}^{M}[g^{M}(\widehat{p})] + \sqrt{2\mathbb{V}_+^{M}[g^{M}(\widehat{p})]} + \sqrt{C(T)T} \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big] \\
&\qquad\quad + \sqrt{2\varepsilon\,\mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] \cdot C(T)T\,\mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big]},
\end{aligned}$$

where the second inequality uses that $D_{\mathsf{H}}^2(\cdot, \cdot) \le 2$ and the last inequality uses (53).

We observe that under the restriction $\varepsilon \le \frac{\gamma}{4TC(T)}$,

$$\begin{aligned}
&\sqrt{2\varepsilon \cdot \mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] \cdot C(T)T\,\mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big]} \\
&\qquad\qquad \le \sqrt{\mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] \cdot \frac{\gamma}{2} \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big]} \\
&\qquad\qquad \le \frac{1}{2}\,\mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] + \frac{\gamma}{4} \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D_{\mathsf{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big],
\end{aligned}$$

where the last line uses AM-GM inequality. The above yields that

$$\mathbb{E}^{\overline{M}}\big[(g^{M}(\widehat{p}) - g^{\overline{M}}(\widehat{p}) - \mathbb{E}^{M}[g^{M}(\widehat{p})])_+\big]$$

$$\leq \mathbb{E}^{M}[g^{M}(\widehat{p})] + \sqrt{2\mathbb{V}^{M}_{+}[g^{M}(\widehat{p})]} + (\sqrt{C(T)T} + \gamma/4) \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] + \frac{1}{2}\,\mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})].$$

Using (49), this implies that

$$\mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] \leq 2\,\mathbb{E}^{M}[g^{M}(\widehat{p})] + \mathbb{E}^{\overline{M}}\big[g^{\overline{M}}(\widehat{p})\big] + \sqrt{2\mathbb{V}^{M}_{+}[g^{M}(\widehat{p})]} + (\sqrt{C(T)T} + \gamma/4) \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] + \frac{1}{2}\,\mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})],$$

and after rearranging,

$$\mathbb{E}^{\overline{M}}[g^{M}(\widehat{p})] \leq 4\,\mathbb{E}^{M}[g^{M}(\widehat{p})] + 2\,\mathbb{E}^{\overline{M}}\big[g^{\overline{M}}(\widehat{p})\big] + \sqrt{8\mathbb{V}^{M}_{+}[g^{M}(\widehat{p})]} + 2(\sqrt{C(T)T} + \gamma/4) \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \tag{54}$$

Completing the proof. Combining (54) with (48) yields the bound

$$\mathsf{dec}_{\gamma} \leq 4\,\mathbb{E}^{M}[g^{M}(\widehat{p})] + 2\,\mathbb{E}^{\overline{M}}\big[g^{\overline{M}}(\widehat{p})\big] + \sqrt{8\mathbb{V}^{M}_{+}[g^{M}(\widehat{p})]} + \Big(2(\sqrt{C(T)T} + \gamma/4) - \gamma\Big) \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big].$$

In particular, whenever $\gamma \geq 4\sqrt{C(T)T}$, this implies that there exists an instance $M' \in \{M, \overline{M}\}$ for which

$$\mathbb{E}^{M'}\big[g^{M'}(\widehat{p})\big] + \sqrt{\mathbb{V}^{M'}_{+}[g^{M'}(\widehat{p})]} \geq 8^{-1} \cdot \mathsf{dec}_{\gamma}.$$

Finally, we observe that $g^{M'}(\widehat{p})$ is identical in law to $\mathbf{Reg}_{\mathsf{DM}}$ under $\mathbb{P}^{M'}$.

□

F.3 Proof of Theorem 2.3

Theorem 2.3. *Suppose there exists $M_0 \in \mathcal{M}$ such that f^{M_0} is a constant function, and that $|\Pi| < \infty$.*

1. *If there exists $\rho > 0$ such that $\lim_{\gamma\to\infty} \mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \cdot \gamma^{\rho} = 0$, then $\lim_{T\to\infty} \frac{\mathfrak{M}(\mathcal{M},T)}{T^{p}} = 0$ for $p < 1$.*
2. *If $\lim_{\gamma\to\infty} \mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \cdot \gamma^{\rho} > 0$ for all $\rho > 0$, then $\lim_{T\to\infty} \frac{\mathfrak{M}(\mathcal{M},T)}{T^{p}} = \infty$ for all $p < 1$.*

The same conclusion holds when $\Pi = \Pi_T$ grows with T, but has $\log|\Pi_T| = O(T^{q})$ for any $q < 1$.

Proof of Theorem 2.3. This proof closely follows that of Theorem 3.5 in Foster et al. [17].

Upper bound. Assume that $\lim_{\gamma\to\infty} \mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \cdot \gamma^{\rho} = 0$ for some $\rho > 0$, and that $\log|\Pi_T| = \widetilde{O}(T^{q})$ for some $q < 1$. Using Theorem 2.1 with $\delta = 1/T$, there exists an algorithm such that for each T, for all adversaries,

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq \widetilde{O}(\mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot \log|\Pi_T|) \leq \widetilde{O}(\mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot T^{q}),$$

with $\widetilde{O}(\cdot)$ hiding factors logarithmic in T. For each T, we set $\gamma = \gamma_T := T^{\frac{1-q}{1+\rho}}$; recall that $1 - q > 0$. The assumption that $\lim_{\gamma\to\infty} \mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \cdot \gamma^{\rho} = 0$, implies that for all $\varepsilon > 0$, there exists $\gamma' > 0$ such that $\mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M})) \leq \varepsilon/\gamma^{\rho}$ for all $\gamma \geq \gamma'$. For T sufficiently large, this implies that for all adversaries

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}\left(\frac{T}{\gamma_T^{\rho}} + \gamma_T \cdot T^{q}\right) = \widetilde{O}(T^{\frac{1+\rho q}{1+\rho}}).$$

Defining $p' := \frac{1}{2}(p+1) < 1$, this establishes that

$$\lim_{T\to\infty} \frac{\mathfrak{M}(\mathcal{M},T)}{T^{p'}} = 0.$$

Lower bound. Assume that $\lim_{\gamma\to\infty} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot \gamma^\rho = \infty$ for all $\rho > 0$ (this is equivalent to assuming that $\lim_{\gamma\to\infty} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot \gamma^\rho > 0$ for all $\rho > 0$, as in the theorem statement). Let $\rho \in (0, 1/2)$ be fixed. Using Theorem 2.2a, we are guaranteed that for any algorithm, there exists an adversary for which $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq 0$ and

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \cdot T} = \widetilde{\Omega}\big(\mathsf{dec}_{\gamma,\varepsilon(\gamma,T)}(\mathrm{co}(\mathcal{M})) \cdot T\big),$$

for all $\gamma = \omega(\sqrt{T\log(T)})$, where $\varepsilon(\gamma, T) := c \cdot \frac{\gamma}{T\log(T)}$ for a sufficiently small numerical constant $c \leq 1$. Since there exists $M_0 \in \mathcal{M}$ such that the function f^{M_0} is constant, Lemma B.1 of Foster et al. [17] further implies that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \cdot T} = \widetilde{\Omega}(\varepsilon(\gamma, T) \cdot \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T).$$

For each T, set $\gamma = \gamma_T := T$. By the assumption that $\lim_{\gamma\to\infty} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot \gamma^\rho = \infty$, it holds that for T sufficiently large, $\mathsf{dec}_{\gamma_T}(\mathrm{co}(\mathcal{M})) \geq \gamma_T^{-\rho}$, which implies that and

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \cdot T} = \widetilde{\Omega}\left(\frac{T}{\gamma_T^\rho}\right),$$

where we have used that $\varepsilon(\gamma_T, T) \propto \frac{1}{\log(T)}$. Rearranging, this implies that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] = \widetilde{\Omega}\big(T^{1-2\rho}\big).$$

Hence, for any $p \in (0, 1)$, by setting $\rho = \frac{1-p}{2} \in (0, 1/2)$, we have

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] = \widetilde{\Omega}(T^p).$$

Applying this argument with $p' = \frac{1}{2}(p + 1) \in (1/2, 1)$ yields

$$\lim_{T\to\infty} \frac{\mathfrak{M}(\mathcal{M}, T)}{T^p} = \infty.$$

□

F.4 Sub-Chebychev Algorithms

Proposition F.1. *Any random variable with $\mathbb{E}\big[X_+^2\big] \leq R$ has*

$$\mathbb{P}(X_+ > t) \leq \frac{R^2}{t^2}, \quad \forall t > 0.$$

Conversely, if $X \in (-\infty, B)$ and has $\mathbb{P}(X_+ > t) \leq \frac{R^2}{t^2}$ $\forall t > 0$, then

$$\mathbb{E}\big[X_+^2\big] \leq R^2(\log(B/R) + 1).$$

Proof of Proposition F.1. For the first direction, note that if $\mathbb{E}\big[X_+^2\big] \leq R$, Chebychev's inequality implies that for all $t > 0$,

$$\mathbb{P}\big(X_+^2 > t\big) \leq \frac{R^2}{t^2}. \tag{55}$$

For the other direction, since $X_+ \in [0, B]$ almost surely, we can bound by

$$\mathbb{E}\big[X_+^2\big] = \int_0^B \mathbb{P}(X_+ > t)tdt \leq R^2 + \int_R^B \mathbb{P}(X_+ > t)tdt \leq R^2 + R^2 \int_R^B \frac{1}{t}dt \leq R^2 + R^2\log(B/R).$$

□

Proposition F.2. *Suppose that for any $\delta > 0$, an algorithm (with δ as a parameter) ensures that with probability at least $1 - \delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq R\log^\rho(\delta^{-1})$$

for some $R \geq 1$ and $\rho > 0$. Then the algorithm, when invoked with parameter $\delta = 1/T^2$, is sub-Chebychev with parameter $5^{1/2}R\log^\rho(T)$.

Proof of Proposition F.2. Set $\delta = 1/T^2$. Then, since $|\mathbf{Reg}_{\mathsf{DM}}| \leq T$, the law of total expectation implies that

$$\mathbb{E}\big[(\mathbf{Reg}_{\mathsf{DM}})_+^2\big] \leq R^2 \log^{2\rho}(T^2) + T^2/T^2 \leq 5R^2 \log^{2\rho}(T),$$

where we have used that $R \geq 1$. Chebychev's inequality now implies that for all $t > 0$

$$\mathbb{P}((\mathbf{Reg}_{\mathsf{DM}})_+ \geq t) \leq \frac{\mathbb{E}\big[(\mathbf{Reg}_{\mathsf{DM}})_+^2\big]}{t^2} \leq \frac{5R^2 \log^{2\rho}(T)}{t}.$$

□

Corollary 2.1. Any regret minimization algorithm with sub-Chebychev parameter $R > 0$ must have

$$R \geq \widetilde{\Omega}(1) \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T - O(T^{1/2}). \tag{17}$$

Proof of Corollary 2.1. This result immediately follows from Proposition F.1, Proposition F.2, and Theorem 2.2. □