On the Complexity of Adversarial Decision Making

Anonymous Author(s)
Affiliation
Address
email

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. With a focus on stochastic environments, a recent line of research provides general structural conditions under which sample-efficient learning is possible, but robust learning guarantees for agnostic or adversarial settings have remained elusive. 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. [18] in the stochastic counterpart to our setting, is both necessary and sufficient for low regret in the adversarial setting. 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 [52] and the Exploration-by-Optimization objective of Lattimore and György [34].

1 Introduction

We consider the problem of robust data-driven decision making in bandits, reinforcement learning, and beyond. The last decade has seen development of data-driven decision algorithms with strong empirical performance in domains including robotics [28, 40], dialogue systems [38], and personalization [2, 57]. Reliably deploying data-driven decision making methods in safety-critical systems requires principled algorithms with provable robustness in the face of dynamic or even adversarial environments. Furthermore, for such algorithms to be applicable, they must effectively take advantage of problem structure as modeled by the practitioner. In high-dimensional problems, this means 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. In particular, with an emphasis on reinforcement learning, a burgeoning body of research identifies specific modeling assumptions under which sample-efficient interactive decision making is possible [13, 60, 22, 44, 6, 29, 15, 39, 14, 63], as well as general structural conditions that aim to unify these settings [50, 21, 56, 59, 16, 23, 18]. For dynamic or adversarial settings, however, comparatively little is known outside of (i) positive results for special cases such as adversarial bandit problems [5, 4, 20, 11, 1, 8, 27, 17, 9, 31], and (ii) a handful of negative results suggesting that online reinforcement learning in agnostic or adversarial settings can actually be statistically intractable

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

[53, 41]. These developments raise the following questions: (a) what are the underlying phenomena that determine the statistical complexity of decision making in adversarial settings? (b) what are the corresponding algorithmic design principles that attain optimal sample complexity?

Contributions. We consider an adversarial variant of the *Decision Making with Structured Observations* (DMSO) framework introduced in Foster et al. [18], where 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 best decision in hindsight. The 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 based on a *convexified* version of the *Decision-Estimation Coefficient* (DEC) complexity measure introduced in Foster et al. [18].
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, 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. [18]. The protocol consists 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 outcomes. This setting subsumes (adversarial) bandit problems, where models consist of reward functions/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), and 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:

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

This formulation generalizes Foster et al. [18], who considered the *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 [5, 4, 20], where $\Pi = \{1, \ldots, A\}$ and $\mathcal{F}_{\mathcal{M}} = \mathbb{R}^A$. Other well-studied examples include adversarial linear bandits [11, 1, 8], bandit convex optimization [27, 17, 9, 31], and nonparametric bandits [27, 7, 43].[1]
- **Reinforcement learning.** The adversarial DMSO framework encompasses finite-horizon, episodic online reinforcement learning, with each round t corresponding to a single episode: $\pi^{(t)}$ is a *policy* (a mapping from state to actions) to play in the episode, $r^{(t)}$ is the cumulative reward in the

[1] 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, semi-stochastic adversaries.

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 [21, 56, 22, 59, 16, 23, 18], but the adversarial setting we study has received less investigation. Examples include the setting in which the adversary chooses a sequence of tabular MDPs, which is known to be intractable [41], and the easier setting in which there is a fixed (known) MDP but rewards are adversarial [45, 64, 46, 24]. See Appendix D for more details.

We refer to Appendix B for additional measure-theoretic details and background, and to Foster et al. [18] for further examples and detailed discussion.[2]

Understanding sample complexity 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 [18] 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 $D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) = \int \big(\sqrt{d\mathbb{P}/d\nu} - \sqrt{d\mathbb{Q}/d\nu}\big)^2$. 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 equivalently 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,*

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \tag{3}$$

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 $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}(T)$ 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. [18] give upper and lower bounds that scale with $\mathsf{dec}_\gamma(\mathcal{M})$ (under appropriate technical assumptions; cf. Section 2.3). Hence, our results show that in general, the gap between optimal regret for stochastic and adversarial settings (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 are convex model classes (where $\mathrm{co}(\mathcal{M}) = \mathcal{M}$), which gives a post-hoc explanation for why these models are tractable in the adversarial setting. Finite state/action Markov decision processes are not a convex 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. [41].

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

1. a "parameterized" variant of the generalized information ratio of Lattimore and György [34].
2. a novel high-probability variant of the *Exploration-by-Optimization* of Lattimore and György [34].

Overall, while our results heavily draw on the work of Foster et al. [18] and Lattimore and György [34], we believe they play a valuable role in bridging these lines of research and formalizing connections.

Our techniques. On the lower bound side, we strengthen the lower bound from Foster et al. [18] with an improved change-of-measure argument (leading to improved results even in the stochastic

[2] 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.

setting), and combine this with the simple idea of choosing a static mixture model as the adversary. On the upper bound side, we extend the powerful Exploration-by-Optimization machinery of Lattimore and György [34] to the DMSO setting, and give a novel high-probability variant of the technique. 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).

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 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 [37, 34], we provide an upper bound on regret via 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), and use these results to give a characterization for learnability (Section 2.3). Finally, we discuss the gap between stochastic and adversarial decision making.

For the sake of keeping 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 substantially (e.g., with covering numbers as in Section 3.4 of Foster et al. [18]).

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: The algorithms and regret bounds based on the DEC that Foster et al. [18] 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 powerful *Exploration-by-Optimization* technique of Lattimore and Szepesvári [37], Lattimore and György [34], which provides an elegant approach to estimating rewards while exploiting the structure of the model class under consideration.

Exploration-by-Optimization—which was introduced in Lattimore and Szepesvári [37] and substantially expanded in Lattimore and György [34]—can be thought of as a generalization of the classical EXP3 algorithm [5], which we recall applies the exponential weights method for full-information online learning to a sequence of unbiased estimators for the rewards (formed via importance weighting). The naive reward estimator used by EXP3 is unsuitable for general model classes because it does 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 linear bandits. The idea behind Exploration-by-Optimization is to solve an optimization problem at each round to find a reward estimator and modified sampling distribution that better exploit the structure of the model class $\mathcal{M}$ for improved regret. Lattimore and György [34] showed that for a general partial monitoring setting (cf. Appendix A), the expected regret of this method—for exponential weights and a more general family of algorithms based on Bregman divergences—is bounded by a generalization of the information ratio [51, 52].

Our development builds on that of Lattimore and György [34], but we pursue *high-probability* guarantees rather than in-expectation guarantees.[3] While high-probability guarantees are useful in their own right, our motivation for studying such guarantees comes from the lower bound in the sequel (Section 2.2), which shows that the convexified Decision-Estimation Coefficient lower bounds the regret for algorithms with "reasonable" tail behavior. To develop high-probability regret bounds and complement this lower bound, we use a novel high-probability variant of the Exploration-by-Optimization objective and a specialized analysis which goes beyond the Bregman divergence framework.

[3] 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 [36].

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:

$$q^{(t)}(\pi) = \exp\Big(\eta \textstyle\sum_{i=1}^{t-1} \widehat{f}^{(i)}(\pi)\Big) \Big/ \sum_{\pi' \in \Pi} \exp\Big(\eta \textstyle\sum_{i=1}^{t-1} \widehat{f}^{(i)}(\pi')\Big). \tag{4}$$

4: Solve *high-probability exploration-by-optimization* objective: // See Eq. (7)

$$(p^{(t)}, g^{(t)}) \leftarrow \operatorname*{arg\,min}_{p \in \Delta(\Pi), g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^\star \in \Pi} \Gamma_{q^{(t)},\eta}(p, g\,; \pi^\star, M). \tag{5}$$

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

6: Form reward estimator:

$$\widehat{f}^{(t)}(\pi) = \frac{g^{(t)}(\pi; \pi^{(t)}, z^{(t)})}{p^{(t)}(\pi^{(t)})}. \tag{6}$$

Our algorithm, ExO^+, is displayed in Algorithm 1. At each round t, the algorithm proceeds by forming 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). Next, for the main step of the algorithm (Line 4), we obtain a sampling distribution $p^{(t)} \in \Delta(\Pi)$ and an *estimation function* $g^{(t)} \in \mathcal{G} := (\Pi \times \Pi \times \mathcal{X} \to \mathbb{R})$ by solving a minimax optimization problem based on a new objective we term *high-probability exploration-by-optmization*:

$$\begin{aligned}\Gamma_{q,\eta}(p, g\,; \pi^\star, M) := {} & \mathbb{E}_{\pi \sim p}[f^M(\pi^\star) - f^M(\pi)] \\ & + \eta^{-1} \cdot \mathbb{E}_{\pi \sim p, z \sim M(\pi)} \mathbb{E}_{\pi' \sim q}\left[\exp\left(\frac{\eta}{p(\pi)}(g(\pi'; \pi, z) - g(\pi^\star; \pi, z))\right) - 1\right].\end{aligned} \tag{7}$$

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 $\widehat{f}^{(t)}(\pi) := g^{(t)}(\pi; \pi^{(t)}, z^{(t)}) \big/ p^{(t)}(\pi^{(t)})$.

The interpretation of the high-probability Exploration-by-Optimization objective (7) 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 and optimal decision, respectively. By solving the minimax problem in (5), 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 $g(\cdot; \pi, z)/p(\pi)$. Here, tail behavior is captured by the MGF-like term in (7), which penalizes the learner for over-estimating rewards under the reference distribution q or under-estimating rewards under $\pi^\star$.

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 $1 - \delta$,*

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

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

$$\mathbf{Reg}_{\mathsf{DM}} \le O(1) \cdot \inf_{\gamma > 0}\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot \log(|\Pi|/\delta)\}. \tag{9}$$

This should be compared to the upper bound for the stochastic setting in Foster et al. [18] (e.g., Theorem 3.3), which takes a similar form, but scales with the weaker quantity $\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$.[4] See also Appendix A for a comparison to Lattimore and Szepesvári [37], Lattimore and György [34].

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

$$\mathsf{exo}_\eta(\mathcal{M}, q) := \inf_{p \in \Delta(\Pi), g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^\star \in \Pi} \Gamma_{q,\eta}(p, g\,; \pi^\star, M), \tag{10}$$

[4] If a proper estimator is available, Foster et al. [18] (Thm. 4.1) gives tighter bounds scaling with $\mathsf{dec}_\gamma(\mathcal{M})$.

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:

$$\mathsf{dec}_{(4\eta)^{-1}}(\mathrm{co}(\mathcal{M})) \le \mathsf{exo}_\eta(\mathcal{M}) \le \mathsf{dec}_{(8\eta)^{-1}}(\mathrm{co}(\mathcal{M})), \quad \forall \eta > 0. \tag{11}$$

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. This result builds on, but goes beyond the Bregman divergence-based framework in Lattimore and György [34], and exploits 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 value of 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*

$$\frac{1}{2} D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) \le \sup_{g:\mathcal{X}\to\mathbb{R}} \left\{1 - \mathbb{E}_{\mathbb{P}}\left[e^{g}\right] \cdot \mathbb{E}_{\mathbb{Q}}\left[e^{-g}\right]\right\} \le D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}). \tag{12}$$

The lower inequality in Lemma 2.1 is proven using a trick similar to one used by Zhang [62] to prove high-probability bounds for maximum likelihood estimation based on Hellinger distance. In the process of proving (11), 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).

Further remarks. The main focus of this work is sample complexity, 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. [18] of obtaining in the frequentist setting a regret bound matching the Bayesian regret bound in their Theorem 3.6.

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

$$\mathcal{M}_\varepsilon(\overline{M}) = \left\{M \in \mathcal{M} : f^{\overline{M}}(\pi_{\overline{M}}) \ge f^{M}(\pi_{M}) - \varepsilon\right\},$$

and let $\mathsf{dec}_{\gamma,\varepsilon}(\mathcal{M}) := \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M})$ be the *localized Decision-Estimation Coefficient*. Let $(x)_+ := \max\{x, 0\}$ and $V(\mathcal{M}) := \sup_{M,M'\in\mathcal{M}} \sup_{\pi\in\Pi} \sup_{A\in\mathscr{A}\otimes\mathscr{O}} \left\{\frac{M(A|\pi)}{M'(A|\pi)}\right\} \vee e$.[5]

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*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] + \sqrt{\mathbb{E}(\mathbf{Reg}_{\mathsf{DM}})^2_+} \ge \Omega(1) \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T - O(T^{1/2}). \tag{13}$$

Theorem 2.2 implies that for any algorithm with "reasonable" tail behavior beyond what is granted by control of the first moment (such as Algorithm 1), 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)**.** *We say that a regret minimization algorithm is sub-Chebychev with parameter R if for all $t > 0$,*

$$\mathbb{P}((\mathbf{Reg}_{\mathsf{DM}})_+ \ge t) \le R^2/t^2. \tag{14}$$

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.

[5] Recall (Appendix B) that $M(\cdot, \cdot \mid \pi)$ is the conditional distribution given π; finiteness of $V(\mathcal{M})$ is not necessary, but removes a $\log(T)$ factor from Theorem 2.2.

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

$$R \geq \widetilde{\Omega}(1) \cdot \sup_{\gamma > \sqrt{2C(T)T}} \mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T - O(T^{1/2}). \tag{15}$$

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 all non-trivial classes.[6] In this case, one can show (cf. Proposition F.2 for a proof) that by setting $\delta = 1/T^2$, Theorem 2.1 implies that Algorithm 1 is sub-Chebychev with parameter

$$R = \widetilde{O}\Big(\inf_{\gamma > 0}\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot \log(|\Pi|)\}\Big) = \widetilde{O}(\mathsf{dec}_{\gamma_{\mathrm{u}}}(\mathrm{co}(\mathcal{M})) \cdot T), \tag{16}$$

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 (15) can be shown to scale with

$$R \geq \widetilde{\Omega}\Big(\mathsf{dec}_{\gamma_\ell,\varepsilon_{\gamma_\ell}}(\mathrm{co}(\mathcal{M})) \cdot T\Big), \tag{17}$$

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.1 that for the multi-armed bandit problem with $\Pi = \{1, \ldots, A\}$, the upper bound in (16) yields $R = O(\sqrt{AT\log A})$, while the lower bound in (17) yields $R = \Omega(\sqrt{AT})$. See Appendix D for additional examples which further illustrate the scaling above.

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. [18]). As shown in Foster et al. [18], localization is inconsequential for essentially all model classes commonly studied in the literature, and 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. [18] 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 [33].

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, stochastic adversary, we believe the more surprising result here is that good behavior of the convexified DEC is also *sufficient* for low regret.

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) in the adversarial setting. This extends the learnability result for the stochastic setting in Foster et al. [18], and follows a long tradition of such characterizations in learning theory [58, 3, 54, 49, 12]. To state the result, we define the minimax regret 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 $\rho > 0$ s.t. $\lim_{\gamma\to\infty} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot \gamma^\rho = 0$, 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 $p < 1$.*

[6] Note that the dominant term $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathrm{co}(\mathcal{M})) \cdot T$ in (13) scales with $\sqrt{T}$ any "non-trivial" class that embeds the two-armed bandit problem, so that the $-O(T^{1/2})$ term can be discarded.

The same conclusion holds when $\Pi = \Pi_T$ grows with T, but has $\log|\Pi_T| = O(T^q)$ for any $q < 1$.[7]

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 to perform rescaling tricks to remove these conditions.

Comparison to stochastic setting. Having shown that the convexified Decision-Estimation Coefficient plays a fundamental role in determining the optimal regret for the adversarial DMSO setting, now is a good time to make comparisons to the stochastic setting. There, Foster et al. [18] obtain upper bounds on regret that have the same form as (9), but scale with the weaker quantity $\max_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$.[8] For classes that are not convex, but where "proper" estimators are available (e.g., tabular MDPs), the upper bounds in Foster et al. [18] 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 MDPs with horizon H, S states, and A actions, Foster et al. [18] show that $\mathsf{dec}_\gamma(\mathcal{M}) = \mathrm{poly}(H, S, A)/\gamma$, and use this to obtain regret $\sqrt{\mathrm{poly}(H, S, A) \cdot T}$. 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 [30, 41]. Indeed, we show (Appendix D) that $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \geq \Omega(A^{\min\{S,H\}})$. This example highlights that in general, the gap between stochastic and adversarial 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 from 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 notion of *generalized information ratio* introduced by Lattimore and György [34], Lattimore [32] which extends the original information ratio of Russo and Van Roy [51, 52]. 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 recall the definition of the generalized information ratio from Lattimore [32], which we state here for a general divergence-like function $D(\cdot \,\|\, \cdot) \to \mathbb{R}^+$ (typically, KL divergence or another Bregman divergence). For a distribution $\mu \in \Delta(\mathcal{M} \times \Pi)$ and decision distribution $p \in \Delta(\Pi)$, define $\mu_{\mathrm{pr}}(\pi') := \mathbb{P}(\pi^\star = \pi')$ and $\mu_{\mathrm{po}}(\pi'; \pi, z) := \mathbb{P}(\pi^\star = \pi' \mid (\pi, z))$, where $\mathbb{P}$ is the law of the process $(M, \pi^\star) \sim \mu, \pi \sim p, z \sim M(\pi)$. μ_{pr} should be thought of as the prior over $\pi^\star$, and μ_{po} as the posterior having observed (z, π); note that the law μ_{po} does not depend on the distribution p. For parameter $\lambda > 1$, Lattimore [33] defines the generalized information ratio for a class $\mathcal{M}$ via[9]

$$\Psi_\lambda(\mathcal{M}) = \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\Delta(\Pi)} \left\{ \frac{(\mathbb{E}_{(M,\pi^\star)\sim\mu}\, \mathbb{E}_{\pi\sim p}[f^M(\pi^\star) - f^M(\pi)])^\lambda}{\mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|p}[D(\mu_{\mathrm{po}}(\cdot; \pi, z) \,\|\, \mu_{\mathrm{pr}})]} \right\}. \tag{18}$$

Here, we have slightly generalized the original definition in Lattimore [33] by incorporating models in $\mathcal{M}$ rather than placing an arbitrary prior over observations z directly. We also use a general divergence, while Lattimore [33] uses KL divergence and Lattimore and György [34] use Bregman divergences.

To understand the connection to the Decision-Estimation Coefficient, it will be helpful introduce another variant of the information ratio that we call the *parameterized information ratio*.

Definition 3.1. *For a divergence $D(\cdot \,\|\, \cdot)$, the* parameterized information ratio *is given by*

$$\begin{aligned} &\mathsf{inf}^D_\gamma(\mathcal{M}) \qquad (19)\\ &= \sup_{\mu\in\Delta(\mathcal{M}\times\Pi)} \inf_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p}\big[\mathbb{E}_{(M,\pi^\star)\sim\mu}[f^M(\pi^\star) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{z|p}[D(\mu_{\mathrm{po}}(\cdot; \pi, z) \,\|\, \mu_{\mathrm{pr}})]\big]. \end{aligned}$$

[7] Allowing Π to grow with T can be used to handle infinite decision spaces using covering arguments.

[8] Theorem 3.1 of Foster et al. [18] attains $\mathbf{Reg}_{\mathsf{DM}} \lesssim \inf_{\gamma>0}\{\max_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) + \gamma \cdot \log|\mathcal{M}|\}$.

[9] Lattimore and György [34] give a slightly different but essentially equivalent definition; cf. Appendix E.

The parameterized information ratio is always bounded by the generalized information ratio in (18); in particular, we have $\mathsf{inf}^D_\gamma(\mathcal{M}) \leq (\Psi_\lambda(\mathcal{M})/\gamma)^{\frac{1}{\lambda-1}}\ \forall\gamma > 0$. All of the regret bounds based on the generalized information ratio that we are aware of [34, 33] 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 (18) is equivalent to requiring that a single distribution p certify a certain bound on the value in (19) 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 $\gamma > 0$, $\mathsf{inf}^{\mathsf{H}}_\gamma(\mathcal{M}) \leq \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \leq \mathsf{inf}^{\mathsf{H}}_{\gamma/4}(\mathcal{M})$.*

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 basic idea is to use Bayes' rule 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.[10]

Proposition 3.1. *For any divergence-like function $D(\cdot \,\|\, \cdot) : \Delta(\Pi) \times \Delta(\Pi) \to \mathbb{R}_+$, 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, and the equivalence of the DEC and Exploration-by-Optimization in (11) 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 seems unlikely that a similar equivalence can be established using the generalized information ratio in (18). We note in passing that one can use similar techniques to lower bound the Bregman divergence-based Exploration-by-Optimization objective in Lattimore and György [34] 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.

Our results add to a growing body of research which shows that online reinforcement learning with agnostic or adversarial outcomes can be statistically intractable [53, 41]. A promising future direction is to extend our techniques to natural semi-adversarial models in which reinforcement learning is tractable (for example, the so-called *adversarially corrupted* setting [42, 19]). Other interesting questions include (i) extending our lower bounds beyond the observable-reward setting and to directly handle expected regret, and (ii) developing computationally efficient algorithms for large decision spaces.

[10] The variants in Lattimore and György [34], Lattimore [33] 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] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, S. Sen, and A. Slivkins. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2016.

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

[19] A. Gupta, T. Koren, and K. Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

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

[28] J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

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

[30] 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.

[31] T. Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *arXiv preprint arXiv:2006.00475*, 2020.

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

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

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

[35] 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.

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

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

[38] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. Deep reinforcement learning for dialogue generation. In *EMNLP*, 2016.

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

[40] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

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

[42] T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

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

[49] 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.

[50] 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.

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

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

[53] 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.

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

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

[56] 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.

[57] A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 2017.

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

[59] 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.

[60] 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.

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

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

[63] 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.

[64] 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]