The Statistical Complexity of Interactive Decision Making

Dylan J. Foster dylanfoster@microsoft.com
Sham M. Kakade sham@seas.harvard.edu
Jian Qian jianqian@mit.edu
Alexander Rakhlin rakhlin@mit.edu

Abstract

A fundamental challenge in interactive learning and decision making, ranging from bandit problems to reinforcement learning, is to provide sample-efficient, adaptive learning algorithms that achieve near-optimal regret. This question is analogous to the classical problem of optimal (supervised) statistical learning, where there are well-known complexity measures (e.g., VC dimension and Rademacher complexity) that govern the statistical complexity of learning. However, characterizing the statistical complexity of interactive learning is substantially more challenging due to the adaptive nature of the problem. The main result of this work provides a complexity measure, the *Decision-Estimation Coefficient*, that is proven to be both *necessary* and *sufficient* for sample-efficient interactive learning. In particular, we provide:

- a lower bound on the optimal regret for *any* interactive decision making problem, establishing the Decision-Estimation Coefficient as a fundamental limit.
- a unified algorithm design principle, *Estimation-to-Decisions* (E2D), which transforms any algorithm for supervised estimation into an online algorithm for decision making. E2D attains a regret bound matching our lower bound, thereby achieving optimal sample-efficient learning as characterized by the Decision-Estimation Coefficient.

Taken together, these results constitute a theory of learnability for interactive decision making. When applied to reinforcement learning settings, the Decision-Estimation Coefficient recovers essentially all existing hardness results and lower bounds. More broadly, the approach can be viewed as a decision-theoretic analogue of the classical Le Cam theory of statistical estimation; it also unifies a number of existing approaches—both Bayesian and frequentist.

Contents

1 Introduction **3**
1.1 Framework: Decision Making with Structured Observations . . . 4
1.2 The Decision-Estimation Coefficient . . . 6
1.3 Contributions . . . 7
1.4 Organization . . . 8

2 Preliminaries **8**
2.1 Minimax Regret: Frequentist and Bayesian . . . 10

3 A Theory of Learnability for Interactive Decision Making **10**
3.1 Lower Bound: The Decision-Estimation Coefficient is a Fundamental Limit . . . 11
3.2 E2D: A Unified Meta-Algorithm for Interactive Decision Making . . . 16
3.3 Learnability . . . 19
3.4 Tighter Guarantees Based on Decision Space Complexity . . . 20
3.5 Discussion . . . 21

4 The E2D Meta-Algorithm: General Toolkit **22**
4.1 Guarantees for General Online Estimation Oracles . . . 23
4.2 Dual Perspective and Connection to Posterior Sampling . . . 24
4.3 General Divergences and Randomized Estimators . . . 26

5 Illustrative Examples **27**
5.1 Multi-Armed Bandits . . . 28

5.2 Tabular Reinforcement Learning . . . 32
5.3 Discussion . . . 39

6 Application to Bandits **39**
6.1 Familiar Examples . . . 39
6.2 Disagreement Coefficient, Star Number, and Eluder Dimension . . . 43

7 Application to Reinforcement Learning **44**
7.1 Bilinear Classes: Basic Results . . . 45
7.2 Bilinear Classes: Refined Guarantees . . . 53
7.3 Lower Bounds . . . 54

8 Incorporating Contextual Information **56**
8.1 The Contextual E2D Algorithm . . . 57
8.2 Application to Contextual Bandits . . . 58
8.3 Application to Contextual Reinforcement Learning . . . 59

9 Related Work **60**
9.1 Statistical Estimation . . . 60
9.2 Instance-Dependent Complexity for Structured Bandits . . . 60
9.3 Posterior Sampling and the Information Ratio . . . 61

10 Discussion **63**

A Technical Tools **72**
A.1 Tail Bounds . . . 72
A.2 Information Theory . . . 73
A.3 Online Learning . . . 77

B Decision-Estimation Coefficient: Structural Results **86**

C Proofs from Section 3 **88**
C.1 Proofs for Lower Bounds . . . 88
C.2 Proofs for Upper Bounds . . . 93
C.3 Proofs for Learnability Results . . . 97
C.4 Additional Proofs . . . 98

D Proofs from Section 4 **99**
D.1 Proofs from Section 4.1 . . . 99
D.2 Proofs from Section 4.2 . . . 102
D.3 Proofs from Section 4.3 . . . 103

E Proofs from Section 6 **103**
E.1 Proofs from Section 6.1 . . . 103
E.2 Proofs from Section 6.2 . . . 107

F Proofs and Additional Results from Section 7 **109**
F.1 Technical Tools . . . 109
F.2 Proofs from Section 7.1 . . . 111
F.3 Proofs from Section 7.2 . . . 116
F.4 Proofs from Section 7.3 . . . 121
F.5 Additional Results . . . 126

G Additional Proofs **128**
G.1 Proofs from Section 1 . . . 128
G.2 Proofs from Section 5 . . . 129
G.3 Proofs from Section 8 . . . 133
G.4 Proofs from Section 9 . . . 133

1 Introduction

Over the last decade, algorithms for data-driven decision making (in particular, contextual bandits and reinforcement learning) have achieved impressive empirical results in application domains ranging from online personalization (Agarwal et al., 2016; Tewari and Murphy, 2017) to game-playing (Mnih et al., 2015; Silver et al., 2016), robotics (Kober et al., 2013; Lillicrap et al., 2015), and dialogue systems (Li et al., 2016). Algorithm design and sample complexity for data-driven decision making have a relatively complete theory for problems with small state and action spaces or short horizon. However, many of the most compelling applications necessitate long-term planning in high-dimensional spaces, where function approximation is required, and where exploration, generalization, and sample efficiency remain major challenges. For real-world problems, where data is limited, it is critical that we bridge this gap and develop sample-efficient decision making methods capable of exploring large state and action spaces.

With a focus on reinforcement learning, a growing body of research identifies specific settings in which sample-efficient interactive decision making is possible, typically under conditions that control the interplay between system dynamics and function approximation (Dean et al., 2020; Yang and Wang, 2019; Jin et al., 2020c; Modi et al., 2020; Ayoub et al., 2020; Krishnamurthy et al., 2016; Du et al., 2019a; Li, 2009; Dong et al., 2019; Zhou et al., 2021). While these results highlight a number of important special cases (e.g., linear function approximation), it is desirable from a practical perspective to develop algorithms that accommodate *general-purpose* function approximation, similar to what one expects in supervised, statistical learning. This leads to significant research challenges:

1. *Sample complexity and fundamental limits.* In statistical learning, the classical Vapnik-Chervonenkis (VC) theory provides complexity measures (e.g., VC dimension and Rademacher complexity) that upper bound the number of samples required to achieve a desired accuracy level, as well as fundamental limits. In data-driven decision making, we lack general tools that can be systematically applied to understand sample complexity for new problem domains. As a result, it is often far from obvious whether existing algorithms are optimal, or to what extent they can be improved.

2. *Algorithmic principles.* Can we design general-purpose algorithms for data-driven decision making that can take any class of models or policies as input and produce accurate decisions out of the box? In statistical learning we have universal algorithmic principles such as empirical risk minimization (taking the function or model that best fits the data) that can be applied to any problem. In data-driven decision making, algorithm design has largely proceeded on a case-by-case basis, and developing reliable, provable algorithms for even the simplest models often requires non-trivial mathematical insights.

Toward a general theory, a recent line of research proposes structural conditions that attempt to unify existing approaches to sample-efficient reinforcement learning (Russo and Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020; Du et al., 2021; Jin et al., 2021). These conditions are not known to be necessary for learning, and they do not recover existing hardness results (Weisz et al., 2021; Wang et al., 2021). In fact, we do not yet have a unified understanding of sample complexity and algorithm design even for the basic bandit problem (that is, reinforcement learning with horizon one) when the action space is structured and high-dimensional.

We address issues (1) and (2) through a two-pronged approach. We introduce a general framework for interactive, online decision making, *Decision Making with Structured Observations*, which subsumes structured (high-dimensional) bandits, reinforcement learning, partially observed Markov decision processes (POMDPs), and beyond. We provide a new complexity measure, the *Decision-Estimation Coefficient* (DEC) and show that it is a fundamental limit that lower bounds the sample complexity for any interactive decision making problem. We complement this result with a universal algorithm design principle, *Estimation-to-Decisions* (E2D), which achieves optimal sample complexity as characterized by the Decision-Estimation Coefficient and leads to new, efficient algorithms. Together, these results provide the first theory of learnability for interactive decision making.

1.1 Framework: Decision Making with Structured Observations

We consider a general framework for interactive decision making, which we refer to as *Decision Making with Structured Observations* (DMSO). The protocol proceeds in T rounds, where for 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 *reward* $r^{(t)} \in \mathcal{R}$ and *observation* $o^{(t)} \in \mathcal{O}$ based on the decision, where $\mathcal{R} \subseteq \mathbb{R}$ is the *reward space* and $\mathcal{O}$ is the *observation space*. The reward and observation are then observed by the learner.

We focus on a stochastic variant of the DMSO framework: At each timestep, the pair $(r^{(t)}, o^{(t)})$ is drawn independently from an unknown distribution $M^\star(\pi^{(t)})$, where $M^\star : \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ is a *model* that maps decisions to distributions over outcomes. To facilitate the use of learning and function approximation, we assume the learner has access to a *model class* $\mathcal{M}$ that attempts to capture the model $M^\star$. Depending on the problem domain, $\mathcal{M}$ might consist of linear models, neural networks, random forests, or other complex function approximators. We make the following standard realizability assumption (Agarwal et al., 2012; Foster and Rakhlin, 2020; Du et al., 2021), which asserts that $\mathcal{M}$ is flexible enough to express the true model.

Assumption 1.1 (Realizability)**.** *The model class $\mathcal{M}$ contains the true model $M^\star$.*

For a model $M \in \mathcal{M}$, let $\mathbb{E}^{M,\pi}[\cdot]$ denote the expectation under $(r, o) \sim M(\pi)$. Further, let $f^M(\pi) := \mathbb{E}^{M,\pi}[r]$ denote the mean reward function and $\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$ denote the decision with the greatest expected reward. Finally, we define $\Pi_\mathcal{M} := \{\pi_M \mid M \in \mathcal{M}\}$ as the induced set of (potentially) optimal decisions and define $\mathcal{F}_\mathcal{M} := \{f^M \mid M \in \mathcal{M}\}$ as the induced class of mean reward functions.

We evaluate the learner's performance in terms of *regret* to the optimal decision for $M^\star$:

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

where we abbreviate $f^\star = f^{M^\star}$ and $\pi^\star = \pi_{M^\star}$, and where $p^{(t)} \in \Delta(\Pi)$ is the learner's distribution over decisions at round t.[1]

In spite of the apparent simplicity, the DMSO framework is general enough to capture most online decision making problems. We focus on two special cases: structured bandits and reinforcement learning.

Example 1.1 (Structured bandits)**.** When there are no observations (i.e., $\mathcal{O} = \{\varnothing\}$), the DMSO framework is equivalent to the well-known *structured bandit* problem (Combes et al., 2017; Degenne et al., 2020; Jun and Zhang, 2020). Here, adopting standard terminology, $\pi^{(t)}$ is referred to as an *action* or *arm* (rather than a decision), and Π is referred to as the *action space*. The structured bandit problem is simplest special case of our framework, but is extremely rich, and includes the following well-studied problems:

- The classical finite-armed bandit problem with A actions, where $\Pi = \{1, \ldots, A\}$ and $\mathcal{F}_\mathcal{M} = \mathbb{R}^A$ (Lai and Robbins, 1985b; Burnetas and Katehakis, 1996; Kaufmann et al., 2016; Garivier et al., 2019).
- The linear bandit problem, in which $\Pi \subseteq \mathbb{R}^d$ and $\mathcal{F}_\mathcal{M}$ is a class of linear functions (Abe and Long, 1999; Auer et al., 2002a; Dani et al., 2008; Chu et al., 2011; Abbasi-Yadkori et al., 2011).
- Bandit convex optimization (or, zeroth order optimization), where $\Pi \subseteq \mathbb{R}^d$ and $\mathcal{F}_\mathcal{M}$ is a class of concave[2] functions (Kleinberg, 2004; Flaxman et al., 2005; Agarwal et al., 2013; Bubeck and Eldan, 2016; Bubeck et al., 2017; Lattimore, 2020).
- Nonparametric bandits, where $\mathcal{F}_\mathcal{M}$ is a class of Lipschitz or Hölder functions over a metric space (Kleinberg, 2004; Auer et al., 2007; Kleinberg et al., 2019; Bubeck et al., 2011; Magureanu et al., 2014; Combes et al., 2017).

[1] Our results are most conveniently stated in terms of this notion of regret, but immediately extend to the empirical regret $\sum_{t=1}^{T} r^{(t)}(\pi^\star) - r^{(t)}(\pi^{(t)})$ via standard concentration arguments.

[2] Here, $\mathcal{F}_\mathcal{M}$ is concave rather than convex because we work with rewards rather than losses.

Despite significant research effort, there is no general theory characterizing what properties of the action space and model class determine the minimax rates for the structured bandit problem. Outside of the asymptotic regime (Combes et al., 2017; Degenne et al., 2020; Jun and Zhang, 2020), progress has proceeded case by case. $\triangleleft$

A more challenging setting is reinforcement learning (with function approximation).

Example 1.2 (Online reinforcement learning)**.** We consider an episodic finite-horizon reinforcement learning setting. With H denoting the horizon, each model $M \in \mathcal{M}$ specifies a non-stationary Markov decision process $M = \left\{\{\mathcal{S}_h\}_{h=1}^H, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\right\}$, where $\mathcal{S}_h$ is the state space for layer h, $\mathcal{A}$ is the action space, $P_h^M : \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$ is the probability transition kernel for layer h, $R_h^M : \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathbb{R})$ is the reward distribution for layer h, and $d_1 \in \Delta(\mathcal{S}_1)$ is the initial state distribution. We allow the reward distribution and transition kernel to vary across models in $\mathcal{M}$, but assume for simplicity that the initial state distribution is fixed.

For a fixed MDP $M \in \mathcal{M}$, each episode proceeds under the following protocol. At the beginning of the episode, the learner selects a randomized, non-stationary *policy* $\pi = (\pi_1, \ldots, \pi_H) \in \Pi_{\mathrm{RNS}}$, where $\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A})$ and Π_{RNS} denotes the set of all such policies. The episode then evolves through the following process, beginning from $s_1 \sim d_1$. For $h = 1, \ldots, H$:

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

For notational convenience, we take s_{H+1} to be a deterministic terminal state. The value for a policy π under M is given by $f^M(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^H r_h\right]$, where $\mathbb{E}^{M,\pi}[\cdot]$ denotes expectation under the process above.

In the online reinforcement learning setting, we interact with an unknown MDP $M^\star \in \mathcal{M}$ for T episodes. We now show how this setting can be captured in the DMSO framework, where we take the reward $r^{(t)}$ to be the cumulative reward in the episode and the observation $o^{(t)}$ to be the observed trajectory. In particular, for each episode $t = 1, \ldots, T$, the learner selects a policy $\pi^{(t)} \in \Pi_{\mathrm{RNS}}$. The policy is executed in the MDP $M^\star$, and the learner observes the resulting trajectory $\tau^{(t)} = (s_1^{(t)}, a_1^{(t)}, r_1^{(t)}), \ldots, (s_H^{(t)}, a_H^{(t)}, r_H^{(t)})$. Their goal is to minimize the total regret $\sum_{t=1}^T \mathbb{E}_{\pi^{(t)} \sim p^{(t)}}[f^{M^\star}(\pi^\star) - f^{M^\star}(\pi^{(t)})]$ against the optimal policy for $M^\star$. This setting immediately falls into the DMSO framework by taking $\Pi = \Pi_{\mathrm{RNS}}$, $r^{(t)} = \sum_{h=1}^H r_h^{(t)}$, and $o^{(t)} = \tau^{(t)}$.

Online reinforcement learning has received extensive attention, in terms of both algorithms and sample complexity bounds for specific model classes (e.g., Dean et al. (2020); Yang and Wang (2019); Jin et al. (2020c); Modi et al. (2020); Ayoub et al. (2020); Krishnamurthy et al. (2016); Du et al. (2019a); Li (2009); Dong et al. (2019)), as well as general structural results (Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020; Du et al., 2021; Jin et al., 2021). However, there is currently no unified understanding of what properties of the class of MDPs $\mathcal{M}$ determine the optimal regret.

This formulation, where the model class $\mathcal{M}$ is taken as a given, may at first seem tailored to *model-based* reinforcement learning, where the underlying transition dynamics are explicitly modeled. However, we can also express *model-free* reinforcement learning problems—where we instead take a class of candidate value functions $\mathcal{Q}$ as given and avoid directly learning the dynamics—by choosing $\mathcal{M}$ to be the set of all MDPs that are consistent with the given value function class. As we show in Section 7, this perspective suffices to recover sample-efficient learning guarantees for model-free RL, as well as other settings that lie between model-free and model-based. $\triangleleft$

Further examples. Beyond these examples, the DMSO framework captures many other canonical decision making problems, including (stochastic) contextual bandits, reinforcement learning in partially observable Markov decision processes (POMDPs) (Even-Dar et al., 2005; Jin et al., 2020a) and reinforcement learning with a generative model (Azar et al., 2013; Yang and Wang, 2019; Agarwal et al., 2020b).

Setting	DEC Lower Bound	Tight?
Multi-Armed Bandit	$\sqrt{AT}$	✓
Multi-Armed Bandit w/ gap	A/Δ	✓
Linear Bandit	$\sqrt{dT}$	✗ $(d\sqrt{T})$
Lipschitz Bandit	$T^{\frac{d+1}{d+2}}$	✓
ReLU Bandit	2^d	✓
Tabular RL	$\sqrt{HSAT}$	✓
Linear MDP	$\sqrt{dT}$	✗ $(d\sqrt{T})$
RL w/ linear $Q^\star$	2^d	✓
Deterministic RL w/ linear $Q^\star$	d	✓

Table 1: Lower bounds for bandits and reinforcement learning recovered by the Decision-Estimation Coefficient, where $A = \#$actions, $\Delta =$ gap, $d =$ feature dim., $H =$ episode horizon, and $S = \#$states. See Sections 5 to 7 for details. The "Tight?" column indicates whether the lower bound is tight up to logarithmic factors and dependence on episode horizon, with the optimal rate stated for ✗ cases.

1.2 The Decision-Estimation Coefficient

We introduce a new complexity measure, the *Decision-Estimation Coefficient* (DEC), defined for a model class $\mathcal{M}$ and nominal model $\overline{M}$ as

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Bigg[\underbrace{f^M(\pi_M) - f^M(\pi)}_{\text{regret of decision}} - \gamma\cdot \underbrace{D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)}_{\text{estimation error for obs.}}\Bigg], \tag{2}$$

where $\gamma > 0$ is a scale parameter and $D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) = \int(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}})^2$ is the Hellinger distance; we further define $\mathsf{dec}_\gamma(\mathcal{M}) = \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$.

The Decision-Estimation Coefficient is the value of a game in which the learner (represented by the min player) aims to find a distribution over decisions such that for a worst-case problem instance (represented by the max player), the *regret* of their decision is controlled by the *estimation error* relative to the nominal model. Conceptually, $\overline{M}$ should be thought of as a guess for the true model, and the learner (the min player) aims to—in the face of an unknown environment (the max player)—optimally balance the regret of their decision with the amount information they acquire. With enough information, the learner can confirm or rule out their guess $\overline{M}$, and scale parameter γ controls how much regret they are willing to incur to do this.

Let us give some intuition as to how the Decision-Estimation Coefficient leads to both *upper* and *lower* bounds on the optimal regret for decision making. On the algorithmic side (upper bounds), the connection between decision making (where the learner's decisions influence what feedback is collected) and estimation (where data is collected passively) may not seem apparent a-priori. Here, the power of the Decision-Estimation Coefficient is that it—*by definition*—provides a bridge. One can select decisions by building an estimate for the model using all of the observations collected so far, then sampling from the distribution p that solves (2) with the estimated model plugged in for $\overline{M}$. Boundedness of the DEC implies that at every round, any learner using this strategy either enjoys small regret or acquires information, with their total regret controlled by the cumulative error for (online/sequential) estimation. This strategy generalizes notable prior approaches—Bayesian (Russo and Van Roy, 2014, 2018) and frequentist (Abe and Long, 1999; Foster and Rakhlin, 2020).

Of course, the perspective above is only useful if the DEC is indeed bounded, which itself is not immediately apparent. However, we show that the DEC is not only bounded (e.g., for the finite-armed bandit problem where $\Pi = \{1, \ldots, A\}$, we have $\mathsf{dec}_\gamma(\mathcal{M}) \propto \frac{A}{\gamma}$), but—turning our focus to lower bounds—quantitatively necessary for sample-efficient learning. Developing lower bounds for a setting as general as the DMSO framework is challenging because any complexity measure needs to capture both i) simple problems like the multi-armed bandit, where the mean rewards serve as a sufficient statistic, and ii) problems with rich, structured feedback (e.g., reinforcement learning), where observations, or even structure in the noise itself, can provide non-trivial information about the underlying problem instance. Here, the information-theoretic nature of the DEC plays a key role. In particular, taking a dual perspective, the DEC can be seen to certify a lower bound on the price (in terms of regret) that *any algorithm* must pay to obtain enough information to distinguish a reference model from the least favorable alternative.

1.3 Contributions

Our main results show that the Decision-Estimation Coefficient captures the statistical complexity of interactive decision making. We provide:

- **A new fundamental limit and lower bound.** We establish (Theorems 3.1 and 3.2) that the Decision-Estimation Coefficient is a fundamental limit for interactive decision making: For any class of models, any algorithm must have

$$\mathbf{Reg}_{\mathsf{DM}} \gtrsim \max_{\gamma>0} \min\left\{\mathsf{dec}^{\mathrm{loc}}_{\gamma}(\mathcal{M}) \cdot T, \gamma\right\}, \tag{3}$$

  where $\mathsf{dec}^{\mathrm{loc}}_{\gamma}(\mathcal{M})$ is a certain localized variant of the DEC (cf. Section 3.1). This lower bound holds for *any model class*, with no assumption on the structure. When applied to bandits and reinforcement learning, it recovers standard minimax lower bounds for canonical problems classes (Sections 6 and 7), as well as strong impossibility results for sample-efficient reinforcement learning with linear function approximation (Weisz et al., 2021); see Table 1 for highlights. In addition, the lower bound has an appealing conceptual interpretation as a decision making analogue of classical modulus of continuity techniques in statistical estimation (Donoho and Liu, 1987, 1991a,b).

- **A unified meta-algorithm.** We provide (Theorem 3.3) a unified meta-algorithm, Estimation-to-Decisions ($\mathsf{E2D}$), which achieves the decision making lower bound in (3) whenever estimation is possible. We show that for any model class $\mathcal{M}$, the regret of $\mathsf{E2D}$ is bounded as

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \min_{\gamma>0} \max\left\{\mathsf{dec}^{\mathrm{loc}}_{\gamma}(\mathcal{M}) \cdot T, \gamma \cdot \mathsf{est}(\mathcal{M})\right\}, \tag{4}$$

  where $\mathsf{est}(\mathcal{M})$ is a certain classical notion of statistical estimation complexity for the class $\mathcal{M}$ (for example, $\mathsf{est}(\mathcal{M}) = \log|\mathcal{M}|$ for finite classes). $\mathsf{E2D}$ is a universal reduction which transforms any algorithm for online estimation with the model class $\mathcal{M}$ into an algorithm for decision making. The algorithm is computationally efficient whenever the minimax program (2) can be solved efficiently. This result recovers classical and contemporary sample complexity guarantees for bandits and reinforcement learning, and leads to new efficient and online algorithms for various classes of interest.

Together, these results provide a theory of learnability for interactive decision making: Whenever $\mathcal{M}$ has non-trivial online estimation complexity, sublinear regret for decision making is possible *if and only if* $\mathsf{dec}_{\gamma}(\mathcal{M})$ decays sufficiently quickly as $\gamma \to \infty$.

Additional contributions. Notable additional features of our results include:

- *Recovering existing frameworks.* Various works have proposed structural conditions that enable sample-efficient reinforcement learning, including Bellman Rank (Jiang et al., 2017), Witness Rank (Sun et al., 2019), (Bellman-) Eluder Dimension (Russo and Van Roy, 2013; Wang et al., 2020; Jin et al., 2021), and Bilinear Classes (Du et al., 2021). We show that the DEC subsumes each of these conditions and, importantly, provides the first such necessary condition. See Figure 1 for a summary.

- *New perspectives on existing algorithms.* Our results tie together and generalize disparate algorithmic approaches across the literature on bandits and reinforcement learning, most notably posterior sampling and information-directed sampling and the information ratio (Russo and Van Roy, 2014, 2018), and inverse gap weighting approaches to bandits and contextual bandits (Abe and Long, 1999; Foster and Rakhlin, 2020; Foster et al., 2020a).

- *Seamlessly incorporating contextual information.* Our algorithms immediately extend to provide regret bounds for a *contextual* variant of the DMSO framework in which the learner observes side information in the form a context (or, covariate) $x^{(t)}$ before making their decision at each round (Section 8). This leads to new oracle-efficient algorithms for contextual bandits with large action spaces (generalizing Foster and Rakhlin (2020); Foster et al. (2020a)) and contextual Markov decision processes (Abbasi-Yadkori and Neu, 2014; Modi et al., 2018; Dann et al., 2019; Modi and Tewari, 2020).

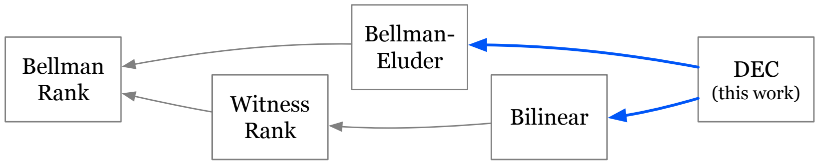

Figure 1: Relationship between Decision-Estimation Coefficient and existing frameworks for generalization in reinforcement learning. An arrow indicates that the head framework is subsumed by the tail framework.

1.4 Organization

The first part of the paper (Sections 3 and 4) presents central results and tools, with preliminaries in Section 2.

- Section 3 contains our main results: a universal lower bound for interactive decision making based on the Decision-Estimation Coefficient, and a nearly-matching upper bound via the Estimation-to-Decisions (E2D) meta-algorithm. Combining these results, we provide a characterization of learnability in the DMSO framework. This section also includes extensive discussion and interpretation of the main results.
- Building on this development, Section 4 provides a toolkit for deriving regret bounds using the E2D algorithm, including background on online estimation, general-purpose regret bounds, a Bayesian variant of the E2D algorithm, and other extensions.

In the second part of this paper (Sections 5 to 7), we apply our general tools to give efficient algorithms and regret bounds for specific examples of interest:

- Section 5 (Illustrative Examples) serves as a warm-up, and contains detailed examples for the basic multi-armed bandit problem and tabular reinforcement learning. For both settings, we carefully show how to upper and lower bound the Decision-Estimation Coefficient and derive efficient algorithms. These examples highlight a number of basic techniques which find use in later sections.
- Section 6 (Application to Bandits) applies our techniques to the structured bandit problem. We recover efficient algorithms and lower bounds for well-known problems (linear bandits, convex bandits, eluder dimension, and more), then derive new guarantees based on a parameter called the *star number*.
- Section 7 (Application to Reinforcement Learning) applies our techniques to reinforcement learning with function approximation. Highlights here include new, efficient algorithms for reinforcement learning with bilinear classes, and lower bounds for learning with linearly realizable value functions.

We conclude with a contextual extension of the E2D algorithm (Section 8), further related work (Section 9), and discussion (Section 10). All proofs are deferred to the appendix unless otherwise stated.

2 Preliminaries

Probability spaces. We briefly formalize the probability spaces associated with the DMSO framework. 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 $(\mathcal{O}, \mathscr{O})$. Let $\mathcal{H}^{(t)} = (\pi^{(1)}, r^{(1)}, o^{(1)}), \ldots, (\pi^{(t)}, r^{(t)}, o^{(t)})$ denote the history up to time 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 $(\Omega^{(t)}, \mathscr{F}^{(t)})$.

Recall that 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 1) For all $x \in \mathcal{X}$, $P(\cdot \mid x)$ is a probability measure, and 2) for all $Y \in \mathscr{Y}$, $x \mapsto P(Y \mid x)$ is measurable. If $P(\cdot \mid x)$ is a σ-finite measure rather than a probability measure, we instead call P a kernel.

Formally, a model $M = M(\cdot,\cdot \mid \cdot) \in \mathcal{M}$ is a probability kernel from $(\Pi, \mathscr{P})$ to $(\mathcal{R}\times\mathcal{O}, \mathscr{R}\otimes\mathscr{O})$. An *algorithm* for horizon T is specified by a sequence of probability kernels $p^{(1)},\ldots,p^{(T)}$, where $p^{(t)}(\cdot \mid \cdot)$ is a probability kernel from $(\Omega^{(t-1)}, \mathscr{F}^{(t-1)})$ to $(\Pi, \mathscr{P})$. We let $\mathbb{P}^{M,p}$ denote the law of $\mathcal{H}^{(T)}$ under the process $\pi^{(t)} \sim p^{(t)}(\cdot \mid \mathcal{H}^{(t-1)})$, $(r^{(t)}, o^{(t)}) \sim M(\cdot,\cdot \mid \pi^{(t)})$. We adopt the shorthand $M(\pi) = M(\cdot,\cdot \mid \pi)$ throughout.

We assume that there exists a common dominating kernel $\nu(\pi) = \nu(\cdot,\cdot \mid \pi)$ from $(\Pi, \mathscr{P})$ to $(\mathcal{R}\times\mathcal{O}, \mathscr{R}\otimes\mathscr{O})$ such that $M(\pi) \ll \nu(\pi)$ for all $M \in \mathcal{M}$, $\pi \in \Pi$. For each $\pi \in \Pi$ we let $m^{M}(\cdot,\cdot \mid \pi)$ denote the density of $M(\cdot,\cdot \mid \pi)$ with respect to $\nu(\cdot,\cdot \mid \pi)$. This assumption facilitates the use of density estimation for our upper bounds, but is not required by our lower bounds.

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

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

The (squared) Hellinger distance is defined as[3]

$$D^2_{\mathrm{H}}(\mathbb{P},\mathbb{Q}) = \int \left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}}\right)^2, \tag{5}$$

and the Kullback-Leibler divergence is defined as

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

Additional reinforcement learning notation. For a given model M and policy π, we define the state-action value function and state value functions via

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^{H} r_{h'} \mid s_h = s, a_h = a\right], \quad \text{and} \quad V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^{H} r_{h'} \mid s_h = s\right].$$

Note that the policy $\pi_M = \arg\max_\pi f^M(\pi)$ may not be uniquely defined. We assume without loss of generality that $\pi_M = (\pi_{M,1},\ldots,\pi_{M,H})$ solves $\pi_{M,h}(s) = \arg\max_{a\in\mathcal{A}} Q_h^{M,\pi_M}(s,a)$ for all $s \in \mathcal{S}$, $h \in [H]$. We adopt the convention that $V_{H+1}^{M,\pi}(s) = 0$ and use $Q_h^{M,\star}(s,a) := Q_h^{M,\pi_M}(s,a)$ and $V_h^{M,\star}(s,a) = V_h^{M,\pi_M}(s,a)$ to denote the optimal value functions.

Let $\mathbb{P}^{M,\pi}(\cdot)$ denote the law of a trajectory evolving under MDP M and policy π. We define state occupancy measures via $d_h^{M,\pi}(s) = \mathbb{P}^{M,\pi}(s_h = s)$ and state-action occupancy measures via $d_h^{M,\pi}(s,a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a)$. Note that we have $d_1^{M,\pi}(s) = d_1(s)$ for all M and π.

We let Π_{RNS} denote the collection of randomized non-stationary policies. For $\pi = (\pi_1,\ldots,\pi_H) \in \Pi_{\mathrm{RNS}}$, $\pi_h(s,a)$ denotes the probability that action a is selected in state s, so that $\pi_h(s) := \pi_h(s,\cdot) \in \Delta(\mathcal{A})$. For a pair of policies π, π', we define $\pi \circ_h \pi'$ as the policy that selects actions $a_1,\ldots,a_{h-1}$ using π and selects all subsequent actions using π'.

For a transition distribution $P(\cdot \mid \cdot,\cdot)$, we define $[Pf](s,a) = \mathbb{E}_{s'\sim P(s,a)}[f(s')]$.

Further notation. 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 some 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 use $f \propto g$ as shorthand for $f = \widetilde{\Theta}(g)$. We use $\lesssim$ only in informal statements to highlight salient elements of an inequality.

[3]If ν is a common dominating measure, then $D^2_{\mathrm{H}}(\mathbb{P},\mathbb{Q}) = \int\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu$, where $\frac{d\mathbb{P}}{d\nu}$ and $\frac{d\mathbb{Q}}{d\nu}$ are Radon-Nikodym derivatives. The notation in (5) (and for other divergences) reflects that this quantity is invariant under the choice of ν.

For a vector $x \in \mathbb{R}^d$, we let $\|x\|_2$ denote the euclidean norm and $\|x\|_\infty$ denote the element-wise ℓ_∞ norm. We let $\mathbf{0}$ denote the all-zeros vector, with dimension clear from context. For an integer $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, \ldots, n\}$. For a set S, we let $\mathrm{unif}(S)$ denote the uniform distribution over all the elements in S. 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 $\mathcal{X}$, and let $\mathrm{star}(\mathcal{X}, x) = \bigcup_{x' \in \mathcal{X}} \mathrm{co}(\{x', x\})$ denote the star hull. We use the convention $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

For a symmetric matrix $A \in \mathbb{R}^{d \times d}$, $A^\dagger$ denotes the pseudoinverse, and when $A \succeq 0$, $\|x\|_A^2 := \langle x, Ax \rangle$.

For a parameter $\mu \in [0, 1]$, $\mathrm{Ber}(\mu)$ denotes a Bernoulli random variable with mean μ. Likewise, for $\mu \in [-1, +1]$, $\mathrm{Rad}(\mu)$ denotes a (biased) Rademacher random variable with mean μ, which has $\mathbb{P}(+1) = \frac{1+\mu}{2}$ and $\mathbb{P}(-1) = \frac{1-\mu}{2}$. For an element x in a measurable space, we let δ_x denote the delta distribution on x.

2.1 Minimax Regret: Frequentist and Bayesian

The main goal of this paper is to characterize what properties of the model class $\mathcal{M}$ determine the *minimax regret*, defined via

$$\mathfrak{M}(\mathcal{M}, T) = \inf_{p^{(1)}, \ldots, p^{(T)}} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)], \tag{6}$$

where we write $\mathbf{Reg}_{\mathsf{DM}}(T)$ to make the dependence on T explicit, and recall that $p^{(t)} = p^{(t)}(\cdot \mid \mathcal{H}^{(t-1)})$. While our focus is on this frequentist notion of regret, we establish certain results by considering the *Bayesian* setting in which the underlying model $M^\star$ is drawn from a prior $\mu \in \Delta(\mathcal{M})$ that is known to the learner.[4] We define the worst-case Bayesian regret via

$$\underline{\mathfrak{M}}(\mathcal{M}, T) = \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p^{(1)}, \ldots, p^{(T)}} \mathbb{E}_{M^\star \sim \mu} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)]. \tag{7}$$

Building on a long line of research in online learning and bandits (Abernethy et al., 2009; Rakhlin et al., 2010; Bubeck et al., 2015; Bubeck and Eldan, 2016; Lattimore and Szepesvári, 2019), one can show that under mild technical conditions (essentially, whenever the prerequisites for Sion's minimax theorem are satisfied), the minimax frequentist regret and Bayesian regret coincide. This allows us to deduce existence of algorithms for the frequentist setting by exhibiting algorithms for the Bayesian setting, which is—in some cases—a simpler task. In particular, we have the following result.

Proposition 2.1. *Suppose that Π is finite and $\mathcal{R}$ is bounded, and that for all $M \in \mathcal{M}$, $\bigcup_{\pi \in \Pi} \mathrm{supp}(M(\pi))$ is countable. Then we have*

$$\mathfrak{M}(\mathcal{M}, T) = \underline{\mathfrak{M}}(\mathcal{M}, T). \tag{8}$$

We emphasize that the condition in Proposition 2.1 was chosen for simplicity and concreteness. We expect that the same conclusion can be proven under far weaker conditions, though we note that minimax theorems for learning with partial information are somewhat more subtle than for classical online learning; see, e.g., Lattimore and Szepesvári (2019). We refer ahead to Section 4.2 for additional discussion of connections between the frequentist and Bayesian setting, and for an analogous Bayesian counterpart to the Decision-Estimation Coefficient.

3 A Theory of Learnability for Interactive Decision Making

In this section we provide our main results: a universal lower bound for interactive decision making based on the Decision-Estimation Coefficient, and a nearly-matching upper bound via the Estimation-to-Decisions meta-algorithm. The section is organized as follows:

- In Section 3.1, we state our main lower bound on regret. We then walk through basic examples that give intuition as to how the Decision-Estimation Coefficient captures the complexity of interactive decision making, and show how the lower bound behaves for each example.

[4] Unless otherwise specified, we assume that $\mathcal{M}$ is equipped with the discrete topology, so that $\Delta(\mathcal{M})$ contains the space of finitely supported probability measures.

- In Section 3.2, we provide our main upper bound on regret and introduce the E2D algorithm, highlighting how these results complement the lower bound.
- Building on these results, in Section 3.3 we provide a characterization for learnability (i.e., achievability of non-trivial regret) based on the Decision-Estimation Coefficient.
- Section 3.4 gives the last major result of the section: a refinement of the main upper bound with weaker dependence on model estimation complexity.

We conclude the section by showing that the upper and lower bounds cannot be improved further without additional assumptions (Section 3.5). All proofs are deferred to Appendix C.

3.1 Lower Bound: The Decision-Estimation Coefficient is a Fundamental Limit

We now present the main lower bound based on the Decision-Estimation Coefficient. The result is proven using a decision-theoretic adaptation of the *local minimax* method (Donoho and Liu, 1987, 1991a,b); the proof uses the DEC to show that, for any nominal model $\overline{M} \in \mathcal{M}$, any algorithm must either experience large regret on $\overline{M}$ or on a worst-case alternative model in the neighborhood of $\overline{M}$. To state the result, we formalize the notion of a neighborhood via a *localized* restriction of the class $\mathcal{M}$.

Definition 3.1 (Localized model class). *For a model class $\mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}$, we define*

$$\mathcal{M}_{\varepsilon}(\overline{M}) = \left\{M \in \mathcal{M} : f^{\overline{M}}(\pi_{\overline{M}}) \geq f^{M}(\pi_{M}) - \varepsilon\right\} \tag{9}$$

as the neighborhood of $\overline{M}$ at radius ε.

The localized model class is tailored to decision making and asserts that no alternative should have mean reward substantially larger than that of the nominal model, but otherwise does not place any restriction on the mean reward function or observation distribution.

Our lower bound takes the worst case over all possible nominal models. We define the shorthand

$$\mathsf{dec}_{\gamma}(\mathcal{M}) = \sup_{\overline{M} \in \mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \quad \text{and} \quad \mathsf{dec}_{\gamma,\varepsilon}(\mathcal{M}) = \sup_{\overline{M} \in \mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}_{\varepsilon}(\overline{M}), \overline{M}). \tag{10}$$

where $\mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$ is defined in (2). We let $V(\mathcal{M})$ denote a bound on the density ratio over all models. Precisely, we define $V(\mathcal{M}) = \sup_{M,M' \in \mathcal{M}} \sup_{\pi \in \Pi} \sup_{A \in \mathscr{R} \otimes \mathscr{O}} \left\{\frac{M(A|\pi)}{M'(A|\pi)}\right\} \vee e$; finiteness of this quantity is not necessary, but leads to tighter guarantees.[5]

The main lower bound is as follows.

Theorem 3.1 (Main lower bound). *Consider a model class $\mathcal{M}$ with $\mathcal{F}_{\mathcal{M}} \subseteq (\Pi \to [0,1])$, and let $\delta \in (0,1)$ and $T \in \mathbb{N}$ be fixed. Define $C(T) := 2^{15} \log(2T \wedge V(\mathcal{M}))$ and $\underline{\varepsilon}_{\gamma} := C(T)^{-1} \frac{\gamma}{T}$. Then for any algorithm, there exists a model in $\mathcal{M}$ for which*

$$\mathbf{Reg}_{\mathsf{DM}} \geq (6C(T))^{-1} \cdot \max_{\gamma > \sqrt{C(T)T}} \min\left\{\left(\mathsf{dec}_{\gamma,\underline{\varepsilon}_{\gamma}}(\mathcal{M}) - 15\delta\right) \cdot T,\ \gamma\right\} \tag{11}$$

with probability at least $\delta/2$.

Theorem 3.1 shows that for any model class, any algorithm has $\mathbf{Reg}_{\mathsf{DM}} \gtrsim \max_{\gamma \gtrsim \sqrt{T}} \min\left\{\mathsf{dec}_{\gamma,\underline{\varepsilon}_{\gamma}}(\mathcal{M}) \cdot T, \gamma\right\}$ with moderate probability. The result serves as a converse to high-probability upper bounds we establish in the sequel, though in general it does not lead to optimal lower bounds on *expected* regret. Our next result is a variant of Theorem 3.1 that provides stronger lower bounds on expected regret using a more restrictive notion of localization. Define $g^{M}(\pi) := f^{M}(\pi_{M}) - f^{M}(\pi)$ and

$$\mathcal{M}^{\infty}_{\varepsilon}(\overline{M}) = \left\{M \in \mathcal{M} : \left|g^{M}(\pi) - g^{\overline{M}}(\pi)\right| \leq \varepsilon \ \ \forall \pi \in \Pi\right\}. \tag{12}$$

[5] We do not require that $V(\mathcal{M}) < \infty$, but whenever this holds we can remove certain $\log(T)$ factors in Theorems 3.1 and 3.2 for tighter results.

Theorem 3.2 (Main lower bound—in-expectation version)**.** *Consider a model class $\mathcal{M}$ with $\mathcal{F}_{\mathcal{M}} \subseteq (\Pi \to [0,1])$. Let $T \in \mathbb{N}$ be fixed and define $C(T) := 2^{11}\log(2T \wedge V(\mathcal{M}))$ and $\varepsilon_\gamma := C(T)^{-1}\frac{\gamma}{T}$. Then for any algorithm, there exists a model in $\mathcal{M}$ for which*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq 6^{-1} \cdot \max_{\gamma>0} \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}^\infty_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \cdot T. \tag{13}$$

Despite using a stronger L_∞ notion of localization, this result suffices to derive all of the lower bounds in Table 1.

3.1.1 Lower Bound and Decision-Estimation Coefficient: Basic Properties and Intuition

To build intuition, we take this opportunity to highlight some basic properties of the Decision-Estimation Coefficient and how these properties influence the behavior of our lower bounds. To keep the examples we consider as simple as possible, we only sketch the proof details. We refer ahead to Sections 5 to 7 for examples with complete calculations. Additional technical aspects of the lower bound are discussed at the end of the section (Section 3.5).

Capturing the complexity of decision making with reward-based feedback. As a starting point, we consider the behavior of the Decision-Estimation Coefficient and the lower bounds in Theorems 3.1 and 3.2 for bandit-type problems (of varying degrees of structure) in which only reward-based feedback is available. Here, a basic property of the Decision-Estimation Coefficient is that it reflects the amount of information that any given decision reveals about other possible decisions, thereby acting as a measure of intrinsic dimension. To illustrate this point, we focus on two problems at opposite ends of the bandit spectrum: Finite-armed bandits and bandits with full information.

Example 3.1 (Finite-armed bandits)**.** For the classical multi-armed bandit problem, we have $\Pi = [A]$, $\mathcal{R} = [0,1]$, and $\mathcal{O} = \{\varnothing\}$, and the model class $\mathcal{M} = \{M : M(\pi) \in \Delta(\mathcal{R})\}$ consists of all possible distributions over $\mathcal{R}$. Let us sketch a lower bound on the Decision-Estimation Coefficient for this setting. Fix $\Delta \in (0, 1/2)$ and define a sub-family of models $\mathcal{M}' = \{M_1, \ldots, M_A\} \cup \{\overline{M}\}$ via $M_i(\pi) := \mathrm{Ber}(1/2 + \Delta\mathbb{I}\{\pi = i\})$ and $\overline{M}(\pi) := \mathrm{Ber}(1/2)$. In other words, each model M_i has $A-1$ arms with suboptimality gap Δ relative to the optimal arm i, while the model $\overline{M}$ has identical rewards for all arms. One can verify that $D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \lesssim \Delta^2 \cdot \mathbb{I}\{\pi = i\}$ (cf. Lemma A.7) and that $f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \Delta \cdot \mathbb{I}\{\pi \neq i\}$. In other words, for each model, only a single decision reveals information about the model's identity and leads to low regret. As a result, we calculate that

$$\begin{aligned}\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) &\geq \inf_{p\in\Delta(\Pi)} \max_{i\in[A]} \mathbb{E}_{\pi\sim p}\big[f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big)\big] \\ &\geq \inf_{p\in\Delta(\Pi)} \mathbb{E}_{i\sim\mathrm{unif}([A])}\mathbb{E}_{\pi\sim p}\big[f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big)\big] \gtrsim \Delta - \gamma \cdot \frac{\Delta^2}{A}. \end{aligned} \tag{14}$$

Crucially, because the decisions that reveal information are disjoint, the negative contribution from the Hellinger divergence scales as A^{-1}. By choosing $\Delta \propto \frac{A}{\gamma}$, we conclude that $\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \gtrsim \frac{A}{\gamma}$, and with more care, we show in Section 5 that this argument in fact implies that (i) $\mathcal{M}' \subseteq \mathcal{M}^\infty_{\varepsilon_\gamma}(\overline{M})$ whenever $\gamma \gtrsim \sqrt{AT}$, and (ii) $V(\mathcal{M}') = O(1)$. Plugging these values into Theorem 3.2, the lower bound in (13) yields

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \gtrsim \max_{\gamma\gtrsim\sqrt{AT}} \left\{\frac{A}{\gamma}\right\} = \Omega\big(\sqrt{AT}\big),$$

which matches the well-known minimax rate (Audibert and Bubeck, 2009). In Section 5 we provide a matching upper bound of the form $\mathsf{dec}_\gamma(\mathcal{M}) \leq \frac{A}{\gamma}$, which certifies that this lower bound on the DEC is optimal. ◁

The finite-armed bandit setting is completely unstructured in the sense that choosing a given arm π reveals no information about the others. The next example, *bandits with full information*, lies at the opposite extreme. Here, each arm reveals the rewards for all other arms, completely removing the need for exploration.

Example 3.2 (Bandits with full information)**.** Consider a full-information variant of the bandit setting. We have $\Pi = [A]$ and $\mathcal{R} = [0,1]$, and for a given decision π we observe a reward r as in Example 3.1, but also receive an observation $o = (r(\pi'))_{\pi' \in [A]}$ consisting of (counterfactual) rewards for every action.

For a given model M, let $M_{\mathcal{R}}(\pi)$ denote the distribution over the reward for π. Then for any decision π, since all rewards are observed, the data processing inequality implies that for all π',

$$D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big) \geq D^2_{\mathrm{H}}\big(M_{\mathcal{R}}(\pi'), \overline{M}_{\mathcal{R}}(\pi')\big). \tag{15}$$

Using this property, we can show that $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \propto \frac{1}{\gamma}$. We sketch a proof of the upper bound; we find this to be more illuminating than the lower bound for this example because it directly highlights how extra information helps to shrink the DEC.

For a given model $\overline{M}$ we choose $p = \delta_{\pi_{\overline{M}}}$ and bound $\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)]$ by

$$\begin{aligned} f^M(\pi_M) - f^M(\pi_{\overline{M}}) &\leq f^M(\pi_M) - f^M(\pi_{\overline{M}}) + f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M) \\ &\leq 2 \cdot \max_{\pi \in \{\pi_M, \pi_{\overline{M}}\}} |f^M(\pi) - f^{\overline{M}}(\pi)| \\ &\leq 2 \cdot \max_{\pi \in \{\pi_M, \pi_{\overline{M}}\}} D_{\mathrm{H}}\big(M_{\mathcal{R}}(\pi), \overline{M}_{\mathcal{R}}(\pi)\big). \end{aligned}$$

We then use the AM-GM inequality, which implies that for any $\gamma > 0$,

$$\max_{\pi \in \{\pi_M, \pi_{\overline{M}}\}} D^2_{\mathrm{H}}\big(M_{\mathcal{R}}(\pi), \overline{M}_{\mathcal{R}}(\pi)\big) \lesssim \gamma \cdot \max_{\pi \in \{\pi_M, \pi_{\overline{M}}\}} D^2_{\mathrm{H}}\big(M_{\mathcal{R}}(\pi), \overline{M}_{\mathcal{R}}(\pi)\big) + \frac{1}{\gamma} \leq \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi_{\overline{M}}), \overline{M}(\pi_{\overline{M}})\big) + \frac{1}{\gamma},$$

where the final inequality uses (15). This certifies that for all $M \in \mathcal{M}$, the choice for p above satisfies

$$\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \lesssim \frac{1}{\gamma},$$

so we have $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \lesssim \frac{1}{\gamma}$. A matching lower bound follows by adapting Example 3.1. Comparing to the finite-armed bandit, we see that the DEC for this example is independent of A, which reflects the information sharing. The final lower bound from Theorem 3.2 takes the form

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega\big(\sqrt{T}\big).$$

◁

In Section 6 we consider structured bandit problems that lie between the extremes above, including linear and convex bandits. These problems typically have $\mathsf{dec}_\gamma(\mathcal{M}) \propto \frac{\mathsf{dim}}{\gamma}$, where—generalizing Examples 3.1 and 3.2—dim is a quantity that reflects the intrinsic problem dimension.

Examples 3.1 and 3.2 can be thought of as bandit problems with "problem-instance agnostic" noise, in the sense that the mean rewards serve as a sufficient statistic, while the additive noise (in the reward distribution) is otherwise uninformative about the problem instance. However, any complexity measure which claims to characterize the complexity of decision making must account for more general (and possibly corner-case) problems, where seemingly irrelevant properties of the noise distribution can encode non-trivial information about the problem instance itself. The following example highlights one such problem.

Example 3.3 (Bandits: hiding information in lower order bits)**.** Consider a family of multi-armed bandit models with $\Pi = [A]$. Following Example 3.1, fix $\Delta \in (0, 1/2)$ and define $\mathcal{M} = \{M_1, \ldots, M_A\} \cup \{\overline{M}\}$ via $M_i(\pi) := \mathcal{N}(1/2 + \Delta\mathbb{I}\{\pi = i\}, 1)$ and $\overline{M}(\pi) := \mathcal{N}(1/2, 1)$. As defined, this family has the same lower bound $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \geq \frac{A}{\gamma}$ as in Example 3.1, because M_i and $\overline{M}$ can only be distinguished by choosing $\pi = i$, and the Hellinger distance for this arm scales as Δ^2. Now, suppose we modify each model M_i such that $M_i(\pi)$ is unchanged for $\pi \neq i$, but for $\pi = i$, the reward r is generated by sampling $r' \sim \mathcal{N}(1/2 + \Delta, 1)$, then setting bits $N, \ldots, N + \lceil \log_2(A) \rceil$ in the reward's infinite-precision binary representation to encode index i, where $N \in \mathbb{N}$ is a parameter. In other words, if decision i is chosen in model $M_i(\pi)$, then the lower order bits of the infinite precision representation of the observed reward will reveal that i is in fact the optimal arm. Taking N large makes the mean reward $f^{M_i}(i)$ arbitrarily close to $1/2 + \Delta$, but we have

$$D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \propto \mathbb{I}\{\pi = i\}$$

for all π when A is sufficiently large. That is, compared to Example 3.1, the Hellinger distance does not scale with the gap Δ due to the auxiliary information encoded in bits $N, \ldots, N + \lceil \log_2(A) \rceil$.[6] Choosing $\Delta = 1/2$ and adapting the argument in Example 3.1 (where we introduce the expectation over $i \sim \mathrm{unif}([A])$ as in (14)), we compute that

$$\mathsf{dec}_\gamma(\mathcal{M}_{1/2}(\overline{M}), \overline{M}) \gtrsim 1 - \gamma \cdot \frac{1}{A} \gtrsim \mathbb{I}\{\gamma \leq A/2\},$$

and Theorem 3.2 yields

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \widetilde{\Omega}(\min\{A, T\}).$$

This is seen to be optimal, since we can detect the index of the underlying instance with constant probability by enumerating over the arms a single time. ◁

Note that for this example, working with Hellinger distance (or another information-theoretic divergence) is critical. A weaker divergence such as the squared distance $D_{\mathrm{Sq}}\big(M_i(\pi), \overline{M}(\pi)\big) := (f^{M_i}(\pi) - f^{\overline{M}}(\pi))^2$ would miss the information hiding in the lower order bits. Further examples with a similar information-theoretic flavor include noiseless multi-armed bandits, as well as "cheating code" problems of the type considered in Jun and Zhang (2020), where a basic multi-armed bandit problem instance is equipped with auxiliary arms that—when queried—directly reveal the identity of the instance.

From reward-based feedback to general observations. Generalizing Example 3.3, one can create structured bandit problems in which arbitrary auxiliary information is hidden in the reward signal. The DMSO framework makes auxiliary information explicit via the observations in the process $(r, o) \sim M(\pi)$, which we find to be more clear conceptually. Our last example considers a setting where accounting for such observations is crucial.

Example 3.4 (Importance of observations for reinforcement learning)**.** In the tabular (finite state/action) reinforcement learning setting, the model class $\mathcal{M}$ is the collection of all non-stationary MDPs with state space $\mathcal{S} = [S]$, action space $\mathcal{A} = [A]$, and horizon H; we have $\Pi = \Pi_{\mathrm{RNS}}$ (recall that Π_{RNS} is the set of randomized non-stationary policies). For a given MDP M, when $(r, o) \sim M(\pi)$, the reward signal r is the total reward for an episode in which the policy π is executed, and the observation o is the resulting trajectory. Here, if only reward information were available, sample complexity $2^{\Omega(H)}$ would be unavoidable. However, because trajectories are observed, the problem admits polynomial sample complexity. We show (Section 5) that

$$\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \propto \frac{\mathrm{poly}(S, A, H)}{\gamma},$$

whenever $\gamma \geq \sqrt{\mathrm{poly}(S, A, H) \cdot T}$, which leads to $\sqrt{\mathrm{poly}(S, A, H) \cdot T}$ upper and lower bounds on regret. ◁

Similar discussion applies to essentially all non-trivial reinforcement learning problems, and highlights the necessity of learning from observations. Because the Decision-Estimation Coefficient is able to account for bandit problems with auxiliary information in the reward signal, it incorporates additional observations seamlessly.

Additional information-theoretic considerations. We conclude by briefly discussing some additional information-theoretic properties of the Decision-Estimation Coefficient.

Example 3.5 (Filtering irrelevant information)**.** Adding observations that are unrelated to the model under consideration never changes the value of the Decision-Estimation Coefficient. In more detail, consider a model class $\mathcal{M}$ with observation space $\mathcal{O}_1$, and consider a class of conditional distributions $\mathcal{D}$ over a secondary

[6] As N and A grow, the probability of observing any given string in the bits $N, \ldots, N + \lceil \log_2(A) \rceil$ becomes arbitrarily small under $\overline{M}$. This causes the total variation distance (and consequently Hellinger distance) to become constant.

observation space $\mathcal{O}_2$, where each $D \in \mathcal{D}$ has the form $D(\pi) \in \Delta(\mathcal{O}_2)$. For $M \in \mathcal{M}$ and $D \in \mathcal{D}$, let $(M \otimes D)(\pi)$ be the model that, given $\pi \in \Pi$, samples $(r, o_1) \sim M(\pi)$ and $o_2 \sim D(\pi)$, then emits $(r, (o_1, o_2))$. Set

$$\mathcal{M} \otimes \mathcal{D} = \{M \otimes D \mid M \in \mathcal{M}, D \in \mathcal{D}\}.$$

Then for all $\overline{M} \in \mathcal{M}$ and $\overline{D} \in \mathcal{D}$,

$$\mathsf{dec}_\gamma(\mathcal{M} \otimes \mathcal{D}, \overline{M} \otimes \overline{D}) = \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}).$$

This can be seen to hold by restricting the supremum in (2) to range over models of the form $M \otimes \overline{D}$. ◁

Example 3.6 (Data processing)**.** Passing observations through a channel never decreases the Decision-Estimation Coefficient. Consider a class of models $\mathcal{M}$ with observation space $\mathcal{O}$. Let $\rho : \mathcal{O} \to \mathcal{O}'$ be given, and define $\rho \circ M$ to be the model that, given decision π, samples $(r, o) \sim M(\pi)$, then emits $(r, \rho(o))$. Let $\rho \circ \mathcal{M} := \{\rho \circ M \mid M \in \mathcal{M}\}$. Then for all $\overline{M} \in \mathcal{M}$, we have

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \leq \mathsf{dec}_\gamma(\rho \circ \mathcal{M}, \rho \circ \overline{M}).$$

This is an immediate consequence of the data processing inequality for Hellinger distance, which implies that $D^2_{\mathrm{H}}\big(\big(\rho \circ M\big)(\pi), \big(\rho \circ \overline{M}\big)(\pi)\big) \leq D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)$. ◁

All of the properties we have discussed up to this point are essential for any complexity measure that aims to characterize the statistical complexity of general decision making problems. With this in mind, we proceed to sketch the proof of Theorem 3.1; the proof for Theorem 3.2 follows similar reasoning.

3.1.2 Theorem 3.1: Proof Sketch

The key idea behind Theorem 3.1 is to show that for any algorithm p and nominal model $\overline{M}$, the Decision-Estimation Coefficient certifies a lower bound on the price (in terms of regret) that the algorithm must pay to obtain enough information to distinguish the nominal model from an adversarially chosen alternative. In particular, we will prove that for any $\gamma \gtrsim \sqrt{T}$ (recall the constraints on γ in Theorem 3.1) and $\overline{M} \in \mathcal{M}$, for any $\varepsilon > 0$ sufficiently small (as a function of γ and T), there exists a model $M \in \mathcal{M}_\varepsilon(\overline{M})$ such that

$$\mathbf{Reg}_{\mathsf{DM}} \gtrsim \min\big\{\big(\mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}) - \delta\big) \cdot T, \gamma\big\} \tag{16}$$

with probability at least $\Omega(\delta)$. From here, the result in (11) follows by maximizing over $\overline{M}$ and γ.

Let $T \in \mathbb{N}$, $\gamma > 0$, and $\varepsilon > 0$ be fixed and consider an algorithm $p = \{p^{(t)}(\cdot \mid \cdot)\}_{t=1}^T$. Recall that $\mathbb{E}^{M,p}[\cdot]$ denotes the expectation over the history $\mathcal{H}^{(T)}$ when M is the underlying problem instance and p is the algorithm. For each model M, let

$$p_M := \mathbb{E}^{M,p}\left[\frac{1}{T}\sum_{t=1}^T p^{(t)}(\cdot \mid \mathcal{H}^{(t-1)})\right]$$

denote the average (conditional) distribution over decisions under M. Fix a nominal model $\overline{M}$. Since $p_{\overline{M}} \in \Delta(\Pi)$, the definition of $\mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M})$ guarantees that

$$\sup_{M \in \mathcal{M}_\varepsilon(\overline{M})} \mathbb{E}_{\pi \sim p_{\overline{M}}}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \geq \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}).$$

In particular, if $M \in \mathcal{M}_\varepsilon(\overline{M})$ attains the supremum above, then by rearranging we have

$$\mathbb{E}_{\pi \sim p_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] \geq \gamma \cdot \mathbb{E}_{\pi \sim p_{\overline{M}}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] + \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}). \tag{17}$$

Note that due to the equality

$$\frac{1}{T} \cdot \mathbb{E}^{M,p}[\mathbf{Reg}_{\mathsf{DM}}] = \mathbb{E}_{\pi \sim p_M}[f^M(\pi_M) - f^M(\pi)], \tag{18}$$

the result would be established if the left-hand side of (17) were to be replaced by $\mathbb{E}_{\pi\sim p_M}[f^M(\pi_M) - f^M(\pi)]$ (recalling that $D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)$ is non-negative). The majority of the technical effort behind the proof is to use a change of measure argument to show that as long as the localization radius ε is sufficiently small as a function of γ and T,

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] \lesssim \mathbb{E}_{\pi\sim p_M}[f^M(\pi_M) - f^M(\pi)] + \gamma \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]. \tag{19}$$

Combining this with (17) and (18) yields

$$\begin{aligned}\mathbb{E}^{M,p}[\mathbf{Reg}_{\mathsf{DM}}] &\gtrsim \left(\mathbb{E}_{\pi\sim p_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]\right) \cdot T \\ &\gtrsim \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}) \cdot T.\end{aligned}$$

The final lower bound scales with $\min\{\mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}) \cdot T, \gamma\}$ (suppressing dependence on δ) because the change of measure argument above requires that $\mathbf{Reg}_{\mathsf{DM}} \lesssim \gamma$ with high probability. We deduce (16) by optimizing over $\overline{M}$ and γ.

□

For Theorem 3.2, the main difference from the proof above is that the L_∞ notion of localization in (12) allows for a stronger change of measure argument.

3.2 E2D: A Unified Meta-Algorithm for Interactive Decision Making

With a fundamental limit based on the Decision-Estimation Coefficient established, we now state the upper bound on regret that serves as its counterpart. Our main regret bound is achieved by the *Estimation-to-Decisions* meta-algorithm (E2D), described in Algorithm 1. The algorithm is based on the primitive of an *online estimation oracle*, which generalizes the notion of online regression oracle used in closely related work on the contextual bandit problem (Foster and Rakhlin, 2020; Foster et al., 2020a; Foster and Krishnamurthy, 2021).

An online estimation oracle, which we denote by $\mathbf{Alg}_{\mathsf{Est}}$, is an algorithm that attempts to estimate the underlying model $M^\star$ from data in a sequential fashion (typically via density estimation), and can be thought of as an (implicit) input to E2D. At each round t, given the observations and rewards collected so far, the estimation oracle produces an estimate

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

for the true model $M^\star$. Using this estimate, the most basic variant of E2D (Option I in Algorithm 1) proceeds by computing the distribution $p^{(t)}$ that achieves the value $\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^{(t)})$ for the Decision-Estimation Coefficient. That is, defining

$$\mathcal{V}^{\widehat{M}}_\gamma(p, M) = \mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \widehat{M}(\pi)\big)\right], \tag{20}$$

we set

$$p^{(t)} = \operatorname*{arg\,min}_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathcal{V}^{\widehat{M}^{(t)}}_\gamma(p, M). \tag{21}$$

E2D then samples the decision $\pi^{(t)}$ from this distribution and moves on to the next round. This approach, where we map an estimate to a decision in a simple per-round fashion, can be viewed as a generalization of the SquareCB algorithm of Foster and Rakhlin (2020).

Option I of E2D leads to regret bounds that scale with $\mathsf{dec}_\gamma(\mathcal{M})$. We achieve a localized regret bound that matches the quantity $\mathsf{dec}_{\gamma,\varepsilon}(\mathcal{M})$ appearing in Theorem 3.1 using a more sophisticated variant of E2D (Option II in Algorithm 1) augments the approach above by first restricting to a Hellinger confidence set $\mathcal{M}^{(t)} \subseteq \mathcal{M}$ based on the estimators $\big\{\widehat{M}^{(i)}\big\}_{i=1}^{t-1}$, then solving the minimization problem in (21) using the restricted set $\mathcal{M}^{(t)}$ rather than the whole model class. We state both variants (Option I and Option II) because the former is simpler and more efficient, yet suffices for essentially all applications we consider. Option II—while important for matching Theorem 3.1 as closely as possible—is mainly of theoretical interest.

Algorithm 1 Estimation to Decision-Making Meta-Algorithm (E2D)

parameters:

  Online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$.

  Exploration parameter $\gamma > 0$.

  Confidence radius $R > 0$. `// Option II only.`

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

  Compute estimate $\widehat{M}^{(t)} = \mathbf{Alg}^{(t)}_{\mathsf{Est}}\Big(\{(\pi^{(i)}, r^{(i)}, o^{(i)})\}_{i=1}^{t-1}\Big)$.

  Option I:

    Let $p^{(t)} = \arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathcal{V}_{\gamma}^{\widehat{M}^{(t)}}(p, M)$. `// Minimizer for` $\mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}^{(t)})$`; cf. Eq. (20).`

  Option II:

    Let $\mathcal{M}^{(t)} = \big\{M \in \mathcal{M}^{(t-1)} \mid \sum_{i=1}^{t-1} \mathbb{E}_{\pi\sim p^{(i)}}\big[D^2_{\mathrm{H}}\big(M(\pi), \widehat{M}^{(i)}(\pi)\big)\big] \le R^2\big\}$, with $\mathcal{M}^{(0)} := \mathcal{M}$.

    Let $p^{(t)} = \arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}^{(t)}} \mathcal{V}_{\gamma}^{\widehat{M}^{(t)}}(p, M)$. `// Minimizer for` $\mathsf{dec}_{\gamma}(\mathcal{M}^{(t)}, \widehat{M}^{(t)})$.

  Sample decision $\pi^{(t)} \sim p^{(t)}$ and update estimation oracle with $(\pi^{(t)}, r^{(t)}, o^{(t)})$.

We show that E2D serves as a universal reduction from estimation to decision making. Whenever the online estimation algorithm $\mathbf{Alg}_{\mathsf{Est}}$ can accurately estimate the true model (with respect to Hellinger distance), E2D enjoys low regret for decision making, with the bound on regret determined by the Decision-Estimation Coefficient. The process bears some similarity to the certainty equivalence principle in control, which also provides a separation between estimation and decision making. The main difference is that rather than simply selecting the optimal decision for the estimated model (which would have poor exploration), we solve the minimax program in (21) with the estimated model to balance exploration and exploitation.

3.2.1 Finite Model Classes

While our most general guarantees for E2D hold for any choice of the estimator $\mathbf{Alg}_{\mathsf{Est}}$, in this section we state specialized guarantees—first for finite, then general model classes—that allow for easy comparison with our lower bounds (Theorems 3.1 and 3.2); general results are deferred to Section 4. To state the guarantees in the simplest form possible, we make the following basic regularity assumption.

Assumption 3.1. *There exists a constant $c_\ell > 1$ such that for all models $\overline{M}$ and all $\gamma > 0$, $\varepsilon > 0$,*

$$\mathsf{dec}_{\gamma}(\mathcal{M}_{(c_\ell\cdot\varepsilon)}(\overline{M}), \overline{M}) \le c_\ell \cdot \mathsf{dec}_{\gamma}(\mathcal{M}_{\varepsilon}(\overline{M}), \overline{M}).$$

Note that if $\mathsf{dec}_{\gamma}(\mathcal{M}_{\varepsilon}(\overline{M}), \overline{M}) \propto \varepsilon^{\rho}$ for $\rho \le 1$, this is assumption is satisfied for all $c_\ell > 1$.

Theorem 3.3 (Main upper bound—finite class version)**.** *Fix $\delta \in (0,1)$. Assume that $\mathcal{R} \subseteq [0,1]$ and Assumption 3.1 holds. Define $C = O(c_\ell^2 \log_{c_\ell}(T))$ and $\bar{\varepsilon}_\gamma := 48\big(\frac{\gamma}{T}\log(|\mathcal{M}|/\delta) + \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma^{-1}\big)$. Then Algorithm 1, with an appropriate choice for parameters and estimation oracle, ensures that with probability at least $1-\delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \le C \cdot \min_{\gamma>0} \max\left\{ \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}_{\bar{\varepsilon}_\gamma}(\overline{M}), \overline{M})) \cdot T,\ \gamma \cdot \log(|\mathcal{M}|/\delta) \right\}. \tag{22}$$

This theorem is a special case of a more general result, Theorem 4.1 in Section 4. The upper bound is seen to have a similar form to the lower bound in Theorem 3.1. In particular, suppressing precise dependence on the failure parameter δ and logarithmic factors, we have

$$\text{Theorem 3.1:}\quad \mathbf{Reg}_{\mathsf{DM}} \gtrsim \max_{\gamma}\min\left\{ \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}_{\underline{\varepsilon}_\gamma}(\overline{M}), \overline{M})) \cdot T, \gamma \right\},$$

$$\text{Theorem 3.3:}\quad \mathbf{Reg}_{\mathsf{DM}} \lesssim \min_{\gamma}\max\left\{ \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}_{\bar{\varepsilon}_\gamma}(\overline{M}), \overline{M})) \cdot T, \gamma \cdot \log|\mathcal{M}| \right\}.$$

The min-max and max-min above can be exchanged under mild conditions on $\mathsf{dec}_{\gamma,\varepsilon}(\mathcal{M})$. The most important difference between the rates is the presence of the term $\log|\mathcal{M}|$ in Theorem 3.3, which serves as an upper bound on the complexity of statistical estimation for the model class $\mathcal{M}$. A secondary difference is that the upper bound in Theorem 3.3 uses the class $\mathrm{co}(\mathcal{M})$, while the lower bound in Theorem 3.1 only considers $\mathcal{M}$.

In terms of failure probability, the lower bound in Theorem 3.1 provides a meaningful converse to Theorem 3.3 in the moderate probability regime where $\delta \lesssim \mathsf{dec}_{\gamma,\underline{\varepsilon}_\gamma}(\mathcal{M})$; setting $\delta \propto 1/\sqrt{T}$ suffices for non-trivial classes.[7] With this choice, the dependence on δ in the lower bound vanishes and the $\log(1/\delta)$ factor in the upper bound becomes $\log(T)$. We conclude that (up to the other differences discussed above), the upper bound cannot be improved beyond this logarithmic factor in this probability regime. In addition, whenever the localized classes $\mathcal{M}_\varepsilon(\overline{M})$ and $\mathcal{M}^\infty_\varepsilon(\overline{M})$ have similar complexity (such is the case for all of the examples in Table 1), Theorem 3.2 provides an in-expectation lower bound of the same order.

These and other points of comparison are discussed in greater detail in Section 3.5.

3.2.2 Theorem 3.3: Proof Sketch

We find it most illuminating to sketch how to prove a non-localized version of Theorem 3.3, which is a nearly immediate consequence of the definition of the Decision-Estimation Coefficient. Consider Algorithm 1 with Option I and parameter $\gamma > 0$. Let $\mathbf{Est}_{\mathsf{H}} := \sum_{t=1}^T \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\big[D^2_{\mathrm{H}}\big(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\big)\big]$ denote the cumulative Hellinger error of the estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$. We write

$$\begin{aligned}
\mathbf{Reg}_{\mathsf{DM}} &= \sum_{t=1}^T \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\big] \\
&= \sum_{t=1}^T \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\big] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\Big[D^2_{\mathrm{H}}\Big(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\Big)\Big] + \gamma\cdot\mathbf{Est}_{\mathsf{H}}.
\end{aligned}$$

For each t, since $M^\star \in \mathcal{M}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\big] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\Big[D^2_{\mathrm{H}}\Big(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\Big)\Big] \\
&\le \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}[f^M(\pi_M) - f^M(\pi^{(t)})] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\Big[D^2_{\mathrm{H}}\Big(M(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\Big)\Big] \\
&= \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma\cdot D^2_{\mathrm{H}}\Big(M(\pi), \widehat{M}^{(t)}(\pi)\Big)\Big] \\
&= \mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^{(t)}).
\end{aligned} \tag{23}$$

We conclude that

$$\mathbf{Reg}_{\mathsf{DM}} \le \max_t \mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^{(t)})\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{H}}.$$

From here, all that remains is to tune γ and show that we can choose the oracle $\mathbf{Alg}_{\mathsf{Est}}$ such that $\mathbf{Est}_{\mathsf{H}} \lesssim \log|\mathcal{M}|$ with high probability and $\widehat{M}^{(t)} \in \mathrm{co}(\mathcal{M})$.

Achieving the localized result in Theorem 3.3 requires more effort, and relies on Option II. The key idea is to use certain properties of the minimax program (21) to relate the value of the Decision-Estimation Coefficient for the confidence sets $\mathcal{M}^{(t)}$ to the value of the localized DEC appearing in (22). □

3.2.3 Infinite Model Classes

In order to incorporate rich, potentially nonparametric classes of models, we provide a generalization of Theorem 3.3 based on covering numbers. This extension is important in our applications to bandits and reinforcement learning (Sections 6 and 7).

[7] Here, "non-trivial" means that the class embeds the two-armed bandit problem.

Definition 3.2 (Model class covering number). *A model class $\mathcal{M}' \subseteq \mathcal{M}$ is said to be an ε-cover for $\mathcal{M}$ if*

$$\forall M \in \mathcal{M} \quad \exists M' \in \mathcal{M}' \quad s.t. \quad \sup_{\pi \in \Pi} D^2_{\mathrm{H}}(M'(\pi), M(\pi)) \leq \varepsilon^2. \tag{24}$$

We let $\mathcal{N}(\mathcal{M}, \varepsilon)$ denote the size of the smallest such cover, and define

$$\mathsf{est}(\mathcal{M}, T) := \inf_{\varepsilon \geq 0} \left\{ \log \mathcal{N}(\mathcal{M}, \varepsilon) + \varepsilon^2 T \right\} \tag{25}$$

as a fundamental complexity parameter associated with $\mathcal{M}$.

Basic examples include parametric models in d dimensions (i.e., $\log \mathcal{N}(\mathcal{M}, \varepsilon) \propto d \log(1/\varepsilon)$) where $\mathsf{est}(\mathcal{M}, T) = \widetilde{O}(d)$, and nonparametric models with $\log \mathcal{N}(\mathcal{M}, \varepsilon) \propto \varepsilon^{-\rho}$ for $\rho > 0$, where $\mathsf{est}(\mathcal{M}, T) = O(T^{\frac{\rho}{2+\rho}})$.

Beyond requiring bounded covering numbers, we place a (fairly weak) regularity condition on the class of densities associated with the model class $\mathcal{M}$, which is standard within the literature on density estimation (Opper and Haussler, 1999; Bilodeau et al., 2021). Recall (cf. Section 2) that $m^M(\cdot, \cdot \mid \cdot)$ denotes the conditional density for M under the common conditional measure $\nu(\cdot, \cdot \mid \cdot)$.

Assumption 3.2 (Bounded densities). *There exists a constant $B \geq e$ such that:*

1. $\nu(\mathcal{R} \times \mathcal{O} \mid \pi) \leq B$ *for all* $\pi \in \Pi$.
2. $\sup_{\pi \in \Pi} \sup_{(r,o) \in \mathcal{R} \times \mathcal{O}} m^M(r, o \mid \pi) \leq B$ *for all* $M \in \mathcal{M}$.

Our result scales with $\log(B)$, which is constant for bandit problems with bounded rewards, logarithmic for reinforcement learning problems with finite state/action spaces, and polynomial in dimension for continuous reinforcement learning problems. See Section 7 for further discussion.

Theorem 3.4 (Main upper bound—general version). *Let $\delta \in (0,1)$ be given, and let c_ℓ be as in Assumption 3.1. Assume that $\mathcal{R} \subseteq [0,1]$ and that Assumption 3.2 holds. Define $C_1 = O(c_\ell^2 \log_{c_\ell}(T) \log^2(BT))$, $C_2 = O(\log^2(BT))$, and $\bar{\varepsilon}_\gamma = C_2\big(\frac{\gamma}{T}(\mathsf{est}(\mathcal{M}, T) + \log(\delta^{-1})) + \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) + \gamma^{-1}\big)$. Then Algorithm 1, with an appropriate choice of parameters and estimation oracle, guarantees that with probability at least $1 - \delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq C_1 \cdot \min_{\gamma > 0} \max \left\{ \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}_{\bar{\varepsilon}_\gamma}(\overline{M}), \overline{M})) \cdot T, \ \gamma \cdot (\mathsf{est}(\mathcal{M}, T) + \log(\delta^{-1})) \right\}. \tag{26}$$

3.3 Learnability

In statistical learning and related settings, a long line of research on *learnability* provides necessary and sufficient conditions under which a hypothesis class under consideration is *learnable* in the sense that there exist algorithms with non-trivial sample complexity (Vapnik, 1995; Alon et al., 1997; Shalev-Shwartz et al., 2010; Rakhlin et al., 2010; Daniely et al., 2011). Equipped with our main upper and lower bounds, we use the Decision-Estimation Coefficient to provide an analogous characterization of learnability (i.e., existence of algorithms with sublinear regret) in the DMSO framework.

Our characterization of learnability applies to model classes $\mathcal{M}$ that are i) convex, and ii) admit non-trivial estimation complexity. In addition, we make the following mild regularity assumption.

Assumption 3.3. *There exists $M_0 \in \mathcal{M}$ such that f^{M_0} is constant.*

This assumption is satisfied for all of the examples considered in this paper. The characterization is as follows.

Theorem 3.5 (Learnability). *Assume that $\mathcal{R} \subseteq [0,1]$, and that Assumption 3.2 and Assumption 3.3 hold. Suppose that $\mathcal{M}$ is convex and has $\mathsf{est}(\mathcal{M}, T) = \widetilde{O}(T^q)$ for some $q < 1$. Then:*

1. *If there exists $\rho > 0$ such that $\lim_{\gamma \to \infty} \mathsf{dec}_\gamma(\mathcal{M}) \cdot \gamma^\rho = 0$, then there exists an algorithm for which*

$$\lim_{T \to \infty} \frac{\mathfrak{M}(\mathcal{M}, T)}{T^p} = 0$$

for some $p < 1$.

2. *If* $\lim_{\gamma\to\infty} \mathsf{dec}_\gamma(\mathcal{M}) \cdot \gamma^\rho > 0$ *for all* $\rho > 0$, *then any algorithm must have*

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

for all $p < 1$.

Theorem 3.5 shows that for any convex model class where the model estimation complexity $\mathsf{est}(\mathcal{M}, T)$ is sublinear, polynomial decay of the Decision-Estimation Coefficient is sufficient to achieve sublinear regret. Conversely, if the Decision-Estimation Coefficient does not decay polynomially, no algorithm can achieve sublinear regret. We emphasize that the latter result (necessity of polynomial decay) applies to any model class, regardless of whether it admits low estimation complexity.

It should be noted that learnability is—by definition—a coarse property, and is best thought of as a basic sanity check. As we show in Sections 5 to 7, our machinery is considerably more precise, and yields quantitative upper and lower bounds that accurately reflect problem-dependent parameters such as dimension.

3.4 Tighter Guarantees Based on Decision Space Complexity

Up to this point, all of the upper bounds we have presented depend on the model estimation complexity $\mathsf{est}(\mathcal{M}, T)$. In general, low model estimation complexity is not required to achieve low regret for decision making. This is because our end goal is to make good *decisions*, so we can give up on accurately estimating the model in regions of the decision space that do not help to distinguish the relative quality of decisions. Our final result for this section provides a tighter bound that replaces the model estimation complexity $\mathsf{est}(\mathcal{M}, T)$ with an analogous estimation complexity parameter for the decision space $\Pi_\mathcal{M}$, albeit at the cost of removing the localization found in Theorem 3.3. This result is non-constructive in nature, and is proven by using the minimax theorem to move to the Bayesian setting where the true model is drawn from a worst-case prior that is known to the learner (cf. Section 2.1), then running $\mathsf{E2D}$ over a special model class constructed with knowledge of the prior. We refer ahead to Section 4.2 for more background on this approach.

Our main upper bound—which applies to both finite and infinite classes—is stated in terms of the following, weaker notion of covering number, which is tailored to decision making.

Definition 3.3 (Decision space covering number)**.** *A decision set* $\Pi' \subseteq \Pi_\mathcal{M}$ *is an* ε*-cover for* $\Pi_\mathcal{M}$ *if*

$$\forall M \in \mathcal{M} \quad \exists \pi' \in \Pi' \quad s.t. \quad f^M(\pi_M) - f^M(\pi') \le \varepsilon. \tag{27}$$

We let $\mathcal{N}(\Pi_\mathcal{M}, \varepsilon)$ *denote the size of the smallest cover, and define*

$$\mathsf{est}(\Pi_\mathcal{M}, T) = \inf_{\varepsilon \ge 0} \{\log \mathcal{N}(\Pi_\mathcal{M}, \varepsilon) + \varepsilon T\} \tag{28}$$

as a fundamental complexity measure associated with $\Pi_\mathcal{M}$.

The result has the same structure as Theorem 3.3, but replaces $\mathsf{est}(\mathcal{M}, T)$ with $\mathsf{est}(\Pi_\mathcal{M}, T)$.

Theorem 3.6. *Whenever the conclusion of Proposition 2.1 holds, there exists an algorithm that ensures that*

$$\begin{aligned}
\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] &\le 2 \cdot \min_{\gamma>0} \max\left\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \;\; \inf_{\varepsilon \ge 0}\{\gamma \cdot \log \mathcal{N}(\Pi_\mathcal{M}, \varepsilon) + \varepsilon \cdot T\}\right\} && (29)\\
&\le 2 \cdot \min_{\gamma>0} \max\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \; (1+\gamma) \cdot \mathsf{est}(\Pi_\mathcal{M}, T)\}. && (30)
\end{aligned}$$

In particular, when $|\Pi_\mathcal{M}| < \infty$, *we have*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \le 2 \cdot \min_{\gamma>0} \max\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \gamma \cdot \log|\Pi_\mathcal{M}|\}. \tag{31}$$

As a concrete example, for the multi-armed bandit, where $\Pi_\mathcal{M} = [A]$ and $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \lesssim \frac{A}{\gamma}$, we have $\mathsf{est}(\mathcal{M}, T) = \widetilde{O}(A)$, while $\mathsf{est}(\Pi_\mathcal{M}, T) = \log A$; the latter bound leads to near-optimal regret $\sqrt{AT \log A}$.

There are some applications for which $\mathsf{est}(\mathcal{M}, T)$ correctly characterizes the precise minimax rate, but in general the estimation complexity for $\mathcal{M}$ can be arbitrarily large relative to $\Pi_{\mathcal{M}}$ (see Section 7 for discussion in the context of reinforcement learning). On the other hand, for all of the applications we are aware of, $\mathsf{est}(\Pi_{\mathcal{M}}, T)$ has a smaller contribution to regret than the Decision-Estimation Coefficient itself after balancing γ (e.g., $\log A$ versus A for the multi-armed bandit).

As a result of Theorem 3.6, we obtain the following strengthening of Theorem 3.5.

Theorem 3.7 (Learnability—refined version). *Suppose that $\mathsf{est}(\Pi_{\mathcal{M}}, T) = \widetilde{O}(T^q)$ for some $q < 1$, and that the conclusion of Proposition 2.1 holds, but place no assumption on $\mathsf{est}(\mathcal{M}, T)$. Then the conclusion of Theorem 3.5 continues to hold.*

3.5 Discussion

We close by highlighting some gaps between our quantitative upper and lower bounds, and opportunities for improvement. We also discuss some gaps in our understanding of regret in the frequentist setting and Bayesian setting. Many of these issues are immaterial from the perspective of understanding learnability, but are important for understanding precise minimax rates.

3.5.1 When are the Upper and Lower Bounds Tight?

The lower and upper bounds from Theorems 3.1 to 3.3 have a nearly identical functional form, with the main differences involving localization, model estimation complexity, and convexity. In what follows, we discuss each of these gaps and the extent to which they can be closed.

Localization. Both Theorem 3.1 and Theorem 3.3 depend on the Decision-Estimation Coefficient for the localized model class $\mathcal{M}_{\varepsilon}(\overline{M})$ rather than the full model class. The localization radius in the upper bound is larger than that of the lower bound (in the non-trivial parameter regime $\gamma \in [1, T]$), and we suspect it can be tightened through more sophisticated arguments.

More broadly, we do not yet fully understand whether localization plays a critical role in determining precise minimax rates. On one hand, it is not hard to construct examples for which localization can significantly improve regret. For example, if $\mathcal{M}$ is large but there is a critical radius ε for which $\mathcal{M}_{\varepsilon}(\overline{M}) = \{\overline{M}\}$, the regret bound in Theorem 4.1 (the general version of Theorem 3.3) is constant. However, for all of the applications in bandits and reinforcement learning that we are aware of, localization seems to only lead to improvements in logarithmic factors, so we have not yet pursued this line of investigation further. We refer interested readers to Appendix B, which shows that for certain "reasonable" model classes, the global and local DEC coincide up to constant factors.

In-expectation lower bounds. Theorem 3.2 provides lower bounds based on the DEC for expected regret, but relies the L_{∞} notion of localization in (12). While this notion suffices to derive the lower bounds in Table 1, it is not difficult to construct examples where L_{∞}-localization is too loose, where the lower bound cannot be achieved (in contrast to the weaker notion of localization in (9), which is achieved by Algorithm 1). In future work we hope to derive an in-expectation version of the lower bound in Theorem 3.1.

Estimation error. The most notable distinction between our upper and lower bounds is the dependence on the complexity of estimation for $\mathcal{M}$ ($\mathsf{est}(\mathcal{M}, T)$ for Theorem 3.3 and $\mathsf{est}(\Pi_{\mathcal{M}}, T)$ for Theorem 3.6). The correct dependence on the estimation complexity is a subtle issue which cannot be fully addressed without introducing additional complexity parameters.

- For the finite-armed bandit problem where $\Pi_{\mathcal{M}} = [A]$, we have $\mathsf{est}(\Pi_{\mathcal{M}}, T) = \log A$. As a result, Theorem 3.6 gives an upper bound scaling with $\sqrt{AT \log A}$, while the lower bounds from Theorems 3.1 and 3.2 scale as $\sqrt{AT}$. In this case, the lower bound coincides with the minimax rate (Audibert and Bubeck, 2009), and the upper bound can be improved by $\sqrt{\log A}$.

- For linear bandits on the unit ball in $\mathbb{R}^d$, $\mathsf{est}(\Pi_{\mathcal{M}}, T)$ scales as $\widetilde{O}(d)$. Here, Theorem 3.6 gives regret $\widetilde{O}(d\sqrt{T})$, while the lower bounds from Theorems 3.1 and 3.2 scale with $\sqrt{dT}$. In this case, the upper bound is tight, and the lower bound can be improved by $\sqrt{d}$ (Lattimore and Szepesvári, 2020).

These observations show that fully resolving the correct dependence on estimation complexity requires additional problem complexity parameters. We leave this issue for future work, but we emphasize that for all the applications we are aware of, the decision space estimation complexity $\mathsf{est}(\Pi_{\mathcal{M}}, T)$ has a smaller contribution to regret than the Decision-Estimation Coefficient itself (after balancing γ), and does not appear to be the deciding factor in whether a problem is learnable.

Finally, we mention in passing that the bound on the estimation complexity in Theorem 3.4 depends on a rather coarse point-wise notion of model covering number, and it would be useful to improve this result to take advantage of refined (e.g., empirical or sequential) covering numbers (Rakhlin and Sridharan, 2012); this is out of scope for the present work.

Convexity. Ignoring localization, the lower bounds in Theorems 3.1 and 3.2 scale with $\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$, while the upper bound in Theorem 3.3 scales with $\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$ and the upper bound in Theorem 3.6 scales with $\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathrm{co}(\mathcal{M}), \overline{M})$. That is, both upper bounds require evaluating the DEC with models in the convex hull of $\mathcal{M}$, which is absent from the lower bound. Similar to the situation with estimation error above, this is a subtle issue. The following proposition highlights an example for which removing convexification in the upper bound is impossible.

Proposition 3.1. *For any $A \in \mathbb{N}$, there exists a structured bandit problem with $|\mathcal{M}| = |\Pi_{\mathcal{M}}| = A$ for which $\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \leq (2\gamma)^{-1}$ for all $\gamma > 0$, yet any algorithm must have regret $\Omega(A)$.*

If we could remove the convexification in our upper bounds, it would lead to regret scaling with $\sqrt{T \log A}$ for this example, contradicting the $\Omega(A)$ lower bound when $T \ll A$. Hence, removing convexification is not possible without introducing additional complexity parameters to the upper bound. Similar issues have been noted in the context of statistical estimation by Donoho and Liu (1991a). We leave the problem of developing a more refined understanding for future research.

Other technical differences. Theorem 3.1 restricts to $\gamma > 8e\sqrt{T}$, whereas Theorem 3.3 allows for any $\gamma > 0$. This is essentially without loss of generality: Whenever $\mathsf{dec}_{\gamma,\underline{\varepsilon}_{\gamma}}(\mathcal{M}) \geq \gamma^{-1}$, which holds for any non-trivial class, the optimal choice for γ satisfies $\gamma \geq \sqrt{T}$.

3.5.2 Regret Bounds: Frequentist vs. Bayesian and Constructive vs. Non-Constructive

Our localized upper bounds, Theorem 3.3 and Theorem 3.4, are proven directly in the DMSO framework described in Section 1 (the "frequentist" setting) by running Algorithm 1 with a particular choice of estimation oracle. Theorem 3.6, which replaces the model estimation complexity $\mathsf{est}(\mathcal{M}, T)$ with the decision space estimation complexity $\mathsf{est}(\Pi_{\mathcal{M}}, T)$, is proven non-constructively by exhibiting an upper bound on regret in the Bayesian setting (cf. Section 2.1). Informally, working in the Bayesian setting is a powerful tool, because it allows one to "average out" models that agree on the optimal action, thereby reducing the estimation complexity. We hope to understand whether there is a frequentist analogue of this principle through future research. Another interesting question is whether one can achieve the dependence on $\mathcal{N}(\Pi_{\mathcal{M}}, \varepsilon)$ in Theorem 3.6 and the localization in Theorem 3.4 simultaneously.

4 The E2D Meta-Algorithm: General Toolkit

In this section we provide a general toolkit for deriving regret bounds and efficient algorithms using E2D meta-algorithm (Algorithm 1). We begin with a regret bound for E2D with generic online estimation oracles (Section 4.1), then provide a dual Bayesian view of the Decision-Estimation Coefficient and E2D algorithm (Section 4.2), and finally close with some simple extensions (Section 4.3).

4.1 Guarantees for General Online Estimation Oracles

The $\mathsf{E2D}$ meta-algorithm can be applied with any user-specified estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$. For general oracles, the decision making performance of $\mathsf{E2D}$ depends on the estimation performance of the oracle, which we measure via cumulative Hellinger error:

$$\mathbf{Est}_{\mathsf{H}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}}\left[D^2_{\mathrm{H}}\left(M^{\star}(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\right)\right]. \tag{32}$$

Our most general theorem shows that for any choice of oracle, $\mathsf{E2D}$ inherits the Hellinger estimation error as a bound on decision making regret, thereby bridging estimation and decision making.

To state the result, we define $\widehat{\mathcal{M}}^{(t)}$ as any set such that $\widehat{M}^{(t)} \in \widehat{\mathcal{M}}^{(t)}$ almost surely for all t, and define $\widehat{\mathcal{M}} = \cup_{t\geq 1}\widehat{\mathcal{M}}^{(t)}$. To provide high probability guarantees, we make the following assumption.

Assumption 4.1. *The estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ guarantees for any $T \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$, $\mathbf{Est}_{\mathsf{H}} \leq \mathbf{Est}_{\mathsf{H}}(T,\delta)$, where $\mathbf{Est}_{\mathsf{H}}(T,\delta)$ is a known upper bound.*

For example, as we show in the sequel, whenever $\mathcal{M}$ is finite, Vovk's aggregating algorithm (Vovk, 1995) ensures that $\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq \log|\mathcal{M}|$, and satisfies Assumption 4.1 with $\mathbf{Est}_{\mathsf{H}}(T,\delta) = 2\log(|\mathcal{M}|/\delta)$.

Theorem 4.1. *Algorithm 1 with* Option I *and exploration parameter $\gamma > 0$ guarantees that*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sup_{\overline{M}\in\widehat{\mathcal{M}}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{H}} \tag{33}$$

almost surely. Furthermore, when Assumption 4.1 holds, Algorithm 1 with Option I *guarantees that for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sup_{\overline{M}\in\widehat{\mathcal{M}}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{H}}(T,\delta). \tag{34}$$

Finally, fix $\delta \in (0,1)$ and consider Algorithm 1 with Option II *and $R^2 = \mathbf{Est}_{\mathsf{H}}(T,\delta)$. Suppose that $\mathbf{Alg}_{\mathsf{Est}}$ has $\widehat{\mathcal{M}}^{(t)} \subseteq \mathrm{co}(\mathcal{M}^{(t)})$ for all t, and that $\mathcal{R} \subseteq [0,1]$. Then for any fixed $T \in \mathbb{N}$ and $\gamma > 0$, with probability at least $1-\delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sum_{t=1}^{T} \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}_{\varepsilon_t}(\overline{M}), \overline{M}) + \gamma \cdot \mathbf{Est}_{\mathsf{H}}(T,\delta), \tag{35}$$

where $\varepsilon_t := 6\frac{\gamma}{t}\mathbf{Est}_{\mathsf{H}}(T,\delta) + \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) + (2\gamma)^{-1}$.

The guarantees in (33) and (34) concern the simpler variant of $\mathsf{E2D}$ (Option I), which is more practical to implement but does not achieve localization. When invoked with Vovk's aggregating algorithm as described above, the regret for Option I scales as roughly $\min_{\gamma>0}\max\{\sup_{\overline{M}\in\widehat{\mathcal{M}}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \cdot T,\ \gamma \cdot \log|\mathcal{M}|\}$, which matches Theorem 3.3 modulo localization. The guarantee in (35) concerns Option II, and constitutes our most general regret guarantee bound on the localized Decision-Estimation Coefficient. When invoked with a variant of the aggregating algorithm designed to satisfy the requisite condition that $\widehat{\mathcal{M}}^{(t)} \subseteq \mathrm{co}(\mathcal{M}^{(t)})$ (Appendix A.3), this result recovers Theorem 3.3 as a special case.

Remark 4.1 (Inexact minimizers)**.** All of the results regarding $\mathsf{E2D}$ (Algorithm 1) are stated for the case where $p^{(t)}$ is chosen to exactly solve the minimax problem

$$\operatorname*{arg\,min}_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}^{(t)}} \mathcal{V}_{\gamma}^{\widehat{M}^{(t)}}(p, M),$$

for a model class $\mathcal{M}^{(t)}$ and estimator $\widehat{M}^{(t)}$. This leads to regret bounds that scale with $\mathsf{dec}_{\gamma}(\mathcal{M}^{(t)}, \widehat{M}^{(t)})$. If we instead use a distribution $p^{(t)}$ that *certifies an upper bound on the DEC* in the sense that

$$\sup_{M\in\mathcal{M}^{(t)}} \mathcal{V}_{\gamma}^{\widehat{M}^{(t)}}(p^{(t)}, M) \leq g_{\gamma}(\mathcal{M}^{(t)}, \widehat{M}^{(t)})$$

for some function $g_{\gamma}(\cdot,\cdot)$, then all of the main theorems (Theorems 3.3, 3.4, 3.6 and 4.1 to 4.3) continue to hold, with occurrences of $\mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$ replaced by $g_{\gamma}(\mathcal{M}, \overline{M})$. We tacitly use this fact in later sections.

4.1.1 Online Estimation Oracles: Density Estimation and Examples

A topic we have not yet addressed is how to go about minimizing the Hellinger estimation error in (32). Building on classical literature in statistical estimation, we show that this problem can generically be solved via *online density estimation*. For each example $(\pi^{(t)}, r^{(t)}, o^{(t)})$, define the logarithmic loss for a model M as

$$\ell_{\log}^{(t)}(M) = \log\left(\frac{1}{m^{M}(r^{(t)}, o^{(t)} \mid \pi^{(t)})}\right), \tag{36}$$

where we recall that $m^{M}(\cdot, \cdot \mid \pi)$ is the conditional density for (r, o) under M (cf. Section 2). As an intermediate quantity, we consider regret under the logarithmic loss:

$$\mathbf{Reg}_{\mathsf{KL}} = \sum_{t=1}^{T} \ell_{\log}^{(t)}(\widehat{M}^{(t)}) - \inf_{M \in \mathcal{M}} \sum_{t=1}^{T} \ell_{\log}^{(t)}(M). \tag{37}$$

Regret minimization with the logarithmic loss (also known as *sequential probability assignment*) is a fundamental problem in online learning. Efficient algorithms are known for model classes of interest (Cover, 1991; Vovk, 1995; Kalai and Vempala, 2002; Hazan and Kale, 2015; Orseau et al., 2017; Rakhlin et al., 2015; Foster et al., 2018; Luo et al., 2018), and this is complemented by theory which provides minimax rates for generic model classes (Shtar'kov, 1987; Opper and Haussler, 1999; Cesa-Bianchi and Lugosi, 1999; Bilodeau et al., 2020). Canonical examples include finite classes, where Vovk's aggregating algorithm guarantees $\mathbf{Reg}_{\mathsf{KL}} \leq \log|\mathcal{M}|$ for every sequence,[8] and linear models (i.e., $M(r, o \mid \pi) = \langle \phi(r, o, \pi), \theta \rangle$ for a fixed feature map in $\phi \in \mathbb{R}^d$), where algorithms with $\mathbf{Reg}_{\mathsf{KL}} = O(d \log(T))$ are known (Rissanen, 1986; Shtar'kov, 1987). All of these algorithms satisfy $\widehat{\mathcal{M}} = \mathrm{co}(\mathcal{M})$.[9] We refer the reader to Chapter 9 of Cesa-Bianchi and Lugosi (2006) for further examples and discussion.

The following result shows that a bound on the log-loss regret immediately yields a bound on the Hellinger estimation error.

Proposition 4.1. *For any estimation algorithm $\mathbf{Alg}_{\mathsf{Est}}$, whenever Assumption 1.1 holds,*

$$\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq \mathbb{E}[\mathbf{Reg}_{\mathsf{KL}}]. \tag{38}$$

Furthermore, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\mathbf{Est}_{\mathsf{H}} \leq \mathbf{Reg}_{\mathsf{KL}} + 2\log(\delta^{-1}). \tag{39}$$

For a proof, refer to Lemma A.14 in Appendix A, which gives a more general version of this result. Further examples of estimation oracles are given throughout Sections 5 to 7.

4.2 Dual Perspective and Connection to Posterior Sampling

The Decision-Estimation Coefficient (2) is a min-max optimization problem, and can be interpreted as a game in which the learner (the "min" player) aims to find a decision distribution p that optimally trades off regret and information acquisition in the face of an adversary (the "max" player) that selects a worst-case model in $\mathcal{M}$. We can define a natural *dual* (or, max-min) analogue of the DEC via

$$\underline{\mathsf{dec}}_{\gamma}(\mathcal{M}, \overline{M}) = \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p \in \Delta(\Pi)} \mathbb{E}_{M \sim \mu} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot D_{\mathrm{H}}^{2}\big(M(\pi), \overline{M}(\pi)\big)\right]. \tag{40}$$

The dual Decision-Estimation Coefficient has the following Bayesian interpretation. The adversary selects a *prior* distribution μ over models in $\mathcal{M}$, and the learner (with knowledge of the prior) finds a decision distribution p that balances the tradeoff between regret and information acquisition when the underlying model is drawn from μ.

[8]See Appendix A.3 for detailed guarantees.

[9]In fact, even for improper estimators that do not satisfy $\widehat{\mathcal{M}} = \mathrm{co}(\mathcal{M})$, it is always possible to project $\widehat{M}^{(t)}$ onto $\mathrm{co}(\mathcal{M})$ while maintaining the estimator's bound on $\mathbb{E}[\mathbf{Est}_{\mathsf{H}}]$.

Algorithm 2 E2D.Bayes

1: **parameters**:

Prior $\mu \in \Delta(\mathcal{M})$.

Exploration parameter $\gamma > 0$.

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

3: Define $\overline{M}^{(t)}(\pi) = \mathbb{E}[M^\star(\pi) \mid \mathcal{H}^{(t-1)}]$ and $\overline{M}^{(t)}_{\pi'}(\pi) = \mathbb{E}[M^\star(\pi) \mid \pi^\star = \pi', \mathcal{H}^{(t-1)}]$ for all $\pi' \in \Pi$.

4: Compute coarsened posterior $\mu^{(t)} \in \Delta(\mathcal{M})$ via $\mu^{(t)}(\{\overline{M}^{(t)}_{\pi}\}) = \mathbb{P}(\pi^\star = \pi \mid \mathcal{H}^{(t-1)})$.

5: Let $p^{(t)} = \arg\min_{p\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu^{(t)}} \mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}^{(t)}(\pi)\big)\big]$.

6: Sample decision $\pi^{(t)} \sim p^{(t)}$ and update $\mathcal{H}^{(t)} = \mathcal{H}^{(t-1)} \cup \{(\pi^{(t)}, r^{(t)}, o^{(t)})\}$.

The connection between the primal and dual DEC is analogous to the connection between primal and dual regret (cf. Section 2.1). As with regret, the primal and dual DEC can be shown to coincide under mild regularity conditions.[10]

Proposition 4.2. *Suppose that Π is finite and $\mathcal{R}$ is bounded. Then for all models $\overline{M}$,*

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}). \tag{41}$$

As a consequence, any bound on the dual DEC immediately yields a bound on the primal DEC. This perspective is useful because it allows us to bring existing tools for Bayesian bandits and reinforcement learning to bear on the primal Decision-Estimation Coefficient. For example, *probability matching* is a well-known Bayesian strategy that—when applied to our setting—uses the distribution p induced by sampling $M \sim \mu$ and selecting π_M. Using analysis from Russo and Van Roy (2014), one can show that this strategy certifies

$$\underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}) \le \frac{|\Pi|}{4\gamma}$$

for the finite-armed bandit; see Section 5 for a proof. Using more sophisticated analysis techniques, we use this approach to bound the Decision-Estimation Coefficient for a general class of structured bandit problems with bounded *star number* in Section 6, and provide bounds for reinforcement learning in Section 7. In fact, many prior results for the Bayesian setting can be viewed as implicitly providing bounds on the dual Decision-Estimation Coefficient (Russo and Van Roy, 2014; Bubeck et al., 2015; Bubeck and Eldan, 2016; Russo and Van Roy, 2018; Lattimore and Szepesvári, 2019; Lattimore, 2020); cf. Section 9.

Bayesian E2D as a generalization of posterior sampling. Beyond the primal and dual Decision-Estimation Coefficient, there are deeper connections between our approach and Bayesian approaches. Consider the Bayesian version of the DMSO framework from Section 2.1, in which the underlying model $M^\star$ is drawn from a known prior $\mu \in \Delta(\mathcal{M})$, and our objective is to minimize the Bayesian regret

$$\mathbb{E}_{M^\star\sim\mu}\,\mathbb{E}^{M^\star}[\mathbf{Reg}_{\mathsf{DM}}].$$

For this setting, the celebrated *posterior sampling* (or, Thompson sampling) algorithm (Thompson, 1933; Agrawal and Goyal, 2012; Russo and Van Roy, 2014) applies the probability matching idea as follows. At each round t, we first compute the conditional law for $M^\star$ given the observed history $\mathcal{H}^{(t-1)}$, which we denote by $\mu^{(t)} = \mu^{(t)}(\cdot \mid \mathcal{H}^{(t-1)})$. We then sample $M \sim \mu^{(t)}$ and play $\pi^{(t)} = \pi_M$. This algorithm leads to Bayesian regret bounds for basic problem settings such as finite-armed bandits, linear bandits, and tabular reinforcement learning (Osband et al., 2013; Russo and Van Roy, 2014; Osband and Van Roy, 2017).

We provide a Bayesian analogue of the E2D algorithm (Algorithm 2), which may be viewed as a generalization of posterior sampling. At each round, the algorithm computes an induced model $\overline{M}^{(t)}(\pi) = \mathbb{E}[M^\star(\pi) \mid \mathcal{H}^{(t-1)}]$,

[10] As with Proposition 2.1, we expect that this result can be proven to hold under weaker conditions.

Algorithm 3 E2D for General Divergences and Randomized Estimators

1: **parameters**:
  Online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$.
  Exploration parameter $\gamma > 0$.
  Divergence $D(\cdot \,\|\, \cdot)$.
2: **for** $t = 1, 2, \cdots, T$ **do**
3:   Compute randomized estimate $\nu^{(t)} = \mathbf{Alg}^{(t)}_{\mathsf{Est}}\Big(\{(\pi^{(i)}, r^{(i)}, o^{(i)})\}_{i=1}^{t-1}\Big)$.
4:   $p^{(t)} \leftarrow \arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M}\sim\nu^{(t)}}\, D\Big(M(\pi) \,\|\, \widehat{M}(\pi)\Big)\Big]$. // Eq. (45).
5:   Sample decision $\pi^{(t)} \sim p^{(t)}$ and update estimation oracle with $(\pi^{(t)}, r^{(t)}, o^{(t)})$.

and a distribution $\mu^{(t)} \in \Delta(\mathcal{M})$ which may be viewed as a coarsened version of the posterior distribution, then solves the optimization problem

$$\arg\min_{p\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu^{(t)}}\, \mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}^{(t)}(\pi)\big)\big]. \tag{42}$$

The algorithm then samples the decision $\pi^{(t)}$ from the resulting distribution. It is straightforward to show that this algorithm has the following guarantee, which specializes Theorem 3.6.[11]

Theorem 4.2. *For parameter $\gamma > 0$, Algorithm 2 guarantees that*

$$\mathbb{E}_{M^\star\sim\mu}\, \mathbb{E}^{M^\star}[\mathbf{Reg}_{\mathsf{DM}}] \leq \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \underline{\mathsf{dec}}_\gamma(\mathrm{co}(\mathcal{M}), \overline{M}) \cdot T + \gamma \cdot \log|\Pi_{\mathcal{M}}|. \tag{43}$$

This algorithm generalizes posterior sampling by replacing the naive probability matching strategy with the optimization problem in (42). It is also closely related to the *information-directed sampling* algorithm of Russo and Van Roy (2018). However, Algorithm 2 can be applied in settings where information directed sampling fails, and—via Theorem 3.5—achieves non-trivial regret whenever non-trivial regret is possible; see Section 9 for further discussion. More broadly, the algorithm illustrates that our methodology extends beyond the frequentist setting.

4.3 General Divergences and Randomized Estimators

In this section we give a generalization of the E2D algorithm that incorporates two extra features: *general divergences* and *randomized estimators*.

General divergences. The Decision-Estimation Coefficient measures estimation error via the Hellinger distance $D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)$. By providing a characterization for learnability (Section 3.3), we show that the choice of Hellinger distance here is fundamental. Nonetheless, for specific applications and model classes, it can be useful to work with alternative distance measures and divergences. For a general divergence $D : \mathrm{co}(\mathcal{M}) \times \mathrm{co}(\mathcal{M}) \to \mathbb{R}_+$, we define

$$\mathsf{dec}^D_\gamma(\mathcal{M}, \overline{M}) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\Big]. \tag{44}$$

This variant of the DEC naturally leads to regret bounds in terms of estimation error under $D(\cdot \,\|\, \cdot)$.

Randomized estimators. The basic version of E2D (Algorithm 1) assumes that at each round, the estimation algorithm $\mathbf{Alg}_{\mathsf{Est}}$ provides a point estimate $\widehat{M}^{(t)}$. In some settings, it useful to consider *randomized*

[11] The proof of Theorem 3.6 (see Eq. (137)) extends this result to infinite classes, and is proven in Appendix C.2.2.

estimators that, at each round, produce a distribution $\nu^{(t)}$ over models. For this setting, we further generalize the DEC by defining

$$\mathsf{dec}_{\gamma}^{D}(\mathcal{M},\nu) = \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot\mathbb{E}_{\overline{M}\sim\nu}\big[D\big(M(\pi)\;\|\;\overline{M}(\pi)\big)\big]\right] \tag{45}$$

for distributions $\nu\in\Delta(\mathcal{M})$.

Algorithm. A generalization of $\mathsf{E2D}$ that incorporates general divergences and randomized estimators is given in Algorithm 3. The algorithm is identical to $\mathsf{E2D}$ with Option I, with the only differences being that i) we play the distribution that solves the minimax problem (45) with the user-specified divergence $D(\cdot\;\|\;\cdot)$ rather than squared Hellinger distance, and ii) we use the randomized estimate $\nu^{(t)}$ rather than a point estimate. Our performance guarantee for this algorithm depends on the estimation performance of the oracle's randomized estimates $\nu^{(1)},\ldots,\nu^{(T)}$ with respect to the given divergence D, which we define as

$$\mathbf{Est}_{\mathsf{D}} := \sum_{t=1}^{T}\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\,\mathbb{E}_{\widehat{M}^{(t)}\sim\nu^{(t)}}\left[D\Big(M^{\star}(\pi^{(t)})\;\|\;\widehat{M}^{(t)}(\pi^{(t)})\Big)\right]. \tag{46}$$

Let $\widehat{\mathcal{M}}$ be any set for which $\widehat{M}^{(t)}\in\widehat{\mathcal{M}}$ for all t almost surely. We have the following guarantee.

Theorem 4.3. *Algorithm 3 with exploration parameter $\gamma>0$ guarantees that*

$$\mathbf{Reg}_{\mathsf{DM}} \le \sup_{\nu\in\Delta(\widehat{\mathcal{M}})}\mathsf{dec}_{\gamma}^{D}(\mathcal{M},\nu)\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{D}} \tag{47}$$

almost surely.

In bandit problems, it is often convenient to work with the divergence $D_{\mathsf{Sq}}\big(M(\pi),\overline{M}(\pi)\big) := (f^{M}(\pi)-f^{\overline{M}}(\pi))^2$, which uses the mean reward function as a sufficient statistic. With this choice of distance, the minimax problem in (44) recovers the action selection strategy used in the $\mathsf{SquareCB}$ contextual bandit algorithm (Abe and Long, 1999; Foster and Rakhlin, 2020). This distance also recovers a generalization of $\mathsf{SquareCB}$ for infinite, linearly-structured action spaces given in Foster et al. (2020a). Note that whenever $\mathcal{F}_{\mathcal{M}}$ has range in $[0,1]$, one has $\mathsf{dec}_{\gamma}^{\mathrm{H}}(\mathcal{M},\overline{M})\le\mathsf{dec}_{\gamma}^{\mathrm{Sq}}(\mathcal{M},\overline{M})$,[12] and in general this inequality is strict, so working with this distance is not always sufficient. The advantage, however, is that for D_{Sq} the estimation error in (89) can be minimized directly using square loss estimation rather than density estimation, which can lead to simpler algorithms and tighter bounds. For example, one can generically obtain $\mathbf{Est}_{\mathsf{Sq}}\le\log|\mathcal{F}_{\mathcal{M}}|$ for finite classes, which can be much tighter than the analogous bound $\mathbf{Est}_{\mathsf{H}}\le\log|\mathcal{M}|$ for Hellinger error. See Foster and Rakhlin (2020) for more background on square loss estimation oracles.

Another advantage of working with general distances is to simplify analysis of the minimax value. For example, some of our lower bounds on the DEC proceed by moving to KL divergence and lower bounding $\mathsf{dec}_{\gamma}^{\mathrm{H}}(\mathcal{M},\overline{M})\ge\mathsf{dec}_{\gamma}^{\mathrm{KL}}(\mathcal{M},\overline{M})$.

We mention without proof that it is also possible to extend Option II of $\mathsf{E2D}$ to accommodate general divergences. In this case, we require the additional assumption that the divergence D is bounded, symmetric, and satisfies the triangle inequality up to a multiplicative constant.

5 Illustrative Examples

In this section, we show how to bound the Decision-Estimation Coefficient and efficiently instantiate the $\mathsf{E2D}$ meta-algorithm for two canonical problem settings: multi-armed bandits and tabular (finite state/action) reinforcement learning. Through the duality introduced in Section 4.2, we showcase the use of both frequentist and Bayesian approaches to bound the DEC. The proofs in this section illustrate key concepts and technical tools that prove useful when we consider richer, more structured settings in Section 6 and Section 7.

[12] Throughout the paper we abbreviate $\mathsf{dec}_{\gamma}^{\mathrm{H}}=\mathsf{dec}_{\gamma}^{D_{\mathrm{H}}^2}$, $\mathsf{dec}_{\gamma}^{\mathrm{KL}}=\mathsf{dec}_{\gamma}^{D_{\mathrm{KL}}}$, $\mathsf{dec}_{\gamma}^{\mathrm{Sq}}=\mathsf{dec}_{\gamma}^{D_{\mathrm{Sq}}}$, and so on.

5.1 Multi-Armed Bandits

For the first example, we provide upper and lower bounds on the Decision-Estimation Coefficient for the classical multi-armed bandit setting. Here, we have $\Pi = [A]$, $\mathcal{R} = [0,1]$, and $\mathcal{O} = \{\varnothing\}$, and the model class $\mathcal{M} = \{M : M(\pi) \in \Delta(\mathcal{R})\}$ consists of all possible distributions over $\mathcal{R}$.

To derive upper bounds, it is simpler to work with the square loss variant of the DEC from Section 4.3,

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

which has $\mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \leq \mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M})$ whenever rewards lie in $[0,1]$. Intuitively, the reason why working with this coarse notion of distance suffices is that—beyond the mean reward for each decision—the noise distribution provides little information about the underlying problem instance for this unstructured setting.

5.1.1 Bayesian Upper Bound via Posterior Sampling

We first show how to bound the Decision-Estimation Coefficient using a simple, yet indirect approach that leverages minimax duality. Recall from Section 4.2 that by the minimax theorem (Proposition 4.2), we have

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) = \underline{\mathsf{dec}}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) = \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot(f^{M}(\pi) - f^{\overline{M}}(\pi))^2\big]. \tag{48}$$

Hence, to bound the DEC, it suffices to show that for any prior $\mu \in \Delta(\mathcal{M})$, we can choose the distribution $p \in \Delta(\Pi)$ such that the quantity in (48) is bounded. We have the following result.

Proposition 5.1. *Consider the multi-armed bandit setting with $\mathcal{R} = \mathbb{R}$. For any $\mu \in \Delta(\mathcal{M})$, the distribution $p(\pi) = \mu(\{M : \pi_M = \pi\})$ satisfies*

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot(f^{M}(\pi) - f^{\overline{M}}(\pi))^2\big] \leq \frac{A}{\gamma},$$

for all $\overline{M}$ and $\gamma > 0$. Consequently, $\underline{\mathsf{dec}}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) \leq \frac{A}{\gamma}$.

Proof of Proposition 5.1. Let μ be given and $\overline{M}$ and $\gamma > 0$ be fixed. Observe that by the AM-GM inequality

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \leq \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi)\big] + \frac{\gamma}{2}\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[(f^{M}(\pi) - f^{\overline{M}}(\pi))^2\big] + (2\gamma)^{-1}.$$

Hence, it suffices to bound the first term above. Since $\pi \sim p$ and π_M under $M \sim \mu$ are identically distributed, we have

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi)\big] = \mathbb{E}_{M\sim\mu}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M})\big].$$

Next, we apply Cauchy-Schwarz to bound by

$$\begin{aligned}\mathbb{E}_{M\sim\mu}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M})\big] &= \mathbb{E}_{M\sim\mu}\left[\frac{p^{1/2}(\pi_{M})}{p^{1/2}(\pi_{M})}(f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M}))\right] \\ &\leq \left(\mathbb{E}_{M\sim\mu}\left[\frac{1}{p(\pi_{M})}\right]\right)^{1/2} \cdot \big(\mathbb{E}_{M\sim\mu}\big[p(\pi_{M})(f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M}))^2\big]\big)^{1/2}.\end{aligned}$$

For the first term, we have

$$\mathbb{E}_{M\sim\mu}\left[\frac{1}{p(\pi_{M})}\right] = \sum_{\pi\in[A]:p(\pi)>0} \frac{\mu(\{M : \pi_M = \pi\})}{p(\pi)} \leq A,$$

while the second term has

$$\mathbb{E}_{M\sim\mu}\big[p(\pi_{M})(f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M}))^2\big] \leq \sum_{\pi\in[A]} p(\pi)\,\mathbb{E}_{M\sim\mu}\big[(f^{M}(\pi) - f^{\overline{M}}(\pi))^2\big] = \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[(f^{M}(\pi) - f^{\overline{M}}(\pi))^2\big].$$

Hence, by the AM-GM inequality,

$$\begin{aligned}\mathbb{E}_{M\sim\mu}\big[f^{M}(\pi_{M})-f^{\overline{M}}(\pi_{M})\big] &\leq \big(A\cdot\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[(f^{M}(\pi)-f^{\overline{M}}(\pi))^{2}\big]\big)^{1/2} \\ &\leq \frac{\gamma}{2}\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[(f^{M}(\pi)-f^{\overline{M}}(\pi))^{2}\big]+\frac{A}{2\gamma}.\end{aligned}$$

□

The key idea above, which originates from Russo and Van Roy (2014), is the use of Cauchy-Schwarz to "decouple" the random variables $M\sim\mu$ and π_M; this idea is generalized in Section 6.2.

Regret upper bound. By plugging the bound on the Decision-Estimation Coefficient from Proposition 5.1 into Theorem 3.6 (and noting that $\mathrm{co}(\mathcal{M})=\mathcal{M}$), we conclude existence of an algorithm with

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}]\leq 2\cdot\min_{\gamma>0}\max\left\{\frac{AT}{\gamma},\gamma\log A\right\}=4\sqrt{AT\log A}. \tag{49}$$

This matches the optimal rate for the multi-armed bandit problem up to the $\log A$ factor (Audibert and Bubeck, 2009). Furthermore, via Algorithm 2, this approach provides an explicit algorithm for the Bayesian setting.

5.1.2 Frequentist Upper Bound via Inverse Gap Weighting

The Bayesian approach in the prequel does not lead to an explicit strategy that achieves the value of the frequentist Decision-Estimation Coefficient. We now give an explicit approach based on the *inverse gap weighting* technique (Abe and Long, 1999; Foster and Rakhlin, 2020) recently popularized by Foster and Rakhlin (2020).

Proposition 5.2. *Consider the multi-armed bandit setting with $\mathcal{R}=\mathbb{R}$. For any $\overline{M}$ and $\gamma>0$, define*

$$p(\pi)=\frac{1}{\lambda+2\gamma(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi))}, \tag{50}$$

where $\lambda\in[1,A]$ is chosen such that $\sum_{\pi}p(\pi)=1$.[13] *This strategy guarantees*

$$\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M})-f^{M}(\pi)-\gamma\cdot(f^{M}(\pi)-f^{\overline{M}}(\pi))^{2}\big]\leq\frac{A}{\gamma},$$

and consequently certifies that $\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M},\overline{M})\leq\frac{A}{\gamma}$.

Conceptually, this result shows that inverse gap weighting can be thought of as a frequentist counterpart to posterior sampling.

Proof of Proposition 5.2. Let $M\in\mathcal{M}$ be fixed. As with the proof of Proposition 5.1, we first apply the AM-GM inequality to bound

$$\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{M}(\pi)]\leq\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M})-f^{\overline{M}}(\pi)\big]+\frac{\gamma}{2}\mathbb{E}_{\pi\sim p}\big[(f^{M}(\pi)-f^{\overline{M}}(\pi))^{2}\big]+(2\gamma)^{-1}.$$

Next, we write

$$\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M})-f^{\overline{M}}(\pi)\big]=f^{M}(\pi_{M})-f^{\overline{M}}(\pi_{\overline{M}})+\mathbb{E}_{\pi\sim p}\big[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\big],$$

and observe that

$$\mathbb{E}_{\pi\sim p}\big[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\big]=\sum_{\pi\neq\pi_{\overline{M}}}\frac{f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)}{\lambda+2\gamma(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi))}\leq\frac{A-1}{2\gamma}.$$

[13] The normalizing constant $\lambda\in[1,A]$ is always guaranteed to exist because we have $\frac{1}{\lambda}\leq\sum_{\pi}p(\pi)\leq\frac{A}{\lambda}$, and because $\lambda\mapsto\sum_{\pi}p(\pi)$ is continuous over $[1,A]$.

To bound the final term, we apply the AM-GM inequality:

$$\begin{aligned}
f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) &= f^M(\pi_M) - f^{\overline{M}}(\pi_M) - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)) \\
&\leq \frac{\gamma}{2} p(\pi_M)(f^M(\pi_M) - f^{\overline{M}}(\pi_M))^2 + \frac{1}{2\gamma p(\pi_M)} - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)) \\
&\leq \frac{\gamma}{2} \mathbb{E}_{\pi\sim p}\big[(f^M(\pi) - f^{\overline{M}}(\pi))^2\big] + \frac{1}{2\gamma p(\pi_M)} - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)).
\end{aligned}$$

We then use the definition of p to bound

$$\frac{1}{2\gamma p(\pi_M)} - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)) = \frac{\lambda + 2\gamma(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))}{2\gamma} - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)) = \frac{\lambda}{2\gamma} \leq \frac{A}{2\gamma}.$$

Combining all of these inequalities, we conclude that

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \leq \gamma\, \mathbb{E}_{\pi\sim p}\big[(f^M(\pi) - f^{\overline{M}}(\pi))^2\big] + \frac{A}{\gamma}.$$

□

Regret upper bound. Applying Theorem 4.3 with the bound on the DEC above, we conclude that $\mathsf{E2D}$ with inverse gap weighting ensures that for any estimation oracle,

$$\mathbf{Reg}_{\mathsf{DM}} \leq \frac{AT}{\gamma} + \gamma \cdot \mathbf{Est}_{\mathsf{Sq}},$$

where

$$\mathbf{Est}_{\mathsf{Sq}} = \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\Big[(f^{\widehat{M}^{(t)}}(\pi^{(t)}) - f^{M^\star}(\pi^{(t)}))^2\Big].$$

As a concrete example, following Foster and Rakhlin (2020), we can estimate the rewards using the Vovk-Azoury-Warmuth algorithm (Azoury and Warmuth, 2001; Vovk, 1998), which has $\mathbb{E}[\mathbf{Est}_{\mathsf{Sq}}] \leq A\log(T)$. The resulting algorithm has $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq A\sqrt{T\log(T)}$, and runs in time $O(A)$ per round. We mention in passing that with a more careful analysis, this approach can recover the $\sqrt{AT\log A}$ rate (49) derived through the Bayesian approach, but we do not pursue this here.

5.1.3 Lower Bound

We conclude this example by proving a lower bound on the Decision-Estimation Coefficient which matches the upper bounds derived above. We use this example to illustrate a general strategy to lower bound the DEC, which is used extensively throughout Sections 6 and 7. The lower bound strategy is based on the following notion of a *hard family of models.*

Definition 5.1 ((α,β,δ)-family)**.** *A reference model $\overline{M} \in \mathcal{M}$ and collection $\{M_1, \dots, M_N\}$ with $N \geq 2$ are said to be an (α,β,δ)-family if the following properties hold.*

1. Regret property. *There exist functions $u_i : \Pi \to [0,1]$, with $\sum_i u_i(\pi) \leq \frac{N}{2}\ \forall\pi$, such that*

$$f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \alpha \cdot (1 - u_i(\pi)).$$

2. Information property. *There exist functions $v_i : \Pi \to [0,1]$, with $\sum_i v_i(\pi) \leq 1\ \forall\pi$, such that*

$$D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \leq \beta \cdot v_i(\pi) + \delta.$$

Informally, any (α,β,δ)-family leads to a difficult decision making problem when N is large because a given action can have low regret or large information gain on at most one model in the family. The following lemma makes this idea precise.

Lemma 5.1. *Let $\mathcal{M}' = \{M_1, \ldots, M_N\} \subseteq \mathcal{M}$ be an (α, β, δ)-family with respect to $\overline{M}$. Then for all $\gamma > 0$*

$$\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \underline{\mathsf{dec}}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{\alpha}{2} - \gamma\left(\frac{\beta}{N} + \delta\right). \tag{51}$$

Proof. Choose $\mu = \mathrm{unif}(\{M_1, \ldots, M_N\})$ as the uniform distribution over models in the family. Then, for any distribution $p \in \Delta(\Pi)$, we have

$$\begin{aligned}
\underline{\mathsf{dec}}_\gamma(\mathcal{M}', \overline{M}) &\geq \mathbb{E}_{i\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big)\big] \\
&\geq \mathbb{E}_{i\sim\mu}\,\mathbb{E}_{\pi\sim p}[\alpha(1 - u_i(\pi)) - \gamma(\beta v_i(\pi) + \delta)] \\
&= \mathbb{E}_{\pi\sim p}\,\mathbb{E}_{i\sim\mu}[\alpha(1 - u_i(\pi)) - \gamma(\beta v_i(\pi) + \delta)].
\end{aligned}$$

Observe that for any fixed decision $\pi \in \Pi$, $\mathbb{E}_{i\sim\mu}[u_i(\pi)] = \frac{1}{N}\sum_{i=1}^N u_i(\pi) \leq 1/2$, and likewise $\mathbb{E}_{i\sim\mu}[v_i(\pi)] \leq 1/N$. Consequently,

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

□

Equipped with this lemma, we prove the following lower bound.

Proposition 5.3. *For the multi-armed bandit problem with $A \geq 2$ actions, $\mathcal{R} = [0, 1]$, and $\overline{M}(\pi) = \mathrm{Ber}(1/2)$,*

$$\mathsf{dec}_\gamma(\mathcal{M}^\infty_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \geq 2^{-6} \cdot \frac{A}{\gamma}$$

for all $\gamma \geq \frac{A}{3}$, where $\varepsilon_\gamma = \frac{A}{12\gamma}$.

Proof of Proposition 5.3. We construct a family of models $\mathcal{M}' = \{M_1, \ldots, M_A\}$ by defining $M_i(\pi) = \mathrm{Ber}(1/2 + \Delta\mathbb{I}\{\pi = i\})$ for $\Delta \in (0, 1/2)$, and setting $u_i(\pi) = v_i(\pi) = \mathbb{I}\{i = \pi\}$. We have

$$f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \Delta(1 - \mathbb{I}\{\pi = i\}), \quad \text{and} \quad D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \leq 3\Delta^2\mathbb{I}\{\pi = i\},$$

where the second inequality is Lemma A.7. Since $\sum_\pi u_i(\pi) = \sum_\pi v_i(\pi) = 1$, this establishes that $\mathcal{M}'$ is a $(\Delta, 3\Delta^2, 0)$-family, so by Lemma 5.1 we have that for any $\gamma > 0$,

$$\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{\Delta}{2} - \gamma\frac{3\Delta^2}{A}.$$

By choosing $\Delta = \frac{A}{12\gamma}$, this gives

$$\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq (48)^{-1}\frac{A}{\gamma}.$$

Furthermore, it is clear that $\mathcal{M}' \subseteq \mathcal{M}^\infty_\Delta(\overline{M})$, and that we may take $V(\mathcal{M}') = O(1)$ in Theorems 3.1 and 3.2 whenever $\Delta \leq 1/4$; it suffices to restrict to $\gamma \geq A/3$. □

Regret lower bound. Applying Theorem 3.2 with the lower bound on the DEC from Proposition 5.3, we are guaranteed that[14]

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq c' \cdot \frac{AT}{\gamma},$$

so long as $\frac{A}{12\gamma} \leq \varepsilon_\gamma = c''\frac{\gamma}{T}$, where $c, c', c'' > 0$ are numerical constants. If we choose $\gamma = C \cdot \sqrt{AT}$ for $C > 0$ sufficiently large, we have that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega(\sqrt{AT}),$$

which matches the minimax rate for the problem (Audibert and Bubeck, 2009).

Note that compared to specialized approaches tailored to the multi-armed bandit setting, the constants in this lower bound are rather loose. This is a consequence of the high generality of our framework.

[14] Recall that the sub-family of models constructed in Proposition 5.3 satisfies $V(\mathcal{M}') = O(1)$ for Theorem 3.2.

5.2 Tabular Reinforcement Learning

As our second example, we show how to bound the Decision-Estimation Coefficient and develop efficient algorithms for reinforcement learning in the tabular (or, finite state/action) setting. Here, $\mathcal{M}$ is the collection of all non-stationary MDPs with state space $\mathcal{S} = [S]$, action space $\mathcal{A} = [A]$, and horizon H, and $\Pi = \Pi_{\mathrm{RNS}}$ is the collection of all randomized, non-stationary Markov policies (cf. Example 1.2). We assume that rewards are normalized such that $\sum_{h=1}^{H} r_h \in [0,1]$ almost surely, and take $\mathcal{R} = [0,1]$ (Jiang and Agarwal, 2018; Zhang et al., 2021). Recall that for each $M \in \mathcal{M}$, $\{P_h^M\}_{h=1}^H$ and $\{R_h^M\}_{h=1}^H$ denote the associated transition kernels and reward distributions.

Our approach to bounding the DEC for this setting parallels the approach to the multi-armed bandit. First, we provide an upper bound on the dual DEC based on a new analysis of posterior sampling, then we provide a new, efficient strategy to bound the frequentist DEC which combines the idea of inverse gap weighting with the notion of a *policy cover*. The latter result shows for the first time how to adapt inverse gap weighting to reinforcement learning, and answers a question raised by Foster and Rakhlin (2020).

5.2.1 Bayesian Upper Bound via Posterior Sampling

The following result shows that posterior sampling leads to a $\mathrm{poly}(S, A, H)$ bound on the dual Bayesian Decision-Estimation Coefficient for tabular reinforcement learning.

Proposition 5.4. *Consider the tabular reinforcement learning setting with $\sum_{h=1}^{H} r_h \in \mathcal{R} := [0,1]$. For any $\mu \in \Delta(\mathcal{M})$, the distribution $p(\pi) = \mu(\{M : \pi_M = \pi\})$ satisfies*

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma\cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \le 26\frac{H^2SA}{\gamma},$$

for all $\overline{M}$ and $\gamma > 0$. Consequently, $\underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}) \le 26\frac{H^2SA}{\gamma}$.

Before proving the result, we state a useful *change of measure* lemma which shows that to bound the DEC for any episodic reinforcement learning setting, it suffices to consider state-action distributions induced by the reference model $\overline{M}$. This lemma is used for many subsequent reinforcement learning results.

Lemma 5.2 (Change of measure for reinforcement learning)**.** *Consider any family of MDPs $\mathcal{M}$ and reference MDP $\overline{M}$, all of which satisfy $\sum_{h=1}^{H} r_h \in [0,1]$. Suppose that $\mu \in \Delta(\mathcal{M})$ and $p \in \Delta(\Pi_{\mathrm{RNS}})$ are such that*

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] \tag{52}$$

$$\le C_1 + C_2\cdot\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D^2_{\mathrm{H}}\big(P^M(s_h,a_h), P^{\overline{M}}(s_h,a_h)\big) + D^2_{\mathrm{H}}\big(R^M(s_h,a_h), R^{\overline{M}}(s_h,a_h)\big)\right], \tag{53}$$

for parameters $C_1, C_2 > 0$. Then for all $\eta > 0$, it holds that

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \le C_1 + (4\eta)^{-1} + (40HC_2 + \eta)\cdot\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \tag{54}$$

See Appendix F.1 for the proof. Equipped with this result, we proceed to prove Proposition 5.4. As in the multi-armed bandit setting (Proposition 5.1), the thrust of the argument is to use Cauchy-Schwarz to decouple the random variables π_M and M under $M \sim \mu$. Compared to the multi-armed bandit setting though, we perform this decoupling by working in the space of *occupancy measures* for the reference MDP $\overline{M}$; this is facilitated by the change of measure lemma. A second point of comparison is that unlike the multi-armed bandit setting, where it was sufficient to work with the square loss Decision-Estimation Coefficient $\mathsf{dec}^{\mathrm{Sq}}_\gamma$, here it is critical to work with an information-theoretic divergence.

Proof of Proposition 5.4. In light of Lemma 5.2, it suffices to bound the quantity

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big]$$

in terms of the quantity on the right-hand side of (53). We start by writing

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] = \mathbb{E}_{M\sim\mu}[f^M(\pi_M)] - \mathbb{E}_{\pi\sim p}\big[f^{\overline{M}}(\pi)\big].$$

By definition, $\pi \sim p$ is identical in law to π_{M} under $M \sim \mu$, so this is equal to

$$\mathbb{E}_{M\sim\mu}[f^{M}(\pi_{M})] - \mathbb{E}_{M\sim\mu}\big[f^{\overline{M}}(\pi_{M})\big] = \mathbb{E}_{M\sim\mu}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M})\big].$$

Next, using a classical simulation lemma (Lemma F.3), we have

$$\begin{aligned}
\mathbb{E}_{M\sim\mu}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M})\big] &\leq \sum_{h=1}^{H} \mathbb{E}_{M\sim\mu}\,\mathbb{E}^{\overline{M},\pi_M}\Big[D_{\mathrm{TV}}\big(P_h^{M}(s_h,a_h), P_h^{\overline{M}}(s_h,a_h)\big) + D_{\mathrm{TV}}\big(R_h^{M}(s_h,a_h), R_h^{\overline{M}}(s_h,a_h)\big)\Big] \\
&= \sum_{h=1}^{H} \mathbb{E}_{M\sim\mu}\,\mathbb{E}^{\overline{M},\pi_M}[\mathrm{err}_h^M(s_h,a_h)],
\end{aligned}$$

where we define $\mathrm{err}_h^M(s,a) := D_{\mathrm{TV}}\big(P^{M}(s,a), \overline{P}(s,a)\big) + D_{\mathrm{TV}}\big(R^{M}(s,a), R^{\overline{M}}(s,a)\big)$. We proceed to bound each term in the sum. Recalling that $d_h^{M,\pi}(s,a)$ denotes the marginal distribution over (s_h,a_h) for policy π under model M, we have

$$\sum_{h=1}^{H} \mathbb{E}_{M\sim\mu}\,\mathbb{E}^{\overline{M},\pi_M}[\mathrm{err}_h^M(s_h,a_h)] = \sum_{h=1}^{H} \mathbb{E}_{M\sim\mu}\left[\sum_{s,a} d_h^{\overline{M},\pi_M}(s,a)\mathrm{err}_h^M(s,a)\right].$$

Define $\bar{d}_h(s,a) = \mathbb{E}_{M\sim\mu}\big[d^{\overline{M},\pi_M}(s,a)\big]$. Then by Cauchy-Schwarz, for each h, we have

$$\begin{aligned}
&\mathbb{E}_{M\sim\mu}\left[\sum_{s,a} d_h^{\overline{M},\pi_M}(s,a)\mathrm{err}_h^M(s,a)\right] \\
&= \mathbb{E}_{M\sim\mu}\left[\sum_{s,a} d_h^{\overline{M},\pi_M}(s,a)\left(\frac{\bar{d}_h(s,a)}{\bar{d}_h(s,a)}\right)^{1/2}\mathrm{err}_h^M(s,a)\right] \\
&\leq \left(\mathbb{E}_{M\sim\mu}\left[\sum_{s,a}\frac{d_h^{\overline{M},\pi_M}(s,a)}{\bar{d}_h(s,a)}\right]\right)^{1/2} \cdot \left(\mathbb{E}_{M\sim\mu}\left[\sum_{s,a}\bar{d}_h(s,a)(\mathrm{err}_h^M(s,a))^2\right]\right)^{1/2}.
\end{aligned}$$

For the first term above, we observe that

$$\mathbb{E}_{M\sim\mu}\left[\sum_{s,a}\frac{d_h^{\overline{M},\pi_M}(s,a)}{\bar{d}_h(s,a)}\right] = \sum_{s,a}\frac{\mathbb{E}_{M\sim\mu}\big[d_h^{\overline{M},\pi_M}(s,a)\big]}{\bar{d}_h(s,a)} = \sum_{s,a}\frac{\bar{d}_h(s,a)}{\bar{d}_h(s,a)} = SA.$$

For the second term, we again use that $\pi \sim p$ and π_{M} under $M \sim \mu$ are identical in law to write

$$\begin{aligned}
\mathbb{E}_{M\sim\mu}\left[\sum_{s,a}\bar{d}_h(s,a)(\mathrm{err}_h^M(s,a))^2\right] &= \sum_{s,a}\bar{d}_h(s,a)\,\mathbb{E}_{M\sim\mu}\big[(\mathrm{err}_h^M(s,a))^2\big] \\
&= \sum_{s,a}\mathbb{E}_{M\sim\mu}\Big[d_h^{\overline{M},\pi_M}(s,a)\Big]\,\mathbb{E}_{M\sim\mu}\big[(\mathrm{err}_h^M(s,a))^2\big] \\
&= \sum_{s,a}\mathbb{E}_{\pi\sim p}\Big[d_h^{\overline{M},\pi}(s,a)\Big]\,\mathbb{E}_{M\sim\mu}\big[(\mathrm{err}_h^M(s,a))^2\big] \\
&= \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\left[\sum_{s,a} d_h^{\overline{M},\pi}(s,a)(\mathrm{err}_h^M(s,a))^2\right] \\
&= \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\big[(\mathrm{err}_h^M(s_h,a_h))^2\big].
\end{aligned}$$

Combining these observations and applying the AM-GM inequality, we have that for any $\eta > 0$,

$$\sum_{h=1}^{H}\mathbb{E}_{M\sim\mu}\,\mathbb{E}^{\overline{M},\pi_M}[\mathrm{err}_h^M(s_h,a_h)] \leq \frac{HSA}{2\eta} + \frac{\eta}{2}\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H}(\mathrm{err}_h^M(s_h,a_h))^2\right].$$

Using that $(x+y)^2 \leq 2(x^2+y^2)$, we conclude that

$$
\begin{aligned}
&\mathbb{E}_{M\sim\mu}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M})\big] \\
&\leq \frac{HSA}{2\eta} + \eta\cdot\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D^2_{\mathrm{TV}}\big(P^{M}_h(s_h,a_h), P^{\overline{M}}_h(s_h,a_h)\big) + D^2_{\mathrm{TV}}\big(R^{M}_h(s_h,a_h), R^{\overline{M}}_h(s_h,a_h)\big)\right].
\end{aligned}
$$

With this result in hand, we apply Lemma 5.2, which implies that for any $\eta' > 0$,

$$
\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \leq \frac{HSA}{2\eta} + (4\eta')^{-1} + (40H\eta+\eta')\cdot\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big)\big].
$$

We conclude by setting $\eta = \eta' = \frac{\gamma}{41H}$.

□

The bound on the Decision-Estimation Coefficient from Proposition 5.4, which scales with $\frac{H^2SA}{\gamma}$, is optimal in terms of dependence on S and A, as we show below. The dependence on horizon can be improved from H^2 to H for MDPs with time-homogeneous dynamics.

In Section 7, we generalize the decoupling idea used in this proof to MDPs with low Bellman rank and, more generally, any *bilinear class*. Indeed, the only property of the tabular RL setting that is essential to the proof above is that the Bellman residuals have rank at most SA.

5.2.2 Frequentist Upper Bound via Policy Cover Inverse Gap Weighting

Algorithm 4 Policy Cover Inverse Gap Weighting (PC-IGW)

1: **parameters**:
Estimated model $\overline{M}$.
Exploration parameter $\eta > 0$.

2: Define *inverse gap weighted policy cover* $\Psi = \{\pi_{h,s,a}\}_{h\in[H],s\in[S],a\in[A]}$ via

$$
\pi_{h,s,a} = \operatorname*{arg\,max}_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d^{\overline{M},\pi}_h(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))}. \tag{55}
$$

3: For each policy $\pi \in \Psi \cup \{\pi_{\overline{M}}\}$, define

$$
p(\pi) = \frac{1}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))}, \tag{56}
$$

where $\lambda \in [1, 2HSA]$ is chosen such that $\sum_{\pi} p(\pi) = 1$.

4: **return** p.

We now provide an explicit, efficiently computable strategy which bounds the frequentist Decision-Estimation Coefficient for tabular reinforcement learning. The strategy, which we call *Policy Cover Inverse Gap Weighting*, is displayed in Algorithm 4. As the name suggests, our approach combines the inverse gap weighting technique introduced in the multi-armed bandit setting with the notion of a *policy cover*—that is, a collection of policies that ensures good coverage on every state (Du et al., 2019a; Misra et al., 2020; Jin et al., 2020b).

Algorithm 4 consists of two steps. First, in (55), we compute the collection of policies $\Psi = \{\pi_{h,s,a}\}_{h\in[H],s\in[S],a\in[A]}$ that constitutes the policy cover. In prior work, this is accomplished by computing the policy $\pi_{h,s,a} = \arg\max_{\pi\in\Pi_{\mathrm{RNS}}} d^{\overline{M},\pi}_h(s,a)$ that maximizes the occupancy measure on the estimated model $\overline{M}$ for each (h,s,a). The main twist here is that we instead consider the policy

$$
\pi_{h,s,a} = \operatorname*{arg\,max}_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d^{\overline{M},\pi}_h(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))}
$$

that maximizes the ratio of the occupancy measure and the regret gap under $\overline{M}$. This *inverse gap weighted policy cover* balances exploration and exploration by trading off coverage with suboptimality, and is critical to deriving a tight bound on the DEC that leads to $\sqrt{T}$-regret.

With the policy cover in hand, the second step of Algorithm 4 computes the exploratory distribution p by simply applying inverse gap weighting to the elements of the cover and the greedy policy $\pi_{\overline{M}}$.

The following result—proven in Appendix G.2—shows that the PC-IGW strategy can be implemented efficiently. Briefly, the idea is to solve (55) by taking a dual approach and optimizing over occupancy measures rather than policies. With this parameterization, (55) becomes a linear-fractional program, which can then be transformed into a standard linear program using classical techniques.

Proposition 5.5. *The PC-IGW algorithm (Algorithm 4) can be implemented in* $\mathrm{poly}(H, S, A, \log(\eta))$ *time via linear programming.*

Our bound on the Decision-Estimation Coefficient for the PC-IGW algorithm is as follows.

Proposition 5.6. *Consider the tabular reinforcement learning setting with* $\sum_{h=1}^{H} r_h \in \mathcal{R} := [0, 1]$*. For any* $\gamma > 0$ *and* $\overline{M} \in \mathcal{M}$*, the PC-IGW strategy in Algorithm 4, with* $\eta = \frac{\gamma}{21H^2}$*, ensures that*

$$\sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \leq 95\frac{H^3SA}{\gamma},$$

and consequently certifies that $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \leq 95\frac{H^3SA}{\gamma}$.

This result shows for the first time how to leverage the inverse gap weighting technique for provable exploration in reinforcement learning. In Section 7, we show to extend this approach to any family of MDPs with *bilinear class* structure.

The proof of Proposition 5.6 follows the structure of the proof of Proposition 5.4 closely; the main difference is that the decoupling step is replaced by an argument based on the inverse gap weighted policy cover in (55).

Proof of Proposition 5.6. We first verify that the strategy in (56) is indeed well-defined, in the sense that a normalizing constant $\lambda \in [1, 2HSA]$ always exists.

Proposition 5.7. *There is a unique choice for* $\lambda > 0$ *such that* $\sum_\pi p(\pi) = 1$*, and its value lies in* $[1, 2HSA]$.

Proof. Let $f(\lambda) = \sum_\pi \frac{1}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))}$. We first observe that if $\lambda > 2HSA$, then $f(\lambda) \leq \sum_\pi \frac{1}{\lambda} = \frac{HSA+1}{\lambda} < 1$. On the other hand for $\lambda \in (0, 1)$, $f(\lambda) \geq \frac{1}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{\overline{M}}))} = \frac{1}{\lambda} > 1$. Hence, since $f(\lambda)$ is continuous and strictly decreasing over $(0, \infty)$, there exists a unique $\lambda^\star \in [1, 2HSA]$ such that $f(\lambda^\star) = 1$. ☐

With this out of the way, we show that the PC-IGW strategy achieves the desired bound on the DEC. Let $M \in \mathcal{M}$ be fixed. As in Proposition 5.4, we bound the quantity

$$\mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big]$$

in terms of the quantity on the right-hand side of (53), then apply change of measure (Lemma 5.2). We begin with the decomposition

$$\mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] = \underbrace{\mathbb{E}_{\pi \sim p}\big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\big]}_{\text{(I)}} + \underbrace{f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}})}_{\text{(II)}}. \tag{57}$$

For the first term (I), which may be thought of as exploration bias, we have

$$\mathbb{E}_{\pi \sim p}\big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\big] = \sum_{\pi \in \Psi \cup \{\pi_{\overline{M}}\}} \frac{f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))} \leq \frac{2HSA}{\eta}, \tag{58}$$

where we have used that $\lambda \geq 0$. We next bound the second term (II), which entails showing that the PC-IGW distribution *explores enough*. We have

$$f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) = f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M}) - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M})). \tag{59}$$

Following Proposition 5.4, we use the simulation lemma to bound

$$\begin{aligned} f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{M}) &\leq \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi_{M}}\left[D_{\mathrm{TV}}\left(P_h^{M}(s_h,a_h), P_h^{\overline{M}}(s_h,a_h)\right) + D_{\mathrm{TV}}\left(R_h^{M}(s_h,a_h), R_h^{\overline{M}}(s_h,a_h)\right)\right] \\ &= \sum_{h=1}^{H}\sum_{s,a} d_h^{\overline{M},\pi_{M}}(s,a)\mathrm{err}_h^{M}(s,a), \end{aligned}$$

where $\mathrm{err}_h^{M}(s,a) := D_{\mathrm{TV}}\left(P^{M}(s,a), P^{\overline{M}}(s,a)\right) + D_{\mathrm{TV}}\left(R^{M}(s,a), R^{\overline{M}}(s,a)\right)$. Define $\bar{d}_h(s,a) = \mathbb{E}_{\pi\sim p}\left[d_h^{\overline{M},\pi}(s,a)\right]$. Then, using the AM-GM inequality, we have that for any $\eta' > 0$,

$$\begin{aligned} \sum_{h=1}^{H}\sum_{s,a} d_h^{\overline{M},\pi_{M}}(s,a)[\mathrm{err}_h^{M}(s,a)] &= \sum_{h=1}^{H}\sum_{s,a} d_h^{\overline{M},\pi_{M}}(s,a)\left(\frac{\bar{d}_h(s,a)}{\bar{d}_h(s,a)}\right)^{1/2}(\mathrm{err}_h^{M}(s,a))^2 \\ &\leq \frac{1}{2\eta'}\sum_{h=1}^{H}\sum_{s,a}\frac{(d_h^{\overline{M},\pi_{M}}(s,a))^2}{\bar{d}_h(s,a)} + \frac{\eta'}{2}\sum_{h=1}^{H}\sum_{s,a}\bar{d}_h(s,a)(\mathrm{err}_h^{M}(s,a))^2 \\ &= \frac{1}{2\eta'}\sum_{h=1}^{H}\sum_{s,a}\frac{(d_h^{\overline{M},\pi_{M}}(s,a))^2}{\bar{d}_h(s,a)} + \frac{\eta'}{2}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[(\mathrm{err}_h^{M}(s_h,a_h))^2\right]. \end{aligned}$$

The second term is exactly the upper bound we want, so it remains to bound the ratio of occupancy measures in the first term. Observe that for each (h,s,a), we have

$$\frac{d_h^{\overline{M},\pi_{M}}(s,a)}{\bar{d}_h(s,a)} \leq \frac{d_h^{\overline{M},\pi_{M}}(s,a)}{d_h^{\overline{M},\pi_{h,s,a}}(s,a)}\cdot\frac{1}{p(\pi_{h,s,a})} \leq \frac{d_h^{\overline{M},\pi_{M}}(s,a)}{d_h^{\overline{M},\pi_{h,s,a}}(s,a)}\left(2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{h,s,a})\right),$$

where the second inequality follows from the definition of p and the fact that $\lambda \leq 2HSA$. Furthermore, since

$$\pi_{h,s,a} = \operatorname*{arg\,max}_{\pi\in\Pi_{\mathrm{RNS}}}\frac{d_h^{\overline{M},\pi}(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))},$$

and since $\pi_{M} \in \Pi_{\mathrm{RNS}}$, we can upper bound by

$$\frac{d_h^{\overline{M},\pi_{M}}(s,a)}{d_h^{\overline{M},\pi_{M}}(s,a)}\left(2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M})\right) = 2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M}). \tag{60}$$

As a result, we have

$$\sum_{h=1}^{H}\sum_{s,a}\frac{(d_h^{\overline{M},\pi_{M}}(s,a))^2}{\bar{d}_h(s,a)} \leq \sum_{h=1}^{H}\sum_{s,a} d_h^{\overline{M},\pi_{M}}(s,a)(2HSA+\eta(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi_{M})) = 2H^2SA+\eta H(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi_{M})).$$

Putting everything together and returning to (59), this establishes that

$$f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) \leq \frac{H^2SA}{\eta'} + \frac{\eta'}{2}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[(\mathrm{err}_h^{M}(s_h,a_h))^2\right] + \frac{\eta H}{2\eta'}(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M})) - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M})).$$

We set $\eta' = \frac{\eta H}{2}$ so that the latter terms cancel and we are left with

$$f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) \leq \frac{2HSA}{\eta} + \frac{\eta H}{4}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[(\mathrm{err}_h^{M}(s_h,a_h))^2\right].$$

Combining this with (57) and (58) gives

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi)\right] \\
&\leq \frac{4HSA}{\eta} + \frac{\eta H}{4}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[(\mathrm{err}_h^{M}(s_h,a_h))^2\right] \\
&\leq \frac{4HSA}{\eta} + \frac{\eta H}{2}\sum_{h=1}^{H}\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[D^2_{\mathrm{TV}}\left(P^{M}(s_h,a_h), P^{\overline{M}}(s_h,a_h)\right) + D^2_{\mathrm{TV}}\left(R^{M}(s_h,a_h), R^{\overline{M}}(s_h,a_h)\right)\right].
\end{aligned}$$

We conclude by applying the change-of-measure lemma (Lemma 5.2), which implies that for any $\eta' > 0$,

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \leq \frac{4HSA}{\eta} + (4\eta')^{-1} + (20H^2\eta + \eta')\cdot\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\left[D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)\right].$$

The result follows by choosing $\eta = \eta' = \frac{\gamma}{21H^2}$. □

5.2.3 Regret Upper Bound

Using an approach in Section 7.1.3, we can obtain an efficient online estimation algorithm for tabular MDPs which guarantees that

$$\mathbf{Est}_{\mathsf{H}} = \sum_{t=1}^{T}\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[D^2_{\mathrm{H}}\left(M^{\star}(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\right)\right] \leq \widetilde{O}(HS^2A).$$

Combining this with the PC-IGW strategy (Proposition 5.6) and E2D (Theorem 4.1), we obtain an efficient algorithm with

$$\mathbf{Reg}_{\mathsf{DM}} \leq \widetilde{O}(\sqrt{H^4S^3A^2T}).$$

Regret bounds for tabular reinforcement learning have received extensive investigation, but this result is exciting because it represents a completely new algorithmic approach. In particular, this is the first frequentist reinforcement learning algorithm we are aware of that does not make use of confidence sets or optimism.

Note that while this bound scales as $\sqrt{\mathrm{poly}(S,A,H)T}$ as desired, it does fall short of the minimax rate, which is $\sqrt{HSAT}$ for our setting (since we consider $\sum_{h=1}^{H} r_h \in [0,1]$ and time-inhomogeneous dynamics). We emphasize that obtaining the tightest possible dependence on problem parameters is not the focus of this work, but it would be interesting to improve this approach to match the minimax rate.

5.2.4 Lower Bound

We close the tabular reinforcement learning example by complementing our upper bounds with a lower bound on the Decision-Estimation Coefficient.

Proposition 5.8. *Let $\mathcal{M}$ be the class of tabular MDPs with $S \geq 2$ states, $A \geq 2$ actions, and $\sum_{h=1}^{H} r_h \in \mathcal{R} := [0,1]$. If $H \geq 2\log_2(S/2)$, then there exists $\overline{M} \in \mathcal{M}$ such that for all $\gamma \geq HSA/24$,*

$$\mathsf{dec}_{\gamma}(\mathcal{M}^{\infty}_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \geq 2^{-9}\cdot\frac{HSA}{\gamma},$$

where $\varepsilon_\gamma = \frac{HSA}{96\gamma}$.

As with the multi-armed bandit example, the proof of this result proceeds by constructing a hard family of models and appealing to Lemma 5.1. We follow a familiar tree MDP construction (e.g., Osband and Van Roy (2016); Domingues et al. (2021)).

Proof of Proposition 5.8. Assume without loss of generality that S is a multiple of 2. Let $\Delta \in (0, 1/2)$ be a parameter. We consider the following class of MDPs.

- Define $H_1 := \log_2(S/2)$ and $H_2 = H - H_1$.
- Let $S' := S/2$. The state space $\mathcal{S}$ is chosen to consist of a depth-H_1 binary tree (which has S' leaves and $\sum_{i=0}^{\log_2(S/2)} 2^i = S - 1$ total states), along with a single terminal state term. We let $\mathcal{S}'$ denote the collection of leaf states.
- All MDPs M in the family have the same (deterministic) dynamics $P^M = P$. The agent begins at the root state in the tree, and for each $h < H_1$, there are two available actions, left and right, which determine whether the next state is the left or right successor node in the tree. For $h \geq H_1$, there are two cases:
  - For $s \in \mathcal{S}'$, the agent can choose to "wait" using action wait, or choose an action from $\mathcal{A}' := [A']$, where $A' := A - 1$. The wait action causes the agent to stay in s (i.e. $P(s \mid s, \mathsf{wait}) = 1) \; \forall s \in \mathcal{S}'$), while actions in $[A']$ cause the agent to immediately transit to term (i.e., $P(\mathsf{term} \mid s, a) = 1 \; \forall s \in \mathcal{S}', a \in [A']$).
  - The terminal state term is self-looping and (i.e. $P(\mathsf{term} \mid \mathsf{term}, \cdot) = 1$).
- Let $\mathcal{H}' = \{H_1, \ldots, H\}$. For each $h^\star \in \mathcal{H}', s^\star \in \mathcal{S}', a^\star \in \mathcal{A}'$, we define an MDP $M_{h^\star, s^\star, a^\star}$ which has the dynamics described above and the following reward functions:
  - $R_h^M(s, a) = 0$ a.s. for all $h < H_1$.
  - $R_h^M(\mathsf{term}, \cdot) = 0$ a.s. for all $h > H_1$
  - $R_h^M(s, \mathsf{wait}) = 0$ a.s. for all $s \in \mathcal{S}'$ for all $h \geq H_1$
  - $R_h^M(s, a) = \mathrm{Ber}\big(\frac{1}{2} + \Delta \cdot \mathbb{I}\{h = h^\star, s = s^\star, a = a^\star\}\big)$ for $s \in \mathcal{S}'$, $a \in [A']$, $h \geq H_1$.

We choose $\mathcal{M}' = \{M_{s,a}\}_{h \in \mathcal{H}', s \in \mathcal{S}', a \in \mathcal{A}'}$. Finally, we take the reference MDP $\overline{M} = (P, R^{\overline{M}})$ to have the same dynamics and rewards as above, except that $R_h^{\overline{M}}(s, a) = \mathrm{Ber}(1/2)$ for all $h \in \mathcal{H}', s \in \mathcal{S}', a \in \mathcal{A}'$.

We claim that $\mathcal{M}'$ is a hard family of MDPs in the sense of Lemma 5.1. To do so, we define

$$u_{h,s,a}(\pi) = v_{h,s,a}(\pi) = \mathbb{P}^{\overline{M},\pi}(s_h = s, a_h = a).$$

We note that for any π,

$$\sum_{s \in \mathcal{S}', a \in \mathcal{A}', h \in \mathcal{H}'} u_{h,s,a}(\pi) = \sum_{s \in \mathcal{S}', a \in \mathcal{A}'} \sum_{h=H_1}^{H} \mathbb{P}^{\overline{M},\pi}(s_h = s, a_h = a) \leq \sum_{s \in \mathcal{S}', a \in \mathcal{A}'} \mathbb{P}^{\overline{M},\pi}(s_{H_1} = s, a_{H_1} = a) \leq 1,$$

where the inequality uses that (i) all $a \in \mathcal{A}'$ transit to the terminal state term for $h \geq H_1$, and (ii) each state $s \in \mathcal{S}'$ is only reachable for $h \geq H_1$ if $s_{H_1} = s$. This verifies that $u_{h,s,a}$ and $v_{h,s,a}$ satisfy the preconditions of Lemma 1.

To proceed, we first observe that the optimal policy for $M_{h,s,a}$ takes the single sequence of actions in the tree that leads to state s, uses action wait until step h, then selects action a at step h for expected reward $\frac{1}{2} + \Delta$. As a result,

$$f^{M_{h,s,a}}(\pi_{h,s,a}) - f^{M_{h,s,a}}(\pi) = \Delta \cdot \mathbb{P}^{M_{h,s,a},\pi}((s_h, a_h) \neq (s, a)) = \Delta(1 - u_{h,s,a}(\pi))$$

where we recall that all MDPs share the same dynamics. Furthermore, since the rewards for $\overline{M}$ and $M_{h,s,a}$ are identical unless $(s_h, a_h) = (s, a)$, we have[15]

$$\begin{aligned} D^2_{\mathrm{H}}\big(M_{h,s,a}(\pi), \overline{M}(\pi)\big) &= \mathbb{P}^{\overline{M},\pi}((s_h, a_h) = (s, a)) \cdot D^2_{\mathrm{H}}(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2 + \Delta)) \\ &\leq \mathbb{P}^{\overline{M},\pi}((s_h, a_h) = (s, a)) \cdot 3\Delta^2 \\ &= v_{h,s,a}(\pi) \cdot 3\Delta^2, \end{aligned}$$

[15]Since squared Hellinger distance is an f-divergence, it satisfies $D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X} \otimes \mathbb{P}_X, \mathbb{Q}_{Y|X} \otimes \mathbb{P}_X\big) = \mathbb{E}_{X \sim \mathbb{P}_X}\big[D^2_{\mathrm{H}}(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X})\big]$.

by Lemma A.7. It follows that $\mathcal{M}'$ is a $(\Delta, 3\Delta^2, 0)$-family, so Lemma 5.1 implies that

$$\underline{\mathsf{dec}}_{\gamma}(\mathcal{M}', \overline{M}) \geq \frac{\Delta}{2} - \gamma\frac{3\Delta^2}{|\mathcal{H}'||\mathcal{S}'||\mathcal{A}'|} \geq \frac{\Delta}{2} - \gamma\frac{24\Delta^2}{HSA}.$$

We set $\Delta = \frac{HSA}{96\gamma}$, which leads to value

$$(384)^{-1}\frac{HSA}{\gamma}.$$

We conclude by noting that $\mathcal{M}' \subseteq \mathcal{M}^{\infty}_{\Delta}(\overline{M})$, and that we may take $V(\mathcal{M}') = O(1)$ in Theorems 3.1 and 3.2 whenever $\Delta \leq 1/4$; it suffices to take $\gamma \geq HSA/24$. $\square$

Regret lower bound. We apply Theorem 3.2 with the lower bound on the DEC from Proposition 5.8, which gives that for any $\gamma \geq c\sqrt{T}$, $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq c' \cdot \frac{HSAT}{\gamma}$, so long as $\frac{HSA}{96\gamma} \leq \underline{\varepsilon}_{\gamma} = c''\frac{\gamma}{T}$, where $c, c', c'' > 0$ are numerical constants.[16] If we choose $\gamma = C \cdot \sqrt{AT}$ for sufficiently large $C > 0$, we have that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega(\sqrt{HSAT}).$$

5.3 Discussion

We showed how to bound the Decision-Estimation Coefficient for multi-armed bandits and tabular reinforcement learning via Bayesian approaches (posterior sampling) and frequentist approaches (inverse gap weighting). The analyses for both techniques parallel each other, and leverage decoupling and change of measure arguments. We build on these ideas in Section 6 and Section 7 to derive algorithms and bounds for more complex settings.

While we have considered posterior sampling (the most ubiquitous Bayesian approach) and inverse gap weighting (a somewhat more recent frequentist approach) and highlighted parallels, upper confidence bound-based approaches have been conspicuously absent up to this point, and do not appear to be sufficient to bound the Decision-Estimation Coefficient. Informally, this is because the DEC considers estimation error under the *learner's own distribution* (i.e., future estimation error after the learner commits to the exploration strategy), while UCB and other confidence-based approaches explore based on estimation error on historical data. Further research is required to better understand whether this distinction is fundamental.

6 Application to Bandits

As a special case of our main results, we obtain algorithms and lower bounds for structured bandits with large action spaces (Example 1.1). In this section, we highlight some well-known instances of the structured bandit problem recovered by our results (Section 6.1), then provide a new guarantee based on a combinatorial parameter called the *star number* (Hanneke and Yang, 2015; Foster et al., 2020b) (Section 6.2). The latter result improves upon previous regret bounds based on the eluder dimension.

6.1 Familiar Examples

We consider three canonical structured bandit settings, linear bandits, convex bandits, and non-parametric bandits, and provide tight upper and lower bounds on the Decision-Estimation Coefficient. We then provide additional lower bounds for bandits with ReLU rewards and for various bandit problems with gaps.

Throughout the section we assume $\mathcal{R} \subseteq [0, 1]$ unless otherwise specified. For a given function class $\mathcal{F} \subseteq (\Pi \to \mathcal{R})$, we define $\mathcal{M}_{\mathcal{F}} = \{M : f^M \in \mathcal{F}\}$ as the induced class of models. We tacitly make use of the fact that all of the lower bound constructions $\mathcal{M}' \subseteq \mathcal{M}$ in this section satisfy $V(\mathcal{M}') = O(1)$ for Theorem 3.1.

[16] As in the multi-armed bandit example, we have $V(\mathcal{M}') = O(1)$ in Theorem 3.2 for the MDP family $\mathcal{M}'$.

6.1.1 Linear Bandits

In the linear bandit setting (Abe and Long, 1999; Auer et al., 2002a; Dani et al., 2008; Chu et al., 2011; Abbasi-Yadkori et al., 2011), we set $\Pi \subseteq \mathbb{R}^d$, define $\mathcal{F} = \{\pi \mapsto \langle \theta, \pi \rangle \mid \theta \in \Theta\}$ for a parameter set $\Theta \subseteq \mathbb{R}^d$, then take $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$ as the induced model class. The following recent result from Foster et al. (2020a) gives an efficient algorithm that leads to upper bounds on the Decision-Estimation Coefficient for this setting.

Proposition 6.1 (Upper bound for linear bandits (Foster et al., 2020a))**.** *Consider the linear bandit setting with $\mathcal{R} = \mathbb{R}$. Let $\overline{M} \in \mathcal{M}$ and $\gamma > 0$ be given, and define*

$$p = \underset{p \in \Delta(\Pi)}{\arg\max} \left\{ \mathbb{E}_{\pi \sim p}\left[f^{\overline{M}}(\pi)\right] + \frac{1}{4\gamma} \log\det(\mathbb{E}_{\pi \sim p}\left[\pi\pi^\top\right]) \right\}.$$

This strategy certifies that

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) \le \frac{d}{4\gamma}.$$

Combining this strategy with Theorem 3.6, we obtain $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \le O(\sqrt{dT \log|\Pi|})$ when $|\Pi| < \infty$, and for infinite action spaces we obtain $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \le \widetilde{O}(d\sqrt{T})$ whenever Θ and Π have bounded diameter; both results are optimal. More generally, using this strategy within the E2D algorithm (Theorem 4.1) yields $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \le O(\sqrt{dT \cdot \mathbf{Est}_{\mathsf{H}}})$.

We now turn our attention to lower bounds. Note that if we take Θ to be the ℓ_∞-ball and Π to be the ℓ_1-ball, a trivial lower bound on the DEC is $\frac{d}{\gamma}$, since this embeds the finite-armed bandit setting. The following result shows that the same lower bound holds under euclidean geometry.

Proposition 6.2 (Lower bound for linear bandits)**.** *Consider the linear bandit setting with $\mathcal{R} = [-1, +1]$, and let $\Pi = \Theta = \left\{v \in \mathbb{R}^d \mid \|v\|_2 \le 1\right\}$. Then for all $d \ge 4$ and $\gamma \ge \frac{2d}{3}$, there exists $\overline{M} \in \mathcal{M}$ such that*

$$\mathsf{dec}_{\gamma}(\mathcal{M}^{\infty}_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \ge \frac{d}{12\gamma},$$

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

Combining this result with Theorem 3.2 leads to a lower bound of the form $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \ge \Omega(\sqrt{dT})$.

6.1.2 Convex Bandits

The convex bandit problem (Kleinberg, 2004; Flaxman et al., 2005; Agarwal et al., 2013; Bubeck and Eldan, 2016; Bubeck et al., 2017; Lattimore, 2020) is a generalization of the linear bandit. We take $\Pi \subseteq \mathbb{R}^d$, define

$$\mathcal{F} = \{f : \Pi \to [0, 1] \mid f \text{ is concave and 1-Lipschitz w.r.t } \ell_2\},$$

and take $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$ as the induced model class.[17] The following recent result of Lattimore (2020) provides a bound on the Decision-Estimation Coefficient for this setting.[18]

Proposition 6.3 (Upper bound for convex bandits (Lattimore (2020), Theorem 3))**.** *For the convex bandit setting with $\mathcal{R} = [0, 1]$, we have*

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) \le O\left(\frac{d^4}{\gamma} \cdot \mathrm{polylog}(d, \mathrm{diam}(\Pi), \gamma)\right)$$

for all $\overline{M} \in \mathcal{M}$ and $\gamma > 0$.

[17] We consider concave rather than concave functions because we work with rewards instead of losses.

[18] The statement presented here requires very slight modifications to the construction in Lattimore (2020). Namely, certain parameters that scale with T in the original construction must be replaced by γ.

Since this setting has $\mathsf{est}(\Pi_{\mathcal{M}}, T) \leq \widetilde{O}(d)$ whenever $\operatorname{diam}(\Pi) = O(1)$, combining the bound above with Theorem 3.6 leads to regret $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}(d^{2.5}\sqrt{T})$.

We mention in passing that previous results in the line of work on Bayesian regret for convex bandits (Bubeck et al., 2015; Bubeck and Eldan, 2016) can also be interpreted as bounds on the Decision-Estimation Coefficient. While the optimal dependence on d for this setting is not yet understood, a lower bound of $\sqrt{dT}$ follows from the result for the linear setting.

6.1.3 Nonparametric Bandits

For the next example, we consider a standard nonparametric bandit problem: Lipschitz bandits in metric spaces (Auer et al., 2007; Kleinberg et al., 2019). We take Π to be a metric space equipped with metric ρ, then take $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$, where we define

$$\mathcal{F} = \{f : \Pi \to [0,1] \mid f \text{ is 1-Lipschitz w.r.t } \rho\}.$$

Our results are stated in terms of covering numbers with respect to the metric ρ. Let us say that $\Pi' \subseteq \Pi$ is an ε-cover with respect to ρ if

$$\forall \pi \in \Pi \quad \exists \pi' \in \Pi' \quad \text{s.t.} \quad \rho(\pi, \pi') \leq \varepsilon,$$

and let $\mathcal{N}_\rho(\Pi, \varepsilon)$ denote the size of the smallest such cover.

Proposition 6.4 (Upper bound for Lipschitz bandits)**.** *Consider the Lipschitz bandit setting with $\mathcal{R} = \mathbb{R}$, and suppose that $\mathcal{N}_\rho(\Pi, \varepsilon) \leq \varepsilon^{-d}$ for $d > 0$. Let $\overline{M}$ and $\gamma \geq 1$ be given and consider the following algorithm:*

1. *Let $\Pi' \subseteq \Pi$ witness the covering number $\mathcal{N}_\rho(\Pi, \varepsilon)$ for a parameter $\varepsilon > 0$.*
2. *Perform the inverse gap weighting strategy in (50) over Π'.*

By setting $\varepsilon = \gamma^{-\frac{1}{d+1}}$, this strategy certifies that

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) \leq 2\gamma^{-\frac{1}{d+1}}.$$

Since $\mathsf{est}(\Pi_{\mathcal{M}}, T) \leq \widetilde{O}(d)$ for this setting, if we apply Theorem 3.6 with this result we obtain

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}(T^{\frac{d+1}{d+2}})$$

which matches the minimax rate derived in Kleinberg et al. (2019). We complement this with a lower bound.

Proposition 6.5 (Lower bound for Lipschitz bandits)**.** *Consider the Lipschitz bandit setting with $\mathcal{R} = [0,1]$. Suppose that $\mathcal{N}_\rho(\Pi, \varepsilon) \geq \varepsilon^{-d}$ for $d \geq 1$. Then for all $\gamma \geq 1$, there exists $\overline{M} \in \mathcal{M}$ such that*

$$\mathsf{dec}_{\gamma}(\mathcal{M}^{\infty}_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \geq 2^{-7}\gamma^{-\frac{1}{d+1}},$$

where $\varepsilon_\gamma = 6^{-2}\gamma^{-\frac{1}{d+1}}$.

Plugging this guarantee into Theorem 3.2, we obtain a lower bound of the form $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \widetilde{\Omega}(T^{\frac{d+1}{d+2}})$, which again recovers the minimax rate. Extending these upper and lower bounds on the DEC to accommodate other (e.g., Hölder) nonparametric bandit problems is straightforward.

6.1.4 ReLU Bandits

We now consider a bandit setting based on the well-known ReLU activation function $\mathsf{relu}(x) = \max\{x, 0\}$. Here, we take $\Pi = \left\{\pi \in \mathbb{R}^d \mid \|\pi\|_2 \leq 1\right\}$ and $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$, where $\mathcal{F}$ is the class of value functions of the form

$$f(\pi) = \mathsf{relu}(\langle \theta, \pi \rangle - b), \tag{61}$$

where $\theta \in \Theta := \left\{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq 1\right\}$ is an unknown parameter vector and $b \geq 0$ is a known bias parameter.

Note that since we work with rewards, this setting would be a special case of bandit convex optimization if we were to replace the $+\mathsf{relu}(\cdot)$ function in (61) with $-\mathsf{relu}(\cdot)$, and in this case it would be possible to appeal to Proposition 6.3 to derive a $\sqrt{\mathrm{poly}(d)T}$ bound on regret. However, the $+\mathsf{relu}(\cdot)$ formulation in (61) cannot be viewed as an instance of bandit convex optimization, and the following proposition shows that this setting is intractable.

Proposition 6.6 (Lower bound for ReLU bandits)**.** *Consider the ReLU bandit setting with $\mathcal{R} = [-1, +1]$. For all $d \geq 16$, there exists $\overline{M} \in \mathcal{M}$ such that for all $\gamma > 0$,*

$$\mathsf{dec}_\gamma(\mathcal{M}^\infty_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \geq \frac{e^{d/8}}{24\gamma} \wedge \frac{1}{8}, \tag{62}$$

where $\varepsilon_\gamma = \frac{e^{d/8}}{6\gamma} \wedge \frac{1}{2}$.

By plugging this result into Theorem 3.1 and setting $\gamma = T$, we conclude that any algorithm must have

$$\mathbf{Reg}_{\mathsf{DM}} \geq \Omega(\min\{e^{\Omega(d)}, T\})$$

with constant probability, which further implies that $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega(\min\{e^{\Omega(d)}, T\})$. This recovers recent impossibility results (Dong et al., 2021; Li et al., 2021).

6.1.5 Gap-Dependent Lower Bounds

In multi-armed bandits, *gap-dependent* regret bounds that adapt to the gap

$$\Delta_{\scriptscriptstyle M} := \min_{\pi \neq \pi_M} \{f^M(\pi_{\scriptscriptstyle M}) - f^M(\pi)\}$$

between the best and second-best action have been the subject of extensive investigation (Lai and Robbins, 1985a; Burnetas and Katehakis, 1996; Garivier et al., 2016, 2019; Kaufmann et al., 2016; Lattimore, 2018; Garivier and Kaufmann, 2016). Gap-dependent guarantees in reinforcement learning with function approximation have also received recent interest as a means to bypass certain intractability results (Du et al., 2019b; Wang et al., 2021). Here we prove lower bounds on the Decision-Estimation Coefficient for finite-armed bandits and linear bandits when the class $\mathcal{M}$ is constrained to have gap Δ. These examples highlight that the DEC leads to meaningful lower bounds even for "easy" problems with low statistical complexity

Proposition 6.7 (Multi-armed bandits with gaps)**.** *Let $\mathcal{M}$ be the class of all multi-armed bandit problems over $\Pi = [A]$ with $\mathcal{R} = [0, 1]$ and gap $\Delta > 0$. For all $\Delta \in (0, 1/8)$, there exists $\overline{M} \in \mathcal{M}$, such that*

$$\mathsf{dec}_\gamma(\mathcal{M}^\infty_\Delta(\overline{M}), \overline{M}) \geq \frac{\Delta}{4}\mathbb{I}\left\{\gamma \leq \frac{A}{48\Delta}\right\}$$

for all $\gamma > 0$.

Applying this result within Theorem 3.2, we are guaranteed that $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq c \cdot \Delta T \mathbb{I}\{\gamma \leq (48)^{-1}\frac{A}{\Delta}\}$, as long as $\Delta \leq \varepsilon_\gamma = c'\frac{\gamma}{T}$, where $c, c' > 0$ are numerical constants. In particular, if we set $\gamma = c_1\frac{A}{\Delta}$ for $c_1 > 0$ sufficiently small, we have $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega(\Delta T)$ for all $T \leq c_2\frac{A}{\Delta^2}$, where $c_2 > 0$ is a numerical constant. Since minimax regret is non-decreasing with T, this implies a lower bound of the form

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega\left(\min\left\{\Delta T, \frac{A}{\Delta}\right\}\right). \tag{63}$$

Up to a $\log(T)$ factor, this matches the usual $\frac{A}{\Delta}$ scaling found in standard gap-dependent lower bounds (Lai and Robbins, 1985a; Burnetas and Katehakis, 1996; Garivier et al., 2016, 2019; Kaufmann et al., 2016; Lattimore, 2018; Garivier and Kaufmann, 2016) when T is sufficiently large. We caution however that these results are *instance-dependent* in nature, and provide gap-dependent lower bounds on the regret for any particular problem instance, whereas (63) is a minimax lower bound over the class of all possible models with gap Δ. This is a consequence of the fact that the Decision-Estimation Coefficient and our associated lower bounds capture minimax complexity for decision making, which is fundamentally different from instance-dependent complexity; see Section 9 for more discussion.

We proceed with an analogous lower bound for the linear setting in Section 6.1.1.

Proposition 6.8 (Linear bandits with gaps)**.** *For every $\Delta \in (0, 1/4)$, there exists a collection $\mathcal{M}$ of linear bandit models with $\mathcal{R} = [-1, +1]$, $\Pi \subseteq \Theta = \{v \in \mathbb{R}^d \mid \|v\|_2 \leq 1\}$, and gap $\Delta > 0$, such that for some $\overline{M} \in \mathcal{M}$,*

$$\mathsf{dec}_\gamma(\mathcal{M}^\infty_\Delta(\overline{M}), \overline{M}) \geq \frac{\Delta}{4}\mathbb{I}\left\{\gamma \leq \frac{d}{12\Delta}\right\}$$

for all $\gamma > 0$.

Using similar calculations to the multi-armed bandit example, we deduce that any algorithm must have

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega\left(\min\left\{\Delta T, \frac{d}{\Delta}\right\}\right). \tag{64}$$

6.2 Disagreement Coefficient, Star Number, and Eluder Dimension

In this section we show that the Decision-Estimation Coefficient can recover regret bounds for bandits based on a well-known combinatorial parameter called the *eluder dimension* (Russo and Van Roy, 2013; Wang et al., 2020; Ayoub et al., 2020; Jin et al., 2021), and then give a new bound based on a closely related but tighter parameter called the *star number* (Hanneke and Yang, 2015; Foster et al., 2020b).

We begin by defining the eluder dimension for a value function class $\mathcal{F}$.

Definition 6.1. *Let $\mathcal{F} : \Pi \to \mathbb{R}$ be given, and define $\check{\mathfrak{e}}(\mathcal{F}, \Delta)$ as the length of the longest sequence of decisions $\pi_1, \ldots, \pi_s \in \Pi$ such that for all i, there exists $f_i \in \mathcal{F}$ such that*

$$|f_i(\pi_i)| > \Delta, \quad \text{and} \quad \sum_{j<i} f_i^2(\pi_j) \leq \Delta^2. \tag{65}$$

The eluder dimension is defined as $\mathfrak{e}(\mathcal{F}, \Delta) = \sup_{\Delta' \geq \Delta} \check{\mathfrak{e}}(\mathcal{F}, \Delta') \vee 1$.

The star number was originally introduced in the context of active learning with binary classifiers by Hanneke and Yang (2015). We work with a scale-sensitive variant introduced by Foster et al. (2020b), which can be thought of as a "non-sequential" analogue of the eluder dimension.

Definition 6.2. *Let $\mathcal{F} : \Pi \to \mathbb{R}$ be given, and define $\check{\mathfrak{s}}(\mathcal{F}, \Delta)$ as the length of the longest sequence of decisions $\pi_1, \ldots, \pi_s \in \Pi$ such that for all i, there exists $f_i \in \mathcal{F}$ such that*

$$|f_i(\pi_i)| > \Delta, \quad \text{and} \quad \sum_{j\neq i} f_i^2(\pi_j) \leq \Delta^2. \tag{66}$$

The star number is defined as $\mathfrak{s}(\mathcal{F}, \Delta) = \sup_{\Delta' \geq \Delta} \check{\mathfrak{s}}(\mathcal{F}, \Delta') \vee 1$.

It is clear from this definition that $\mathfrak{s}(\mathcal{F}, \Delta) \leq \mathfrak{e}(\mathcal{F}, \Delta)$. In general, the star number can be arbitrarily small compared to the eluder dimension (Foster et al., 2020b).

The following result shows that boundedness of star number and eluder dimension always implies boundedness of the Decision-Estimation Coefficient.

Theorem 6.1. *Consider any class $\mathcal{M}$ with $\mathcal{R} = [0, 1]$. Suppose the conclusion of Proposition 4.2 holds. Then for all $\gamma \geq e$, we have*

$$\mathsf{dec}^{\mathrm{Sq}}_\gamma(\mathcal{M}, \overline{M}) \leq 2^6 \inf_{\Delta>0}\left\{\Delta + \frac{\mathfrak{s}^2(\mathcal{F}_\mathcal{M} - f^{\overline{M}}, \Delta)\log^2(\gamma)}{\gamma}\right\} + (2\gamma)^{-1} \tag{67}$$

and

$$\mathsf{dec}^{\mathrm{Sq}}_\gamma(\mathcal{M}, \overline{M}) \leq 2^6 \inf_{\Delta>0}\left\{\Delta + \frac{\mathfrak{e}(\mathcal{F}_\mathcal{M} - f^{\overline{M}}, \Delta)\log^2(\gamma)}{\gamma}\right\} + (2\gamma)^{-1}. \tag{68}$$

An upper bound in terms of the star number immediately implies an upper bound in terms of the eluder dimension, but we present separate bounds for each parameter because Eq. (67) has quadratic dependence on the star number, while Eq. (68) has linear dependence on the eluder dimension. Abbreviating $\mathfrak{e} \equiv \sup_{f\in\mathcal{F}_{\mathcal{M}}} \mathfrak{e}(\mathcal{F}_{\mathcal{M}} - f, 1/T)$ and $\mathfrak{s} \equiv \sup_{f\in\mathcal{F}_{\mathcal{M}}} \mathfrak{s}(\mathcal{F}_{\mathcal{M}} - f, 1/T)$, this result recovers the $\widetilde{O}(\sqrt{\mathfrak{e}T})$ regret bound derived in Russo and Van Roy (2013) as a special case by appealing to Theorem 3.6, but also implies a regret bound of the form $\widetilde{O}(\sqrt{\mathfrak{s}^2 T})$, which can be arbitrarily tighter.

Compared to the earlier examples in this section, one should not expect to derive a matching lower bound on the Decision-Estimation Coefficient. Unlike the DEC itself, the star number and eluder dimension only provide sufficient conditions for sample-efficient learning, and neither parameter plays a fundamental role in determining the minimax regret. For example, both parameters are exponential in the dimension for bandit convex optimization (Li et al., 2021), while Proposition 6.3 shows that the DEC is polynomial.

Theorem 6.1 is a corollary of a more general result that provides a *prior-dependent* upper bound on the dual Bayesian Decision-Estimation Coefficient when posterior sampling is applied. For a given prior $\mu \in \Delta(\mathcal{M})$ and distance $D(\cdot \,\|\, \cdot)$, define

$$\underline{\mathsf{dec}}^{D}_{\gamma}(\mu, \overline{M}) = \inf_{p\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\Big[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot D\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\Big], \tag{69}$$

so that $\underline{\mathsf{dec}}^{D}_{\gamma}(\mathcal{M}, \overline{M}) = \sup_{\mu\in\Delta(\mathcal{M})} \underline{\mathsf{dec}}^{D}_{\gamma}(\mu, \overline{M})$. Our result is stated in terms of a parameter called the (scale-sensitive) disagreement coefficient, introduced in Foster et al. (2020b).

Definition 6.3. *For a distribution $\rho \in \Delta(\Pi)$, the disagreement coefficient is defined as*

$$\boldsymbol{\theta}(\mathcal{F}, \Delta_0, \varepsilon_0; \rho) = \sup_{\Delta\geq\Delta_0, \varepsilon\geq\varepsilon_0} \left\{ \frac{\Delta^2}{\varepsilon^2} \cdot \mathbb{P}_{\pi\sim\rho}\big(\exists f \in \mathcal{F} : |f(\pi)| > \Delta, \mathbb{E}_{\pi\sim\rho}\big[f^2(\pi)\big] \leq \varepsilon^2\big) \right\} \vee 1. \tag{70}$$

Our main result is as follows.

Theorem 6.2. *For any class $\mathcal{F} \subseteq (\Pi \to [0,1])$ and prior $\mu \in \Delta(\mathcal{M})$, the posterior sampling strategy, which plays $\rho_\mu(\pi) := \mu(\{\pi_M = \pi\})$, certifies that*

$$\underline{\mathsf{dec}}^{\mathrm{Sq}}_{\gamma}(\mu, \overline{M}) \leq \inf_{\Delta>0} \left\{ 2\Delta + 12\frac{\boldsymbol{\theta}(\mathcal{F}_{\mathcal{M}} - f^{\overline{M}}, \Delta, \gamma^{-1}; \rho_\mu)\log^2(\gamma)}{\gamma} \right\} + (2\gamma)^{-1} \tag{71}$$

for all $\gamma \geq e$.

Theorem 6.2 generalizes the decoupling argument used to prove the upper bound on the Decision-Estimation Coefficient for multi-armed bandits in Section 5, with the disagreement coefficient providing a bound on the price of decoupling. Theorem 6.1 follows immediately from this theorem, along with the following technical result from Foster et al. (2020b).

Lemma 6.1 (Foster et al. (2020b))**.** *For all $\rho \in \Delta(\Pi)$ and $\Delta, \varepsilon > 0$, we have $\boldsymbol{\theta}(\mathcal{F}, \Delta, \varepsilon; \rho) \leq 4(\mathfrak{s}(\mathcal{F}, \Delta))^2$ and $\boldsymbol{\theta}(\mathcal{F}, \Delta, \varepsilon; \rho) \leq 4\mathfrak{e}(\mathcal{F}, \Delta)$.*

7 Application to Reinforcement Learning

We now turn our focus to episodic reinforcement learning with function approximation (Example 1.2), providing efficient algorithms and upper bounds (via E2D) and lower bounds. Our main results are as follows.

1. First, in Section 7.1, we show how to extend the PC-IGW method of Section 5 to obtain new, efficient algorithms for *bilinear classes* (Du et al., 2021), a large class of reinforcement learning problems which captures many settings where sample-efficient reinforcement learning is possible. Our results here are applicable when the class of models has moderate model estimation complexity, and when a proper estimation algorithm is available.

2. Next, in Section 7.2, we provide tighter guarantees for bilinear classes that scale only with the estimation complexity for the underlying class of value functions. This result recovers a broader set of sample-efficient learning guarantees for bilinear classes but, unlike the first result, is not computationally efficient.
3. Finally, in Section 7.3, we derive lower bounds for reinforcement learning. As a highlight, we show that the Decision-Estimation Coefficient recovers exponential lower bounds for reinforcement learning with linearly realizable function approximation (Wang et al., 2021).

All of the results in this section take $\Pi = \Pi_{\mathrm{RNS}}$ and $\mathcal{R} \subseteq [0,1]$ (that is, $\sum_{h=1}^{H} r_h \in [0,1]$) unless otherwise specified (Jiang and Agarwal, 2018; Zhang et al., 2021).

Reinforcement learning: Model-based, model-free, and beyond. Recall that for reinforcement learning, each model $M \in \mathcal{M}$ consists of a collection of probability transition functions $P_1^M, \ldots, P_H^M$ and reward distributions $R_1^M, \ldots, R_H^M$. This formulation can be viewed as an instance of *model-based* reinforcement learning, where one uses function approximation to directly model the dynamics of the environment. What is perhaps less obvious is that this formulation also suffices to capture model-free methods and direct policy search methods.

- For model-free (or, value function approximation) methods, one typically assumes that we are given a class of Q-value functions $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_H$ that is *realizable* in the sense that it contains the optimal Q-function for every problem instance under consideration. This is captured in the DMSO framework by taking
$$\mathcal{M}_{\mathcal{Q}} = \left\{ M \mid Q_h^{M,\star} \in \mathcal{Q}_h \;\; \forall h \in [H] \right\} \tag{72}$$
as the induced class of models, and hence we can derive upper and lower bounds for this setting. A well-known special case is that of *linearly realizable* function approximation, where each class $\mathcal{Q}_h$ is linear; this setting is addressed in Section 7.3. Naturally, one can modify the definition in (72) to incorporate commonly used additional assumptions such as low rank structure (Jin et al., 2020c) or completeness under Bellman backups.
- Direct policy search methods do not model the dynamics or value functions, and instead work directly with a given class of policies Π (specified via function approximation). Here, a natural notion of realizability (e.g., Mou et al. (2020)) is to assume the policy class contains the optimal policy for all problem instances under consideration. This is captured by taking
$$\mathcal{M}_{\Pi} = \{ M \mid \pi_M \in \Pi \}$$
as the induced class of models. We do not focus on this setting here, as few positive results are known.

7.1 Bilinear Classes: Basic Results

In this section we use the E2D algorithm to provide efficient guarantees for reinforcement learning with bilinear classes (Du et al., 2021). The bilinear class framework generalizes a number of previous structural conditions, most notably Bellman rank (Jiang et al., 2017), and captures most known settings where sample-efficient reinforcement learning is possible. The following is an adaptation of the definition from Du et al. (2021).

Definition 7.1 (Bilinear class). *A model class $\mathcal{M}$ is said to be bilinear relative to reference model $\overline{M}$ if:*

1. *There exist functions $W_h(\cdot\,; \overline{M}) : \mathcal{M} \to \mathbb{R}^d$, $X_h(\cdot\,; \overline{M}) : \mathcal{M} \to \mathbb{R}^d$ such that for all $M \in \mathcal{M}$ and $h \in [H]$,*
$$\left| \mathbb{E}^{\overline{M},\pi_M}\left[Q_h^{M,\star}(s_h, a_h) - r_h - V_h^{M,\star}(s_{h+1}) \right] \right| \leq |\langle W_h(M; \overline{M}), X_h(M; \overline{M}) \rangle|. \tag{73}$$
We assume that $W_h(\overline{M}; \overline{M}) = 0$.
2. *Let $z_h = (s_h, a_h, r_h, s_{h+1})$. There exists a collection of estimation policies $\{\pi_M^{\mathrm{est}}\}_{M \in \mathcal{M}}$ and estimation functions $\{\ell_M^{\mathrm{est}}(\cdot\,;\cdot)\}_{M \in \mathcal{M}}$ such that for all $M, M' \in \mathcal{M}$ and $h \in [H]$,*
$$\langle X_h(M; \overline{M}), W_h(M'; \overline{M}) \rangle = \mathbb{E}^{\overline{M}, \pi_M \circ_h \pi_M^{\mathrm{est}}}\left[\ell_M^{\mathrm{est}}(M'; z_h) \right]. \tag{74}$$
If $\pi_M^{\mathrm{est}} = \pi_M$, we say that estimation is on-policy.

If $\mathcal{M}$ is bilinear relative to all $\overline{M} \in \mathcal{M}$, we say that $\mathcal{M}$ is a bilinear class. We let $d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$ denote the minimal dimension d for which the bilinear class property holds relative to $\overline{M}$, and define $d_{\mathrm{bi}}(\mathcal{M}) = \sup_{\overline{M}\in\mathcal{M}} d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$. We let $L_{\mathrm{bi}}(\mathcal{M};\overline{M}) \geq 1$ denote any almost sure upper bound on $|\ell^{\mathrm{est}}_{M}(M'; z_h)|$ under $\overline{M}$, and let $L_{\mathrm{bi}}(\mathcal{M}) = \sup_{\overline{M}\in\mathcal{M}} L_{\mathrm{bi}}(\mathcal{M};\overline{M})$.

For the remainder of this section, we define concatenated factorizations $X(M;\overline{M}), W(M;\overline{M}) \in \mathbb{R}^{dH}$ via

$$X(M;\overline{M}) = (X_1(M;\overline{M}), \ldots, X_H(M;\overline{M})), \quad \text{and} \quad W(M;\overline{M}) = (W_1(M;\overline{M}), \ldots, W_H(M;\overline{M})). \tag{75}$$

Basic examples of bilinear classes (cf. Du et al. (2021)) include:

- Linear MDPs (Yang and Wang, 2019; Jin et al., 2020c).
- Block MDPs and reactive POMDPs (Krishnamurthy et al., 2016; Du et al., 2019a).
- FLAMBE/feature selection in low rank MDPs (Agarwal et al., 2020a).
- MDPs with Linear $Q^\star$ and $V^\star$ (Du et al., 2021).
- MDPS with Low Occupancy Complexity (Du et al., 2021).
- Linear mixture MDPs (Modi et al., 2020; Ayoub et al., 2020).
- Linear dynamical systems (LQR) (Dean et al., 2020).

Further examples include $Q^\star$-irrelevant state aggregation (Li, 2009; Dong et al., 2019) and classes with low Bellman rank (Jiang et al., 2017) or Witness rank (Sun et al., 2019).

The results in this subsection are applicable to any bilinear class $\mathcal{M}$ for which i) a proper online estimation algorithm is available (i.e., an estimator for which $\widehat{M}^{(t)} \in \mathcal{M}$),[19] and ii) the model estimation complexity $\mathsf{est}(\mathcal{M}, T)$ is non-trivial. The general setting (without these assumptions) is handled in the Section 7.2, albeit not with an efficient algorithm. We proceed as follows:

- Section 7.1.1 gives a generalization of the PC-IGW algorithm for bilinear classes, and shows that it leads to a bound on the DEC. We also provide an analogous guarantee for posterior sampling.
- Section 7.1.2 shows how to implement the PC-IGW strategy efficiently given access to a planning oracle.
- Section 7.1.3 gives general tools to design (proper) online estimation algorithms.
- Section 7.1.4 puts these results together to give end-to-end algorithms for bilinear classes of interest.

7.1.1 Bounding the Decision-Estimation Coefficient: PC-IGW and Posterior Sampling

We now show how to bound the Decision-Estimation Coefficient for reinforcement learning with bilinear classes. Our development here parallels that of the tabular setting in Section 5.2. We first provide a bound on the Bayesian DEC via posterior sampling, then provide an efficient algorithm that leads to a bound on the frequentist DEC by adapting the inverse gap weighting technique.

Bounding the DEC with Posterior Sampling. For $\alpha \in [0,1]$, let π^{α}_{M} be the randomized policy that—for each h—plays $\pi_{M,h}$ with probability $1-\alpha/H$ and $\pi^{\mathrm{est}}_{M,h}$ with probability α/H. Our guarantee for (modified) posterior sampling is as follows.

Theorem 7.1. *Let $\mathcal{M}$ be a bilinear class and let $\overline{M} \in \mathcal{M}$. Let $\mu \in \Delta(\mathcal{M})$ be given, and consider the modified posterior sampling strategy that samples $M \sim \mu$ and plays π^{α}_{M}, where $\alpha \in [0,1]$ is a parameter.*

- *If $\pi^{\mathrm{est}}_{M} = \pi_{M}$ (i.e., estimation is on-policy), this strategy with $\alpha = 0$ certifies that*

$$\underline{\mathsf{dec}}_{\gamma}(\mathcal{M}, \overline{M}) \leq \frac{4H^2 L^2_{\mathrm{bi}}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}$$

  for all $\gamma > 0$.

[19] This restriction arises due to technical issues in bounding $\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$ (versus $\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$).

- *For general estimation policies, this strategy with* $\alpha = \left(\frac{8H^3 L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}\right)^{1/2}$ *certifies that*

$$\underline{\mathsf{dec}}_{\gamma}(\mathcal{M}, \overline{M}) \le \left(\frac{32H^3 L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}\right)^{1/2}$$

  whenever $\gamma \ge 32H^3 L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$.

Focusing on dimension, the DEC bound in the on-policy case where $\pi_M^{\mathrm{est}} = \pi_M$ scales as $\frac{d_{\mathrm{bi}}(\mathcal{M})}{\gamma}$, which leads to regret

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{d_{\mathrm{bi}}(\mathcal{M}) \cdot T \cdot \mathbf{Est}_{\mathsf{H}}}.$$

In the general case, the DEC bound scales as $\sqrt{\frac{d_{\mathrm{bi}}(\mathcal{M})}{\gamma}}$ due to the use of forced exploration, which leads to

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim (d_{\mathrm{bi}}(\mathcal{M}) \cdot \mathbf{Est}_{\mathsf{H}})^{1/3} \cdot T^{2/3}.$$

This matches the regret bound implied by the results in Du et al. (2021), which also rely on forced exploration; algorithms with $\sqrt{T}$-regret for general bilinear classes are not currently known.

See Appendix F.5 for an extension of this result which covers the closely related setting of Bellman-eluder dimension (Jin et al., 2021).

Bounding the DEC with PC-IGW. Our frequentist algorithm, PC-IGW.Bilinear (Algorithm 5), is an adaptation of the PC-IGW algorithm used in the tabular setting. The algorithm is based on the primitive of *G-optimal design*, which we use to generalize the notion of policy cover (i.e., a collection of policies that maximizes the visitation probability for any given state-action pair) used in Algorithm 4.

Definition 7.2 (G-optimal design)**.** *Let a set* $\mathcal{X} \subseteq \mathbb{R}^d$ *be given. A distribution* $p \in \Delta(\mathcal{X})$ *is said to be a G-optimal design with approximation factor* $C_{\mathrm{opt}} \ge 1$ *if*

$$\sup_{x \in \mathcal{X}} \langle \Sigma_p^{\dagger} x, x \rangle \le C_{\mathrm{opt}} \cdot d, \tag{76}$$

where $\Sigma_p := \mathbb{E}_{x \sim p}\left[x x^{\top}\right]$.

The following result guarantees existence of an exact optimal design with $C_{\mathrm{opt}} = 1$; we consider designs with $C_{\mathrm{opt}} > 1$ for computational reasons.

Fact 7.1 (Kiefer and Wolfowitz (1960))**.** *For any compact* $\mathcal{X} \subseteq \mathbb{R}^d$*, there exists an optimal design with* $C_{\mathrm{opt}} = 1$.

The PC-IGW.Bilinear algorithm (Algorithm 5) combines inverse gap weighting with optimal design. Let $\overline{M} \in \mathcal{M}$ be the estimated model. For each layer h, the algorithm computes an (approximate) G-optimal design for the collection of vectors $\{Y_h(M; \overline{M})\}_{M \in \mathcal{M}}$, where

$$Y_h(M; \overline{M}) := \frac{X_h(M; \overline{M})}{\sqrt{1 + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))}}.$$

We denote resulting optimal design by $q_h^{\mathrm{opt}} \in \Delta(\mathcal{M})$. Analogous to the tabular setting, the optimal design for the unweighted factors $\left\{X_h(M; \overline{M})\right\}_{M \in \mathcal{M}}$ would suffice to ensure good exploration, but the optimal design for the weighted factors $\left\{Y_h(M; \overline{M})\right\}_{M \in \mathcal{M}}$ balances exploration and regret, and is critical for deriving $\sqrt{T}$-type regret bounds.

For the next step, we mix the optimal designs for each layer via $q := \frac{1}{2H}\sum_{h=1}^{H} q_h^{\mathrm{opt}} + \frac{1}{2}\delta_{\overline{M}}$; we also mix in the estimated model $\overline{M}$. Finally, we compute a distribution over policies via inverse gap weighting:

$$p(\pi_M^{\alpha}) = \frac{q(M)}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))}.$$

We use the mixed policies π_{M}^{α} to allow for a small amount of forced exploration (controlled by the parameter α) in the off-policy case where $\pi_{M}^{\mathrm{est}} \neq \pi_{M}$.

Algorithm 5 is efficient whenever i) we can compute an approximate optimal design for the factors $\{Y_h(M; \overline{M})\}_{M \in \mathcal{M}}$ efficiently, and ii) the design has small support. The final guarantee for the algorithm scales linearly with the approximation factor C_{opt}. In the sequel, we show that both desiderata can be achieved (with $C_{\mathrm{opt}} = O(d)$) whenever the learner has access to a certain planning oracle, leading to an efficient algorithm.

The main guarantee for the PC-IGW.Bilinear algorithm is as follows.

Theorem 7.2. *Let $\mathcal{M}$ be a bilinear class. Let $\gamma > 0$ and $\overline{M} \in \mathcal{M}$ be given, and consider the PC-IGW.Bilinear strategy in Algorithm 5. Suppose the optimal design solver in Line 5 has approximation factor $C_{\mathrm{opt}} \geq 1$.*

- *If $\pi_{M}^{\mathrm{est}} = \pi_{M}$ (i.e., estimation is on-policy), this strategy with $\eta = \frac{\gamma}{3H^3 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}$ and $\alpha = 0$ certifies that*
$$\mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \leq \frac{9H^3 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}.$$
- *For general estimation policies, choosing $\eta = \frac{\gamma}{6H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}$ and $\alpha = \left(\frac{18H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}\right)^{1/2}$ certifies that*
$$\mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \leq \left(\frac{72H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}\right)^{1/2},$$
whenever $\gamma \geq 72H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$.

This guarantee matches the bound on the DEC for posterior sampling (Theorem 7.1) up to a factor of H. In particular, it leads to $\mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{d_{\mathrm{bi}}(\mathcal{M}) \cdot T \cdot \mathbf{Est}_{\mathsf{H}}}$ in the on-policy case. As with the tabular variant of PC-IGW, this approach is unique compared to previous algorithms for bilinear classes and low Bellman rank, in that it does not explicitly rely on confidence sets or optimism.

7.1.2 Computational Efficiency

We now show that it is possible to efficiently implement the optimal design step required by Algorithm 5 whenever we have access to an efficient oracle for planning with a fixed model. For the results in this subsection, we assume that for all $\overline{M} \in \mathcal{M}$, each set $\{X_h(M; \overline{M})\}_{M \in \mathcal{M}}$ is compact and has full dimension. We adopt the convention that for a collection of vectors $x_1, \ldots, x_d \in \mathbb{R}^d$, $\det(x_1, \ldots, x_d)$ denotes the determinant of the matrix $(x_1, \ldots, x_d) \in \mathbb{R}^{d \times d}$.

Oracle and assumptions. Our main computational primitive is as follows.

Definition 7.3 (Bilinear planning oracle)**.** *For a bilinear class $\mathcal{M}$, a bilinear planning oracle takes as input a model $\overline{M} \in \mathcal{M}$ with bilinear dimension d and vector $\theta \in \mathbb{R}^{dH}$ with $\|\theta\|_2 \leq 1$ and returns*
$$\mathbf{Alg}_{\mathsf{Plan}}(\theta; \overline{M}) = \operatorname*{arg\,max}_{M \in \mathcal{M}} \langle X(M; \overline{M}), \theta \rangle. \tag{78}$$

Informally, the planning oracle Definition 7.3 asserts that we can efficiently perform linear optimization when the underlying model $\overline{M}$ is held fixed. For intuition recall that for most bilinear classes, we have $X_h(M; \overline{M}) = X_h(\pi_M; \overline{M})$, and that $X_h(\pi_M; \overline{M})$ typically represents a sufficient statistic for the roll-in distribution induced by running π_M in $\overline{M}$. In this case, the bilinear planning oracle precisely corresponds to planning (i.e., maximizing a known reward function) with a known model.

Fact 7.2. *The following models admit efficient bilinear planning oracles:*

- *For Linear MDPs, Linear Bellman Complete MDPs, Block MDPs, FLAMBE/Feature Selection, and MDPs with Linear $Q^\star/V^\star$, the oracle $\mathbf{Alg}_{\mathsf{Plan}}$ can be implemented efficiently whenever we can efficiently plan with known dynamics in $\mathcal{M}$ and any linear reward function.*

Algorithm 5 PC-IGW.Bilinear

// Exploration for bilinear classes via inverse gap weighting.

1: **parameters**:

Model class $\mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}$ with bilinear dimension d.

Learning rate $\eta > 0$.

Forced exploration parameter $\alpha > 0$.

2: For each $M \in \mathcal{M}$, let π_{M}^{α} be the randomized policy that for each h plays $\pi_{M,h}$ with probability $1-\alpha/H$ and $\pi_{M,h}^{\mathrm{est}}$ with probability α/H.

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

4: For each $M \in \mathcal{M}$ and $h \in [H]$, define

$$Y_h(M;\overline{M}) = \frac{X_h(M;\overline{M})}{\sqrt{1+\eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))}}. \tag{77}$$

// The distribution q_h^{opt} is assumed to solve (76) with approximation factor C_{opt}.

5: For each h, obtain $q_h^{\mathrm{opt}} \in \Delta(\mathcal{M})$ from optimal design solver (e.g., Algorithm 6) for $\left\{Y_h(M;\overline{M})\right\}_{M\in\mathcal{M}}$.

6: Define $q = \frac{1}{2}\sum_{h=1}^{H} q_h^{\mathrm{opt}} + \frac{1}{2}\delta_{\overline{M}}$.

7: For each $M \in \mathrm{supp}(q)$, let

$$p(\pi_M^{\alpha}) = \frac{q(M)}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))},$$

where $\lambda \in [1/2, 1]$ is chosen such that $\sum_{M\in\mathcal{M}} p(M) = 1$.

8: **return** p.

- *For the Low Occupancy Complexity setting,* $\mathbf{Alg}_{\mathsf{Plan}}$ *can be implemented efficiently whenever we can efficiently plan with known dynamics in* $\mathcal{M}$ *and an arbitrary reward function.*[20]

Beyond a planning oracle, we require a (mild) additional assumption on the bilinear representation for $\mathcal{M}$.

Assumption 7.1. *For all* $\overline{M} \in \mathcal{M}$*, there exists* $\theta(\overline{M}) \in \mathbb{R}^{dH}$ *such that for all* $M \in \mathcal{M}$*,*

$$f^{\overline{M}}(\pi_M) = \left\langle X(M;\overline{M}), \theta(\overline{M})\right\rangle \tag{79}$$

This assumption implies that the bilinear planning oracle can be used to find an optimal policy when the model is known to the learner, which is a fairly minimal requirement if one wishes to develop efficient algorithms when the dynamics are *unknown*.

Fact 7.3. *The following models satisfy Assumption 7.1: Linear/Linear Bellman Complete MDPs,*[21] *Block MDPs, FLAMBE/Feature Selection, Linear* $Q^\star/V^\star$*, Low Occupancy Complexity, and Linear Mixture MDPs.*[22]

Algorithm. The algorithm we consider, IGW-Spanner (Algorithm 6), is an adaptation of an algorithm for contextual bandits with linearly structured actions from a consecutive paper (Zhu et al., 2021). The idea behind the algorithm is to replace the notion of optimal design with that of a *barycentric spanner* (Awerbuch and Kleinberg, 2008).

Definition 7.4 (Barycentric spanner)**.** *For a set* $\mathcal{X} \subseteq \mathbb{R}^d$*, a subset* $\mathcal{C} = \{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ *is said to be a* C-approximate barycentric spanner *(for* $C \geq 1$*) if every element* $x \in \mathcal{X}$ *can be expressed as a linear combination of elements in* $\mathcal{C}$ *with coefficients in* $[-C, C]$.

[20] For low occupancy complexity, we require that the feature map has $\dim(\{\phi(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}) = d$.

[21] For the Linear Bellman Complete setting, we require that the all-zero function is contained in the value function class of interest.

[22] For linear mixture MDPs, we must artificially expand the bilinear class construction in Du et al. (2021) from dimension d to dimension $2d$.

The following well-known result shows that any barycentric spanner yields a d-approximate optimal design.

Fact 7.4 (Awerbuch and Kleinberg (2008); Dani et al. (2007)). *Let $\mathcal{C} = \{x_1, \ldots, x_d\}$ be a C-approximate barycentric spanner for $\mathcal{X} \subseteq \mathbb{R}^d$. Then $p := \mathrm{unif}(\mathcal{C})$ is a $C \cdot d$-approximate optimal design.*

Awerbuch and Kleinberg (2008) show that for any compact set $\mathcal{X} \subseteq \mathbb{R}^d$, one can efficiently find a 2-approximate barycentric spanner using $O(d^2 \log(d))$ calls to a linear optimization oracle capable of solving $\arg\max_{x \in \mathcal{X}} \langle x, \theta \rangle$ for any $\|\theta\|_2 \leq 1$. Algorithm 6 is an adaptation of their algorithm. The key challenge in applying this technique is that the bilinear planning oracle (Definition 7.3) can solve $\arg\max_{M \in \mathcal{M}} \langle X_h(M; \overline{M}), \theta \rangle$, but our aim is to find a bayrcentric spanner for the reweighted factors $\{Y_h(M; \overline{M})\}_{M \in \mathcal{M}}$. Implementing a linear optimization oracle for the reweighted factors entails solving

$$\arg\max_{M \in \mathcal{M}} \frac{\langle X_h(M; \overline{M}), \theta \rangle}{\sqrt{1 + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))}}, \tag{80}$$

which is not a linear function and hence cannot be solved by $\mathbf{Alg}_{\mathsf{Plan}}$ directly. To overcome this challenge, Algorithm 6 appeals to a subroutine, IGW-ArgMax (Algorithm 7), which uses binary search to reduce the optimization problem in (80) to a series of linear subproblems which can be solved by the bilinear planning oracle. In total, $\widetilde{O}(d)$ oracle calls are required to implement (80). The final guarantee for the algorithm is as follows.

Theorem 7.3 (Efficiency of IGW-Spanner (Zhu et al., 2021)). *Let $\mathcal{M}$ have bilinear dimension d relative to $\overline{M} \in \mathcal{M}$, and let Assumption 7.1 hold. Consider Algorithm 6 with $\eta \geq 1$, and suppose the initial collection $M_1, \ldots, M_d$ has*

$$\left|\det\left(X(M_1; \overline{M}), \ldots, X(M_d; \overline{M})\right)\right| \geq r^d$$

for some $r \in (0, 1)$, and that $\sup_{M \in \mathcal{M}} \|X_h(M; \overline{M})\|_2 \leq 1$. Then Algorithm 6 computes a 2-approximate barycentric spanner (and consequently a $2d$-approximate optimal design) for the collection $\{Y_h(M; \overline{M})\}_{M \in \mathcal{M}}$ using $O(d^3 \log(d/r) \log(\eta/r))$ calls to the bilinear planning oracle $\mathbf{Alg}_{\mathsf{Plan}}$.

Applying this result, we can implement PC-IGW.Bilinear using $\widetilde{O}(d^3_{\mathrm{bi}}(\mathcal{M}))$ calls to the bilinear planning oracle, at the cost an extra $d_{\mathrm{bi}}(\mathcal{M})$ factor on the DEC bound due to the approximation factor paid by the barycentric spanner. For example, the final bound on the Decision-Estimation Coefficient is

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \lesssim \frac{d^2_{\mathrm{bi}}(\mathcal{M})}{\gamma}$$

in the on-policy case.

7.1.3 Online Estimation

The bounds on the Decision-Estimation Coefficient in Theorems 7.1 and 7.2 are proven under the assumption that $\mathcal{M}$ is a bilinear class relative to the reference model $\overline{M}$. In order to apply the E2D algorithm (via Theorem 4.1), we need to bound $\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^{(t)})$ for the sequence of estimators $\widehat{M}^{(1)}, \ldots, \widehat{M}^{(T)}$ produced by the online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$. This provides a slight technical challenge because most online estimation algorithms are improper and produce estimates in $\mathrm{co}(\mathcal{M})$,[23] but it is not clear whether $\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$ enjoys the bounds on $\sup_{\overline{M} \in \mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$ from Theorems 7.1 and 7.2.

To address this issue, we sketch an approach based on layer-wise estimators.[24] Suppose for simplicity that rewards are known, and let $\mathcal{P}_h = \{P_h^M \mid M \in \mathcal{M}\}$ be the class of transition kernels for layer h. We make the following assumption.

Assumption 7.2. *The class $\mathcal{M}$ has product structure $\mathcal{M}_1 \times \cdots \times \mathcal{M}_H$. Moreover, each layer-wise class $\mathcal{P}_h$ is convex.*

[23] Recall that an estimation algorithm is said to be proper if it produces estimates that lie inside $\mathcal{M}$.

[24] This issue can be addressed more directly under various technical conditions, but we leave this for a future version of this work.

Algorithm 6 IGW-Spanner (Zhu et al., 2021)

// Find barycentric spanner for collection $\left\{ \frac{X_h(M;\overline{M})}{\sqrt{1+\eta(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi_M))}} \right\}_{M\in\mathcal{M}}$ using bilinear planning oracle.

1: **parameters**:
  Class $\mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}$ with bilinear dimension d.
  Layer $h \in [H]$.
  Learning rate $\eta > 0$.
  Initial collection $M_1, \ldots M_d \in \mathcal{M}$ with $\left|\det\left(X_h(M_1;\overline{M}), \ldots, X_h(M_d;\overline{M})\right)\right| \geq r^d$ for $r > 0$.
2: Set $\mathcal{C} = \{Y_h(M_1;\overline{M}), \ldots, Y_h(M_d;\overline{M})\}$. // $Y_h(M;\overline{M})$ is defined in (77).
3: **repeat**
4: **for** $i = 1, \ldots, d$ **do**
5:   Let $\theta \in \mathbb{R}^d$ represent the linear function $Y \mapsto \det(Y, \mathcal{C}_{-i})$. // $\det(Y, \mathcal{C}_{-i}) = \langle \theta, Y \rangle$ for all Y.
6:   Let $M \leftarrow$ IGW-ArgMax$(\theta; \overline{M}, h, \eta, r)$. // Algorithm 7.
7:   **if** $\left|\det(Y_h(M;\overline{M}), \mathcal{C}_{-i})\right| \geq 2^{1/2}|\det(\mathcal{C})|$ **then**
8:    Replace M_i with M in $\mathcal{C}$. // That is, $\mathcal{C} \leftarrow \{Y_h(M;\overline{M})\} \cup \mathcal{C}_{-i}$.
9:    **continue** to Line 3.
10: **break**
11: **return** unif$(M_1, \ldots M_d)$ // $2d$-approximate optimal design.

Instead of directly working with an estimator for the entire model M, we assume access to layer-wise estimators $\mathbf{Alg}_{\mathsf{Est};1}, \ldots, \mathbf{Alg}_{\mathsf{Est};H}$. At each round t, given the history $\{(\pi^{(i)}, r^{(i)}, o^{(i)})\}_{i=1}^{t-1}$, the layer-$h$ estimator $\mathbf{Alg}_{\mathsf{Est};h}$ produces an estimate $\widehat{P}_h^{(t)}$ for the true transition kernel $P_h^{M^\star}$. We measure performance of the estimator via layer-wise Hellinger error:

$$\mathbf{Est}_{\mathsf{H};h} := \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \mathbb{E}^{M^\star, \pi^{(t)}}\left[D^2_{\mathrm{H}}\left(P_h^{M^\star}(s_h, a_h), \widehat{P}_h^{(t)}(s_h, a_h)\right)\right]. \tag{81}$$

We obtain an estimation algorithm $\mathbf{Alg}_{\mathsf{Est}}$ for the full model by taking $\widehat{M}^{(t)}$ as the MDP that has $\widehat{P}_h^{(t)}$ as the transition kernel for each layer h. This algorithm has the following guarantee.

Proposition 7.1. *The estimator* $\mathbf{Alg}_{\mathsf{Est}}$ *described above has*

$$\mathbf{Est}_{\mathsf{H}} \leq O(\log(H)) \cdot \sum_{h=1}^{H} \mathbf{Est}_{\mathsf{H};h}.$$

Moreover, if $\widehat{P}^{(t)} \in \mathrm{co}(\mathcal{P}_h)$ *for all* h *and Assumption 7.2 is satisfied, then* $\widehat{M}^{(t)} \in \mathcal{M}$.

Proof. Immediate consequence of Lemma A.13. □

For example, whenever Assumption 7.2 is satisfied, we can use this reduction with the tools in Appendix A.3 to generically provide proper estimators with

$$\mathbf{Est}_{\mathsf{H}} \leq \widetilde{O}(H \cdot \max_h \mathsf{est}(\mathcal{P}_h, T)),$$

where $\mathsf{est}(\mathcal{P}_h, T)$ is the covering number-based complexity measure introduced in Definition A.1.

7.1.4 Examples

We highlight some basic examples in which one can put all of the results in this section together to obtain efficient algorithms for reinforcement learning.

Algorithm 7 IGW-ArgMax (Zhu et al., 2021)

// Approximately solve $\arg\max_{M\in\mathcal{M}} \frac{|\langle X_h(M;\overline{M}),\theta\rangle|}{\sqrt{1+\eta(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi_M))}}$ with bilinear planning oracle (by grid search).

1: **parameters**:
   Bilinear planning oracle $\mathbf{Alg}_{\mathsf{Plan}}$ for $\mathcal{M}$. // See (78).
   Parameter $\theta \in \mathbb{R}^d$.
   Reference model $\overline{M} \in \mathcal{M}$.
   Layer $h \in [H]$, learning rate $\eta > 0$, scale parameter $r \in (0,1)$.
2: Let $N = \lceil \log_{\frac{4}{3}}(\frac{4}{3}r^{-d})\rceil$ and define $\mathcal{E} = \left\{(\frac{3}{4})^i\right\}_{i=1}^N \cup \left\{-(\frac{3}{4})^i\right\}_{i=1}^N$.
3: $\widehat{\mathcal{M}} = \{\varnothing\}$. // Candidate argmax models.
4: **for** each $\varepsilon \in \mathcal{E}$ **do**
5: Define $\tilde{\theta} \in \mathbb{R}^{dH}$ via $\tilde{\theta} = \varepsilon \cdot (\theta \otimes e_h) + \eta\varepsilon^2 \cdot \theta(\overline{M})$.
6: $M \leftarrow \mathbf{Alg}_{\mathsf{Plan}}(\tilde{\theta}; \overline{M})$. // Call bilinear planning oracle (78) with $\tilde{\theta}$.
7: Add M to $\widehat{\mathcal{M}}$.
8: **return** $\arg\max_{M\in\widehat{\mathcal{M}}} \frac{|\langle X_h(M;\overline{M}),\theta\rangle|}{\sqrt{1+\eta(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi_M))}}$. // Enumeration over $\widetilde{O}(d)$ candidates.

Example 7.1 (Linear MDP)**.** In the linear MDP setting (Jin et al., 2020c), all $M \in \mathcal{M}$ have $P_h^M(s' \mid s,a) = \langle \phi_h(s,a), \mu_h^M(s')\rangle$, where $\phi_h(s,a) \in \mathbb{R}^d$ is a known feature map and $\mu_h^M(s) \in \mathbb{R}^d$ is a model-dependent feature map; we assume $\|\phi_h(s,a)\|_2, \|\mu^M(s')\|_2 \le 1$. In addition, the reward distribution $R_h^M(s,a)$ is assumed to have mean $\langle \phi(s,a), w_h^M\rangle$. This setting has $d_{\mathrm{bi}}(\mathcal{M}) = d$, $L_{\mathrm{bi}}(\mathcal{M}) = 1$, and $\pi_M^{\mathrm{est}} = \pi_M$. One can verify that the bilinear planning oracle is equivalent to planning in a linear MDP with fixed dynamics and rewards, which is computationally efficient, and that Assumption 7.1 is satisfied. As a result, PC-IGW.Bilinear with the IGW-Spanner subroutine is efficient, and certifies that

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \le O\left(\frac{H^3 d^2}{\gamma}\right),$$

and—when used within the E2D algorithm—ensures that

$$\mathbf{Reg}_{\mathsf{DM}} \le O\big(\sqrt{H^3 d^2 T \cdot \mathbf{Est}_{\mathsf{H}}}\big).$$

Since this class has the layer-wise structure in Assumption 7.2, we can obtain $\mathbf{Est}_{\mathsf{H}} \le \widetilde{O}(H \max_h \mathsf{est}(\mathcal{P}_h, T))$. ◁

Example 7.2 (FLAMBE/Feature Selection)**.** The FLAMBE/feature selection setting (Agarwal et al., 2020a) is similar to the linear MDP setting, except that the feature map is unknown. That is, all $M \in \mathcal{M}$ take the form $P_h^M(s' \mid s,a) = \langle \phi_h^M(s,a), \mu_h^M(s')\rangle$, where $\phi_h^M(s,a), \mu_h^M(s) \in \mathbb{R}^d$ are model-dependent feature maps with $\|\phi_h^M(s,a)\|_2, \|\mu^M(s')\|_2 \le 1$. This setting has $d_{\mathrm{bi}}(\mathcal{M}) = d$, $L_{\mathrm{bi}}(\mathcal{M}) = |\mathcal{A}|$ and $\pi_M^{\mathrm{est}} = \mathrm{unif}(\mathcal{A})$ (i.e., estimation is not on-policy).

As in the linear MDP setting, the bilinear planning oracle is equivalent to planning in a fixed linear MDP, which is computationally efficient, and Assumption 7.1 is satisfied. Hence, PC-IGW.Bilinear with the IGW-Spanner subroutine is efficient, and certifies that

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \le O\left(\sqrt{\frac{H^4 |\mathcal{A}|^2 d^2}{\gamma}}\right).$$

When used within the E2D algorithm, this gives

$$\mathbf{Reg}_{\mathsf{DM}} \le O\Big(\big(H^4 |\mathcal{A}|^2 d^2 \cdot \mathbf{Est}_{\mathsf{H}}\big)^{1/3} T^{2/3}\Big).$$

Whenever each layer-wise dynamics class $\mathcal{P}_h$ is convex, this setting satisfies Assumption 7.2, and we can obtain $\mathbf{Est}_{\mathsf{H}} \le \widetilde{O}(H \max_h \mathsf{est}(\mathcal{P}_h, T))$.

$\triangleleft$

In both of these examples, the final regret bound depends on the estimation complexity for the class of models, which is not required information-theoretically (Jin et al., 2020c; Agarwal et al., 2020a). This is an instance of one of the main gaps between our generic upper and lower bounds discussed in Section 3.5. We show how to remove this issue in the sequel, but the resulting algorithm is not efficient.

7.2 Bilinear Classes: Refined Guarantees

We now use the E2D framework to give refined regret bounds that sharpen our results for bilinear classes in the prequel. Instead of scaling with the estimation error for the model class $\mathcal{M}$, our results here scale only with the estimation complexity for the class of induced Q-functions

$$\mathcal{Q}_{\mathcal{M}} = \{Q^{M,\star} \mid M \in \mathcal{M}\},$$

which makes them better suited to reinforcement learning settings of the model-free variety. We emphasize that the results here are specialized to bilinear classes, and require modifications to the basic framework in Section 3. Deriving tighter regret bounds as a direct corollary of our general results is an important topic for further research.

Our refined results require the following modifications to the basic bilinear class framework:

- We assume that for all $M \in \mathcal{M}$, we have $X_h(M; \overline{M}) = X_h(\pi_M; \overline{M})$.
- We assume that $\pi_M^{\mathrm{est}} = \pi_M$ for all $M \in \mathcal{M}$ (i.e., estimation is on-policy), and replace Property 2 of Definition 7.1 with the following condition: For all $M, M' \in \mathcal{M}$, $h \in [H]$

$$\langle X_h(M; \overline{M}), W_h(M'; \overline{M})\rangle^2 \le L_{\mathrm{bi}}^2(\mathcal{M}) \cdot \mathbb{E}^{\overline{M},\pi_M}\left[\left([\mathcal{T}_h^{M'} V_{h+1}^{M',\star}](s_h, a_h) - [\mathcal{T}_h^{\overline{M}} V_{h+1}^{M',\star}](s_h, a_h)\right)^2\right]. \tag{82}$$

Finally, we make a Bellman completeness assumption. For a given MDP M, let $\mathcal{T}_h^M$ denote the Bellman operator for layer h, defined via $\mathcal{T}_h^M[f](s,a) = \mathbb{E}^M[r_h + \max_{a'} f(s_{h+1}, a') \mid s_h = s, a_h = a]$. In addition, let $\mathcal{Q}_{\mathcal{M};h} := \{Q_h^{M,\star} \mid M \in \mathcal{M}\}$.

Assumption 7.3 (Completeness). *For all h and $M \in \mathcal{M}$, $[\mathcal{T}_h^{M^\star} V_{h+1}^{M,\star}] \in \mathcal{Q}_{\mathcal{M};h}$.*

The first two properties above are satisfied in most standard model-free settings, including Linear and Linear Bellman Complete MDPs, Low Occupancy Complexity MDPs, and Linear $Q^\star/V^\star$ MDPs. Assumption 7.3 is satisfied for Linear and Linear Bellman Complete MDPs, but in general may or may not be satisfied for the problem under consideration. One can extend the results we provide here to the off-policy estimation case and the case where Assumption 7.3 does not hold using similar arguments, at the cost of a more complicated analysis and worse dependence on T. We do not pursue this here, since our goal is only to give a taste for how the E2D framework can recover existing results.

We derive tighter guarantees for bilinear classes by replacing the Hellinger divergence found in the definition of $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$ with the following divergence tailored to the bilinear class setting:

$$D_{\mathrm{bi}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big) := \sum_{h=1}^{H} \big\langle X_h(\pi; \overline{M}), W_h(M; \overline{M})\big\rangle^2. \tag{83}$$

This quantity is always upper bounded by Hellinger distance, but leads to tighter rates because it only depends on the models under consideration through their Bellman residuals.

Following the development for general divergences described in Section 4.3, we consider a variant of E2D.Bayes (Algorithm 2) tailored to the bilinear divergence $D_{\mathrm{bi}}(\cdot \,\|\, \cdot)$. For each time t:

- Compute posterior $\mu^{(t)} \in \Delta(\mathcal{M})$ given $\mathcal{H}^{(t-1)}$.

- Solve

$$\arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma\cdot\mathbb{E}_{\overline{M}\sim\mu^{(t)}}\big[D_{\mathrm{bi}}\big(M(\pi)\parallel\overline{M}(\pi)\big)\big]\Big], \tag{84}$$

  which corresponds to the optimization problem defining the generalized Decision-Estimation Coefficient $\mathsf{dec}_{\gamma}^{D_{\mathrm{bi}}}(\mathcal{M},\mu^{(t)})$ in (45).

We abbreviate $\mathsf{dec}_{\gamma}^{\mathrm{bi}}(\mathcal{M},\nu)\equiv\mathsf{dec}_{\gamma}^{D_{\mathrm{bi}}}(\mathcal{M},\nu)$ going forward. The following result—which is proven using the same approach as Theorem 7.1—shows that by moving to $D_{\mathrm{bi}}(\cdot\parallel\cdot)$, the DEC is bounded by the bilinear dimension.

Proposition 7.2. *Let $\mathcal{M}$ be a bilinear class. Then for all $\gamma>0$ and $\nu\in\Delta(\mathcal{M})$,*

$$\underline{\mathsf{dec}}_{\gamma}^{\mathrm{bi}}(\mathcal{M},\nu)\le\frac{H\cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma}.$$

Equipped with a bound on the Decision-Estimation Coefficient, we proceed to give a regret bound for the E2D variant above. Our regret bound depends on the following notion of estimation complexity for the value function class $\mathcal{Q}_{\mathcal{M}}$.

Definition 7.5. *Models $M_1,\ldots,M_N$ are said to be an ε-cover for $\mathcal{Q}_{\mathcal{M}}$ if for all $M\in\mathcal{M}$, there exists $i\in[N]$ such that*

$$\max_{h\in[H]}\sup_{s\in\mathcal{S},a\in\mathcal{A}}\left|Q_h^{M_i,\star}(s,a)-Q_h^{M,\star}(s,a)\right|\le\varepsilon.$$

Let $\mathcal{N}(\mathcal{Q}_{\mathcal{M}},\varepsilon)$ denote the size of the smallest such cover, and define $\mathsf{est}(\mathcal{Q}_{\mathcal{M}},T)=\inf_{\varepsilon>0}\big\{\log\mathcal{N}(\mathcal{Q}_{\mathcal{M}},\varepsilon)+\varepsilon^2T\big\}$.

Note that for any setting in which the optimal Q-functions are linear functions in dimension d (e.g., Linear and Linear Bellman Complete MDPs, Linear $Q^\star/V^\star$), we have $\mathsf{est}(\mathcal{Q}_{\mathcal{M}},T)=\widetilde{O}(dH)$.

Our refined regret bound for bilinear classes is as follows.

Theorem 7.4. *Suppose that Assumption 7.3 holds, and that estimation is on-policy (i.e., $\pi_M^{\mathrm{est}}=\pi_M$). Consider the* E2D.Bayes *algorithm with the optimization problem given in (84). This algorithm guarantees that for any prior $\mu\in\Delta(\mathcal{M})$ and $\gamma>0$,*

$$\mathbb{E}_{M^\star\sim\mu}\,\mathbb{E}^{M^\star}[\mathbf{Reg}_{\mathsf{DM}}]\le\widetilde{O}\left(\sqrt{H^2L_{\mathrm{bi}}^2(\mathcal{M})\cdot d_{\mathrm{bi}}^2(\mathcal{M})\cdot T\cdot\mathsf{est}(\mathcal{Q}_{\mathcal{M}},T)}\right).$$

Consequently, we have $\mathfrak{M}(\mathcal{M},T)\le\widetilde{O}\big(\sqrt{H^2L_{\mathrm{bi}}^2(\mathcal{M})\cdot d_{\mathrm{bi}}^2(\mathcal{M})\cdot T\cdot\mathsf{est}(\mathcal{Q}_{\mathcal{M}},T)}\big)$.

As a concrete example, this leads to $\sqrt{\mathrm{poly}(d,H)\cdot T}$ regret for Linear and Linear Bellman Complete MDPs. The main idea behind the proof is as follows. Using the definition of the DEC, we show that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}]\lesssim\sup_{\nu\in\Delta(\mathcal{M})}\underline{\mathsf{dec}}_{\gamma}^{\mathrm{bi}}(\mathcal{M},\nu)\cdot T+\gamma\cdot\sum_{t=1}^{T}\mathbb{E}\Big[D_{\mathrm{bi}}\Big(M^\star(\pi^{(t)})\parallel\widehat{M}^{(t)}(\pi^{(t)})\Big)\Big].$$

The first term above is bounded by Proposition 7.2, and we bound the estimation error term involving $D_{\mathrm{bi}}(\cdot\parallel\cdot)$ using an argument based on the elliptic potential (Du et al., 2021; Jin et al., 2021).

7.3 Lower Bounds

We close this section by complementing our upper bounds on the Decision-Estimation Coefficient with lower bounds. We first give a basic lower bound for bilinear classes, then give lower bounds for learning with linearly realizable Q-functions.

7.3.1 Lower Bound for Bilinear Classes

The upper bounds for bilinear classes in Theorems 7.1 and 7.2 scale with $\frac{d}{\gamma}$, where d is the bilinear dimension. The following result shows that this dependence is unavoidable.

Proposition 7.3 (Lower bound for linear MDPs)**.** *For all $d \geq 4$ there exists a bilinear class $\mathcal{M}$ with dimension d, horizon $H = 1$, $\mathcal{R} = [-1, +1]$, and $\|X(M; \overline{M})\|_2, \|W(M; \overline{M})\|_2 \leq 1$ for all $M, \overline{M} \in \mathcal{M}$, such that for all $\gamma \geq \frac{d}{3}$,*

$$\mathsf{dec}_\gamma(\mathcal{M}^\infty_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \geq \frac{d}{12\gamma}$$

for some $\overline{M} \in \mathcal{M}$, where $\varepsilon_\gamma = \frac{d}{3\gamma}$.

Proof of Proposition 7.3. Immediate consequence of Proposition 6.2. □

Invoking this result within Theorem 3.2 leads to a lower bound on regret of the form $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \Omega(\sqrt{dT})$, which matches the upper bound in Section 7.1 in terms of dependence on d and T.

7.3.2 Lower Bound for Linearly Realizable MDPs

Learning with *linearly realizable* function approximation is a well-studied problem in reinforcement learning (Weisz et al., 2021; Wang et al., 2021). Here, we are given a feature map $\phi(s, a) \in \mathbb{R}^d$ satisfying $\|\phi(s, a)\|_2 \leq 1$, and consider a class of value functions $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_H$ given by

$$\mathcal{Q}_h = \left\{(s, a) \mapsto \langle \theta, \phi(s, a)\rangle \mid \theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1\right\}. \tag{85}$$

We assume that $\mathcal{Q}$ is realizable in the sense that it contains the optimal Q-function for all problem instances under consideration, which corresponds to choosing $\mathcal{M}$ to be

$$\mathcal{M}_{\mathcal{Q}} = \left\{M \mid Q_h^{M,\star} \in \mathcal{Q}_h \; \forall h\right\}.$$

Weisz et al. (2021) provide an exponential lower bound which establishes that sample-efficient reinforcement learning is not possible in this setting, and Wang et al. (2021) show that this lower bound continues to hold even when the instances under consideration have constant suboptimality gap. The following result shows that the Decision-Estimation Coefficient is exponential for this setting, thereby recovering these results.

Proposition 7.4 (Lower bound for linearly realizable MDPs)**.** *For all $d \geq 2^9$ and $H \geq 2$, there exists a family of linearly realizable MDPs $\mathcal{M}$ with $\mathcal{R} = [-1, +1]$ such that for all $\gamma > 0$, there exists $\overline{M} \in \mathcal{M}$ for which*

$$\mathsf{dec}_\gamma(\mathcal{M}_{1/2}(\overline{M}), \overline{M}) \geq \frac{1}{48}\mathbb{I}\left\{\gamma \leq 2^{-9}\min\{2^H, \exp(2^{-10}d)\}\right\}. \tag{86}$$

Using Theorem 3.1 with $\gamma \propto T\log(T)$, we conclude that any algorithm for the linearly realizable setting must have

$$\mathbf{Reg}_{\mathsf{DM}} \geq \widetilde{\Omega}\Big(\min\Big\{2^H, 2^{\Omega(d)}, T\Big\}\Big)$$

with constant probability, which in turn implies that $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \widetilde{\Omega}\big(\min\{2^H, 2^{\Omega(d)}, T\}\big)$.

7.3.3 Lower Bound for Linearly Realizable MDPs with Deterministic Dynamics and Gap

While Proposition 7.4 shows that linear realizability is not sufficient for sample-efficient reinforcement learning, it is known that linear realizability *does* suffice if one restricts to deterministic dynamics and rewards. In particular, Wen and Van Roy (2017) provide an algorithm for this setting that has

$$\mathbf{Reg}_{\mathsf{DM}} \leq O(dH)$$

whenever the feature dimension is d and $\sum_{h=1}^H r_h \in [0, 1]$.[25] The following result provides a complementary lower bound.

[25] See Du et al. (2019b); Wang et al. (2021) for generalizations to various types of nearly-deterministic systems.

Proposition 7.5 (Lower bound for deterministic linearly realizable MDPs)**.** *There exists a collection $\mathcal{M}$ of linearly realizable MDPs $\mathcal{M}$ with $\mathcal{R} = [0, 1]$, $H = 1$, constant suboptimality gap, and deterministic rewards and dynamics, such that for all $\gamma > 0$, there exists $\overline{M} \in \mathcal{M}$ such that*

$$\mathsf{dec}_\gamma(\mathcal{M}^{\infty}_{1/3}(\overline{M}), \overline{M}) \geq \frac{1}{12}\mathbb{I}\left\{\gamma \leq \frac{d}{48}\right\}.$$

By Theorem 3.1, this result implies that any algorithm for the linearly realizable setting with deterministic dynamics and rewards must have

$$\mathbf{Reg}_{\mathsf{DM}} \geq \widetilde{\Omega}(\min\{d, T\})$$

with constant probability, which in turn implies that $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \widetilde{\Omega}(\min\{d, T\})$.

8 Incorporating Contextual Information

In this section we consider a *contextual* variant of the Decision Making with Structured Observations framework in which the learner is given additional side information (in the form of a covariate or context) before each decision is made. This setting encompasses the well-studied contextual bandit problem (Auer et al., 2002b; Langford and Zhang, 2008; Chu et al., 2011; Beygelzimer et al., 2011; Agarwal et al., 2014; Foster and Rakhlin, 2020) as well, as well as various contextual reinforcement learning problems (Abbasi-Yadkori and Neu, 2014; Modi et al., 2018; Dann et al., 2019; Modi and Tewari, 2020). We show that the E2D paradigm seamlessly extends to incorporate contextual information, leading to new, efficient algorithms.

Setting. In the contextual DMSO framework, we adopt the following protocol for T rounds, where for each round $t = 1, \ldots, T$:

1. Nature provides the learner with a *context* $x^{(t)} \in \mathcal{X}$, where $\mathcal{X}$ is the *context space*.
2. The learner selects a decision $\pi^{(t)} \in \Pi$.
3. Nature selects a reward $r^{(t)} \in \mathcal{R}$ and observation $o^{(t)} \in \mathcal{O}$ based on the decision, which are then observed by the learner.

We allow each context $x^{(t)}$ to be chosen in an arbitrary, potentially adaptive fashion but—following the development so far—assume that rewards and observations are stochastic. We work with models of the form $M(\cdot, \cdot)$, where $M(x, \pi)$ denotes the distribution over (r, o) when x is the context and decision π is selected. We assume that at each timestep, given $x^{(t)}$, the pair $(r^{(t)}, o^{(t)})$ is drawn independently from an unknown model $M^\star(x^{(t)}, \pi^{(t)})$. As before, we assume access to a class of models $\mathcal{M}$ that contains the true model $M^\star$.

Assumption 8.1. *The model class $\mathcal{M}$ contains the true model $M^\star$.*

For each model $M \in \mathcal{M}$, let $f^M(x, \pi) := \mathbb{E}_{r \sim M(x,\pi)}[r(\pi)]$ denote the mean reward function and let $\boldsymbol{\pi}_M(x) := \arg\max_{\pi \in \Pi} f^M(x, \pi)$ denote the decision-making policy with the greatest expected reward under M. Finally, we define $\mathcal{F}_{\mathcal{M}} = \{f^M \mid M \in \mathcal{M}\}$ as the induced class of mean reward functions.[26]

We evaluate the learner's performance in terms of regret to the optimal decision-making policy for $M^\star$:

$$\mathbf{Reg}_{\mathsf{DM}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}}[f^\star(x^{(t)}, \boldsymbol{\pi}^\star(x^{(t)})) - f^\star(x^{(t)}, \pi^{(t)})], \tag{87}$$

where we abbreviate $f^\star = f^{M^\star}$ and $\boldsymbol{\pi}^\star = \boldsymbol{\pi}_{M^\star}$.

The advantage of this formulation—which generalizes similar formulations for the contextual bandit problem (Chu et al., 2011; Beygelzimer et al., 2011; Foster and Rakhlin, 2020)—is that we accommodate arbitrarily generated sequences of contexts, which may correspond to, e.g., users arriving at a website as they please. Note that the fully stochastic contextual bandit problem, in which $x^{(1)}, \ldots, x^{(T)}$ are i.i.d., is a special case of the basic DMSO framework in Section 1 (with policies as decisions), and does not require dedicated treatment.

[26] We use the notation $\boldsymbol{\pi}_M$ (compared to π_M in the non-contextual setting) to distinguish *decisions* from *policies that map contexts to decisions.*

8.1 The Contextual E2D Algorithm

Algorithm 8 Contextual E2D

1: **parameters**:

Online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$.

Exploration parameter $\gamma > 0$.

Divergence $D(\cdot \,\|\, \cdot)$.

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

3: Receive context $x^{(t)}$.

4: Compute estimate $\widehat{M}^{(t)} = \mathbf{Alg}^{(t)}_{\mathsf{Est}}\Big(\{(x^{(i)}, \pi^{(i)}, r^{(i)}, o^{(i)})\}_{i=1}^{t-1}\Big)$.

5: Define

$$p^{(t)} = \arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^{M}(x^{(t)}, \boldsymbol{\pi}_{M}(x^{(t)})) - f^{M}(x^{(t)}, \pi) - \gamma \cdot D\Big(M(x^{(t)}, \pi) \,\|\, \widehat{M}^{(t)}(x^{(t)}, \pi)\Big)\Big].$$

6: Sample decision $\pi^{(t)} \sim p^{(t)}$ and update estimation oracle with $(x^{(t)}, \pi^{(t)}, r^{(t)}, o^{(t)})$.

Algorithm 8 is a contextual generalization of the E2D meta-algorithm. The algorithm has the same structure as the basic (non-contextual) E2D algorithm (Option I), with the main difference being that we use the context $x^{(t)}$ to form the minimax problem solved at each step. Following Section 4.3, the algorithm is stated in terms of an arbitrary user-specified divergence $D(\cdot \,\|\, \cdot)$ for added flexibility.

In more detail, at each time t, the algorithm first receives the context $x^{(t)}$, then obtains an estimated model $\widehat{M}^{(t)}(\cdot, \cdot)$ from the estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$; here, unlike the non-contextual setting, the oracle can make use of previous contexts to form the estimate. Given the estimator and context, the algorithm solves the optimization problem

$$p^{(t)} = \arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^{M}(x^{(t)}, \boldsymbol{\pi}_{M}(x^{(t)})) - f^{M}(x^{(t)}, \pi) - \gamma \cdot D\Big(M(x^{(t)}, \pi) \,\|\, \widehat{M}^{(t)}(x^{(t)}, \pi)\Big)\Big], \tag{88}$$

which corresponds to the non-contextual optimization problem (44) applied to the projected class $\mathcal{M}|_{x^{(t)}}$, where

$$\mathcal{M}|_{x} := \{M(x, \cdot) \mid M \in \mathcal{M}\}.$$

Finally, the algorithm samples $\pi^{(t)} \sim p^{(t)}$ and updates the estimation oracle with the example $(x^{(t)}, \pi^{(t)}, r^{(t)}, o^{(t)})$. Notably, the per-round computational complexity is exactly the same as in the non-contextual setting.

Conceptually, Algorithm 8 can be interpreted as a universal generalization of the SquareCB algorithm of Foster and Rakhlin (2020) from finite-action contextual bandits to arbitrary contextual decision making problems. Rather than using the inverse gap weighting strategy in SquareCB, which is tailored to contextual bandits with finite actions, we simply solve the optimization problem that defines the Decision-Estimation Coefficient for the current context, thereby accommodating any learnable contextual decision making problem.

The performance guarantee for the contextual E2D algorithm depends on the estimation performance of the oracle $\mathbf{Alg}_{\mathsf{Est}}$ with respect to the divergence D, on the *observed sequence of contexts*:

$$\mathbf{Est}_{\mathsf{D}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\Big[D\Big(M^{\star}(x^{(t)}, \pi^{(t)}) \,\|\, \widehat{M}^{(t)}(x^{(t)}, \pi^{(t)})\Big)\Big]. \tag{89}$$

Let $\widehat{\mathcal{M}}$ be any set such that $\widehat{M}^{(t)} \in \widehat{\mathcal{M}}$ for all t almost surely, and recall that in the non-contextual setting, $\mathsf{dec}^{D}_{\gamma}(\mathcal{M}, \widehat{\mathcal{M}}) := \sup_{\overline{M}\in\widehat{\mathcal{M}}} \mathsf{dec}^{D}_{\gamma}(\mathcal{M}, \overline{M})$. We have the following regret bound, generalizing Theorem 4.3.

Theorem 8.1. *Algorithm 8 with exploration parameter $\gamma > 0$ guarantees that*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sup_{x\in\mathcal{X}} \mathsf{dec}^{D}_{\gamma}(\mathcal{M}|_{x}, \widehat{\mathcal{M}}|_{x}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{D}} \tag{90}$$

almost surely.

This result shows that any interactive decision making problem that is learnable in the non-contextual setting is also learnable in the presence of arbitrarily selected contexts, as long as estimation is feasible.

8.2 Application to Contextual Bandits

The contextual bandit problem (Chu et al., 2011; Beygelzimer et al., 2011; Foster and Rakhlin, 2020) is the most basic special case of the contextual DMSO setting, and corresponds to the case in which there are no auxiliary observations (i.e., $\mathcal{O} = \{\varnothing\}$). The contextual bandit problem with finite actions is quite well-studied but contextual bandits with continuous, structured action spaces are comparatively under-explored. As a consequence of the bounds on the Decision-Estimation Coefficient from Section 6, we show that Algorithm 8 leads to efficient contextual algorithms for structured action spaces.

- *Finite actions.* In the finite-action setting where $\Pi = [A]$, we have $\sup_{x\in\mathcal{X}} \mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}|_x, \widehat{\mathcal{M}}|_x) \leq \frac{A}{\gamma}$, and this is achieved efficiently through the inverse gap weighting strategy (cf. Proposition 5.2). In this case, Algorithm 8 reduces to the $\mathsf{SquareCB}$ algorithm of Foster and Rakhlin (2020), which achieves
$$\mathbf{Reg}_{\mathsf{DM}} \leq O\big(\sqrt{AT \cdot \mathbf{Est}_{\mathsf{Sq}}}\big)$$
after tuning γ. For finite classes where $|\mathcal{F}_{\mathcal{M}}| < \infty$ and $\mathcal{R} = [0, 1]$, this leads to $\mathbf{Reg}_{\mathsf{DM}} \leq O\big(\sqrt{AT \log|\mathcal{F}_{\mathcal{M}}|}\big)$ for an appropriate choice of estimation oracle (Foster and Rakhlin, 2020).

- *Linear action spaces.* Suppose actions are linearly structured in the sense that $\Pi \subseteq \mathbb{R}^d$, and each $f^M \in \mathcal{F}_{\mathcal{M}}$ factorizes as
$$f^M(x, \pi) = \langle g^M(x), \pi\rangle$$
for some $g^M : \mathcal{X} \to \mathbb{R}^d$. Here, we can efficiently achieve $\sup_{x\in\mathcal{X}} \mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}|_x, \widehat{\mathcal{M}}|_x) \leq \frac{d}{\gamma}$ using the strategy from Proposition 6.1. In this case, Algorithm 8 provides an efficient algorithm with
$$\mathbf{Reg}_{\mathsf{DM}} \leq \widetilde{O}\big(\sqrt{dT \cdot \mathbf{Est}_{\mathsf{Sq}}}\big)$$
after tuning γ. This recovers the result from Foster et al. (2020a).

- *Continuous actions with concave rewards.* If $\Pi \subseteq \mathbb{R}^d$ and the function $\pi \mapsto f^M(x, \pi)$ is concave for all $x \in \mathcal{X}$ and $M \in \mathcal{M}$, Proposition 6.3 implies that $\sup_{x\in\mathcal{X}} \mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}|_x, \widehat{\mathcal{M}}|_x) \leq \widetilde{O}(\frac{\mathrm{poly}(d)}{\gamma})$, so that Algorithm 8 enjoys
$$\mathbf{Reg}_{\mathsf{DM}} \leq \widetilde{O}\big(\sqrt{\mathrm{poly}(d)T \cdot \mathbf{Est}_{\mathsf{Sq}}}\big).$$
This yields the first oracle-efficient algorithm for contextual bandits with continuous actions and concave (resp. convex) rewards, though we emphasize that more research is required to understand when the optimization problem (88) can be solved efficiently for this setting.

- *Action spaces with bounded eluder dimension and star number.* Let $\mathcal{F}_{\mathcal{M}}|_x = \{f^M(x, \cdot) \mid M \in \mathcal{M}\}$, and suppose that for all x, the action space has bounded eluder dimension (cf. Definition 6.1):
$$\mathfrak{e} \equiv \sup_{x\in\mathcal{X}} \sup_{f\in\mathcal{F}_{\mathcal{M}}} \mathfrak{e}(\mathcal{F}_{\mathcal{M}}|_x - f(x, \cdot), 1/T) < \infty.$$
Then, via Theorem 6.1, Algorithm 8 guarantees that
$$\mathbf{Reg}_{\mathsf{DM}} \leq \widetilde{O}\big(\sqrt{\mathfrak{e}T \cdot \mathbf{Est}_{\mathsf{Sq}}}\big)$$
after tuning γ. This provides the first algorithm for contextual bandits with continuous action spaces for which regret scales only with the eluder dimension of the action space. The results of Russo and Van Roy (2013), when applied to this setting, scale with the eluder dimension of the entire context-action space (i.e., the eluder dimension of the function class $(x, \pi) \mapsto f^M(x, \pi)$), which can be arbitrarily large in comparison, even if the action space itself is finite. Moreover, our algorithm is efficient whenever the min-max optimization problem in (88) can be solved efficiently.

Of course, another consequence of Theorem 6.1 is that if the star number $\mathfrak{s} \equiv \sup_{x\in\mathcal{X}} \sup_{f\in\mathcal{F}_{\mathcal{M}}} \mathfrak{s}(\mathcal{F}_{\mathcal{M}}|_x - f(x,\cdot), 1/T)$ is bounded, Algorithm 8 guarantees that

$$\mathbf{Reg}_{\mathsf{DM}} \leq \widetilde{O}\left(\sqrt{\mathfrak{s}^2 T \cdot \mathbf{Est}_{\mathsf{Sq}}}\right).$$

This guarantee is also new, and can be arbitrarily small compared to the guarantee based on the eluder dimension.

Beyond these examples, the appeal of Algorithm 8 is that it allows one to immediately translate any future bounds on the Decision-Estimation Coefficient for structured bandit problems into oracle-efficient algorithms for contextual bandits with structured action spaces.

8.3 Application to Contextual Reinforcement Learning

Contextual reinforcement learning (sometimes referred to as the contextual MDP problem) is setting that generalizes both contextual bandits and reinforcement learning (Abbasi-Yadkori and Neu, 2014; Modi et al., 2018; Dann et al., 2019; Modi and Tewari, 2020). At each time, the learner receives a context $x^{(t)}$, selects a policy $\pi^{(t)}$, then executes the policy in an (unknown) finite-horizon MDP and observes a trajectory. Formally, this setting is simply a special case of the contextual DMSO framework in which $M(x,\cdot)$ is a finite-horizon MDP for all $M \in \mathcal{M}$ and $x \in \mathcal{X}$. Similar to the contextual bandit problem, the underlying MDP changes from round based on the context, so it is essential to—via modeling and function approximation—generalize across similar contexts.

Example: Contextual reinforcement learning with finite states and actions. The most basic and well-studied contextual reinforcement learning problem is the setting where, given the context x, each model $M(x,\cdot)$ is a finite state/action MDP with $\mathcal{S} = [S]$ and $\mathcal{A} = [A]$ (Abbasi-Yadkori and Neu, 2014; Modi et al., 2018; Dann et al., 2019; Modi and Tewari, 2020). Here, we can apply the PC-IGW strategy (Algorithm 4) which, via Proposition 5.6, certifies that

$$\sup_{x\in\mathcal{X}} \mathsf{dec}_{\gamma}(\mathcal{M}|_x, \widehat{\mathcal{M}}|_x) \leq O\left(\frac{H^3 SA}{\gamma}\right),$$

as long as the estimator class $\widehat{\mathcal{M}}$ has the same state-action space. With this choice, the contextual E2D algorithm guarantees that for any estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$,

$$\mathbf{Reg}_{\mathsf{DM}} \leq O\left(\sqrt{H^3 SAT \cdot \mathbf{Est}_{\mathsf{Sq}}}\right),$$

after tuning γ. This result is computationally efficient—beyond updating the estimation oracle, the only computational overhead at each round is to run the PC-IGW algorithm with the estimated model—and constitutes the first universal reduction from contextual reinforcement learning to supervised online learning.

Note that while we require that each MDP has finite states and actions, we place no assumption on the context space $\mathcal{X}$, and no assumption on how each model $x \mapsto M(x,\cdot)$ varies as a function of the context. In particular, one can leverage rich, flexible function approximation (e.g., neural networks or kernels) to learn the mapping from contexts to MDPs by simply choosing an appropriate estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$. In constrast, previous approaches are limited to linear or generalized linear function approximation.

Example: Contextual reinforcement learning with bilinear classes. The advantage of our general approach is that we can easily provide efficient learning guarantees for contextual reinforcement learning beyond the finite state/action setting. For example, suppose that for each x, $\mathcal{M}|_x$ is a bilinear class (Definition 7.1) with $\sup_x d_{\mathrm{bi}}(\mathcal{M}|_x) \leq d$, $\sup_x L_{\mathrm{bi}}(\mathcal{M}|_x) \leq L$, and $\pi_M^{\mathrm{est}} = \pi_M$, and that a proper estimation oracle with $\widehat{\mathcal{M}}|_x \subseteq \mathcal{M}|_x$ is available. In this case, the PC-IGW.Bilinear algorithm (Algorithm 5) guarantees that

$$\sup_{x\in\mathcal{X}} \mathsf{dec}_{\gamma}(\mathcal{M}|_x, \widehat{\mathcal{M}}|_x) \leq O\left(\frac{H^3 L^2 d}{\gamma}\right),$$

which leads to the following regret bound for Algorithm 8:

$$\mathbf{Reg}_{\mathsf{DM}} \leq O\Big(\sqrt{H^3 L^2 d T \cdot \mathbf{Est}_{\mathsf{H}}}\Big).$$

This result can be achieved efficiently (at the cost of an extra d factor) whenever an efficient bilinear planning oracle is available for each $\mathcal{M}|_x$; see Section 7.1.2. Basic examples include linear and low rank MDPs.

9 Related Work

We now highlight some relevant lines of research not already covered by our discussion.

9.1 Statistical Estimation

Our results build on a long line of work on minimax lower bounds for statistical estimation (Le Cam, 1973; Ibragimov and Has'Minskii, 1981; Assouad, 1983; Birgé and Massart, 1995). Here, the seminal work of Donoho and Liu (1987, 1991a,b) shows that for a large class of nonparametric estimation problems, the local (and in some cases, global) minimax rates are characterized by a local *modulus of continuity* with respect to Hellinger distance, defined via:

$$\omega_{\varepsilon}(\mathcal{M}, \overline{M}) := \sup_{M \in \mathcal{M}} \big\{ \big\| f^{M} - f^{\overline{M}} \big\| : D^2_{\mathrm{H}}\big(M, \overline{M}\big) \leq \varepsilon^2 \big\}, \tag{91}$$

for an appropriate norm $\|\cdot\|$. An important special case concerns parametric models, where—under mild regularity conditions—the modulus of continuity (91) asymptotically coincides with the Fisher information (Le Cam and Yang, 2000; Van der Vaart, 2000).

One can view the Decision-Estimation Coefficient (2) as an interactive decision making analogue of the modulus of continuity. To make the connection more apparent, consider a regularized (as opposed to constrained) variant of (91):

$$\omega_{\gamma}(\mathcal{M}, \overline{M}) := \sup_{M \in \mathcal{M}} \big\{ \big\| f^{M} - f^{\overline{M}} \big\| - \gamma \cdot D^2_{\mathrm{H}}\big(M, \overline{M}\big) \big\}. \tag{92}$$

Like the Decision-Estimation Coefficient, the modulus of continuity (92) captures a worst-case tradeoff between risk and information gain relative to a reference model $\overline{M}$, with the scale parameter $\gamma > 0$ controlling the tradeoff. The key difference is that because the classical estimation setting is purely passive, the modulus of continuity (92) does not involve decisions made by the learner, and hence is expressed as a "max" rather than a "min-max".

9.2 Instance-Dependent Complexity for Structured Bandits

An important special case of our general decision making framework is the problem of structured bandits with large action spaces (Section 6). The pioneering work of Graves and Lai (1997) gives a characterization for the optimal asymptotic instance-dependent rates for this problem, as well as a broader class of problems called *controlled Markov chains*. In detail, for a model class $\mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}$, consider the complexity measure

$$c(\mathcal{M}, \overline{M}) := \inf_{w \in \mathbb{R}^{\Pi}_{+}} \left\{ \sum_{\pi \in \Pi} w_{\pi} (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)) \mid \forall M \in \mathcal{C}(\mathcal{M}, \overline{M}) : \sum_{\pi \in \Pi} w_{\pi} D_{\mathrm{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) \geq 1 \right\}, \tag{G\&L}$$

where $\mathcal{C}(\mathcal{M}, \overline{M}) := \big\{ M \in \mathcal{M} : f^{M}(\pi_{\overline{M}}) = f^{\overline{M}}(\pi_{\overline{M}}), \pi_{M} \neq \pi_{\overline{M}} \big\}$ is a set of "confusing" alternatives. Graves and Lai (1997) show that for every problem instance $M^{\star}$, any "uniformly consistent" algorithm must incur regret $(1 - o(1)) \cdot c(\mathcal{M}, M^{\star}) \log(T)$ asymptotically, and that regret $(1 + o(1)) \cdot c(\mathcal{M}, M^{\star}) \log(T)$ is asymptotically achievable. A more recent line of work attempts to achieve this fundamental limit non-asymptotically (Combes et al., 2017; Degenne et al., 2020; Jun and Zhang, 2020), and Ok et al. (2018) extends these results to the reinforcement learning setting under strong ergodicity assumptions.

The Graves-Lai complexity measure (G&L) has evident structural similarities to the Decision-Estimation Coefficient, which can be made especially clear by considering the following regularized variant:

$$c_\gamma(\mathcal{M},\overline{M}) := \inf_{w\in\mathbb{R}_+^\Pi}\sup_{M\in\mathcal{C}(\mathcal{M},\overline{M})}\left\{\sum_{\pi\in\Pi}w_\pi(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi))-\gamma\cdot\left(\sum_{\pi\in\Pi}w_\pi D_{\mathrm{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)-1\right)\right\}. \quad (93)$$

Note that $w\in\mathbb{R}_+^\Pi$ may be interpreted as an unnormalized distribution over decisions. With this perspective, the most important difference between this complexity measure and the Decision-Estimation Coefficient is that (93) only considers regret under the nominal model $\overline{M}$, while the DEC considers regret under a worst-case model selected by nature. We believe this difference is a fundamental consequence of considering minimax regret under finite samples rather than asymptotic instance-dependent regret. In particular, all existing results that provide finite-sample guarantees based on the Graves-Lai complexity measure (Combes et al., 2017; Degenne et al., 2020; Jun and Zhang, 2020) require strong assumptions on the problem structure (e.g., finite actions) in order to control the error incurred by evaluating (G&L) with a plug-in estimator for the true model $M^\star$. This suggests that (G&L) alone may not be sufficient to capture optimal instance-dependent guarantees with finite samples. In contrast, we avoid similar assumptions because the DEC incorporates uncertainty in a stronger fashion.

As an aside, we mention that it is possible to bound the complexity measure (G&L) in terms of the DEC using the method of Lagrange multipliers, but the result is difficult to interpret.

9.3 Posterior Sampling and the Information Ratio

In the structured bandit setting, Russo and Van Roy (2014, 2018) introduce a parameter measure known as the *information ratio* which bounds the Bayesian regret for posterior sampling and a related strategy called information-directed sampling. For a given distribution $\mu\in\Delta(\mathcal{M})$ (typically the posterior distribution at a given round), estimator $\overline{M}$, and action distribution p, the information ratio is given by

$$\frac{\left(\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{M\sim\mu}[f^M(\pi)-f^M(\pi_M)]\right)^2}{\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{M\sim\mu}\left[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\right]}.$$

Note that we consider the original definition from Russo and Van Roy (2018), which uses KL divergence, but the concept of information ratio readily extends to other divergences such as Hellinger distance (a-la Section 4.3).

Russo and Van Roy (2014) show that when the distribution p is chosen via posterior sampling (i.e., sample $M'\sim\mu$ and follow $\pi_{M'}$), the information ratio is bounded by $|\Pi|$ for finite-armed bandits; similar bounds hold for linear bandits. A more general strategy, *information-directed sampling* (Russo and Van Roy, 2018) directly optimizes the information ratio for the given prior and estimator:

$$p_{\mathrm{ids}} = \operatorname*{arg\,min}_{p\in\Delta(\Pi)}\frac{\left(\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{M\sim\mu}[f^M(\pi)-f^M(\pi_M)]\right)^2}{\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{M\sim\mu}\left[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\right]}.$$

To relate these concepts and techniques to our own results, we consider a natural "worst-case" complexity measure for the Bayesian setting based on the information ratio:

$$\mathcal{I}_{\mathrm{B}}(\mathcal{M},\overline{M}) = \max_{\mu\in\Delta(\mathcal{M})}\min_{p\in\Delta(\Pi)}\frac{\left(\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{M\sim\mu}[f^M(\pi)-f^M(\pi_M)]\right)^2}{\mathbb{E}_{\pi\sim p}\,\mathbb{E}_{M\sim\mu}\left[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\right]}. \quad (94)$$

We show that boundedness of this parameter implies boundedness of the KL divergence variant of the Decision-Estimation Coefficient.

Proposition 9.1. *For all $\overline{M}\in\mathcal{M}$ and $\gamma>0$,*

$$\underline{\mathsf{dec}}_\gamma^{\mathrm{KL}}(\mathcal{M},\overline{M}) \le \frac{\mathcal{I}_{\mathrm{B}}(\mathcal{M},\overline{M})}{4\gamma}.$$

Defining $\mathcal{I}_{\mathrm{B}}(\mathcal{M}) = \sup_{\overline{M}\in\mathcal{M}} \mathcal{I}_{\mathrm{B}}(\mathcal{M}, \overline{M})$, the results of Russo and Van Roy (2018) imply that information-directed sampling attains Bayesian regret $O\big(\sqrt{\mathcal{I}_{\mathrm{B}}(\mathrm{co}(\mathcal{M}))\cdot T\log|\Pi|}\big)$. By combining Proposition 9.1 with Theorem 3.6, we recover this result.

Proof of Proposition 9.1. Recall that

$$\underline{\mathsf{dec}}^{\mathrm{KL}}_{\gamma}(\mathcal{M}, \overline{M}) = \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\big].$$

For all $\mu\in\Delta(\mathcal{M})$ and $p\in\Delta(\Pi)$, we can apply the AM-GM inequality to bound

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^{M}(\pi_{M}) - f^{M}(\pi)]\\
&= \mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^{M}(\pi_{M}) - f^{M}(\pi)]\cdot\frac{\big(\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\big]\big)^{1/2}}{\big(\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\big]\big)^{1/2}}\\
&\leq \frac{1}{4\gamma}\cdot\frac{\left(\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^{M}(\pi) - f^{M}(\pi_{M})]\right)^{2}}{\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\big]} + \gamma\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\big].
\end{aligned}$$

Since this holds uniformly for all distributions p, we can choose p to minimize the information ratio. □

The proof of Proposition 9.1 shows that the information ratio and the DEC are closely related, and suggests that the information ratio might be thought of as a parameter-free analogue of the dual Bayesian DEC. However, the following proposition shows that in general, the information ratio can be arbitrarily large compared to DEC.

Proposition 9.2. *Consider the Lipschitz bandit problem in which* $\Pi = [0,1]^{d}$, $\mathcal{F}_{\mathcal{M}} = \big\{f : [0,1]^{d}\to[0,1] \mid |f(x) - f(y)| \leq \|x-y\|_{\infty}\big\}$, $\mathcal{O} = \{\varnothing\}$, *and* $\mathcal{M} = \big\{\pi\mapsto\mathcal{N}(f(\pi),1) \mid f\in\mathcal{F}_{\mathcal{M}}\big\}$. *For this setting, the information ratio is infinite for all* $d\geq 1$*:*

$$\mathcal{I}_{\mathrm{B}}(\mathcal{M}) = +\infty.$$

On the other hand, we have $\mathsf{dec}_{\gamma}(\mathcal{M}) \leq \widetilde{O}(\gamma^{-\frac{1}{d+1}})$, *and consequently Theorem 3.6 recovers the optimal regret bound* $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}(T^{\frac{d+1}{d+2}})$.

This result continues to hold if the KL divergence in $\mathcal{I}_{\mathrm{B}}(\mathcal{M})$ is replaced by squared error, TV distance, or Hellinger distance. The proof of the result indicates that ratios—which inherently suffer from boundedness issues and numerical instability—may not lead to fundamental complexity measures. We note that in spite of this limitation, the information ratio was an important influence on the present work.

Beyond the issues highlighted above, the information ratio is tied to the Bayesian bandit setting, and does not immediately yield algorithms for the frequentist setting that is the focus of this paper. To address this issue, one might consider the following frequentist analogue:

$$\mathcal{I}_{\mathrm{F}}(\mathcal{M}, \overline{M}) := \min_{p\in\Delta(\Pi)}\max_{\mu\in\Delta(\mathcal{M})}\frac{\left(\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^{M}(\pi) - f^{M}(\pi_{M})]\right)^{2}}{\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)\big]}.$$

While it always holds that

$$\mathcal{I}_{\mathrm{B}}(\mathcal{M}, \overline{M}) \leq \mathcal{I}_{\mathrm{F}}(\mathcal{M}, \overline{M}),$$

we show that this inequality is strict in general: even for the multi-armed bandit, the frequentist information ratio $\mathcal{I}_{\mathrm{F}}(\mathcal{M}, \overline{M})$ can be arbitrarily large compared to the Bayesian version.

Proposition 9.3. *Consider the multi-armed bandit setting with* $\Pi = [A]$ *and* $\mathcal{M} = \big\{M(\pi) := \mathcal{N}(f(\pi), 1/2) : f\in[0,1]^{A}\big\}$. *For any* $\overline{M}\in\mathcal{M}$ *for which* $f^{\overline{M}}\in\mathrm{int}([0,1]^{A})$ *and* $\min_{\pi\neq\pi_{\overline{M}}}\{f^{\overline{M}}(\pi) - f^{\overline{M}}(\pi_{\overline{M}})\} > 0$, *we have*

$$\mathcal{I}_{\mathrm{F}}(\mathcal{M}, \overline{M}) = +\infty.$$

On the other hand, this example has $\mathsf{dec}_{\gamma}(\mathcal{M}) \leq \mathcal{I}_{\mathrm{B}}(\mathcal{M}) \leq O(A/\gamma)$ *for all* $\gamma > 0$.

This result should be contrasted with the situation for the DEC, where the "min-max" and "max-min" variants coincide under mild regularity conditions (cf. Section 4.2).

Further Bayesian results. Building on Russo and Van Roy (2014, 2018), subsequent works have generalized their information-theoretic approach to provide minimax regret bounds for various settings, including bandit convex optimization (Bubeck et al., 2015; Bubeck and Eldan, 2016; Lattimore, 2020, 2021) and partial monitoring (Lattimore and Szepesvári, 2019; Kirschner et al., 2020). Notably, Lattimore (2020), building on Bubeck et al. (2015); Bubeck and Eldan (2016), analyzes a variant of information-directed sampling using a parameter that can be interpreted as a variant of the information ratio or the DEC:[27]

$$\widetilde{\mathsf{dec}}_{\gamma}(\mathcal{M},\overline{M}) = \sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\left\{\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{M}(\pi)] - \sqrt{\beta\cdot\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}[(f^{M}(\pi)-f^{\overline{M}}(\pi))^{2}]}\right\},$$

where $\beta > 0$ is a scale parameter. This quantity can always be bounded in terms of information ratio $\mathcal{I}_{\mathrm{B}}(\mathcal{M})$. Likewise, it is straightforward to see that it can be used to bound the dual DEC: Whenever $\widetilde{\mathsf{dec}}_{\gamma}(\mathcal{M},\overline{M}) \leq 0$, we have $\underline{\mathsf{dec}}^{\mathrm{Sq}}_{\gamma}(\mathcal{M},\overline{M}) \lesssim \frac{\beta}{\gamma}$ for all $\gamma > 0$. We denote the quantity by $\widetilde{\mathsf{dec}}_{\gamma}(\mathcal{M},\overline{M})$ to highlight that—similar to the information ratio—it can be viewed as a parameter-free analogue of the (dual) DEC. This can be made apparent by using that $\sqrt{x} = \inf_{\gamma>0}\{\gamma x + \frac{1}{4\gamma}\}$ to re-write the definition as

$$\widetilde{\mathsf{dec}}_{\gamma}(\mathcal{M},\overline{M}) = \sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\sup_{\gamma>0}\left\{\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M})-f^{M}(\pi)-\gamma\cdot(f^{M}(\pi)-f^{\overline{M}}(\pi))^{2}\big]-\frac{\beta}{4\gamma}\right\}. \quad (95)$$

As with the information ratio, the lack of a fixed parameter $\gamma > 0$ means that—unlike the DEC—the minimax theorem cannot be applied to the quantity $\widetilde{\mathsf{dec}}_{\gamma}(\mathcal{M},\overline{M})$, and it can be shown that the frequentist analogue given by

$$\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\sup_{\gamma>0}\left\{\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M})-f^{M}(\pi)-\gamma\cdot(f^{M}(\pi)-f^{\overline{M}}(\pi))^{2}\big]-\frac{\beta}{4\gamma}\right\}.$$

can be arbitrarily large, even for the multi-armed bandit.

We mention in passing that concurrent work of Zhang (2021) provides frequentist regret bounds for a modified posterior sampling strategy using a variant of the quantity in (95).

10 Discussion

We have developed a theory of learnability and unified algorithm design principle for reinforcement learning and interactive decision making. Our results provide a solid foundation on which to build the theory of data-driven decision making going forward, and we are excited about many directions for future research.

- *Computation.* While E2D provides a unified algorithm design principle for decision making, we have mainly focused on statistical rather than computational aspects of the algorithm in this paper, outside of special cases. Going forward, we intend to fully explore when and how the algorithm can be implemented efficiently, which we believe to have strong practical implications.
- *Beyond reinforcement learning.* The examples in this paper focus on bandits and reinforcement learning, but the DMSO framework is substantially more general, and encompasses rich settings such as POMDPs. It remains to fully understand the implications of our results for these settings.

Beyond these questions, we anticipate many natural extensions to our framework and results, including adaptive or instance-dependent guarantees, and incorporating constraints such as safety.

We close by mentioning some technical questions. First, while our upper bounds generally achieve polynomial sample complexity whenever this is achievable, more work is required to understand to what extent the estimation complexity in our bounds can be tightened or removed, and to achieve the sharpest possible guarantees. Second, a natural question is whether our algorithmic results can be extended to support offline oracles for estimation, in the same vein as Simchi-Levi and Xu (2021).

[27] See Eq. (2) of Lattimore (2020). Note that the statement in Lattimore (2020) replaces $\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{M}(\pi)]$ by $\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{\overline{M}}(\pi)]$. These notions are equivalent up to constant factors.

Acknowledgements

We thank Ayush Sekhari and Karthik Sridharan for helpful discussions, and 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 NSF through award DMS-2031883, and from the ARO through award W911NF-21-1-0328.

References

Yasin Abbasi-Yadkori and Gergely Neu. Online learning in MDPs with side information. *arXiv preprint arXiv:1406.6812*, 2014.

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

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 3–11. Morgan Kaufmann Publishers Inc., 1999.

Jacob Abernethy, Alekh Agarwal, Peter Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 1–10. School of Computer Science, McGill University, 2009.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, 2012.

Alekh Agarwal, Dean P Foster, Daniel Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Aleksandrs Slivkins. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2016.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. *Neural Information Processing Systems (NeurIPS)*, 2020a.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020b.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

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

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, June 2001.

Keith Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

Blair Bilodeau, Dylan J Foster, and Daniel Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, 2020.

Blair Bilodeau, Dylan J Foster, and Daniel M Roy. Minimax rates for conditional density estimation via empirical entropy. *arXiv preprint arXiv:2109.10461*, 2021.

Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

V.I. Bogachev. *Measure Theory*, volume 2. Springer, 2007. ISBN 3540345132.

Sébastien Bubeck and Ronen Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589, 2016.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

Sébastien Bubeck, Ofer Dekel, Tomer Koren, and Yuval Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Conference on Learning Theory*, pages 266–278, 2015.

Sébastien Bubeck, Yin Tat Lee, and Ronen 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.

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

Nicolò Cesa-Bianchi and Gábor Lugosi. Minimax regret under log loss for general classes of experts. In *Conference on Computational Learning Theory*, 1999.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

Abraham Charnes and William W Cooper. Programming with linear fractional functionals. *Naval Research logistics quarterly*, 9(3-4):181–186, 1962.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 2011.

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

Thomas M Cover. Universal portfolios. *Mathematical Finance*, 1991.

Varsha Dani, Thomas P Hayes, and Sham Kakade. The price of bandit information for online optimization. 2007.

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

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 207–232, 2011.

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

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

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

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

Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Neural Information Processing Systems (NeurIPS)*, 2021.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Provably efficient reinforcement learning with aggregated states. *arXiv preprint arXiv:1912.06366*, 2019.

David L Donoho and Richard C Liu. Geometrizing rates of convergence. *Annals of Statistics*, 1987.

David L Donoho and Richard C Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, pages 633–667, 1991a.

David L Donoho and Richard C Liu. Geometrizing rates of convergence, III. *The Annals of Statistics*, pages 668–701, 1991b.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019a.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8060–8070, 2019b.

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

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Reinforcement learning in POMDPs without resets. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 690–695, 2005.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan 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.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Neural Information Processing Systems (NeurIPS)*, 2021.

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

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. *Conference on Learning Theory*, 2018.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020a.

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

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

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

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

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

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *The Journal of Machine Learning Research*, 16(1):3487–3602, 2015.

Elad Hazan and Satyen Kale. An online portfolio selection algorithm with regret logarithmic in price variation. *Mathematical Finance*, 2015.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has'Minskii. *Statistical estimation: asymptotic theory*. Springer Science & Business Media, 1981.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.

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

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete POMDPs. *Advances in Neural Information Processing Systems*, 33, 2020a.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020b.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020c.

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

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

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

Adam Kalai and Santosh Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, 2002.

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

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369. PMLR, 2020.

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

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2):245–272, 2010.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

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

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985a.

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

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

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

Tor Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *arXiv preprint arXiv:2006.00475*, 2020.

Tor Lattimore. Minimax regret for bandit convex optimisation of ridge functions. *arXiv preprint arXiv:2106.00444*, 2021.

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

Yin Tat Lee and Aaron Sidford. Solving linear programs with sqrt(rank) linear system solves. *arXiv preprint arXiv:1910.08033*, 2019.

Gene Li, Pritish Kamath, Dylan J Foster, and Nathan Srebro. Eluder dimension and generalized rank. *arXiv preprint arXiv:2104.06970*, 2021.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *EMNLP*, 2016.

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

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

Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, 2018.

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

Nishant A Mehta. Fast rates with high probability in exp-concave statistical learning. *International Conference on Artificial Intelligence and Statistics*, 2017.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

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

Aditya Modi and Ambuj Tewari. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 829–838. PMLR, 2020.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

Wenlong Mou, Zheng Wen, and Xi Chen. On the sample complexity of reinforcement learning with policy space generalization. *arXiv preprint arXiv:2008.07353*, 2020.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8874–8882, 2018.

Manfred Opper and David Haussler. Worst case prediction over sequences under log loss. In *The Mathematics of Information Coding, Extraction and Distribution*, 1999.

Laurent Orseau, Tor Lattimore, and Shane Legg. Soft-bayes: Prod for mixtures of experts with log-loss. In *International Conference on Algorithmic Learning Theory*, 2017.

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

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26:3003–3011, 2013.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. 2014.

Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction, 2012. Available at `http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf`.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems 23*, pages 1984–1992, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

Jorma Rissanen. Complexity of strings in the class of markov sources. *IEEE Transactions on Information Theory*, 32(4):526–532, 1986.

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

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability, and uniform convergence. *Journal of Machine Learning Research (JMLR)*, 2010.

Yurii Mikhailovich Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 1987.

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

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958.

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

Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 2017.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Vladimir N Vapnik. The nature of statistical learning theory. 1995.

Vladimir Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on computational learning theory*, pages 51–60. ACM, 1995.

Vladimir Vovk. Competitive on-line linear regression. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 364–370, Cambridge, MA, USA, 1998. MIT Press.

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

Yuanhao Wang, Ruosong Wang, and Sham M Kakade. An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. *Neural Information Processing Systems (NeurIPS)*, 2021.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

Zheng Wen and Benjamin Van Roy. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

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

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

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *arXiv preprint arXiv:2110.00871*, 2021.

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

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

Yinglun Zhu, Dylan J. Foster, Paul Mineiro, and John Langford. Making contextual bandits with large action spaces practical. *In preparation*, 2021.

A Technical Tools

In this section of the appendix we provide a collection of basic technical results used throughout the paper: Tail bounds for sequences of random variables (Appendix A.1), inequalities for information-theoretic divergences (Appendix A.2), and regret bounds for online learning (Appendix A.3).

A.1 Tail Bounds

Lemma A.1 (Azuma-Hoeffding)**.** *Let $(X_t)_{t\leq T}$ be a sequence of real-valued random variables adapted to a filtration $(\mathscr{F}_t)_{t\leq T}$. If $|X_t| \leq R$ almost surely, then with probability at least $1-\delta$,*

$$\left|\sum_{t=1}^{T} X_t - \mathbb{E}_{t-1}[X_t]\right| \leq R \cdot \sqrt{8T\log(2\delta^{-1})}.$$

Lemma A.2 (Freedman's inequality (e.g., Agarwal et al. (2014)))**.** *Let $(X_t)_{t\leq T}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathscr{F}_t)_{t\leq T}$. If $|X_t| \leq R$ almost surely, then for any $\eta \in (0, 1/R)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \leq \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}\left[X_t^2\right] + \frac{\log(\delta^{-1})}{\eta}.$$

The following result is an immediate consequence of Lemma A.2.

Lemma A.3. *Let $(X_t)_{t\leq T}$ be a sequence of random variables adapted to a filtration $(\mathscr{F}_t)_{t\leq T}$. If $0 \leq X_t \leq R$ almost surely, then with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \leq \frac{3}{2}\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] + 4R\log(2\delta^{-1}),$$

and

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] \leq 2\sum_{t=1}^{T} X_t + 8R\log(2\delta^{-1}).$$

Lemma A.4. *For any sequence of real-valued random variables $(X_t)_{t\leq T}$ adapted to a filtration $(\mathscr{F}_t)_{t\leq T}$, it holds that with probability at least $1-\delta$, for all $T' \leq T$,*

$$\sum_{t=1}^{T'} X_t \leq \sum_{t=1}^{T'} \log\left(\mathbb{E}_{t-1}\left[e^{X_t}\right]\right) + \log(\delta^{-1}). \tag{96}$$

Proof of Lemma A.4. We claim that the sequence

$$Z_\tau := \exp\left(\sum_{t=1}^{\tau} X_t - \log\left(\mathbb{E}_{t-1}\left[e^{X_t}\right]\right)\right)$$

is a nonnegative supermartingale with respect to the filtration $(\mathscr{F}_\tau)_{\tau\leq T}$. Indeed, for any choice of τ, we have

$$\begin{aligned}
\mathbb{E}_{\tau-1}[Z_\tau] &= \mathbb{E}_{\tau-1}\left[\exp\left(\sum_{t=1}^{\tau} X_t - \log\left(\mathbb{E}_{t-1}\left[e^{X_t}\right]\right)\right)\right] \\
&= \exp\left(\sum_{t=1}^{\tau-1} X_t - \log\left(\mathbb{E}_{t-1}\left[e^{X_t}\right]\right)\right) \cdot \mathbb{E}_{\tau-1}\left[\exp\left(X_\tau - \log\left(\mathbb{E}_{\tau-1}\left[e^{X_\tau}\right]\right)\right)\right] \\
&= \exp\left(\sum_{t=1}^{\tau-1} X_t - \log\left(\mathbb{E}_{t-1}\left[e^{X_t}\right]\right)\right) \\
&= Z_\tau.
\end{aligned}$$

Since $Z_0 = 1$, Ville's inequality (e.g., Howard et al. (2020)) implies that for all $\lambda > 0$,

$$\mathbb{P}_0(\exists \tau : Z_\tau > \lambda) \leq \frac{1}{\lambda}.$$

The result now follows by the Chernoff method. $\square$

A.2 Information Theory

We begin with some basic inequalities between divergences.

Lemma A.5 (e.g., Polyanskiy and Wu (2014))**.** *The following inequalities hold:*

- $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q})}$ *and* $D^2_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \leq 2D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$.
- $D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \leq D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q})$.

Lemma A.6 (Divergence for Gaussians distributions)**.** *We have* $D_{\mathrm{KL}}\big(\mathcal{N}(\mu_1, \sigma^2) \,\|\, \mathcal{N}(\mu_2, \sigma^2)\big) = \frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2$ *and* $D^2_{\mathrm{H}}\big(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\big) = 1 - \exp\Big(-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}\Big)$ *for all* $\mu_1, \mu_2 \in \mathbb{R}$.

Lemma A.7 (Divergence for Bernoulli distributions)**.** *For all* $p, q \in [0, 1]$,

$$D^2_{\mathrm{H}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) \leq (p - q)^2 \cdot \left(\frac{1}{p + q} + \frac{1}{1 - p + 1 - q}\right).$$

In particular, for all $\Delta \in (0, 1/2)$*, we have*

$$D^2_{\mathrm{H}}(\mathrm{Ber}(1/2 + \Delta), \mathrm{Ber}(1/2)) \leq 3\Delta^2,$$

and for all $\Delta \in (0, 1/4)$*, we have*

$$D^2_{\mathrm{H}}(\mathrm{Ber}(1/2 + \Delta), \mathrm{Ber}(1/2 + 2\Delta)) \leq 5\Delta^2.$$

Lemma A.8 (Divergence for Rademacher distributions)**.** *For all* $\mu_1, \mu_2 \in [-1, +1]$*, we have*

$$D^2_{\mathrm{H}}(\mathrm{Rad}(\mu_1), \mathrm{Rad}(\mu_2)) \leq \frac{(\mu_1 - \mu_2)^2}{2}\left(\frac{1}{2 + \mu_1 + \mu_2} + \frac{1}{2 - \mu_1 - \mu_2}\right).$$

In particular, for all $\mu_1, \mu_2 \in [-1/2, +1/2]$,

$$D^2_{\mathrm{H}}(\mathrm{Rad}(\mu_1), \mathrm{Rad}(\mu_2)) \leq (\mu_1 - \mu_2)^2,$$

and for all $\mu \in [-1, +1]$,

$$D^2_{\mathrm{H}}(\mathrm{Rad}(\mu), \mathrm{Rad}(0)) \leq \frac{3}{4}\mu^2.$$

Proof of Lemma A.8. This is an immediate consequence of Lemma A.7, since $D^2_{\mathrm{H}}(\mathrm{Rad}(\Delta), \mathrm{Rad}(0)) = D^2_{\mathrm{H}}(\mathrm{Ber}((1+\Delta)/2), \mathrm{Ber}(1/2))$. $\square$

Lemma A.9. *For any pair of random variables* (X, Y),

$$\mathbb{E}_{X \sim \mathbb{P}_X}\big[D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\big)\big] \leq 4D^2_{\mathrm{H}}(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}).$$

Proof of Lemma A.9. Since squared Hellinger distance is an f-divergence, we have

$$\mathbb{E}_{X \sim \mathbb{P}_X}\big[D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\big)\big] = D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X} \otimes \mathbb{P}_X, \mathbb{Q}_{Y|X} \otimes \mathbb{P}_X\big).$$

Next, using that Hellinger distance satisfies the triangle inequality, along with the elementary inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have,

$$\begin{aligned}\mathbb{E}_{X\sim\mathbb{P}_X}\left[D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\big)\right] &\leq 2D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X}\otimes\mathbb{P}_X, \mathbb{Q}_{Y|X}\otimes\mathbb{Q}_X\big) + 2D^2_{\mathrm{H}}\big(\mathbb{Q}_{Y|X}\otimes\mathbb{P}_X, \mathbb{Q}_{Y|X}\otimes\mathbb{Q}_X\big)\\ &= 2D^2_{\mathrm{H}}(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}) + 2D^2_{\mathrm{H}}(\mathbb{P}_X, \mathbb{Q}_X)\\ &\leq 4D^2_{\mathrm{H}}(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}),\end{aligned}$$

where the final line follows from the data processing inequality. $\square$

Lemma A.10 (Lemma 4 of Yang and Barron (1998)[28])**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions over a measurable space $(\mathcal{X}, \mathscr{F})$. If $\sup_{F\in\mathscr{F}} \frac{\mathbb{P}(F)}{\mathbb{Q}(F)} \leq V$, then*

$$D_{\mathrm{KL}}(\mathbb{P}\,\|\,\mathbb{Q}) \leq (2+\log(V))D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}). \tag{97}$$

Next, we state a "multiplicative" variant of Pinsker's inequality, which provides faster rates at the cost of multiplicative rather than additive error.

Lemma A.11 (Multiplicative Pinsker-type inequality for Hellinger distance)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathscr{F})$. For all $h : \mathcal{X} \to \mathbb{R}$ with $0 \leq h(X) \leq R$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have*

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2R(\mathbb{E}_{\mathbb{P}}[h(X)] + \mathbb{E}_{\mathbb{Q}}[h(X)]) \cdot D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q})}. \tag{98}$$

In particular,

$$\mathbb{E}_{\mathbb{P}}[h(X)] \leq 3\,\mathbb{E}_{\mathbb{Q}}[h(X)] + 4RD^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}). \tag{99}$$

Proof of Lemma A.11. Let a measurable event A be fixed. Let $p = \mathbb{P}(A)$ and $q = \mathbb{Q}(A)$. Then we have

$$\frac{(p-q)^2}{2(p+q)} \leq (\sqrt{p}-\sqrt{q})^2 \leq D^2_{\mathrm{H}}((p, 1-p), (q, 1-q)) \leq D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}),$$

where the second inequality is the data-processing inequality. It follows that

$$|p-q| \leq \sqrt{2(p+q)D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q})},$$

To deduce the final result for $R = 1$, we observe that $\mathbb{E}_{\mathbb{P}}[h(X)] = \int_0^1 \mathbb{P}(h(X) > t)dt$ and likewise for $\mathbb{E}_{\mathbb{Q}}[h(X)]$, then apply Jensen's inequality. The result for general R follows by rescaling.

The inequality in (99) follows by applying the AM-GM inequality to (98) and rearranging.

$\square$

Lemma A.12. *Let $\mathbb{P}$ and $\mathbb{Q}$ be a pair of probability measures for a random variable $(\mathcal{Z}, \mathscr{F})$. Suppose we can write $Z = X - Y$, where $X, Y \geq 0$ and $|X - Y| \leq \varepsilon$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$. Then*

$$|\mathbb{E}_{\mathbb{P}}[Z] - \mathbb{E}_{\mathbb{Q}}[Z]| \leq \sqrt{8\varepsilon \cdot (\mathbb{E}_{\mathbb{P}}[X+Y] + \mathbb{E}_{\mathbb{Q}}[X+Y]) \cdot D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q})}. \tag{100}$$

Proof of Lemma A.12. Let $Z_+ = [Z]_+$ and $Z_- = [-Z]_+$, so that $Z = Z_+ - Z_-$. We have

$$|\mathbb{E}_{\mathbb{P}}[Z] - \mathbb{E}_{\mathbb{Q}}[Z]| \leq |\mathbb{E}_{\mathbb{P}}[Z_+] - \mathbb{E}_{\mathbb{Q}}[Z_+]| + |\mathbb{E}_{\mathbb{P}}[Z_-] - \mathbb{E}_{\mathbb{Q}}[Z_-]|.$$

[28] This result is stated in Yang and Barron (1998) in terms of densities, but the variant here is an immediate consequence.

Note that $0 \le Z_+, Z_- \le \varepsilon$ almost surely, so by Lemma A.11, we have

$$|\mathbb{E}_{\mathbb{P}}[Z_+] - \mathbb{E}_{\mathbb{Q}}[Z_+]| \le \sqrt{2\varepsilon(\mathbb{E}_{\mathbb{P}}[Z_+] + \mathbb{E}_{\mathbb{Q}}[Z_+]) \cdot D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q})}$$

and

$$|\mathbb{E}_{\mathbb{P}}[Z_-] - \mathbb{E}_{\mathbb{Q}}[Z_-]| \le \sqrt{2\varepsilon(\mathbb{E}_{\mathbb{P}}[Z_-] + \mathbb{E}_{\mathbb{Q}}[Z_-]) \cdot D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q})}$$

To conclude, we use that $Z_+ \le X + Y$ and $Z_- \le X + Y$ to simplify.

□

Finally, we provide an approximate chain rule-type inequality for the squared Hellinger distance, which allows the distance between two joint distributions to be decomposed into a sum of distances between conditional distributions.

Lemma A.13 (Subadditivity for squared Hellinger distance)**.** *Let $(\mathcal{X}_1, \mathscr{F}_1), \ldots, (\mathcal{X}_n, \mathscr{F}_n)$ be a sequence of measurable spaces, and let $\mathcal{X}^{(i)} = \prod_{i=t}^{i} \mathcal{X}_t$ and $\mathscr{F}^{(i)} = \bigotimes_{t=1}^{i} \mathscr{F}_t$. For each i, let $\mathbb{P}^{(i)}(\cdot \mid \cdot)$ and $\mathbb{Q}^{(i)}(\cdot \mid \cdot)$ be probability kernels from $(\mathcal{X}^{(i-1)}, \mathscr{F}^{(i-1)})$ to $(\mathcal{X}_i, \mathscr{F}_i)$. Let $\mathbb{P}$ and $\mathbb{Q}$ be the laws of $X_1, \ldots, X_n$ under $X_i \sim \mathbb{P}^{(i)}(\cdot \mid X_{1:i-1})$ and $X_i \sim \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1})$ respectively. Then it holds that*

$$D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \le 10^2 \log(n) \cdot \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right]. \tag{101}$$

Furthermore, if there exists a constant $V \ge e$ such that $\sup_{x_{1:i-1} \in \mathcal{X}^{(i-1)}} \sup_{A \in \mathscr{F}_i} \frac{\mathbb{P}^{(i)}(A | x_{1:i-1})}{\mathbb{Q}^{(i)}(A | x_{1:i-1})} \le V$ for all i, then

$$D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \le 3 \log(V) \cdot \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right]. \tag{102}$$

Proof of Lemma A.13. We first prove the result in (101). For a parameter $\lambda \in (0, e^{-1})$, define a probability kernel

$$\mathbb{P}^{(i)}_\lambda(\cdot \mid x_{1:i-1}) = (1-\lambda)\mathbb{P}^{(i)}(\cdot \mid x_{1:i-1}) + \lambda \mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}).$$

Let $\mathbb{P}_\lambda$ be the law of $X_1, \ldots, X_n$ under $X_i \sim \mathbb{P}^{(i)}_\lambda(\cdot \mid X_{1:i-1})$. Since Hellinger distance satisfies the triangle inequality, we have

$$D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \le \underbrace{D_{\mathrm{H}}(\mathbb{P}_\lambda, \mathbb{Q})}_{\mathrm{I}} + \underbrace{D_{\mathrm{H}}(\mathbb{P}, \mathbb{P}_\lambda)}_{\mathrm{II}}.$$

We proceed to bound each term individually.

Term I. We have

$$D^2_{\mathrm{H}}(\mathbb{P}_\lambda, \mathbb{Q}) \le D_{\mathrm{KL}}(\mathbb{Q} \,\|\, \mathbb{P}_\lambda) = \mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^{n} D_{\mathrm{KL}}\big(\mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}) \,\|\, \mathbb{P}^{(i)}_\lambda(\cdot \mid X_{1:i-1})\big)\right],$$

where the right-hand expression follows from the chain rule for KL divergence. We proceed by appealing to Lemma A.10. Since $\mathbb{P}^{(i)}_\lambda = (1-\lambda)\mathbb{P}^{(i)} + \lambda \mathbb{Q}^{(i)}$, we have that for any $x_{1:i-1} \in \mathcal{X}^{(i-1)}$ and measurable event $A \in \mathscr{F}_i$,

$$\frac{\mathbb{Q}^{(i)}(A \mid x_{1:i-1})}{\mathbb{P}^{(i)}_\lambda(A \mid x_{1:i-1})} \le \frac{1}{\lambda}.$$

As a result, Lemma A.10 implies that for all $x_{1:i-1} \in \mathcal{X}^{(i-1)}$,

$$\begin{aligned} D_{\mathrm{KL}}\big(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}) \,\|\, \mathbb{P}^{(i)}_\lambda(\cdot \mid x_{1:i-1})\big) &\le (2 + \log(1/\lambda)) D^2_{\mathrm{H}}\big(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}_\lambda(\cdot \mid x_{1:i-1})\big) \\ &\le 3 \log(1/\lambda) D^2_{\mathrm{H}}\big(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}_\lambda(\cdot \mid x_{1:i-1})\big), \end{aligned}$$

since $\lambda \leq 2/e$. Moreover, by convexity of squared Hellinger distance,

$$\begin{aligned} D^2_{\mathrm{H}}\big(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}_{\lambda}(\cdot \mid x_{1:i-1})\big) &\leq (1-\lambda) D^2_{\mathrm{H}}(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}(\cdot \mid x_{1:i-1})) + \lambda D^2_{\mathrm{H}}(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1})) \\ &\leq D^2_{\mathrm{H}}(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}(\cdot \mid x_{1:i-1})). \end{aligned}$$

We conclude that

$$D^2_{\mathrm{H}}(\mathbb{P}_{\lambda}, \mathbb{Q}) \leq 3\log(1/\lambda) \cdot \mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}))\right].$$

Term II. We begin in a similar fashion as the first term and bound

$$D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{P}_{\lambda}) \leq D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{P}_{\lambda}) = \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D_{\mathrm{KL}}\big(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}) \,\|\, \mathbb{P}^{(i)}_{\lambda}(\cdot \mid X_{1:i-1})\big)\right].$$

For any measurable event $A \in \mathscr{F}_i$ and $x_{1:i-1} \in \mathcal{X}^{(i-1)}$, we have

$$\frac{\mathbb{P}^{(i)}(A \mid x_{1:i-1})}{\mathbb{P}^{(i)}_{\lambda}(A \mid x_{1:i-1})} \leq \frac{1}{1-\lambda},$$

so that Lemma A.10 yields

$$\begin{aligned} D_{\mathrm{KL}}\big(\mathbb{P}^{(i)}(\cdot \mid x_{1:i-1}) \,\|\, \mathbb{P}^{(i)}_{\lambda}(\cdot \mid x_{1:i-1})\big) &\leq (2 + \log(1/(1-\lambda))) D^2_{\mathrm{H}}\big(\mathbb{P}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}_{\lambda}(\cdot \mid x_{1:i-1})\big) \\ &\leq 3 D^2_{\mathrm{H}}\big(\mathbb{P}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}_{\lambda}(\cdot \mid x_{1:i-1})\big) \\ &\leq 3\lambda D^2_{\mathrm{H}}(\mathbb{Q}^{(i)}(\cdot \mid x_{1:i-1}), \mathbb{P}^{(i)}(\cdot \mid x_{1:i-1})), \end{aligned}$$

where the last inequality uses convexity of squared Hellinger distance.

Altogether we have

$$D^2_{\mathrm{H}}(\mathbb{P}_{\lambda}, \mathbb{P}) \leq 3\lambda \cdot \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right].$$

Since the sum inside the expectation on the right-hand side is bounded by $2n$ with probability 1, Lemma A.11 gives the following upper bound:

$$9\lambda \cdot \mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right] + 24\lambda n \cdot D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}).$$

Completing the proof. Combining our bounds on terms I and II and using the elementary fact that $(x+y)^2 \leq 2(x^2+y^2)$, we have

$$D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \leq (6\log(1/\lambda) + 18\lambda)\, \mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right] + 48\lambda n D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}).$$

We set $\lambda = (96n)^{-1}$ (so that $\lambda \leq 1/e$ as required) which, after rearranging, gives

$$\frac{1}{2} D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \leq \left(6\log(96n) + \frac{18}{96n}\right) \mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right].$$

To conclude, we note that $2\big(6\log(96n) + \frac{18}{96n}\big) \leq 10^2 \log(n)$. This establishes (101).

To prove (102), we simply write

$$D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q}) \le D_{\mathrm{KL}}(\mathbb{P}\,\|\,\mathbb{Q}) \le \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^n D_{\mathrm{KL}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1})\,\|\,\mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right]$$
$$\le 3\log(V)\,\mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^n D_{\mathrm{H}}^2(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}),\mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right].$$

where the first inequality is the chain rule and the second inequality uses Lemma A.10 along with the assumption that $V \ge e$.

□

A.3 Online Learning

In this section, we provide results for online learning in a conditional density estimation setting. The setup is as follows. Let $(\mathcal{X},\mathscr{X})$ be the *covariate/context space* and $(\mathcal{Y},\mathscr{Y})$ be the *outcome space*. Let ν be an unnormalized kernel from $(\mathcal{X},\mathscr{X})$ to $(\mathcal{Y},\mathscr{Y})$ (i.e., for every $A \in \mathscr{Y}$, $x \mapsto \nu(A \mid x)$ is $\mathscr{X}$-measurable, and for all $x \in \mathcal{X}$, $\nu(\cdot \mid x)$ is a σ-finite measure; cf. Section 2), and let $\mathcal{K}(\mathcal{X},\mathcal{Y})$ denote the collection of all regular conditional densities with respect to ν. That is, each $g \in \mathcal{K}(\mathcal{X},\mathcal{Y})$ is a jointly measurable map $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ such that for all $x \in \mathcal{X}$

$$y \mapsto g(y \mid x)$$

is a probability density with respect to $\nu(\cdot \mid x)$.

A.3.1 Online Learning, Regret, and Estimation Error

We consider the following online learning process: For $t = 1,\ldots,T$:

- Learner predicts $\widehat{g}^{(t)} \in \mathcal{K}(\mathcal{X},\mathcal{Y})$.
- Nature reveals $x^{(t)} \in \mathcal{X}$ and $y^{(t)} \in \mathcal{Y}$ and learner suffers loss

$$\ell_{\log}^{(t)}(\widehat{g}^{(t)}) := \log\left(\frac{1}{\widehat{g}^{(t)}(y^{(t)} \mid x^{(t)})}\right).$$

Define $\mathcal{H}^{(t)} = (x^{(1)},y^{(1)}),\ldots,(x^{(t)},y^{(t)})$ and let $\mathscr{F}^{(t)} = \sigma(\mathcal{H}^{(t)})$. Let $\mathcal{I}$ be an abstract "index" set. We consider a so-called *expert setting* in which we are given a class of history-dependent functions

$$\mathcal{G} = \{g_i\}_{i\in\mathcal{I}},$$

where each "expert" $g_i = (g_i^{(1)},\ldots,g_i^{(T)})$ is a sequence of functions of the form

$$g_i^{(t)}(\cdot \mid \cdot\,; \mathcal{H}^{(t-1)}) \in \mathcal{K}(\mathcal{X},\mathcal{Y}).$$

Each such function can be thought of as a conditional density that becomes known to the learner after observing the history $\mathcal{H}^{(t-1)}$ (i.e., at the beginning of round t). We abbreviate $g_i^{(t)}(\cdot \mid \cdot) = g_i^{(t)}(\cdot \mid \cdot\,; \mathcal{H}^{(t-1)})$ when the history is clear from context and define $\mathcal{G}^{(t)} = \{g_i^{(t)}\}_{i\in\mathcal{I}}$. We occasionally overload the density $g_i^{(t)}(x) \equiv g_i^{(t)}(\cdot \mid x)$ with its induced probability measure.

We define regret to the expert class $\mathcal{G}$ via

$$\mathbf{Reg}_{\mathsf{KL}} = \sum_{t=1}^T \ell_{\log}^{(t)}(\widehat{g}^{(t)}) - \inf_{i\in\mathcal{I}} \sum_{t=1}^T \ell_{\log}^{(t)}(g_i^{(t)}). \tag{103}$$

For the main results in this section, we make the following realizability assumption.

Assumption A.1. *There exists $g_\star := g_{i^\star} \in \mathcal{G}$ such that for all $t \in [T]$,*

$$y^{(t)} \sim g_\star^{(t)}(\cdot \mid x^{(t)}) \mid x^{(t)}, \mathcal{H}^{(t-1)}.$$

Under Assumption A.1, our main object of interest will be the Hellinger estimation error

$$\mathbf{Est}_{\mathsf{H}} := \sum_{t=1}^{T} \mathbb{E}_{t-1}\left[D_{\mathrm{H}}^2(\widehat{g}^{(t)}(x^{(t)}), g_\star^{(t)}(x^{(t)}))\right], \tag{104}$$

where $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathscr{F}^{(t)}]$. The following lemma shows that minimizing the log-loss regret (103) suffices to minimize the Hellinger estimation error.

Lemma A.14. *For any estimation algorithm, whenever Assumption A.1 holds,*

$$\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq \mathbb{E}[\mathbf{Reg}_{\mathsf{KL}}]. \tag{105}$$

Furthermore, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathbf{Est}_{\mathsf{H}} \leq \mathbf{Reg}_{\mathsf{KL}} + 2\log(\delta^{-1}). \tag{106}$$

Finally, consider a sequence of $\{0,1\}$-valued random variables $(\mathbb{I}_t)_{t\leq T}$, where $\mathbb{I}_t$ is $\mathscr{F}^{(t-1)}$-measurable. For any $\delta \in (0,1)$, we have that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[D_{\mathrm{H}}^2(\widehat{g}^{(t)}(x^{(t)}), g_\star^{(t)}(x^{(t)}))\right]\mathbb{I}_t \leq \sum_{t=1}^{T}\Big(\ell_{\log}^{(t)}(\widehat{g}^{(t)}) - \ell_{\log}^{(t)}(g_\star^{(t)})\Big)\mathbb{I}_t + 2\log(\delta^{-1}). \tag{107}$$

A.3.2 Finite Classes

We begin with a basic guarantee for finite expert classes. We assume $|\mathcal{I}| = |\mathcal{G}|$ without loss of generality.

Lemma A.15 (Cesa-Bianchi and Lugosi (2006))**.** *Consider Vovk's aggregating algorithm, which predicts via*

$$\widehat{g}^{(t)} = \mathbb{E}_{i\sim q^{(t)}}\left[g_i^{(t)}\right], \quad \textit{where} \quad q^{(t)}(i) \propto \exp\left(-\sum_{s=1}^{t-1}\ell_{\log}^{(s)}(g_i^{(s)})\right).$$

This algorithm guarantees that

$$\mathbf{Reg}_{\mathsf{KL}} \leq \log|\mathcal{G}|.$$

Consequently, we have $\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq \log|\mathcal{G}|$ and $\mathbf{Est}_{\mathsf{H}} \leq \log|\mathcal{G}| + 2\log(\delta^{-1})$ with probability at least $1-\delta$.

A.3.3 Infinite Classes

We now provide results for infinite expert classes based on covering numbers.

Definition A.1. *We say that $\mathcal{J} \subseteq \mathcal{I}$ is an ε-cover for $\mathcal{G}$ if*

$$\forall i \in \mathcal{I} \quad \exists j \in \mathcal{J} \quad \textit{s.t.} \quad \sup_{t\leq T}\sup_{\mathcal{H}^{(t-1)}}\sup_{x\in\mathcal{X}} D_{\mathrm{H}}^2\big(g_j^{(t)}(x\,;\mathcal{H}^{(t-1)}), g_i^{(t)}(x\,;\mathcal{H}^{(t-1)})\big) \leq \varepsilon^2.$$

We let $\mathcal{N}(\mathcal{G},\varepsilon)$ denote the size of the smallest such cover and define

$$\mathsf{est}(\mathcal{G},T) := \inf_{\varepsilon\geq 0}\left\{\log\mathcal{N}(\mathcal{G},\varepsilon) + \varepsilon^2 T\right\} \tag{108}$$

as an associated complexity parameter.

We require the following mild regularity assumption.

Assumption A.2. *There exists a constant $B \geq e$ such that:*

1. *$\nu(\mathcal{Y} \mid x) \leq B$ for all $x \in \mathcal{X}$.*
2. *$\sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} g_i^{(t)}(y \mid x) \leq B$.*

Algorithm. We consider the following algorithm. Define $B_x = \nu(\mathcal{Y} \mid x)$ for each $x \in \mathcal{X}$. Let $\alpha \in (0,1)$ be a parameter, and for each i define a smoothed conditional density

$$h_i^{(t)}(y \mid x) = (1-\alpha) g_i^{(t)}(y \mid x) + \alpha B_x^{-1}.$$

One can verify that this is indeed a probability density with respect to ν, since for all x, $\int B_x^{-1} \nu(dy \mid x) = 1$. Let $\mathcal{J} \subseteq \mathcal{I}$ witness the covering number $\mathcal{N}(\mathcal{G}, \varepsilon)$, and let $j^\star \in \mathcal{J}$ be the corresponding cover element for $g_\star$; note that $|\mathcal{J}| \leq \mathcal{N}(\mathcal{G}, \varepsilon)$. We run the aggregating algorithm over $\{h_j\}_{j \in \mathcal{J}}$: we set

$$q^{(t)}(j) \propto \exp\left(-\sum_{s=1}^{t-1} \ell_{\log}^{(s)}(h_j^{(s)}) \right)$$

for each $j \in \mathcal{J}$, then predict with $\widehat{g}^{(t)} = \mathbb{E}_{j \sim q^{(t)}}\left[g_j^{(t)} \right]$.

Lemma A.16. *Suppose Assumption A.2 holds. Then the algorithm above, with an appropriate setting for α and ε, guarantees that*

$$\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq 34 \cdot \inf_{\varepsilon > 0} \left\{ b_T \cdot \varepsilon^2 T + \log \mathcal{N}(\mathcal{G}, \varepsilon) \right\} = O(b_T \cdot \mathsf{est}(\mathcal{G}, T)), \tag{109}$$

where $b_T := \log(2B^2 T)$. Furthermore, the algorithm satisfies $\widehat{g}^{(t)} \in \mathrm{co}(\mathcal{G}^{(t)})$.

The same algorithm guarantees that for all $\delta \in (0, e^{-1})$, with probability at least $1-\delta$,

$$\begin{aligned} \mathbf{Est}_{\mathsf{H}} &\leq 40 \cdot \inf_{\varepsilon > 0} \left\{ b_T \cdot \varepsilon^2 T + \log \mathcal{N}(\mathcal{G}, \varepsilon) \right\} + 424 b_T^2 \log(\delta^{-1}) \\ &\leq O(b_T \cdot \mathsf{est}(\mathcal{G}, T) + b_T^2 \log(\delta^{-1})). \end{aligned} \tag{110}$$

A.3.4 Expert Classes with Time-Varying Availability

We now add an additional twist to the setting in considered in the prequel At each timestep t, we receive a subset $\mathcal{I}^{(t)} \subseteq \mathcal{I}$, which is a measurable function of $\mathcal{H}^{(t-1)}$, and we are required to respect the constraint that $\widehat{g}^{(t)} \in \mathrm{co}\big(\{g_i^{(t)}\}_{i \in \mathcal{I}}\big)$. $\mathcal{I}^{(t)}$ may be thought of as a set of "available" experts. In particular, we use the results in this section to prove Theorem 3.3, where $\mathcal{I}^{(t)}$ corresponds to a subset of models in a confidence set constructed based on the data observed so far. Here, satisfying the constraint that $\widehat{g}^{(t)} \in \mathrm{co}\big(\{g_i^{(t)}\}_{i \in \mathcal{I}}\big)$ allows us to appeal to the regret bound in Eq. (35) of Theorem 4.1.

As in the previous section, we give a result for arbitrary, infinite expert classes based on covering numbers.

Algorithm. We give algorithm that builds on the algorithm for infinite classes in Lemma A.16, but incorporates tricks from the *sleeping experts* literature (Cesa-Bianchi et al., 1997; Kleinberg et al., 2010). With parameters $\varepsilon > 0$, $\alpha \in (0,1)$, we proceed as follows:

- Let $\mathcal{J} \subseteq \mathcal{I}$ witness the covering number $\mathcal{N}(\mathcal{G}, \varepsilon)$. Let $\iota : \mathcal{I} \to \mathcal{J}$ be any function that maps an index $i \in \mathcal{I}$ to a covering element $\iota(i)$ such that $D_{\mathrm{H}}^2\Big(g_i^{(t)}(x; \mathcal{H}^{(t-1)}), g_{\iota(i)}^{(t)}(x; \mathcal{H}^{(t-1)})\Big) \leq \varepsilon^2$ for all x, $\mathcal{H}^{(t-1)}$.
- For each $1 \leq t \leq T$:
  - Let $\mathcal{J}^{(t)} = \{\iota(i)\}_{i \in \mathcal{I}^{(t)}}$.
  - Define a modified class of experts $\check{\mathcal{G}} := \{\check{g}_j^{(t)}\}_{j \in \mathcal{J}}$ as follows: For each $j \in \mathcal{J}$:
    - If $j \notin \mathcal{J}^{(t)}$ or $j \in \mathcal{J}^{(t)} \cap \mathcal{I}^{(t)}$, set $\check{g}_j^{(t)} = g_j^{(t)}$.
    - If $j \in \mathcal{J}^{(t)}$ but $j \notin \mathcal{I}^{(t)}$, take $\check{g}_j^{(t)} = g_k^{(t)}$, where $k \in \mathcal{I}^{(t)}$ is any element such that

    $$D_{\mathrm{H}}^2\big(g_k^{(t)}(x; \mathcal{H}^{(t-1)}), g_j^{(t)}(x; \mathcal{H}^{(t-1)})\big) \leq \varepsilon^2 \quad \forall x. \tag{111}$$

    Such an element is guaranteed to exist, or else we would not have $j \in \mathcal{J}^{(t)}$ to begin with. Furthermore, since $\mathcal{I}^{(t)}$ is a measurable function of $\mathcal{H}^{(t-1)}$, the resulting class $\check{\mathcal{G}} = \big\{\check{g}_j^{(t)}\big\}_{j \in \mathcal{J}}$ is itself a valid time-varying expert class.

- For each $j \in \mathcal{J}$, define $h_j^{(t)}(y \mid x) = (1-\alpha)\check{g}_j^{(t)}(y \mid x) + \alpha B_x^{-1}$.
- Let $q^{(t)}$ denote the distribution produced by running the aggregating algorithm on the expert class $\{\bar{h}_j^{(t)}\}_{j \in \mathcal{J}}$, which we define inductively as follows:
  * $q^{(t)}(j) \propto \exp\Big(-\sum_{i=1}^{t-1} \ell_{\log}^{(t)}(\bar{h}^{(i)})\Big) \;\; \forall j \in \mathcal{J}$.
  * Define $\bar{q}^{(t)}(j) := \frac{q^{(t)}(j)\mathbb{I}\{j \in \mathcal{J}^{(t)}\}}{\sum_{j \in \mathcal{J}^{(t)}} q^{(t)}(j)}$ and $v^{(t)}(y \mid x) := \sum_{j \in \mathcal{J}} \bar{q}^{(t)}(j) h_j^{(t)}(y \mid x)$.
  * Set $\bar{h}_j^{(t)} = h_j^{(t)}$ if $j \in \mathcal{J}^{(t)}$ and $\bar{h}_j^{(t)} = v^{(t)}$ otherwise. Note that $q^{(t)}$ depends only on $\{\bar{h}_j^{(1)}, \ldots, \bar{h}_j^{(t-1)}\}_{j \in \mathcal{J}}$, and hence all of these quantities are well-defined and $\{\bar{h}_j\}_{j \in \mathcal{J}}$ is a valid class of time-varying experts.
- Predict via $\widehat{g}^{(t)} = \mathbb{E}_{j \sim \bar{q}^{(t)}}\big[\check{g}_j^{(t)}\big]$.

Lemma A.17. *Suppose Assumption A.2 holds. Then the algorithm above, with an appropriate setting for α and ε, guarantees that for all $\delta \in (0, e^{-1})$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(\widehat{g}^{(t)}(x^{(t)}), g_\star^{(t)}(x^{(t)}))\big]\mathbb{I}\{i^\star \in \mathcal{I}^{(t)}\} \le 160 \cdot \inf_{\varepsilon > 0}\big\{b_T \cdot \varepsilon^2 T + \log \mathcal{N}(\mathcal{G}, \varepsilon)\big\} + 424 b_T^2 \log(\delta^{-1}), \quad (112)$$

where $b_T := \log(2B^2T)$. Furthermore, the algorithm satisfies $\widehat{g}^{(t)} \in \mathrm{co}\big(\{g_i^{(t)}\}_{i \in \mathcal{I}^{(t)}}\big)$.

For finite classes, we prove a slightly tighter version of this guarantee that does not depend on the parameter B. We consider the following simplified version of the algorithm above.

- For each $1 \le t \le T$:
  - Let $q^{(t)}$ denote the distribution produced by running the aggregating algorithm on the expert class $\{\bar{g}_i^{(t)}\}_{i \in \mathcal{I}}$, which we define inductively as follows:
    * $q^{(t)}(i) \propto \exp\Big(-\sum_{i=1}^{t-1} \ell_{\log}^{(t)}(\bar{g}^{(i)})\Big) \;\; \forall i \in \mathcal{I}$.
    * Define $\bar{q}^{(t)}(i) := \frac{q^{(t)}(i)\mathbb{I}\{i \in \mathcal{I}^{(t)}\}}{\sum_{i \in \mathcal{I}^{(t)}} q^{(t)}(i)}$ and $v^{(t)}(y \mid x) := \sum_{i \in \mathcal{I}} \bar{q}^{(t)}(i) h_i^{(t)}(y \mid x)$.
    * Set $\bar{g}_i^{(t)} = g_i^{(t)}$ if $i \in \mathcal{I}^{(t)}$ and $\bar{g}_i^{(t)} = v^{(t)}$ otherwise. As before, $q^{(t)}$ depends only on $\{\bar{g}_i^{(1)}, \ldots, \bar{g}_i^{(t-1)}\}_{i \in \mathcal{I}}$, and hence all of these quantities are well-defined and $\{\bar{g}_i\}_{i \in \mathcal{I}}$ is a valid class of time-varying experts.
  - Predict via $\widehat{g}^{(t)} = \mathbb{E}_{i \sim \bar{q}^{(t)}}\big[g_i^{(t)}\big]$.

Lemma A.18. *The algorithm above guarantees that for all $\delta \in (0, e^{-1})$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(\widehat{g}^{(t)}(x^{(t)}), g_\star^{(t)}(x^{(t)}))\big]\mathbb{I}\{i^\star \in \mathcal{I}^{(t)}\} \le \log|\mathcal{G}| + 2\log(\delta^{-1}). \quad (113)$$

Furthermore, the algorithm satisfies $\widehat{g}^{(t)} \in \mathrm{co}\big(\{g_i^{(t)}\}_{i \in \mathcal{I}^{(t)}}\big)$.

A.3.5 Proofs

Proof of Lemma A.14. We first-prove the in-expectation result. From (103), we have that

$$\sum_{t=1}^{T} \ell_{\log}^{(t)}(\widehat{g}^{(t)}) - \sum_{t=1}^{T} \ell_{\log}^{(t)}(g_{i^\star}^{(t)}) \le \mathbf{Reg}_{\mathsf{KL}}.$$

Taking expectations, Assumption A.1 implies that

$$\sum_{t=1}^{T} \mathbb{E}\big[D_{\mathrm{KL}}\big(g_{i^\star}^{(t)}(x^{(t)}) \,\|\, \widehat{g}^{(t)}(x^{(t)})\big)\big] \le \mathbb{E}[\mathbf{Reg}_{\mathsf{KL}}].$$

The result now follows from Lemma A.5, which states that $D^2_{\mathrm{H}}\big(g^{(t)}_{i^\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)})\big) \le D_{\mathrm{KL}}\big(g^{(t)}_{i^\star}(x^{(t)}) \,\|\, \widehat{g}^{(t)}(x^{(t)})\big)$.

We now prove the high-probability result in (107); the result in (106) is a special case. Define $X_t = \frac{1}{2}(\ell^{(t)}_{\log}(\widehat{g}^{(t)}) - \ell^{(t)}_{\log}(g^{(t)}_{i^\star}))$ and $Z_t = X_t \cdot \mathbb{I}_t$. Applying Lemma A.4 with the sequence $(-Z_t)_{t\le T}$, we are guaranteed that with probability at least $1-\delta$,

$$\sum_{t=1}^T -\log\big(\mathbb{E}_{t-1}\big[e^{-Z_t}\big]\big) \le \sum_{t=1}^T Z_t + \log(\delta^{-1}) = \frac{1}{2}\sum_{t=1}^T \Big(\ell^{(t)}_{\log}(\widehat{g}^{(t)}) - \ell^{(t)}_{\log}(g^{(t)}_{\star})\Big)\mathbb{I}_t + \log(\delta^{-1})$$

Let t be fixed, and note that since $\mathbb{I}_t$ is $\mathscr{F}^{(t-1)}$-measurable,

$$\mathbb{E}_{t-1}\big[e^{-Z_t}\big] = (1-\mathbb{I}_t) + \mathbb{I}_t\,\mathbb{E}_{t-1}\big[e^{-X_t}\big].$$

Next, we observe that

$$\begin{aligned}
\mathbb{E}_{t-1}\big[e^{-X_t} \mid x^{(t)}\big] &= \mathbb{E}_{t-1}\left[\sqrt{\frac{\widehat{g}^{(t)}(y^{(t)} \mid x^{(t)})}{g^{(t)}_{i^\star}(y^{(t)} \mid x^{(t)})}} \mid x^{(t)}\right] \\
&= \int g^{(t)}_{i^\star}(y \mid x^{(t)})\sqrt{\frac{\widehat{g}^{(t)}(y \mid x^{(t)})}{g^{(t)}_{i^\star}(y \mid x^{(t)})}}\nu(dy \mid x^{(t)}) \\
&= \int \sqrt{g^{(t)}_{i^\star}(y \mid x^{(t)})\widehat{g}^{(t)}(y \mid x^{(t)})}\nu(dy \mid x^{(t)}) = 1 - \frac{1}{2}D^2_{\mathrm{H}}\big(g^{(t)}_{i^\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)})\big).
\end{aligned}$$

Hence,

$$\mathbb{E}_{t-1}\big[e^{-Z_t}\big] = 1 - \frac{1}{2}D^2_{\mathrm{H}}\big(g^{(t)}_{i^\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)})\big)\mathbb{I}_t$$

and, since $-\log(1-x) \ge x$ for $x \in [0,1]$, we conclude that

$$\frac{1}{2}\sum_{t=1}^T \mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}\big(g^{(t)}_{i^\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)})\big)\big]\mathbb{I}_t \le \frac{1}{2}\sum_{t=1}^T \Big(\ell^{(t)}_{\log}(\widehat{g}^{(t)}) - \ell^{(t)}_{\log}(g^{(t)}_{\star})\Big)\mathbb{I}_t + \log(\delta^{-1}).$$

□

Proof of Lemma A.16. Recall that we run the aggregating algorithm over $\{h_j\}_{j\in\mathcal{J}}$ which, per Lemma A.15 guarantees that

$$\sum_{t=1}^T \ell^{(t)}_{\log}(\widehat{h}^{(t)}) - \min_{j\in\mathcal{J}} \sum_{t=1}^T \ell^{(t)}_{\log}(h^{(t)}_j) \le \log|\mathcal{J}|,$$

where $\widehat{h}^{(t)} := \mathbb{E}_{j\sim q^{(t)}}\big[h^{(t)}_j\big]$. In particular, we have

$$\sum_{t=1}^T \ell^{(t)}_{\log}(\widehat{h}^{(t)}) - \sum_{t=1}^T \ell^{(t)}_{\log}(h^{(t)}_{j^\star}) \le \log|\mathcal{J}|.$$

In-expectation bound. Let $X_t = \ell^{(t)}_{\log}(\widehat{h}^{(t)}) - \ell^{(t)}_{\log}(h^{(t)}_{j^\star})$, which has

$$|X_t| \le \log(2B^2/\alpha) =: R$$

by Assumption A.2. Our starting point is to write

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{t-1}[X_t]\right] \le \log|\mathcal{J}|.$$

Observe that

$$\begin{aligned}\mathbb{E}_{t-1}[X_t] &= \mathbb{E}_{t-1}\Big[D_{\mathrm{KL}}\Big(g_\star^{(t)}(x^{(t)}) \,\|\, \widehat{h}^{(t)}(x^{(t)})\Big)\Big] - \mathbb{E}_{t-1}\big[D_{\mathrm{KL}}\big(g_\star^{(t)}(x^{(t)}) \,\|\, h_{j^\star}^{(t)}(x^{(t)})\big)\big] \\ &\geq \mathbb{E}_{t-1}\Big[D^2_{\mathrm{H}}\Big(g_\star^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big] - \mathbb{E}_{t-1}\big[D_{\mathrm{KL}}\big(g_\star^{(t)}(x^{(t)}) \,\|\, h_{j^\star}^{(t)}(x^{(t)})\big)\big],\end{aligned}$$

where the second inequality uses Lemma A.5. Since $g_\star^{(t)}(y \mid x)/h_{j^\star}^{(t)}(y \mid x) \leq B^2/\alpha$, Lemma A.10 gives

$$\begin{aligned}\mathbb{E}_{t-1}\big[D_{\mathrm{KL}}\big(g_\star^{(t)}(x^{(t)}) \,\|\, h_{j^\star}^{(t)}(x^{(t)})\big)\big] &\leq 3R\,\mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}\big(g_\star^{(t)}(x^{(t)}), h_{j^\star}^{(t)}(x^{(t)})\big)\big] \\ &\leq 3R\,\mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}\big(g_\star^{(t)}(x^{(t)}), g_{j^\star}^{(t)}(x^{(t)})\big)\big] + 6R\alpha \\ &\leq 6R(\alpha + \varepsilon^2).\end{aligned}$$

Furthermore, recalling that $\widehat{g}^{(t)} := \mathbb{E}_{j \sim q^{(t)}}\big[g_j^{(t)}\big]$ we have

$$D^2_{\mathrm{H}}(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)})) \leq 2D^2_{\mathrm{H}}\Big(g_\star^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big) + 2D^2_{\mathrm{H}}\Big(\widehat{g}^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big) \leq 2D^2_{\mathrm{H}}\Big(g_\star^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big) + 4\alpha.$$

Combining all of these results, we have that

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big] \leq 2\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] + 16R(\alpha + \varepsilon^2)T. \tag{114}$$

and

$$\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq 2\log|\mathcal{J}| + 16R(\alpha + \varepsilon^2)T.$$

To conclude, we set $\alpha = (T\log(2B^2T))$, so that

$$R\alpha T = \frac{\log(2B^2T\log(2B^2T))}{\log(2B^2T)} \leq 2,$$

which allows us to upper bound

$$\mathbb{E}[\mathbf{Est}_{\mathsf{H}}] \leq 34\log|\mathcal{J}| + 32\log(2B^2T)\varepsilon^2 T.$$

High-probability bound. Applying Lemma A.2 to the sequence $X_t' := \mathbb{E}_{t-1}[X_t] - X_t$, we are guaranteed that for any $\eta \in (0, R^{-1})$, with probability at least $1 - \delta$,

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] &\leq \sum_{t=1}^{T} X_t + \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}\big[(X_t - \mathbb{E}_{t-1}[X_t])^2\big] + \eta^{-1}\log(\delta^{-1}) \\ &\leq \log|\mathcal{J}| + \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}\big[X_t^2\big] + \eta^{-1}\log(\delta^{-1}).\end{aligned} \tag{115}$$

Let t be fixed. Let $\widetilde{\mathbb{E}}_t$ denote the conditional expectation for a modified process in which we draw $y^{(t)} \sim h_{j^\star}^{(t)}(\cdot \mid x^{(t)}; \mathcal{H}^{(t-1)}) \mid x^{(t)}, \mathcal{H}^{(t-1)}$, rather than $y^{(t)} \sim g_\star(\cdot \mid \cdots)$. Using Lemma A.11, we have

$$\mathbb{E}_{t-1}\big[X_t^2\big] \leq 3\widetilde{\mathbb{E}}_{t-1}\big[X_t^2\big] + 4R^2\,\mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}\big(g_\star^{(t)}(x^{(t)}), h_{j^\star}^{(t)}(x^{(t)})\big)\big].$$

Furthermore, by convexity and boundedness of squared Hellinger distance along with the covering property for $\mathcal{J}$, we have

$$\mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}\big(g_\star^{(t)}(x^{(t)}), h_{j^\star}^{(t)}(x^{(t)})\big)\big] \leq 2\alpha + \mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}\big(g_\star^{(t)}(x^{(t)}), g_{j^\star}^{(t)}(x^{(t)})\big)\big] \leq 2\alpha + \varepsilon^2.$$

We now appeal to the following lemma.

Lemma A.19 (Central-to-Bernstein (Mehta, 2017))**.** *Let X be a random variable taking values in $[-R, R]$. If $\mathbb{E}\big[e^{-X}\big] \leq 1$, then $\mathbb{E}\big[X^2\big] \leq 4(1 + R)\,\mathbb{E}[X]$.*

Observe that

$$\widetilde{\mathbb{E}}_{t-1}\big[e^{-X_t} \mid x^{(t)}\big] = \widetilde{\mathbb{E}}_{t-1}\left[\frac{\widehat{h}^{(t)}(y^{(t)} \mid x^{(t)})}{h^{(t)}_{j^\star}(y^{(t)} \mid x^{(t)})} \mid x^{(t)}\right] = \int h_{j^\star}(y \mid x^{(t)})\frac{\widehat{h}^{(t)}(y \mid x^{(t)})}{h^{(t)}_{j^\star}(y \mid x^{(t)})}\nu(dy \mid x^{(t)}) = 1,$$

so Lemma A.19 implies that

$$\widetilde{\mathbb{E}}_{t-1}\big[X_t^2\big] \leq 8R\widetilde{\mathbb{E}}_{t-1}[X_t] = 8R\,\mathbb{E}_{t-1}\Big[D_{\mathrm{KL}}\Big(h^{(t)}_{j^\star}(x^{(t)}) \,\|\, \widehat{h}^{(t)}(x^{(t)})\Big)\Big].$$

Furthermore, since $h^{(t)}_{j^\star}(y \mid x)/\widehat{h}^{(t)}(y \mid x) \leq 2B^2/\alpha$, Lemma A.10 implies that

$$\mathbb{E}_{t-1}\Big[D_{\mathrm{KL}}\Big(h^{(t)}_{j^\star}(x^{(t)}) \,\|\, \widehat{h}^{(t)}(x^{(t)})\Big)\Big] \leq (2+\log(2B^2/\alpha))\,\mathbb{E}_{t-1}\Big[D^2_{\mathrm{H}}\Big(h^{(t)}_{j^\star}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big] \leq 3R\,\mathbb{E}_{t-1}\Big[D^2_{\mathrm{H}}\Big(h^{(t)}_{j^\star}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big].$$

Putting everything together, we have that

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_{t-1}[X_t^2] &\leq 24R^2 \sum_{t=1}^T \mathbb{E}_{t-1}\Big[D^2_{\mathrm{H}}\Big(h^{(t)}_{j^\star}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big] + 4R^2(2\alpha+\varepsilon^2)T \\
&\leq 24R^2 \sum_{t=1}^T \mathbb{E}_{t-1}\Big[D^2_{\mathrm{H}}\Big(g^{(t)}_{\star}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big] + 56R^2(\alpha+\varepsilon^2)T,
\end{aligned}$$

where the second inequality uses convexity of squared Hellinger distance. We set $\eta = 1/96R^2 \leq R^{-1}$ and plug this inequality into (115) to get

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_{t-1}[X_t] &\leq \log|\mathcal{J}| + \frac{1}{4}\sum_{t=1}^T \mathbb{E}_{t-1}\Big[D^2_{\mathrm{H}}\Big(g^{(t)}_{\star}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big] + (\alpha+\varepsilon^2)T + 96R^2\log(\delta^{-1}) \\
&\leq \log|\mathcal{J}| + \frac{1}{4}\sum_{t=1}^T \mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}(g^{(t)}_{\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big] + 2(\alpha+\varepsilon^2)T + 96R^2\log(\delta^{-1}). \qquad (116)
\end{aligned}$$

Combining this with (114), we have

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}(g^{(t)}_{\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big] \\
&\leq 2\sum_{t=1}^T\sum_{t=1}^T \mathbb{E}_{t-1}[X_t] + 16R(\alpha+\varepsilon^2)T \\
&\leq 2\log|\mathcal{J}| + \frac{1}{2}\sum_{t=1}^T \mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}(g^{(t)}_{\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big] + 20R(\alpha+\varepsilon^2)T + 192R^2\log(\delta^{-1}).
\end{aligned}$$

Rearranging gives

$$\sum_{t=1}^T \mathbb{E}_{t-1}\big[D^2_{\mathrm{H}}(g^{(t)}_{\star}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big] \leq 4\log|\mathcal{J}| + 40R(\alpha+\varepsilon^2)T + 384R^2\log(\delta^{-1}).$$

We set $\alpha = 1/T$ and use that $\delta \leq e^{-1}$ to upper bound

$$\mathbf{Est}_{\mathsf{H}} \leq 4\log|\mathcal{J}| + 40R\varepsilon^2 T + 424R^2\log(\delta^{-1}).$$

$\square$

Proof of Lemma A.17. Let $j^\star = \iota(i^\star)$, and note that $i^\star \in \mathcal{I}^{(t)}$ implies that $j^\star \in \mathcal{J}^{(t)}$. We first observe that since $\{\bar{h}_j\}_{j\in\mathcal{J}}$ is a valid class of experts and $q^{(t)}$ is selected by running the aggregating algorithm on this class, Lemma A.15 guarantees that

$$\sum_{t=1}^{T} \ell_{\log}^{(t)}(\mathbb{E}_{j\sim q^{(t)}}[\bar{h}_j^{(t)}]) - \min_{j\in\mathcal{J}} \sum_{t=1}^{T} \ell_{\log}^{(t)}(\bar{h}_j^{(t)}) \leq \log|\mathcal{J}|$$

and in particular,

$$\sum_{t=1}^{T} \ell_{\log}^{(t)}(\mathbb{E}_{j\sim q^{(t)}}[\bar{h}_j^{(t)}]) - \sum_{t=1}^{T} \ell_{\log}^{(t)}(\bar{h}_{j^\star}^{(t)}) \leq \log|\mathcal{J}|. \tag{117}$$

Next, we use the definition of $\bar{h}_j^{(t)}$, which implies that

$$\begin{aligned}
\mathbb{E}_{j\sim q^{(t)}}[\bar{h}_j^{(t)}] &= \sum_{j\in\mathcal{J}^{(t)}} q^{(t)}(j) h_j^{(t)} + \sum_{j\notin\mathcal{J}^{(t)}} q^{(t)}(j) v^{(t)} \\
&= \left(\sum_{j\in\mathcal{J}^{(t)}} q^{(t)}(j)\right) \sum_{j\in\mathcal{J}^{(t)}} \bar{q}^{(t)}(j) h_j^{(t)} + \left(1 - \sum_{j\in\mathcal{J}^{(t)}} q^{(t)}(j)\right) \sum_{j\in\mathcal{J}^{(t)}} \bar{q}^{(t)}(j) h_j^{(t)} \\
&= \mathbb{E}_{j\sim\bar{q}^{(t)}}\big[h_j^{(t)}\big].
\end{aligned}$$

Likewise, we have

$$\bar{h}_{j^\star}^{(t)} = \begin{cases} h_{j^\star}^{(t)}, & j^\star \in \mathcal{J}^{(t)}, \\ \mathbb{E}_{j\sim\bar{q}^{(t)}}\big[h_j^{(t)}\big], & j^\star \notin \mathcal{J}^{(t)}. \end{cases}$$

As a result, (117) implies that

$$\sum_{t=1}^{T} \Big(\ell_{\log}^{(t)}(\mathbb{E}_{j\sim\bar{q}^{(t)}}[h_j^{(t)}]) - \ell_{\log}^{(t)}(h_{j^\star}^{(t)})\Big)\mathbb{I}\{j^\star \in \mathcal{J}^{(t)}\} \leq \log|\mathcal{J}|. \tag{118}$$

From here, the proof closely follows that of Lemma A.16, so we omit details for certain steps. We will use the following fact.

Proposition A.1. *The set $\{\check{g}_j\}_{j\in\mathcal{J}}$ is a $\check{\varepsilon} := 2\varepsilon$-cover for $\mathcal{G}$, and has $\{\check{g}_j^{(t)}\}_{j\in\mathcal{J}^{(t)}} \subseteq \{g_i^{(t)}\}_{i\in\mathcal{I}^{(t)}}$ for all t.*

Proof. Consider a fixed step t and let $i \in \mathcal{I}$ and $j = \iota(i)$. For the first case, if $j \notin \mathcal{J}^{(t)}$ or $j \in \mathcal{J}^{(t)} \cap \mathcal{I}^{(t)}$, we have $\check{g}_j^{(t)} = g_j^{(t)}$, so $\sup_x D^2_{\mathrm{H}}(\check{g}_j^{(t)}(x), g_i^{(t)}(x)) \leq \varepsilon^2$ by the cover property of $\mathcal{J}$. Furthermore, if $j \in \mathcal{J}^{(t)}$, we have $\check{g}_j^{(t)} \in \{g_i^{(t)}\}_{i\in\mathcal{I}^{(t)}}$ by definition. For the second case, if $j \in \mathcal{J}^{(t)} \setminus \mathcal{I}^{(t)}$, we have $\check{g}_j^{(t)} = g_k^{(t)}$ for some $k \in \mathcal{I}^{(t)}$ such that $\sup_x D^2_{\mathrm{H}}(g_j^{(t)}(x), g_k^{(t)}(x)) \leq \varepsilon^2$. In this case it is immediate that $\check{g}_j^{(t)} \in \{g_i^{(t)}\}_{i\in\mathcal{I}^{(t)}}$, and by the triangle inequality we have that for all x,

$$D_{\mathrm{H}}(\check{g}_j^{(t)}(x), g_i^{(t)}(x)) \leq D_{\mathrm{H}}(g_j^{(t)}(x), g_i^{(t)}(x)) + D_{\mathrm{H}}(g_j^{(t)}(x), g_k^{(t)}(x)) \leq 2\varepsilon.$$

□

Let $\mathbb{I}_t := \mathbb{I}\{j^\star \in \mathcal{J}^{(t)}\}$, and let $X_t = \Big(\ell_{\log}^{(t)}(\mathbb{E}_{j\sim\bar{q}^{(t)}}[h_j^{(t)}]) - \ell_{\log}^{(t)}(h_{j^\star}^{(t)})\Big)$ and $Z_t = X_t \cdot \mathbb{I}_t$, which has $|Z_t| \leq |X_t| \leq \log(2B^2/\alpha) := R$. Applying Lemma A.2 to the sequence $Z_t' := \mathbb{E}_{t-1}[Z_t] - Z_t$, we are guaranteed that for any $\eta \in (0, R^{-1})$, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t]\mathbb{I}_t \leq \log|\mathcal{J}| + \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}\big[X_t^2\big]\mathbb{I}_t + \eta^{-1}\log(\delta^{-1}), \tag{119}$$

where we have used that $\mathbb{I}_t$ is $\mathscr{F}^{(t-1)}$-measurable. Defining $\widehat{h}^{(t)} = \mathbb{E}_{j\sim\bar{q}^{(t)}}\big[h_j^{(t)}\big]$, we have

$$\begin{aligned}
\mathbb{E}_{t-1}[X_t] &= \mathbb{E}_{t-1}\Big[D_{\mathrm{KL}}\Big(g_\star^{(t)}(x^{(t)}) \,\|\, \widehat{h}^{(t)}(x^{(t)})\Big)\Big] - \mathbb{E}_{t-1}\big[D_{\mathrm{KL}}\big(g_\star^{(t)}(x^{(t)}) \,\|\, h_{j^\star}^{(t)}(x^{(t)})\big)\big] \\
&\geq \mathbb{E}_{t-1}\Big[D^2_{\mathrm{H}}\Big(g_\star^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big] - \mathbb{E}_{t-1}\big[D_{\mathrm{KL}}\big(g_\star^{(t)}(x^{(t)}) \,\|\, h_{j^\star}^{(t)}(x^{(t)})\big)\big].
\end{aligned}$$

Following Lemma A.16, we compute

$$\begin{aligned}\mathbb{E}_{t-1}\big[D_{\mathrm{KL}}\big(g_\star^{(t)}(x^{(t)}) \,\|\, h_{j^\star}^{(t)}(x^{(t)})\big)\big] &\le 3R\,\mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2\big(g_\star^{(t)}(x^{(t)}), \check{g}_{j^\star}^{(t)}(x^{(t)})\big)\big] + 6R\alpha \\ &\le 6R(\alpha + \check{\varepsilon}^2).\end{aligned}$$

Recalling that $\widehat{g}^{(t)} := \mathbb{E}_{j\sim\bar{q}^{(t)}}\big[\check{g}_j^{(t)}\big]$, we have

$$D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)})) \le 2D_{\mathrm{H}}^2\Big(g_\star^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big) + 2D_{\mathrm{H}}^2\Big(\widehat{g}^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big) \le 2D_{\mathrm{H}}^2\Big(g_\star^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big) + 4\alpha.$$

We conclude that

$$\sum_{t=1}^T \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big]\mathbb{I}_t \le 2\sum_{t=1}^T\sum_{t=1}^T \mathbb{E}_{t-1}[X_t]\mathbb{I}_t + 16R(\alpha + \check{\varepsilon}^2)T. \tag{120}$$

Next, using the same calculation as in Lemma A.16 (with $g^{(t)}$ replaced by $\check{g}^{(t)}$), we have that

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_{t-1}[X_t^2]\mathbb{I}_t &\le 24R^2\sum_{t=1}^T \mathbb{E}_{t-1}\Big[D_{\mathrm{H}}^2\Big(g_\star^{(t)}(x^{(t)}), \widehat{h}^{(t)}(x^{(t)})\Big)\Big]\mathbb{I}_t + 56R^2(\alpha + \check{\varepsilon}^2)T \\ &\le 24R^2\sum_{t=1}^T \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big]\mathbb{I}_t + 104R^2(\alpha + \check{\varepsilon}^2)T.\end{aligned}$$

We set $\eta = 1/96R^2 \le R^{-1}$ and plug this inequality into (119) to get

$$\sum_{t=1}^T \mathbb{E}_{t-1}[X_t]\mathbb{I}_t \le \log|\mathcal{J}| + \frac{1}{4}\sum_{t=1}^T \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big]\mathbb{I}_t + 2(\alpha + \check{\varepsilon}^2)T + 96R^2\log(\delta^{-1}). \tag{121}$$

Combining this with (120), we have

$$\begin{aligned}&\sum_{t=1}^T \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big]\mathbb{I}_t \\ &\le 2\sum_{t=1}^T\sum_{t=1}^T \mathbb{E}_{t-1}[X_t]\mathbb{I}_t + 16R(\alpha + \check{\varepsilon}^2)T \\ &\le 2\log|\mathcal{J}| + \frac{1}{2}\sum_{t=1}^T \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big]\mathbb{I}_t + 20R(\alpha + \check{\varepsilon}^2)T + 192R^2\log(\delta^{-1}),\end{aligned}$$

which, after rearranging, gives

$$\sum_{t=1}^T \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big]\mathbb{I}_t \le 4\log|\mathcal{J}| + 40R(\alpha + \check{\varepsilon}^2)T + 384R^2\log(\delta^{-1}).$$

Since $i^\star \in \mathcal{I}^{(t)}$ implies $j^\star \in \mathcal{J}^{(t)}$ by construction, this implies that

$$\sum_{t=1}^T \mathbb{E}_{t-1}\big[D_{\mathrm{H}}^2(g_\star^{(t)}(x^{(t)}), \widehat{g}^{(t)}(x^{(t)}))\big]\mathbb{I}\{i^\star \in \mathcal{I}^{(t)}\} \le 4\log|\mathcal{J}| + 40R(\alpha + \check{\varepsilon}^2)T + 384R^2\log(\delta^{-1}).$$

We set $\alpha = 1/T$ and use that $\delta \le e^{-1}$ to upper bound

$$\mathbf{Est}_{\mathsf{H}} \le 4\log|\mathcal{J}| + 40R\check{\varepsilon}^2T + 424R^2\log(\delta^{-1}) \le 4\log|\mathcal{J}| + 160R\varepsilon^2T + 424R^2\log(\delta^{-1}).$$

Since $\widehat{g}^{(t)} = \mathbb{E}_{i\sim\bar{q}^{(t)}}\big[\check{g}_j^{(t)}\big]$ and $\bar{q}^{(t)}$ is supported on $\mathcal{J}^{(t)}$, Proposition A.1 implies that $\widehat{g}^{(t)} \in \mathrm{co}(\{g_i^{(t)}\}_{i\in\mathcal{I}^{(t)}})$. $\square$

Proof of Lemma A.18. We first observe that since $\{\bar{g}_j\}_{i\in\mathcal{I}}$ is a valid class of experts and $q^{(t)}$ is selected by running the aggregating algorithm on this class, Lemma A.15 guarantees that

$$\sum_{t=1}^{T}\ell_{\log}^{(t)}(\mathbb{E}_{i\sim q^{(t)}}[\bar{g}_i^{(t)}]) - \sum_{t=1}^{T}\ell_{\log}^{(t)}(\bar{g}_{i^\star}^{(t)}) \leq \log|\mathcal{I}|. \tag{122}$$

Next, we use the definition of $\bar{g}_i^{(t)}$, which implies that

$$\begin{aligned}
\mathbb{E}_{j\sim q^{(t)}}[\bar{g}_i^{(t)}] &= \sum_{i\in\mathcal{I}^{(t)}} q^{(t)}(i)g_i^{(t)} + \sum_{i\notin\mathcal{I}^{(t)}} q^{(t)}(i)v^{(t)} \\
&= \left(\sum_{i\in\mathcal{I}^{(t)}} q^{(t)}(i)\right)\sum_{i\in\mathcal{I}^{(t)}}\bar{q}^{(t)}(i)g_i^{(t)} + \left(1-\sum_{i\in\mathcal{I}^{(t)}} q^{(t)}(i)\right)\sum_{i\in\mathcal{I}^{(t)}}\bar{q}^{(t)}(i)g_i^{(t)} \\
&= \mathbb{E}_{i\sim\bar{q}^{(t)}}\big[g_i^{(t)}\big].
\end{aligned}$$

Likewise, we have

$$\bar{g}_{i^\star}^{(t)} = \begin{cases} g_{i^\star}^{(t)}, & i^\star\in\mathcal{I}^{(t)}, \\ \mathbb{E}_{i\sim\bar{q}^{(t)}}\big[g_i^{(t)}\big], & i^\star\notin\mathcal{I}^{(t)}. \end{cases}$$

As a result, (122) implies that

$$\sum_{t=1}^{T}\Big(\ell_{\log}^{(t)}(\mathbb{E}_{i\sim\bar{q}^{(t)}}[g_i^{(t)}]) - \ell_{\log}^{(t)}(g_{i^\star}^{(t)})\Big)\mathbb{I}\{i^\star\in\mathcal{I}^{(t)}\} \leq \log|\mathcal{I}|. \tag{123}$$

The result now follows from Lemma A.14. $\square$

B Decision-Estimation Coefficient: Structural Results

In this section we provide conditions under which one can lower bound the localized Decision-Estimation Coefficient by the global version.

Lemma B.1 (Local-to-global lemma). *Let $\mathcal{M}$ have $\mathcal{F}_{\mathcal{M}} \subseteq (\Pi\to[0,1])$. Suppose there exists $M_0\in\mathcal{M}$ such that $\mathrm{star}(\mathcal{M},M_0)\subseteq\mathcal{M}$ and f^{M_0} is constant. Then*

$$\sup_{\overline{M}\in\mathcal{M}}\mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}),\overline{M}) \geq \varepsilon\cdot\sup_{\overline{M}\in\mathcal{M}}\mathsf{dec}_\gamma(\mathcal{M},\overline{M}) \tag{124}$$

for all $\varepsilon\in[0,1]$ and $\gamma\geq 0$.

Proof of Lemma B.1. For any pair $M,\overline{M}\in\mathcal{M}$, define $M' = (1-\varepsilon)M_0+\varepsilon M$ and $\overline{M}' = (1-\varepsilon)M_0+\varepsilon\overline{M}$, and note that $M',\overline{M}'\in\mathrm{star}(\mathcal{M},M_0)\subseteq\mathcal{M}$. Furthermore, observe that

$$\begin{aligned}
f^{\overline{M}'}(\pi_{\overline{M}'}) - f^{M'}(\pi_{M'}) &= \varepsilon(f^{\overline{M}}(\pi_{\overline{M}}) - f^{M}(\pi_{M})) + (1-\varepsilon)(f^{M_0}(\pi_{\overline{M}}) - f^{M_0}(\pi_{M})) \\
&= \varepsilon(f^{\overline{M}}(\pi_{\overline{M}}) - f^{M}(\pi_{M})) \geq -\varepsilon,
\end{aligned}$$

since f^{M_0} is constant and $\mathcal{F}_{\mathcal{M}}$ takes values in $[0,1]$. Hence, $M'\in\mathcal{M}_\varepsilon(\overline{M}')$. It follows that for any $\overline{M}\in\mathcal{M}$,

$$\begin{aligned}
\sup_{\overline{M}\in\mathcal{M}}\mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}),\overline{M}) &\geq \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}'),\overline{M}') \\
&\geq \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\big[f^{M'}(\pi_{M'}) - f^{M'}(\pi) - \gamma\cdot D^2_{\mathrm{H}}\big(\overline{M}'(\pi),M'(\pi)\big)\big]. \\
&= \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\big[\varepsilon\cdot(f^{M}(\pi_{M}) - f^{M}(\pi)) - \gamma\cdot D^2_{\mathrm{H}}\big(\overline{M}'(\pi),M'(\pi)\big)\big],
\end{aligned}$$

where we again use that f^{M_0} is constant. Next, using the joint convexity of squared Hellinger distance, we have

$$D^2_{\mathrm{H}}\big(\overline{M}'(\pi), M'(\pi)\big) \le \varepsilon D^2_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big) + (1-\varepsilon) D^2_{\mathrm{H}}(M_0(\pi), M_0(\pi)) = \varepsilon D^2_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big).$$

It follows that

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

Since this holds uniformly for all $\overline{M}\in\mathcal{M}$, the result is established. □

Next, we prove a stronger local-to-global lemma from a *structured Gaussian bandit* setting in which there are no observations (i.e., $\mathcal{O} = \{\varnothing\}$), and rewards are Gaussian in the sense that $M(\pi) = \mathcal{N}(f^M(\pi), \sigma^2)$ for all $M\in\mathcal{M}$. We place no assumption on structure of the mean reward function class itself.

Lemma B.2 (Local-to-global lemma for Gaussian bandits). *For any structured Gaussian bandit class $\mathcal{M}$ with $\mathrm{star}(\mathcal{F}_\mathcal{M}, 0) \subseteq \mathcal{F}_\mathcal{M} \subseteq (\Pi \to [0,1])$ and $\sigma^2 \ge 1/8$, we have*

$$\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}) \ge \varepsilon \cdot \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{4e(\varepsilon\gamma)}(\mathcal{M}, \overline{M}) \tag{125}$$

for all $\varepsilon\in[0,1]$ and $\gamma \ge 0$.

For most of the applications considered in this paper, $\mathsf{dec}_\gamma(\mathcal{M}) \propto \frac{C}{\gamma}$, where C is a problem-dependent parameter such as dimension. In this case, the dependence on ε on the right-hand side of (125) vanishes, and we are left with

$$\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}) \ge \frac{1}{4e} \cdot \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}).$$

Proof of Lemma B.2. For any $M, \overline{M}\in\mathcal{M}$, define $M', \overline{M}'$ via $f^{M'} = \varepsilon f^M$ and $f^{\overline{M}'} = \varepsilon f^{\overline{M}}$, and note that $f^{M'}, f^{\overline{M}'} \in \mathrm{star}(\mathcal{F}_\mathcal{M}, 0) \subseteq \mathcal{F}_\mathcal{M}$. Observe that $f^{\overline{M}'}(\pi_{\overline{M}'}) - f^{M'}(\pi_{M'}) \ge -\varepsilon$ (since $\mathcal{F}_\mathcal{M}$ takes values in $[0,1]$) and hence $M' \in \mathcal{M}_\varepsilon(\overline{M}')$. It follows that for any fixed $\overline{M}\in\mathcal{M}$,

$$\begin{aligned}
\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}) &\ge \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}'), \overline{M}') \\
&\ge \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\big[f^{M'}(\pi_{M'}) - f^{M'}(\pi) - \gamma\cdot D^2_{\mathrm{H}}\big(\overline{M}'(\pi), M'(\pi)\big)\big]. \\
&= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\big[\varepsilon\cdot(f^M(\pi_M) - f^M(\pi)) - \gamma\cdot D^2_{\mathrm{H}}\big(\overline{M}'(\pi), M'(\pi)\big)\big].
\end{aligned}$$

Next, we compute

$$D^2_{\mathrm{H}}\big(\overline{M}'(\pi), M'(\pi)\big) \le D_{\mathrm{KL}}\big(\overline{M}'(\pi) \,\|\, M'(\pi)\big) = \frac{1}{2\sigma^2}(f^{\overline{M}'}(\pi) - f^{M'}(\pi))^2 = \frac{\varepsilon^2}{2\sigma^2}(f^{\overline{M}}(\pi) - f^M(\pi))^2$$

and

$$D^2_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big) = 1 - \exp\left(-\frac{(f^{\overline{M}}(\pi) - f^M(\pi))^2}{8\sigma^2}\right) \ge \frac{1}{8e\sigma^2}(f^{\overline{M}}(\pi) - f^M(\pi))^2,$$

where the inequality uses that $e^{-x} \le 1 - \frac{x}{e}$ for $x \le 1$ along with the assumption that $f^M, f^{\overline{M}} \in [0,1]$ and $\sigma^2 \ge 1/8$. It follows that

$$\begin{aligned}
&\inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\big[\varepsilon\cdot(f^M(\pi_M) - f^M(\pi)) - \gamma\cdot D^2_{\mathrm{H}}\big(\overline{M}'(\pi), M'(\pi)\big)\big] \\
&\ge \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\big[\varepsilon\cdot(f^M(\pi_M) - f^M(\pi)) - 4e\varepsilon^2\gamma\cdot D^2_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big)\big] \\
&= \varepsilon\cdot \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - 4e\varepsilon\gamma\cdot D^2_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big)\big] \\
&\ge \varepsilon\cdot \mathsf{dec}_{4e\varepsilon\gamma}(\mathcal{M}, \overline{M}).
\end{aligned}$$

Since this holds uniformly for all $\overline{M}$, the result is established.

$\square$

C Proofs from Section 3

C.1 Proofs for Lower Bounds

C.1.1 Proof of Theorem 3.1

Theorem 3.1 (Main lower bound)**.** *Consider a model class $\mathcal{M}$ with $\mathcal{F}_{\mathcal{M}} \subseteq (\Pi \to [0,1])$, and let $\delta \in (0,1)$ and $T \in \mathbb{N}$ be fixed. Define $C(T) := 2^{15} \log(2T \wedge V(\mathcal{M}))$ and $\underline{\varepsilon}_\gamma := C(T)^{-1}\frac{\gamma}{T}$. Then for any algorithm, there exists a model in $\mathcal{M}$ for which*

$$\mathbf{Reg}_{\mathsf{DM}} \geq (6C(T))^{-1} \cdot \max_{\gamma > \sqrt{C(T)T}} \min\left\{\left(\mathsf{dec}_{\gamma,\underline{\varepsilon}_\gamma}(\mathcal{M}) - 15\delta\right) \cdot T,\ \gamma\right\} \tag{11}$$

with probability at least $\delta/2$.

Proof of Theorem 3.1. We will prove the following slightly more refined result: For any $\gamma > \sqrt{C(T)T}$ and nominal model $\overline{M} \in \mathcal{M}$ and any algorithm, there exists a model $M \in \mathcal{M}_{\underline{\varepsilon}_\gamma}(\overline{M})$ such that

$$\mathbf{Reg}_{\mathsf{DM}} \geq (6C(T))^{-1} \cdot \min\left\{\left(\mathsf{dec}_\gamma(\mathcal{M}_{\underline{\varepsilon}_\gamma}(\overline{M}), \overline{M}) - 15\delta\right) \cdot T, \gamma\right\} \tag{126}$$

with probability at least $\delta/2$. The result in (11) is an immediate corollary.

Let $T \in \mathbb{N}$ be fixed, and let $\varepsilon = c_1\frac{\gamma}{T}$, where c_1 is a free parameter to be specified later. Let $\gamma \geq \sqrt{T}/c_2$ be given, where c_2 is another free parameter.

Consider a fixed algorithm $p = \{p^{(t)}(\cdot \mid \cdot)\}_{t=1}^T$. Recall that $\mathbb{P}^{M,p}$ is the law of $\mathcal{H}^{(T)}$ when M is the underlying model and p is the algorithm. We abbreviate this to $\mathbb{P}^M$, and let $\mathbb{E}^M[\cdot]$ denote the corresponding expectation.

Let $\overline{M} \in \mathcal{M}$ be fixed. Let $\widehat{p} := \frac{1}{T}\sum_{t=1}^T p^{(t)}(\cdot \mid \mathcal{H}^{(t-1)}) \in \Delta(\Pi)$, which is a random variable, and let $p_{\overline{M}} = \mathbb{E}^{\overline{M}}[\widehat{p}]$. Since $p_{\overline{M}} \in \Delta(\Pi)$, the definition of $\mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M})$ guarantees that

$$\sup_{M \in \mathcal{M}_\varepsilon(\overline{M})} \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \geq \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}).$$

Let $M \in \mathcal{M}_\varepsilon(\overline{M})$ attain the supremum above.[29] Then, rearranging, we have

$$\mathbb{E}_{\pi \sim p_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] \geq \gamma \cdot \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] + \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}). \tag{127}$$

For the remainder of the proof, we abbreviate $\mathsf{dec}_\gamma \equiv \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M})$. We define $\Delta_M = \mathbb{E}_{\pi \sim \widehat{p}}[f^M(\pi_M) - f^M(\pi)]$ and $\Delta_{\overline{M}} = \mathbb{E}_{\pi \sim \widehat{p}}\left[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\right]$, which are random variables. Note that we have $\Delta_M = \frac{1}{T}\mathbf{Reg}_{\mathsf{DM}}$ when $\mathcal{H}^{(T)} \sim \mathbb{P}^M$, and likewise for $\Delta_{\overline{M}}$. With this notation, we can rewrite (127) as

$$\mathbb{E}^{\overline{M}}[\Delta_M] \geq \gamma \cdot \mathbb{E}^{\overline{M}}\left[\mathbb{E}_{\pi \sim \widehat{p}}\left[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]\right] + \mathsf{dec}_\gamma. \tag{128}$$

Let $c_3 > 0$ be a final free parameter. We consider two cases. First, if either $\mathbb{P}^M(\Delta_M > c_3\frac{\gamma}{T}) > \delta$ or $\mathbb{P}^{\overline{M}}(\Delta_{\overline{M}} > c_3\frac{\gamma}{T}) > \delta$, the main result is implied whenever c_3 is sufficiently large, since Δ_M is equal in law to $\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}}$ when M is the underlying problem instance. Hence, for the second case, we define $\mathcal{E}^M = \left\{\Delta_M \leq c_3\frac{\gamma}{T}\right\}$ and assume that $\mathbb{P}^{\overline{M}}[\mathcal{E}^{\overline{M}}], \mathbb{P}^M[\mathcal{E}^M] \geq 1 - \delta$.

The strategy for the remainder of the proof is as follows. Given the assumption that $\mathbb{P}^{\overline{M}}[\mathcal{E}^{\overline{M}}], \mathbb{P}^M[\mathcal{E}^M] \geq 1 - \delta$, we will prove a lower bound on the regret for model M in terms of dec_γ, using the lower bound in (128) as a

[29] If the supremum is not attained, one can consider a limit sequence. We omit the details.

starting point. The main challenge is to relate the quantity $\mathbb{E}^{\overline{M}}[\Delta_M]$ on the left-hand side of (128) to the quantity $\mathbb{E}^{M}[\Delta_M]$, which corresponds to regret when actions are selected under the law induced by M. We address this issue using change of measure arguments, which take advantage of the localization property and the assumption that $\mathbb{P}^{\overline{M}}[\mathcal{E}^{\overline{M}}], \mathbb{P}^{M}[\mathcal{E}^{M}] \geq 1-\delta$.

To begin, we recall the following result from Appendix A.

Lemma A.11 (Multiplicative Pinsker-type inequality for Hellinger distance). *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathscr{F})$. For all $h : \mathcal{X} \to \mathbb{R}$ with $0 \leq h(X) \leq R$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have*

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2R(\mathbb{E}_{\mathbb{P}}[h(X)] + \mathbb{E}_{\mathbb{Q}}[h(X)]) \cdot D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q})}. \tag{98}$$

In particular,

$$\mathbb{E}_{\mathbb{P}}[h(X)] \leq 3\,\mathbb{E}_{\mathbb{Q}}[h(X)] + 4R D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}). \tag{99}$$

We apply (99) with $X = \mathcal{H}^{(T)}$, $h(X) = \Delta_M \mathbb{I}\{\mathcal{E}^{M}\}$, $\mathbb{P} = \mathbb{P}^{\overline{M}}$, and $\mathbb{Q} = \mathbb{P}^{M}$, which gives

$$\mathbb{E}^{\overline{M}}[\Delta_M \mathbb{I}\{\mathcal{E}^{M}\}] \leq 3 \cdot \mathbb{E}^{M}[\Delta_M \mathbb{I}\{\mathcal{E}^{M}\}] + 4c_3 \frac{\gamma}{T} \cdot D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big),$$

where we have used that $0 \leq \Delta_M \mathbb{I}\{\mathcal{E}^{M}\} \leq c_3 \frac{\gamma}{T}$ almost surely. We can further bound by

$$\mathbb{E}^{\overline{M}}[\Delta_M \mathbb{I}\{\mathcal{E}^{M}\}] \leq 3 \cdot \mathbb{E}^{M}[\Delta_M] + 3\delta + 4c_3 \frac{\gamma}{T} \cdot D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) \tag{129}$$

using that $f^{M} \in [0,1]$ and $\mathbb{P}^{M}[\neg\mathcal{E}^{M}] \leq \delta$. Next, we observe that

$$\mathbb{E}^{\overline{M}}[\Delta_M \mathbb{I}\{\mathcal{E}^{M}\}] = \mathbb{E}^{\overline{M}}[\Delta_M] - \mathbb{E}^{\overline{M}}[\Delta_M \mathbb{I}\{\neg\mathcal{E}^{M}\}]. \tag{130}$$

Combining (128), (129), and (130), we have

$$\begin{aligned}
\mathbb{E}^{M}[\Delta_M] &\geq \frac{1}{3}\,\mathbb{E}^{\overline{M}}[\Delta_M] - \frac{1}{3}\,\mathbb{E}^{\overline{M}}[\Delta_M \mathbb{I}\{\neg\mathcal{E}^{M}\}] - \delta - \frac{4c_3}{3}\frac{\gamma}{T} \cdot D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) \\
&\geq \frac{1}{3}\mathsf{dec}_\gamma - \frac{1}{3}\,\mathbb{E}^{\overline{M}}[\Delta_M \mathbb{I}\{\neg\mathcal{E}^{M}\}] - \delta + \frac{\gamma}{3} \cdot \mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big] - \frac{4c_3}{3}\frac{\gamma}{T} \cdot D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big).
\end{aligned}$$

We bound the error term involving $\mathbb{I}\{\neg\mathcal{E}^{M}\}$ using the following technical lemma.

Lemma C.1. *When $\mathbb{P}^{\overline{M}}[\mathcal{E}^{\overline{M}}], \mathbb{P}^{M}[\mathcal{E}^{M}] \geq 1-\delta$, we have*

$$\mathbb{E}^{\overline{M}}[\Delta_M \mathbb{I}\{\neg\mathcal{E}^{M}\}] \leq (7c_1 + 14c_3) \cdot \frac{\gamma}{T} D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) + \sqrt{14 D^2_{\mathrm{H}}(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}})\,\mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big]} + 7\delta.$$

Applying Lemma C.1 and rearranging, we have

$$\begin{aligned}
\mathbb{E}^{M}[\Delta_M] \geq \frac{1}{3}\mathsf{dec}_\gamma - 4\delta + \frac{\gamma}{3} \cdot \mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big] - 3^{-1}(7c_1 + 18c_3)\frac{\gamma}{T} \cdot D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) \\
- \frac{1}{3}\sqrt{14 D^2_{\mathrm{H}}(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}})\,\mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big]}.
\end{aligned}$$

Next, we invoke Lemma A.13, which implies that

$$D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) \leq C_T \sum_{t=1}^{T} \mathbb{E}^{\overline{M}}\big[D^2_{\mathrm{H}}\big(M(\pi^{(t)}), \overline{M}(\pi^{(t)})\big)\big] = C_T T\,\mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big],$$

where $C_T := 2^8(\log(2T) \wedge \log(V^{\overline{M}}(\mathcal{M})))$ and $V^{\overline{M}}(\mathcal{M}) := \sup_{M\in\mathcal{M}} \sup_{\pi\in\Pi} \sup_{A\in\mathscr{R}\otimes\mathscr{O}} \frac{\overline{M}(A|\pi)}{M(A|\pi)} \leq V(\mathcal{M})$ (cf. Theorem 3.1).[30] This gives

$$\begin{aligned}
\mathbb{E}^M[\Delta_M] &\geq \frac{1}{3}\mathsf{dec}_\gamma - 4\delta + 3^{-1}\gamma(1 - 7C_Tc_1 - 18C_Tc_3)\cdot\mathbb{E}^{\overline{M}}\left[\mathbb{E}_{\pi\sim\widehat{p}}\left[D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\right] \\
&\quad - 3^{-1}\sqrt{14C_TT}\,\mathbb{E}^{\overline{M}}\left[\mathbb{E}_{\pi\sim\widehat{p}}\left[D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\right] \\
&\geq \frac{1}{3}\mathsf{dec}_\gamma - 4\delta + 3^{-1}\gamma(1 - 7C_Tc_1 - \sqrt{14C_T}c_2 + 18C_Tc_3)\cdot\mathbb{E}^{\overline{M}}\left[\mathbb{E}_{\pi\sim\widehat{p}}\left[D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\right],
\end{aligned}$$

where the final line uses the assumption that $\gamma \geq \sqrt{T}/c_2$. It follows that if we (conservatively) choose $c_1 = c_3 = (126C_T)^{-1}$ and $c_2 = (126C_T)^{-1/2}$, we have

$$\frac{1}{T}\,\mathbb{E}^M[\mathbf{Reg}_{\mathsf{DM}}] = \mathbb{E}^M[\Delta_M] \geq 3^{-1}(\mathsf{dec}_\gamma - 12\delta).$$

This establishes a lower bound on the expected regret. To conclude, we show that this implies a lower bound in probability.

Lemma C.2. *For any real-valued random variable Z with $\mathbb{E}[Z] \geq \mu$ and $Z \leq R \in \mathbb{R}_+$ almost surely, $\mathbb{P}[Z > \mu/2] \geq \frac{\mu}{2R}$.*

Lemma C.2 implies that $\mathbb{P}^M\left[\mathbf{Reg}_{\mathsf{DM}} > 6^{-1}(\mathsf{dec}_\gamma - 12\delta)T\right] \geq 6^{-1}\frac{(\mathsf{dec}_\gamma - 12\delta)T}{\mathbb{E}^M[\mathbf{Reg}_{\mathsf{DM}}]} \geq 6^{-1}(\mathsf{dec}_\gamma - 12\delta)$, since $\mathbf{Reg}_{\mathsf{DM}} \leq T$. We consider two cases. First, if $\mathsf{dec}_\gamma < 15\delta$, $\mathbb{P}^M\left[\mathbf{Reg}_{\mathsf{DM}} > 6^{-1}(\mathsf{dec}_\gamma - 15\delta)T\right] = 1$ trivially. Otherwise, we have

$$\mathbb{P}^M\left[\mathbf{Reg}_{\mathsf{DM}} > 6^{-1}(\mathsf{dec}_\gamma - 15\delta)T\right] \geq \mathbb{P}^M\left[\mathbf{Reg}_{\mathsf{DM}} > 6^{-1}(\mathsf{dec}_\gamma - 12\delta)T\right] \geq 6^{-1}(15\delta - 12\delta) = \delta/2.$$

□

C.1.2 Proofs for Auxiliary Lemmas (Theorem 3.1)

Proof of Lemma C.1. We first note that by Lemma A.11, we have

$$\mathbb{P}^{\overline{M}}[\neg\mathcal{E}^M] \leq 3\mathbb{P}^M[\neg\mathcal{E}^M] + 4D^2_{\mathrm{H}}\big(\mathbb{P}^M,\mathbb{P}^{\overline{M}}\big) \leq 3\delta + 4D^2_{\mathrm{H}}\big(\mathbb{P}^M,\mathbb{P}^{\overline{M}}\big). \tag{131}$$

We consider two cases. First, if $D^2_{\mathrm{H}}\big(\mathbb{P}^M,\mathbb{P}^{\overline{M}}\big) \leq \delta$, then $\mathbb{P}^{\overline{M}}[\neg\mathcal{E}^M] \leq 7\delta$ by (131). Since $f^M \in [0,1]$, this implies that $\mathbb{E}^{\overline{M}}[\Delta_M\mathbb{I}\{\neg\mathcal{E}^M\}] \leq \mathbb{P}^{\overline{M}}[\neg\mathcal{E}^M] \leq 7\delta$ as desired.

For the second case where $D^2_{\mathrm{H}}\big(\mathbb{P}^M,\mathbb{P}^{\overline{M}}\big) \geq \delta$, (131) implies that

$$\mathbb{P}^{\overline{M}}[\neg\mathcal{E}^M] \leq 7\cdot D^2_{\mathrm{H}}\big(\mathbb{P}^M,\mathbb{P}^{\overline{M}}\big), \tag{132}$$

and we proceed by breaking the error into three terms:

$$\mathbb{E}^{\overline{M}}[\Delta_M\mathbb{I}\{\neg\mathcal{E}^M\}] = \underbrace{\mathbb{E}^{\overline{M}}[\Delta_{\overline{M}}\mathbb{I}\{\neg\mathcal{E}^M\}]}_{\mathrm{Err_I}} + \underbrace{(f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}))\mathbb{P}^{\overline{M}}[\neg\mathcal{E}^M]}_{\mathrm{Err_{II}}} + \underbrace{\mathbb{E}^{\overline{M}}\left[\mathbb{E}_{\pi\sim\widehat{p}}\left[f^{\overline{M}}(\pi) - f^M(\pi)\right]\mathbb{I}\{\neg\mathcal{E}^M\}\right]}_{\mathrm{Err_{III}}}.$$

Term I. We express this quantity as

$$\mathbb{E}^{\overline{M}}[\Delta_{\overline{M}}\mathbb{I}\{\neg\mathcal{E}^M\}] = \mathbb{E}^{\overline{M}}\left[\Delta_{\overline{M}}\mathbb{I}\{\mathcal{E}^{\overline{M}} \wedge \neg\mathcal{E}^M\}\right] + \mathbb{E}^{\overline{M}}\left[\Delta_{\overline{M}}\mathbb{I}\{\neg\mathcal{E}^{\overline{M}} \wedge \neg\mathcal{E}^M\}\right].$$

For the first term above, we have

$$\mathbb{E}^{\overline{M}}\left[\Delta_{\overline{M}}\mathbb{I}\{\mathcal{E}^{\overline{M}} \wedge \neg\mathcal{E}^M\}\right] \leq c_3\frac{\gamma}{T}\mathbb{P}^{\overline{M}}[\neg\mathcal{E}^M] \leq 7c_3\cdot\frac{\gamma}{T}D^2_{\mathrm{H}}\big(\mathbb{P}^M,\mathbb{P}^{\overline{M}}\big),$$

[30] We apply Lemma A.13 with the sequence $X_1,\ldots,X_{2T}$, where $X_t = \pi^{(t)}$, $X_{t+1} = (r^{(t)}, o^{(t)})$, and use that the conditional distribution of $\pi^{(t)}$ is identical under $\mathbb{P}^M$ and $\mathbb{P}^{\overline{M}}$.

where we have used that $\Delta_{\overline{M}} \leq c_3\frac{\gamma}{T}$ under $\mathcal{E}^{\overline{M}}$, along with (132). For the second term above, since $\mathbb{I}\{A \wedge B\} \leq \frac{1}{2}(\mathbb{I}\{A\} + \mathbb{I}\{B\})$, we have

$$\begin{aligned}
\mathbb{E}^{\overline{M}}\big[\Delta_{\overline{M}}\mathbb{I}\{\neg\mathcal{E}^{\overline{M}} \wedge \neg\mathcal{E}^{M}\}\big] &\leq \frac{1}{2}\,\mathbb{E}^{\overline{M}}[\Delta_{\overline{M}}\mathbb{I}\{\neg\mathcal{E}^{M}\}] + \frac{1}{2}\,\mathbb{E}^{\overline{M}}\big[\Delta_{\overline{M}}\mathbb{I}\{\neg\mathcal{E}^{\overline{M}}\}\big] \\
&\leq \frac{1}{2}\,\mathbb{E}^{\overline{M}}[\Delta_{\overline{M}}\mathbb{I}\{\neg\mathcal{E}^{M}\}] + \frac{1}{2}\delta,
\end{aligned}$$

since we have $\mathbb{P}^{\overline{M}}[\neg\mathcal{E}^{\overline{M}}] \leq \delta$ by assumption, and $f^{\overline{M}} \in [0,1]$. After putting both bounds together and rearranging, we have

$$\mathrm{Err}_{\mathrm{I}} \leq 14c_3 \cdot \frac{\gamma}{T} D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) + \delta.$$

Term II. From the definition of $\mathcal{M}_{\varepsilon}(\overline{M})$, we have $f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) \leq \varepsilon = c_1\frac{\gamma}{T}$, so that

$$\mathrm{Err}_{\mathrm{II}} \leq 7c_1 \cdot \frac{\gamma}{T} D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big).$$

Term III. By Cauchy-Schwarz, we have

$$\mathrm{Err}_{\mathrm{III}} = \mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[f^{\overline{M}}(\pi) - f^{M}(\pi)\big]\mathbb{I}\{\neg\mathcal{E}^{M}\}\big] \leq \sqrt{\mathbb{E}^{\overline{M}}\Big[\big(\mathbb{E}_{\pi\sim\widehat{p}}[f^{\overline{M}}(\pi) - f^{M}(\pi)]\big)^2\Big]\mathbb{P}^{\overline{M}}[\neg\mathcal{E}^{M}]}.$$

Furthermore, we have

$$\begin{aligned}
\mathbb{E}^{\overline{M}}\Big[\big(\mathbb{E}_{\pi\sim\widehat{p}}\big[f^{\overline{M}}(\pi) - f^{M}(\pi)\big]\big)^2\Big] &= \mathbb{E}^{\overline{M}}\left[\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim p^{(t)}}\big[f^{\overline{M}}(\pi) - f^{M}(\pi)\big]\right)^2\right] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}^{\overline{M}}\Big[\big(\mathbb{E}_{\pi\sim p^{(t)}}\big[f^{\overline{M}}(\pi) - f^{M}(\pi)\big]\big)^2\Big].
\end{aligned}$$

We can further bound the right-hand side by

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}^{\overline{M}}\Big[\big(f^{\overline{M}}(\pi^{(t)}) - f^{M}(\pi^{(t)})\big)^2\Big] &\leq \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}^{\overline{M}}\big[D^2_{\mathrm{TV}}\big(M(\pi^{(t)}), \overline{M}(\pi^{(t)})\big)\big] \\
&\leq \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}^{\overline{M}}\big[D^2_{\mathrm{H}}\big(M(\pi^{(t)}), \overline{M}(\pi^{(t)})\big)\big] \\
&= 2\,\mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big].
\end{aligned}$$

where the last inequality is from Lemma A.5. Combining this with (132), we have

$$\mathrm{Err}_{\mathrm{III}} \leq \sqrt{14D^2_{\mathrm{H}}(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}})\,\mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big]}.$$

Putting everything together. Combining the bounds on $\mathrm{Err}_{\mathrm{I}}$, $\mathrm{Err}_{\mathrm{II}}$, and $\mathrm{Err}_{\mathrm{III}}$, we have

$$\mathbb{E}^{\overline{M}}[\Delta_{M}\mathbb{I}\{\neg\mathcal{E}^{M}\}] \leq (7c_1 + 14c_3) \cdot \frac{\gamma}{T} D^2_{\mathrm{H}}\big(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}}\big) + \sqrt{14D^2_{\mathrm{H}}(\mathbb{P}^{M}, \mathbb{P}^{\overline{M}})\,\mathbb{E}^{\overline{M}}\big[\mathbb{E}_{\pi\sim\widehat{p}}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big]} + \delta.$$

for the second case. □

Proof of Lemma C.2. By the law of total expectation,

$$\begin{aligned}
\mu \leq \mathbb{E}[Z] &= \mathbb{E}[Z \mid Z > \mu/2]\mathbb{P}[Z > \mu/2] + \mathbb{E}[Z \mid Z \leq \mu/2]\mathbb{P}[Z \leq \mu/2] \\
&\leq R \cdot \mathbb{P}[Z > \mu/2] + \mu/2.
\end{aligned}$$

Rearranging yields the result. □

C.1.3 Proof of Theorem 3.2

Theorem 3.2 (Main lower bound—in-expectation version)**.** *Consider a model class $\mathcal{M}$ with $\mathcal{F}_{\mathcal{M}} \subseteq (\Pi \to [0,1])$. Let $T \in \mathbb{N}$ be fixed and define $C(T) := 2^{11}\log(2T \wedge V(\mathcal{M}))$ and $\varepsilon_\gamma := C(T)^{-1}\frac{\gamma}{T}$. Then for any algorithm, there exists a model in $\mathcal{M}$ for which*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq 6^{-1} \cdot \max_{\gamma>0} \sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}^\infty_{\varepsilon_\gamma}(\overline{M}), \overline{M}) \cdot T. \tag{13}$$

Proof of Theorem 3.2. This proof follows the same structure as Theorem 3.1, with the main difference being that the stronger notion of localization allows for a stronger (as well as simpler) change of measure argument.

Let $T \in \mathbb{N}$ be fixed, and let $\varepsilon = c_1\frac{\gamma}{T}$, where c_1 is a free parameter to be specified later. Consider a fixed algorithm $p = \{p^{(t)}(\cdot \mid \cdot)\}_{t=1}^T$, and let $\mathbb{P}^M(\cdot)$ and $\mathbb{E}^M[\cdot]$ denote the law and expectation when the algorithm is run on model M. Define $p_M = \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^T p^{(t)}(\cdot \mid \mathcal{H}^{(t-1)})\right] \in \Delta(\Pi)$.

Fix $\overline{M} \in \mathcal{M}$. since $p_{\overline{M}} \in \Delta(\Pi)$, the definition of $\mathsf{dec}_\gamma(\mathcal{M}^\infty_\varepsilon(\overline{M}), \overline{M})$ guarantees that

$$\sup_{M\in\mathcal{M}^\infty_\varepsilon(\overline{M})} \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \geq \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}).$$

Let $M \in \mathcal{M}_\varepsilon(\overline{M})$ attain the supremum above, or consider a limit sequence of the supremum is not attained. Then, rearranging, we have

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] \geq \gamma \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] + \mathsf{dec}_\gamma(\mathcal{M}_\varepsilon(\overline{M}), \overline{M}). \tag{133}$$

We abbreviate $\mathsf{dec}_\gamma \equiv \mathsf{dec}_\gamma(\mathcal{M}^\infty_\varepsilon(\overline{M}), \overline{M})$ for the remainder of the proof. Recalling that $g^M(\pi) := f^M(\pi_M) - f^M(\pi)$, it follows from (133) that

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[g^{\overline{M}}(\pi)\right] \\
&= \frac{1}{3}\big(\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[g^{\overline{M}}(\pi)\right]\big) \\
&\quad + \frac{2}{3}\big(\mathbb{E}_{\pi\sim p_M}\left[g^{\overline{M}}(\pi)\right] + \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi)] + (\mathbb{E}_{\pi\sim p_M}[g^M(\pi) - g^{\overline{M}}(\pi)] - \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi) - g^{\overline{M}}(\pi)])\big) \qquad (134) \\
&\geq \frac{1}{3}\big(\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[g^{\overline{M}}(\pi)\right] + \mathbb{E}_{\pi\sim p_M}\left[g^{\overline{M}}(\pi)\right] + \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi)]\big) + \frac{1}{3}\mathsf{dec}_\gamma \\
&\quad + \frac{\gamma}{3}\,\mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] - \frac{2}{3}\big|\mathbb{E}_{\pi\sim p_M}[g^M(\pi) - g^{\overline{M}}(\pi)] - \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi) - g^{\overline{M}}(\pi)]\big|. \qquad (135)
\end{aligned}$$

Recall from the definition of $\mathcal{M}^\infty_\varepsilon(\overline{M})$ that $\left|g^M(\pi) - g^{\overline{M}}(\pi)\right| \leq \varepsilon$ for all π. Hence, by Lemma A.12, we have that

$$\begin{aligned}
&\left|\mathbb{E}_{\pi\sim p_M}[g^M(\pi) - g^{\overline{M}}(\pi)] - \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi) - g^{\overline{M}}(\pi)]\right| \\
&\leq \sqrt{8\varepsilon \cdot \big(\mathbb{E}_{\pi\sim p_M}[g^M(\pi) + g^{\overline{M}}(\pi)] + \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi) + g^{\overline{M}}(\pi)]\big) \cdot D^2_{\mathrm{H}}(\mathbb{P}^M, \mathbb{P}^{\overline{M}})} \\
&\leq 4\varepsilon D^2_{\mathrm{H}}\big(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\big) + \frac{1}{2}\big(\mathbb{E}_{\pi\sim p_M}\left[g^M(\pi) + g^{\overline{M}}(\pi)\right] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[g^M(\pi) + g^{\overline{M}}(\pi)\right]\big). \qquad (136)
\end{aligned}$$

Furthermore, Lemma A.13 implies that

$$D^2_{\mathrm{H}}\big(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\big) \leq C_T \sum_{t=1}^T \mathbb{E}^{\overline{M}}\left[D^2_{\mathrm{H}}\big(M(\pi^{(t)}), \overline{M}(\pi^{(t)})\big)\right] = C_T T \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\right],$$

where $C_T := 2^8(\log(2T) \wedge \log(V^{\overline{M}}(\mathcal{M})))$ and $V^{\overline{M}}(\mathcal{M}) := \sup_{M\in\mathcal{M}}\sup_{\pi\in\Pi}\sup_{A\in\mathscr{R}\otimes\mathscr{O}} \frac{\overline{M}(A|\pi)}{M(A|\pi)} \leq V(\mathcal{M})$. If we set $\varepsilon \leq \frac{\gamma}{8C_T T}$, then after combining this inequality with (135) and (136) and rearranging, we are guaranteed that

$$\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] + \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[g^{\overline{M}}(\pi)\right] \geq \frac{1}{3}\mathsf{dec}_\gamma.$$

Since $\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] = \mathbb{E}^M[\mathbf{Reg}_{\mathsf{DM}}]$, this completes the proof.

$\square$

C.2 Proofs for Upper Bounds

C.2.1 Proof of Theorem 3.3 and Theorem 3.4

We prove Theorem 3.3 and Theorem 3.4 as consequences of Theorem 4.1 (cf. Section 4).

Theorem 3.3 (Main upper bound—finite class version). *Fix $\delta \in (0,1)$. Assume that $\mathcal{R} \subseteq [0,1]$ and Assumption 3.1 holds. Define $C = O(c_\ell^2 \log_{c_\ell}(T))$ and $\bar{\varepsilon}_\gamma := 48\big(\frac{\gamma}{T}\log(|\mathcal{M}|/\delta) + \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M},\overline{M}) + \gamma^{-1}\big)$. Then Algorithm 1, with an appropriate choice for parameters and estimation oracle, ensures that with probability at least $1-\delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \le C \cdot \min_{\gamma>0} \max\left\{ \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}_{\bar{\varepsilon}_\gamma}(\overline{M}),\overline{M}) \cdot T,\ \gamma \cdot \log(|\mathcal{M}|/\delta) \right\}. \tag{22}$$

Theorem 3.4 (Main upper bound—general version). *Let $\delta \in (0,1)$ be given, and let c_ℓ be as in Assumption 3.1. Assume that $\mathcal{R} \subseteq [0,1]$ and that Assumption 3.2 holds. Define $C_1 = O(c_\ell^2 \log_{c_\ell}(T)\log^2(BT))$, $C_2 = O(\log^2(BT))$, and $\bar{\varepsilon}_\gamma = C_2\big(\frac{\gamma}{T}(\mathsf{est}(\mathcal{M},T) + \log(\delta^{-1})) + \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M},\overline{M}) + \gamma^{-1}\big)$. Then Algorithm 1, with an appropriate choice of parameters and estimation oracle, guarantees that with probability at least $1-\delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \le C_1 \cdot \min_{\gamma>0} \max\left\{ \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}_{\bar{\varepsilon}_\gamma}(\overline{M}),\overline{M}) \cdot T,\ \gamma \cdot (\mathsf{est}(\mathcal{M},T) + \log(\delta^{-1})) \right\}. \tag{26}$$

Proof of Theorem 3.3 and Theorem 3.4. The proof proceeds identically for Theorem 3.3 and Theorem 3.4, with the only difference being the choice of the estimation algorithm $\mathbf{Alg}_{\mathsf{Est}}$. For Theorem 3.3, we appeal to an algorithm from Lemma A.18 (which handles finite classes), while for Theorem 3.4 we appeal to an algorithm from Lemma A.17 (which handles infinite classes). In particular, we are guaranteed that with probability at least $1-\delta$, for any realization of the sequence of confidence sets $\{\mathcal{M}^{(t)}\}_{t=1}^T$,

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}_{\pi\sim p^{(t)}}\Big[D^2_{\mathrm{H}}\Big(M^\star(\pi), \widehat{M}^{(t)}(\pi)\Big)\Big]\mathbb{I}\{M^\star \in \mathcal{M}^{(t)}\} \\
&\le R^2 := \begin{cases} \log|\mathcal{M}| + 2\log(\delta^{-1}), & \text{Finite class case (Lemma A.18)}, \\ 2^9\log^2(2B^2T)\cdot\big(\mathsf{est}(\mathcal{M},T) + \log(\delta^{-1})\big), & \text{Infinite class case (Lemma A.17)}. \end{cases}
\end{aligned}$$

This is accomplished by choosing $\mathcal{I}^{(t)} = \mathcal{M}^{(t)}$ in each lemma, and with this choice both estimators ensure that $\widehat{M}^{(t)} \in \mathrm{co}(\mathcal{M}^{(t)})$. Hence, since the conditions of Theorem 4.1a are satisfied, it follows that if we run Algorithm 1 using Option II with the radius R^2 set as above, we are guaranteed that with probability at least $1-\delta$,

$$\mathbf{Reg}_{\mathsf{DM}} \le \sum_{t=1}^T \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}_{\varepsilon_t}(\overline{M}),\overline{M}) + \gamma\cdot R^2,$$

where $\varepsilon_t := 6\frac{\gamma}{t}R^2 + \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\mathsf{dec}_\gamma(\mathcal{M},\overline{M}) + (2\gamma)^{-1}$. To simplify this result, we use to the following lemma.

Lemma C.3. *Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be any non-decreasing function for which there exists $C > 1$ such that $\psi(C\varepsilon) \le C\psi(\varepsilon)$ for all $\varepsilon > 0$. Define $\varepsilon_t = \frac{a}{t} + b$ for $a, b \ge 0$. Then*

$$\sum_{t=1}^T \psi(\varepsilon_t) \le C^2\lceil\log_C(T)\rceil \cdot \psi(4\varepsilon_T)T.$$

Proof of Lemma C.3. Let T_0 be the largest index t for which $b \leq \frac{a}{t}$. We break into two cases. First, for $t > T_0$, we have

$$\varepsilon_t = b + \frac{a}{t} \leq 2b \leq 2\varepsilon_T,$$

so that

$$\sum_{t=T_0+1}^{T} \psi(\varepsilon_t) \leq (T - T_0)\psi(2\varepsilon_T).$$

We now handle the sum to T_0. Define $N = \lceil \log_C(T_0) \rceil$ and $\tau_n = C^{n-1}$ for $n \in [N]$. Let $\tau_{N+1} := T_0$, and note that $\tau_N \leq T_0 = \tau_{N+1}$. We write

$$\sum_{t=1}^{T_0} \psi(\varepsilon_t) = \sum_{n=1}^{N} \sum_{t=\tau_n}^{\tau_{n+1}-1} \psi(\varepsilon_t) \leq \sum_{n=1}^{N} \sum_{t=\tau_n}^{\tau_{n+1}-1} \psi(\varepsilon_{\tau_n}) \leq \sum_{n=1}^{N} \sum_{t=\tau_n}^{\tau_{n+1}-1} \psi(2a/\tau_n),$$

where the least inequality uses that $b \leq a/t$ for $t \leq T_0$. Applying the growth property for ψ recursively, we have

$$\psi(2a/\tau_n) = \psi(2aC^{-(n-1)}) = \psi(C^{N-n+1} \cdot 2aC^{-N}) \leq C^{N-n+1}\psi(2aC^{-N}) \leq C^{N-n+1}\psi(2a/T_0).$$

Furthermore, we have

$$\frac{a}{T_0} \leq 2\frac{a}{T_0 + 1} \leq 2b \leq 2\varepsilon_T.$$

Hence

$$\sum_{t=1}^{T_0} \psi(\varepsilon_t) \leq \psi(4\varepsilon_T) \sum_{n=1}^{N} \sum_{t=\tau_n}^{\tau_{n+1}-1} C^{N-n+1} = \psi(4\varepsilon_T)(C-1) \sum_{n=1}^{N} C^N \leq 2\psi(4\varepsilon_T)T_0 C^2 \lceil \log_C(T_0) \rceil.$$

The result follows by summing the two cases (using that $C > 1$).

□

Applying Lemma C.3 and simplifying for the respective values for R yields both theorems. □

C.2.2 Proof of Theorem 3.6

Algorithm 9 E2D.Bayes for Infinite Classes

1: **parameters**:
  Prior $\mu \in \Delta(\mathcal{M})$.
  Exploration parameter $\gamma > 0$.
  Cover scale $\varepsilon > 0$.
2: Let $\mathrm{P} \subseteq \Pi_{\mathcal{M}}$ be a cover that witnesses $\mathcal{N}(\Pi_{\mathcal{M}}, \varepsilon)$.
3: Let $\iota : \Pi_{\mathcal{M}} \to \mathrm{P}$ be any map that takes $\pi \in \Pi_{\mathcal{M}}$ to a corresponding cover element, and let $\rho^\star := \iota(\pi_{M^\star})$.
4: **for** $t = 1, 2, \cdots, T$ **do**
5:   Define $\overline{M}^{(t)}(\pi) = \mathbb{E}_{t-1}[M^\star(\pi)]$.
6:   Define $\overline{M}^{(t)}_\rho(\pi) = \mathbb{E}_{t-1}[M^\star(\pi) \mid \rho^\star = \rho]$ for each $\rho \in \mathrm{P}$, and set $\mathcal{M}^{(t)} = \{\overline{M}^{(t)}_\rho\}_{\rho \in \mathrm{P}}$.
7:   Let $p^{(t)} = \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}^{(t)}} \mathcal{V}^{\overline{M}^{(t)}}_\gamma(p, M)$. `// Minimizer for` $\mathsf{dec}_\gamma(\mathcal{M}^{(t)}, \overline{M}^{(t)})$`; cf. Eq. (20).`
8:   Sample decision $\pi^{(t)} \sim p^{(t)}$ and update $\mathcal{H}^{(t)} = \mathcal{H}^{(t-1)} \cup \{(\pi^{(t)}, r^{(t)}, o^{(t)})\}$.

Theorem 3.6. *Whenever the conclusion of Proposition 2.1 holds, there exists an algorithm that ensures that*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq 2 \cdot \min_{\gamma > 0} \max\left\{ \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \;\; \inf_{\varepsilon \geq 0} \{\gamma \cdot \log \mathcal{N}(\Pi_{\mathcal{M}}, \varepsilon) + \varepsilon \cdot T\} \right\} \tag{29}$$

$$\leq 2 \cdot \min_{\gamma > 0} \max\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \; (1 + \gamma) \cdot \mathsf{est}(\Pi_{\mathcal{M}}, T)\}. \tag{30}$$

In particular, when $|\Pi_{\mathcal{M}}| < \infty$*, we have*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq 2 \cdot \min_{\gamma>0} \max\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T, \gamma \cdot \log|\Pi_{\mathcal{M}}|\}. \tag{31}$$

Proof of Theorem 3.6. Per the discussion in Section 2.1, it suffices to prove the regret bound for the Bayesian setting in which $M^\star \sim \mu$, where $\mu \in \Delta(\mathcal{M})$ is a known prior. We will show that the regret bound on the right-hand side of (29) holds uniformly for all choices of prior:

$$\mathbb{E}_{M^\star\sim\mu}\,\mathbb{E}^{M^\star}[\mathbf{Reg}_{\mathsf{DM}}] \leq 2 \cdot \min_{\gamma>0} \max\left\{\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T,\ \inf_{\varepsilon\geq 0}\{\gamma \cdot \log \mathcal{N}(\Pi_{\mathcal{M}}, \varepsilon) + \varepsilon \cdot T\}\right\}. \tag{137}$$

This implies that the there exists an algorithm with the same regret bound for the frequentist setting whenever the conclusion of Proposition 2.1 holds. Going forward, we use $\mathbb{E}[\cdot]$ to denote expectation with respect to the joint law over $(M^\star, \mathcal{H}^{(T)})$ when $M^\star \sim \mu$.

We consider the Bayesian variant of $\mathsf{E2D}$ described in Algorithm 9. The algorithm begins by building with a cover P that witnesses the covering number $\mathcal{N}(\Pi_{\mathcal{M}}, \varepsilon)$. Let $\iota : \Pi_{\mathcal{M}} \to \mathrm{P}$ be any fixed map that takes $\pi \in \Pi_{\mathcal{M}}$ to a corresponding ε-covering element in P (in the sense of (27)). We let $\rho^\star := \iota(\pi_{M^\star})$ denote the covering element for the optimal decision $\pi_{M^\star}$, which is a random variable.

At each round t, the algorithm computes $\overline{M}^{(t)}(\pi) := \mathbb{E}_{t-1}[M^\star(\pi)]$, where we recall that $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathscr{F}^{(t)}]$. This may be thought of as the Bayesian analogue of the estimator $\widehat{M}^{(t)}$ used in Algorithm 1. The algorithm then computes a collection of mixture models $\mathcal{M}^{(t)} = \{\overline{M}^{(t)}_\rho\}_{\rho\in\mathrm{P}}$, where

$$\overline{M}^{(t)}_\rho(\pi) := \mathbb{E}_{t-1}[M^\star(\pi) \mid \rho^\star = \rho].$$

Note that we have $\overline{M}^{(t)} \in \mathrm{co}(\mathcal{M})$ and $\mathcal{M}^{(t)} \subseteq \mathrm{co}(\mathcal{M})$. The algorithm proceeds to computes the distribution

$$p^{(t)} = \operatorname*{arg\,min}_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}^{(t)}} \mathcal{V}^{\overline{M}^{(t)}}_\gamma(p, M),$$

which achieves the value $\mathsf{dec}_\gamma(\mathcal{M}^{(t)}, \overline{M}^{(t)})$, and then samples $\pi^{(t)} \sim p^{(t)}$ and proceeds to the next round.

We begin by expressing the expected regret as

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

We first prove an elementary bound on each conditional expectation term above. Note that

$$f^{\overline{M}^{(t)}}(\pi) = \mathbb{E}_{t-1}[\mathbb{E}_{t-1}[r(\pi) \mid M^\star]] = \mathbb{E}_{t-1}\left[f^{M^\star}(\pi)\right].$$

Hence, since $f^{M^\star}$ is conditionally independent of $\pi^{(t)}$ given $\mathscr{F}^{(t-1)}$, we have

$$\mathbb{E}_{t-1}\left[f^{M^\star}(\pi^{(t)})\right] = \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}[f^{\overline{M}^{(t)}}(\pi^{(t)})].$$

Next, we write

$$\mathbb{E}_{t-1}\left[f^{M^\star}(\pi_{M^\star})\right] = \mathbb{E}_{t-1}\left[\mathbb{E}_{t-1}\left[f^{M^\star}(\pi_{M^\star}) \mid \rho^\star\right]\right] = \sum_{\rho\in\mathrm{P}} \mathbb{P}_{t-1}(\rho^\star = \rho)\,\mathbb{E}_{t-1}\left[f^{M^\star}(\pi_{M^\star}) \mid \rho^\star = \rho\right].$$

We bound

$$\begin{aligned}\mathbb{E}_{t-1}\left[f^{M^\star}(\pi_{M^\star}) \mid \rho^\star = \rho\right] &= \mathbb{E}_{t-1}\left[f^{M^\star}(\rho) \mid \rho^\star = \rho\right] + \mathbb{E}_{t-1}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\rho) \mid \rho^\star = \rho\right] \\ &\leq \mathbb{E}_{t-1}\left[f^{M^\star}(\rho) \mid \rho^\star = \rho\right] + \varepsilon,\end{aligned}$$

using that $\rho^\star$ is the corresponding ε-cover element for $\pi_{M^\star}$. We simplify the expectation as

$$\begin{aligned}
\mathbb{E}_{t-1}\left[f^{M^\star}(\rho) \mid \rho^\star = \rho\right] &= \mathbb{E}_{t-1}\left[\mathbb{E}_{t-1}\left[f^{M^\star}(\rho) \mid M^\star, \rho^\star = \rho\right] \mid \rho^\star = \rho\right] \\
&= \mathbb{E}_{t-1}\left[\mathbb{E}_{t-1}\left[f^{M^\star}(\rho) \mid M^\star\right] \mid \rho^\star = \rho\right] \\
&= \mathbb{E}_{t-1}[\mathbb{E}_{t-1}[r(\rho) \mid M^\star] \mid \rho^\star = \rho] \\
&= f^{\overline{M}_\rho^{(t)}}(\rho).
\end{aligned}$$

Finally, noting that

$$\sum_{\rho \in \mathrm{P}} \mathbb{P}_{t-1}(\rho^\star = \rho) f^{\overline{M}_\rho^{(t)}}(\rho) = \mathbb{E}_{t-1}\left[f^{\overline{M}_{\rho^\star}^{(t)}}(\rho^\star)\right],$$

we have the upper bound.

$$\mathbb{E}_{t-1}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] \leq \mathbb{E}_{t-1}\left[\mathbb{E}_{\pi^{(t)} \sim p^{(t)}}\left[f^{\overline{M}_{\rho^\star}^{(t)}}(\rho^\star) - f^{\overline{M}^{(t)}}(\pi^{(t)})\right]\right] + \varepsilon. \tag{138}$$

For the next step, it follows from the definition of $p^{(t)}$ in Line 7 that for every possible realization of $\rho^\star$,

$$\begin{aligned}
\mathbb{E}_{\pi^{(t)} \sim p^{(t)}}\left[f^{\overline{M}_{\rho^\star}^{(t)}}(\rho^\star) - f^{\overline{M}^{(t)}}(\pi^{(t)})\right] &\leq \mathsf{dec}_\gamma(\mathcal{M}^{(t)}, \overline{M}^{(t)}) + \gamma \cdot \mathbb{E}_{\pi^{(t)} \sim p^{(t)}}\left[D^2_{\mathrm{H}}\left(\overline{M}_{\rho^\star}^{(t)}(\pi^{(t)}), \overline{M}^{(t)}(\pi^{(t)})\right)\right] \\
&\leq \mathsf{dec}_\gamma(\mathcal{M}^{(t)}, \overline{M}^{(t)}) + \gamma \cdot \mathbb{E}_{\pi^{(t)} \sim p^{(t)}}\left[D_{\mathrm{KL}}\left(\overline{M}_{\rho^\star}^{(t)}(\pi^{(t)}) \,\|\, \overline{M}^{(t)}(\pi^{(t)})\right)\right],
\end{aligned}$$

where the second inequality is Lemma A.5. We conclude that

$$\begin{aligned}
\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] &\leq \gamma \cdot \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{t-1}\,\mathbb{E}_{\pi^{(t)} \sim p^{(t)}}\left[D_{\mathrm{KL}}\left(\overline{M}_{\rho^\star}^{(t)}(\pi^{(t)}) \,\|\, \overline{M}^{(t)}(\pi^{(t)})\right)\right]\right] + \mathbb{E}\left[\sum_{t=1}^T \mathsf{dec}_\gamma(\mathcal{M}^{(t)}, \overline{M}^{(t)})\right] + \varepsilon T \\
&\leq \gamma \cdot \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{t-1}\left[D_{\mathrm{KL}}\left(\overline{M}_{\rho^\star}^{(t)}(\pi^{(t)}) \,\|\, \overline{M}^{(t)}(\pi^{(t)})\right)\right]\right] + \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}), \overline{M}) \cdot T + \varepsilon T.
\end{aligned} \tag{139}$$

It remains to bound the first sum, for which we closely follow Russo and Van Roy (2014). Consider a fixed timestep t and decision $\pi \in \Pi$. Observe that $\overline{M}^{(t)}(\pi)$ is identical to the law $\mathbb{P}_{(r^{(t)}, o^{(t)}) \mid \pi^{(t)} = \pi, \mathscr{F}^{(t-1)}}$, while $\overline{M}_\rho^{(t)}(\pi)$ is identical to the law $\mathbb{P}_{(r^{(t)}, o^{(t)}) \mid, \rho^\star = \rho, \pi^{(t)} = \pi, \mathscr{F}^{(t-1)}}$. Hence, for any π, we have

$$\mathbb{E}_{t-1}\left[D_{\mathrm{KL}}\left(\overline{M}_{\rho^\star}^{(t)}(\pi) \,\|\, \overline{M}^{(t)}(\pi)\right)\right] = I_{t-1}(\rho^\star\,; (r^{(t)}, o^{(t)}) \mid \pi^{(t)} = \pi),$$

where $I_{t-1}(X\,; Y \mid Z)$ denotes the conditional mutual information of (X, Y) given Z (under $\mathscr{F}^{(t-1)}$).[31] Since $\pi^{(t)}$ and $\rho^\star$ are conditionally independent given $\mathscr{F}^{(t-1)}$, we further have

$$\begin{aligned}
\mathbb{E}_{t-1}\left[D_{\mathrm{KL}}\left(\overline{M}_{\rho^\star}^{(t)}(\pi^{(t)}) \,\|\, \overline{M}^{(t)}(\pi^{(t)})\right)\right] &= I_{t-1}(\rho^\star\,; (r^{(t)}, o^{(t)}) \mid \pi^{(t)}) \\
&= I_{t-1}(\rho^\star\,; (r^{(t)}, o^{(t)}) \mid \pi^{(t)}) + I_{t-1}(\rho^\star\,; \pi^{(t)}) \\
&= I_{t-1}(\rho^\star\,; (\pi^{(t)}, r^{(t)}, o^{(t)})).
\end{aligned}$$

Summing and using the chain rule for mutual information, we have

$$\mathbb{E}\left[\sum_{t=1}^T I_{t-1}(\rho^\star\,; (\pi^{(t)}, r^{(t)}, o^{(t)}))\right] = \sum_{t=1}^T I(\rho^\star\,; (\pi^{(t)}, r^{(t)}, o^{(t)}) \mid \mathcal{H}^{(t-1)}) = I(\rho^\star\,; \mathcal{H}^{(T)}).$$

Finally, we have

$$I(\rho^\star\,; \mathcal{H}^{(T)}) \leq H(\rho^\star) \leq \log|\mathrm{P}| = \log \mathcal{N}(\Pi_{\mathcal{M}}, \varepsilon).$$

Putting everything together gives

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \gamma \cdot \log \mathcal{N}(\Pi_{\mathcal{M}}, \varepsilon) + \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathrm{co}(\mathcal{M}), \overline{M}) \cdot T + \varepsilon T,$$

and the result follows by optimizing over $\varepsilon \geq 0$ and $\gamma > 0$. □

[31] For random variables X, Y, Z, we define $I(X\,; Y \mid Z) = \mathbb{E}_Z\left[D_{\mathrm{KL}}\left(\mathbb{P}_{(X,Y)|Z} \,\|\, \mathbb{P}_{X|Z} \otimes \mathbb{P}_{Y|Z}\right)\right] = \mathbb{E}_{Y,Z}\left[D_{\mathrm{KL}}\left(\mathbb{P}_{X|Y,Z} \,\|\, \mathbb{P}_{X|Z}\right)\right]$. $I_t(X\,; Y \mid Z)$ denotes the same quantity, conditioned on the outcome for $\mathscr{F}^{(t)}$.

C.3 Proofs for Learnability Results

Theorem 3.5 (Learnability)**.** *Assume that* $\mathcal{R} \subseteq [0,1]$*, and that Assumption 3.2 and Assumption 3.3 hold. Suppose that* $\mathcal{M}$ *is convex and has* $\mathsf{est}(\mathcal{M},T) = \widetilde{O}(T^q)$ *for some* $q<1$*. Then:*

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

Theorem 3.7 (Learnability—refined version)**.** *Suppose that* $\mathsf{est}(\Pi_\mathcal{M},T) = \widetilde{O}(T^q)$ *for some* $q<1$*, and that the conclusion of Proposition 2.1 holds, but place no assumption on* $\mathsf{est}(\mathcal{M},T)$*. Then the conclusion of Theorem 3.5 continues to hold.*

Proof of Theorem 3.5 and Theorem 3.7. The proofs for Theorem 3.5 and Theorem 3.7 differ only in how we derive the upper bound on regret.

We begin with the upper bound. Assume that $\lim_{\gamma\to\infty}\mathsf{dec}_\gamma(\mathcal{M})\cdot\gamma^\rho = 0$ for some $\rho>0$. For Theorem 3.5, we assume $\mathsf{est}(\mathcal{M},T) = \widetilde{O}(T^q)$ for some $q<1$ and apply Theorem 3.4, which gives that for any $T\in\mathbb{N}$ and $\gamma>0$, $\mathsf{E2D}$ with an appropriate oracle has[32]

$$\sup_{M\in\mathcal{M}}\mathbb{E}^{M}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}(\mathsf{dec}_\gamma(\mathcal{M})\cdot T + \gamma\cdot\mathsf{est}(\mathcal{M},T)) \leq \widetilde{O}(\mathsf{dec}_\gamma(\mathcal{M})\cdot T + \gamma\cdot T^q),$$

where $\widetilde{O}(\cdot)$ hides factors logarithmic in T and B. For Theorem 3.7, we assume $\mathsf{est}(\Pi_\mathcal{M},T) = \widetilde{O}(T^q)$ for some $q<1$ and apply Theorem 3.6, which implies that there exists an algorithm with

$$\sup_{M\in\mathcal{M}}\mathbb{E}^{M}[\mathbf{Reg}_{\mathsf{DM}}] \leq O(\mathsf{dec}_\gamma(\mathcal{M})\cdot T + \gamma\cdot\mathsf{est}(\Pi_\mathcal{M},T)) \leq \widetilde{O}(\mathsf{dec}_\gamma(\mathcal{M})\cdot T + \gamma\cdot T^q).$$

For both cases, we set $\gamma_T = T^{\frac{1-q}{1+\rho}}$, where we recall that $1-q>0$. Since $\lim_{\gamma\to\infty}\mathsf{dec}_\gamma(\mathcal{M})\cdot\gamma^\rho = 0$, we have that for any $\varepsilon>0$, there exists $\gamma'>0$ such that $\mathsf{dec}_\gamma(\mathcal{M}) \leq \varepsilon/\gamma^\rho$ for all $\gamma\geq\gamma'$. In particular, for T sufficiently large, we have

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

Since $p := \frac{1+\rho q}{1+\rho} < 1$, this establishes the result. In particular, defining $p' := \frac{1}{2}(p+1) < 1$, we have

$$\lim_{T\to\infty}\frac{\sup_{M\in\mathcal{M}}\mathbb{E}^{M}[\mathbf{Reg}_{\mathsf{DM}}]}{T^{p'}} = 0.$$

We now proceed with the lower bound. Suppose that $\lim_{\gamma\to\infty}\mathsf{dec}_\gamma(\mathcal{M})\cdot\gamma^\rho = \infty$ for all $\rho>0$, and consider any fixed $\rho\in(0,1/2)$.[33] Applying Theorem 3.1, we are guaranteed that for any algorithm and $\delta\in(0,1)$, for any $\gamma = \omega(\sqrt{T\log(T)})$, with probability at least $\delta/2$,

$$\mathbf{Reg}_{\mathsf{DM}} = \widetilde{\Omega}\big(\min\big\{\mathsf{dec}_{\gamma,\varepsilon(\gamma,T)}(\mathcal{M})\cdot T, \gamma\big\} - \delta T\big),$$

[32] Theorem 3.4 is stated as a high-probability guarantee, but since $\mathcal{R}\subseteq[0,1]$ we can deduce this in-expectation guarantee by setting $\delta = 1/T$.

[33] Assuming that $\lim_{\gamma\to\infty}\mathsf{dec}_\gamma(\mathcal{M})\cdot\gamma^\rho = \infty$ for all $\rho>0$ is equivalent to assuming that $\lim_{\gamma\to\infty}\mathsf{dec}_\gamma(\mathcal{M})\cdot\gamma^\rho > 0$ for all $\rho>0$.

where $\varepsilon(\gamma, T) := c \cdot \frac{\gamma}{T\log(T)}$ for a numerical constant $c \leq 1$. Hence, since Assumption 3.3 is satisfied, whenever $\varepsilon(\gamma, T) \leq 1$ we can apply the local-to-global lemma (Lemma B.1) to lower bound by

$$\mathbf{Reg}_{\mathsf{DM}} \geq \widetilde{\Omega}(\min\{\varepsilon(\gamma, T) \cdot \mathsf{dec}_{\gamma}(\mathcal{M}) \cdot T, \gamma\} - \delta T).$$

Let $\gamma_T = T$ (which has $\gamma_T = \omega(\sqrt{T\log(T)})$) and $\delta_T = c_T \cdot T^{-\rho}$, where $c_T := 1/\mathrm{polylog}(T)$ is sufficiently small. Since $\lim_{\gamma\to\infty} \mathsf{dec}_{\gamma}(\mathcal{M}) \cdot \gamma^{\rho} = \infty$, we have that for all $C > 0$, there exists $\gamma' > 0$ such that $\mathsf{dec}_{\gamma}(\mathcal{M}) \geq C\gamma^{-\rho}$ for all $\gamma \geq \gamma'$. Hence, for T sufficiently large, we have $\mathsf{dec}_{\gamma_T}(\mathcal{M}) \geq \gamma_T^{-\rho}$ and

$$\mathbf{Reg}_{\mathsf{DM}} = \widetilde{\Omega}\left(\min\left\{\frac{T}{\gamma_T^{\rho}}, \gamma_T\right\} - \delta_T T\right) = \widetilde{\Omega}\big(\min\{T^{1-\rho}, T\} - \delta_T T\big) = \widetilde{\Omega}(T^{1-\rho}),$$

since $\varepsilon(\gamma_T, T) \propto \frac{1}{\log(T)}$. Furthermore, since $\mathbf{Reg}_{\mathsf{DM}}$ is non-negative, the law of total expectation implies that

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

It follows that for any $p \in (0, 1)$, if we set $\rho = \frac{1-p}{2} \in (0, 1/2)$, we have

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

In particular, if we apply this argument with $p' = \frac{1}{2}(p + 1) \in (1/2, 1)$, we are guaranteed that

$$\lim_{T\to\infty} \frac{\sup_{M\in\mathcal{M}} \mathbb{E}^{M}[\mathbf{Reg}_{\mathsf{DM}}]}{T^{p}} = \infty.$$

□

C.4 Additional Proofs

Proposition 3.1. *For any $A \in \mathbb{N}$, there exists a structured bandit problem with $|\mathcal{M}| = |\Pi_{\mathcal{M}}| = A$ for which $\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \leq (2\gamma)^{-1}$ for all $\gamma > 0$, yet any algorithm must have regret $\Omega(A)$.*

Proof of Proposition 3.1. Let $\Delta \in (0, 1/2)$ be a free parameter. Let $\Pi = [A]$ and consider the class of models $\{M_i\}_{i\in[A]}$ defined via

$$f^{M_i}(\pi) := \frac{1}{2} + \Delta\mathbb{I}\{\pi = i\},$$

and $M_i(\pi) = \mathcal{N}(f^{M_i}(\pi), 1)$ (recall that $\mathcal{O} = \{\varnothing\}$). We state two lemmas, with proofs deferred to the end. The first lemma shows that the value of the Decision-Estimation Coefficient is independent of the number of actions.

Lemma C.4. *We have $\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \leq (4\gamma)^{-1}$ for all $\gamma > 0$.*

While $\sup_{\overline{M}\in\mathcal{M}} \mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M})$ is independent of the number of actions A, the following lemma shows that the regret of any algorithm must depend linearly on A.

Lemma C.5. *For any algorithm, there exists $M^{\star} \in \mathcal{M}$ such that $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \geq \frac{\Delta}{6} \cdot \min\{A, T\}$.*

This establishes the result. □

Proof of Lemma C.4. Let $\overline{M} \in \mathcal{M}$ be given. Choose $p = \delta_{\pi_{\overline{M}}}$ as the distribution over decisions. Then, since $f^{M}(\pi_{M}) = \frac{1}{2} + \Delta$ for all $M \in \mathcal{M}$, we have that for any such M,

$$\begin{aligned}
\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] &= f^{M}(\pi_{M}) - f^{M}(\pi_{\overline{M}}) \\
&= f^{\overline{M}}(\pi_{\overline{M}}) - f^{M}(\pi_{\overline{M}}) \\
&\leq \frac{1}{4\gamma} + \gamma(f^{\overline{M}}(\pi_{\overline{M}})) - f^{M}(\pi_{\overline{M}}))^{2} \\
&= \frac{1}{4\gamma} + \gamma\,\mathbb{E}_{\pi\sim p}\big[(f^{\overline{M}}(\pi) - f^{M}(\pi))^{2}\big].
\end{aligned}$$

Hence, since $|f^{\overline{M}}(\pi) - f^M(\pi)| \leq D_{\mathrm{TV}}\big(\overline{M}(\pi), M(\pi)\big) \leq D_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big)$, the minimax value $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$ is at most $(4\gamma)^{-1}$. $\square$

Proof of Lemma C.5. Assume that $T \geq A$ without loss of generality. Let $S = \{\pi^{(t)}\}_{t=1}^{A/4}$. For any choice of $M^\star \in \mathcal{M}$ and any algorithm, we have

$$\sum_{t=1}^{T} f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)}) \geq \frac{\Delta}{4} A \cdot \mathbb{I}\{\pi_{M^\star} \notin S\}.$$

Next, we observe that

$$\mathbb{I}\{\pi_{M^\star} \in S\} \leq \sum_{t=1}^{A/4} \mathbb{I}\{\pi^{(t)} = \pi_{M^\star}, \pi^{(1)}, \ldots, \pi^{(t-1)} \neq \pi_{M^\star}\}.$$

Hence, if we let $\mathbb{P}[\cdot]$ denote the law of $\mathcal{H}^{(T)}$ under $M^\star \sim \mathrm{unif}(\mathcal{M})$, we have

$$\mathbb{P}(\pi_{M^\star} \in S) \leq \mathbb{E}\left[\sum_{t=1}^{A/4} \mathbb{P}_{t-1}(\pi^{(t)} = \pi_{M^\star} \mid \pi^{(1)}, \ldots, \pi^{(t-1)} \neq \pi_{M^\star})\right],$$

where $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid \mathscr{F}^{(t-1)}]$. Since, $\pi^{(t)}$ is conditionally independent of $\pi_{M^\star}$ given $\mathscr{F}^{(t-1)}$, and since $\pi_{M^\star}$ is uniformly distributed over $[A] \setminus \{\pi^{(1)}, \ldots, \pi^{(t-1)}\}$ given $\pi_{M^\star} \notin \{\pi^{(1)}, \ldots, \pi^{(t-1)}\}$, the quantity above is bounded by

$$\sum_{t=1}^{A/4} \frac{1}{A - t + 1} \leq \frac{A}{4} \cdot \frac{4}{3A} = \frac{1}{3}.$$

It follows that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] = \mathbb{E}\left[\sum_{t=1}^{T} f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] \geq \frac{\Delta}{6} A.$$

$\square$

D Proofs from Section 4

D.1 Proofs from Section 4.1

In this section we prove Theorem 4.1. We prove the localized regret bound (35) in Theorem 4.1 under the following, slightly more general, version of Assumption 4.1, which will be useful for applications.

Assumption D.1. *The online estimation algorithm $\mathbf{Alg}_{\mathsf{Est}}$ guarantees that for a given $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}}\left[D^2_{\mathrm{H}}\Big(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\Big)\right]\mathbb{I}\{M^\star \in \mathcal{M}^{(t)}\} \leq \widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta), \tag{140}$$

where $\widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta)$ is a known upper bound. We further assume that $\widehat{M}^{(t)} \in \mathrm{co}(\mathcal{M}^{(t)})$.

Theorem 4.1a. *Fix $\delta \in (0,1)$ and consider Algorithm 1 with* Option II *and $R^2 = \widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta)$. Suppose Assumption D.1, that $\mathcal{R} \subseteq [0,1]$, and that $\mathbf{Alg}_{\mathsf{Est}}$ ensures that $\widehat{\mathcal{M}}^{(t)} \subseteq \mathrm{co}(\mathcal{M}^{(t)})$ for all t. Then for any fixed $T \in \mathbb{N}$ and $\gamma > 0$, with probability at least $1 - \delta$,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sum_{t=1}^{T} \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}_{\varepsilon_t}(\overline{M}), \overline{M}) + \gamma \cdot \widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta), \tag{141}$$

where $\varepsilon_t := 6\frac{\gamma}{t}\widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta) + \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) + (2\gamma)^{-1}$.

Proof of Theorem 4.1 and Theorem 4.1a. We first prove the basic non-localized results in (33) and (34). We write

$$\begin{aligned}\mathbf{Reg}_{\mathsf{DM}} &= \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[D^2_{\mathrm{H}}\left(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\right)\right] + \gamma\cdot\mathbf{Est}_{\mathsf{H}}.\end{aligned}$$

For each t, since $M^\star\in\mathcal{M}$, we have

$$\begin{aligned}&\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[D^2_{\mathrm{H}}\left(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\right)\right] \\ &\leq \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M}(\pi_{M}) - f^{M}(\pi^{(t)})\right] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[D^2_{\mathrm{H}}\left(M(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\right)\right] \\ &= \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(t)}(\pi)\right)\right] \\ &= \mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^{(t)}). \end{aligned} \tag{142}$$

We conclude that

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sup_{\overline{M}\in\widehat{\mathcal{M}}} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{H}},$$

which establishes (33). Since this bound holds almost surely, the result in (34) follows from Assumption 4.1.

Localized upper bound (Theorem 4.1a). We now prove the localized regret bound in (141) under Assumption D.1. Let $\mathcal{E}$ denote the high-probability event in Assumption D.1. Recall that we set the confidence radius used by Algorithm 1 to $R^2 = \widetilde{\mathbf{Est}}_{\mathsf{H}}(T,\delta)$. We first prove that under $\mathcal{E}$, the confidence sets contain $M^\star$ and the Hellinger error is controlled.

Lemma D.1. *Under $\mathcal{E}$, we have $M^\star\in\mathcal{M}^{(t)}$ for all $t\leq T$, and consequently*

$$\mathbf{Est}_{\mathsf{H}} = \sum_{t=1}^{T}\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[D^2_{\mathrm{H}}\left(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\right)\right] \leq \widetilde{\mathbf{Est}}_{\mathsf{H}}(T,\delta).$$

Proof of Lemma D.1. We prove the result by induction. For the base case, we have $M^\star\in\mathcal{M}^{(1)}$ with probability 1. Now, fix $t\geq 2$ and suppose $M^\star\in\mathcal{M}^{(i)}$ for all $i<t$. Then (140) implies that

$$\begin{aligned}&\sum_{i=1}^{t-1}\mathbb{E}_{\pi^{(i)}\sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M^\star(\pi^{(i)}), \widehat{M}^{(i)}(\pi^{(i)})\right)\right] \\ &= \sum_{i=1}^{t-1}\mathbb{E}_{\pi^{(i)}\sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M^\star(\pi^{(i)}), \widehat{M}^{(i)}(\pi^{(i)})\right)\right]\mathbb{I}\{M^\star\in\mathcal{M}^{(i)}\} \leq \widetilde{\mathbf{Est}}_{\mathsf{H}}(T,\delta).\end{aligned}$$

We conclude from the definition of $\mathcal{M}^{(t)}$ in Line 7 that $M^\star\in\mathcal{M}^{(t)}$. □

We next state the key technical lemma for the proof, which uses properties of the minimax problem that defines the Decision-Estimation Coefficient (i.e., (21)) to relate the localization radius (in the sense of $\mathcal{M}_\varepsilon$) to Hellinger estimation error.

Lemma D.2. *Assume that $\mathcal{R}\subseteq[0,1]$. Consider any model M such that $M\in\mathcal{M}^{(i)}$ for all $i\leq t$. Then for any model $\overline{M}$ (not necessarily in $\mathcal{M}$),*

$$\begin{aligned}&f^{M}(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) \\ &\leq (2\gamma)^{-1} + \frac{1}{t}\sum_{i=1}^{t}\left(\mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + 2\gamma\,\mathbb{E}_{\pi\sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right] + \gamma\,\mathbb{E}_{\pi\sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(\overline{M}(\pi), \widehat{M}^{(i)}(\pi)\right)\right]\right).\end{aligned} \tag{143}$$

Proof of Lemma D.2. Consider $i \leq t$. From the definition of the minimax program (21), we have that for all $M \in \mathcal{M}^{(i)}$,

$$\mathbb{E}_{\pi \sim p^{(i)}}[f^M(\pi_M) - f^M(\pi)] \leq \mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + \gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right].$$

By the AM-GM inequality, this implies that

$$\begin{aligned}
&\mathbb{E}_{\pi \sim p^{(i)}}\left[f^M(\pi_M) - f^{\overline{M}}(\pi)\right] \\
&\leq \mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + \gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right] + \frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{\pi \sim p^{(i)}}\left[(f^M(\pi) - f^{\overline{M}}(\pi))^2\right] \\
&\leq \mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + \gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right] + \frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big)\right],
\end{aligned}$$

where the second inequality uses that $|f^M(\pi) - f^{\overline{M}}(\pi)| \leq D_{\mathrm{TV}}\big(\overline{M}(\pi), M(\pi)\big) \leq D_{\mathrm{H}}\big(\overline{M}(\pi), M(\pi)\big)$ when rewards lie in $[0,1]$. Since $D_{\mathrm{H}}(\cdot,\cdot)$ satisfies the triangle inequality, we can use the elementary inequality $(x+y)^2 \leq 2(x^2+y^2)$ to upper bound by

$$\begin{aligned}
&\mathbb{E}_{\pi \sim p^{(i)}}\left[f^M(\pi_M) - f^{\overline{M}}(\pi)\right] \\
&\leq \mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + 2\gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right] + \frac{1}{2\gamma} + \gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(\overline{M}(\pi), \widehat{M}^{(i)}(\pi)\right)\right].
\end{aligned}$$

Since $\mathbb{E}_{\pi \sim p^{(i)}}[f^{\overline{M}}(\pi)] \leq f^{\overline{M}}(\pi_{\overline{M}})$, it follows immediately that

$$f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) \leq \mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + \frac{1}{2\gamma} + 2\gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right] + \gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(\overline{M}(\pi), \widehat{M}^{(i)}(\pi)\right)\right].$$

We obtain the final result by averaging this inequality over all $i \leq t$. □

We now proceed as follows. Condition on the event $\mathcal{E}$. By Lemma D.1, this implies that $M^\star \in \mathcal{M}^{(t)}$ for all $t \leq T$. Hence, proceeding as in (142), we are guaranteed that

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sum_{t=1}^T \mathsf{dec}_\gamma(\mathcal{M}^{(t)}, \widehat{M}^{(t)}) + \gamma \cdot \mathbf{Est}_{\mathsf{H}} \leq \sum_{t=1}^T \mathsf{dec}_\gamma(\mathcal{M}^{(t)}, \widehat{M}^{(t)}) + \gamma \cdot \widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta).$$

We now relate the confidence sets $\mathcal{M}^{(t)}$ to localized sets of the form $\mathcal{M}_{\varepsilon_t}(\widehat{M}^{(t)})$. We trivially have $\mathcal{M}^{(1)} \subseteq \mathcal{M}_{\varepsilon_1}(\widehat{M}^{(1)})$, so consider $t \geq 2$. Since $\mathcal{M}^{(t)} \subseteq \mathcal{M}^{(t-1)} \subseteq \cdots \subseteq \mathcal{M}^{(1)}$ by definition, Lemma D.2 guarantees that for all $M \in \mathcal{M}^{(t)}$,

$$\begin{aligned}
&f^M(\pi_M) - \widehat{f}^{(t)}(\widehat{\pi}^{(t)}) \\
&\leq (2\gamma)^{-1} + \frac{1}{t-1}\sum_{i=1}^{t-1}\left(\mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + 2\gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right] + \gamma\, \mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(\widehat{M}^{(t)}(\pi), \widehat{M}^{(i)}(\pi)\right)\right]\right),
\end{aligned}$$

where we define $\widehat{f}^{(t)} := f^{\widehat{M}^{(t)}}$ and $\widehat{\pi}^{(t)} := \pi_{\widehat{M}^{(t)}}$. Using the definition of $\mathcal{M}^{(t)}$ from Line 7 of Algorithm 1, we have that for all $M \in \mathcal{M}^{(t)}$,

$$\sum_{i=1}^{t-1}\mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right] \leq \widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta).$$

Furthermore, since $\widehat{M}^{(t)} \in \mathrm{co}(\mathcal{M}^{(t)})$, if we let $q^{(t)} \in \Delta(\mathcal{M}^{(t)})$ be such that $\widehat{M}^{(t)} = \mathbb{E}_{M \sim q^{(t)}}[M]$, we have

$$\begin{aligned}
\sum_{i=1}^{t-1}\mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(\widehat{M}^{(t)}(\pi), \widehat{M}^{(i)}(\pi)\right)\right] &= \sum_{i=1}^{t-1}\mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(\mathbb{E}_{M \sim q^{(t)}}[M(\pi)], \widehat{M}^{(i)}(\pi)\right)\right] \\
&\leq \mathbb{E}_{M \sim q^{(t)}}\left[\sum_{i=1}^{t-1}\mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right]\right] \\
&\leq \sup_{M \in \mathcal{M}^{(t)}}\left\{\sum_{i=1}^{t-1}\mathbb{E}_{\pi \sim p^{(i)}}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^{(i)}(\pi)\right)\right]\right\} \leq \widetilde{\mathbf{Est}}_{\mathsf{H}}(T, \delta).
\end{aligned}$$

Hence, since $\frac{1}{t-1} \leq \frac{2}{t}$, we have

$$\begin{aligned}
f^M(\pi_M) - \widehat{f}^{(t)}(\widehat{\pi}^{(t)}) &\leq (2\gamma)^{-1} + \frac{2}{t}\sum_{i=1}^{t-1} \mathsf{dec}_\gamma(\mathcal{M}^{(i)}, \widehat{M}^{(i)}) + 6\frac{\gamma}{t}\widetilde{\mathbf{Est}}_{\mathsf{H}}(T,\delta) \\
&\leq (2\gamma)^{-1} + 2\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) + 6\frac{\gamma}{t}\widetilde{\mathbf{Est}}_{\mathsf{H}}(T,\delta) =: \varepsilon_t
\end{aligned}$$

It follows that $\mathcal{M}^{(t)} \subseteq \mathcal{M}_{\varepsilon_t}(\widehat{M}^{(t)})$ and

$$\begin{aligned}
\mathbf{Reg}_{\mathsf{DM}} &\leq \sum_{t=1}^{T} \mathsf{dec}_\gamma(\mathcal{M}_{\varepsilon_t}(\widehat{M}^{(t)}), \widehat{M}^{(t)}) + \gamma \cdot \widetilde{\mathbf{Est}}_{\mathsf{H}}(T,\delta) \\
&\leq \sum_{t=1}^{T} \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\gamma(\mathcal{M}_{\varepsilon_t}(\overline{M}), \overline{M}) + \gamma \cdot \widetilde{\mathbf{Est}}_{\mathsf{H}}(T,\delta).
\end{aligned}$$

□

D.2 Proofs from Section 4.2

Proposition 4.2. *Suppose that Π is finite and $\mathcal{R}$ is bounded. Then for all models $\overline{M}$,*

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}). \tag{41}$$

Proof of Proposition 4.2. We follow the approach in Lattimore and Szepesvári (2019). Since $\underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}) \leq \mathsf{dec}_\gamma(\mathcal{M}, \overline{M})$ by definition, it suffices to prove the opposite direction of the inequality.

Recall that for a topological space $\mathcal{Z}$, let $\Delta(\mathcal{Z})$ denotes the space of Radon probability measures over $\mathcal{Z}$ when $\mathcal{Z}$ is equipped with the Borel σ-algebra. In addition, recall that the weak* topology on $\Delta(\mathcal{Z})$ is the coarsest topology such that the function $\mu \mapsto \int f d\mu$ is continuous for all bounded, continuous functions $f : \mathcal{Z} \to \mathbb{R}$.

Let Π be equipped with the discrete topology, and recall that $\mathcal{M}$ is equipped with the discrete topology as well. Since Π is finite, it is compact with respect to the discrete topology. Hence, Theorem 8.9.3 of Bogachev (2007), $\Delta(\Pi)$ is weak*-compact; $\Delta(\Pi)$ is also convex.

Let $\mathcal{Q}$ be the space of finitely supported probability measures on $\mathcal{M}$, which is a convex subset of $\Delta(\mathcal{M})$ when $\mathcal{M}$ is equipped with the discrete topology. We equip $\mathcal{Q}$ with the weak*-topology.

Consider any function $g : \Delta(\Pi) \times \mathcal{Q} \to \mathbb{R}$ that is linear and continuous with respect to both arguments in their respective topologies (i.e. $p \mapsto g(p, \mu)$ is continuous for all $\mu \in \mathcal{Q}$ and $\mu \mapsto g(p, \mu)$ is continuous for all $p \in \Delta(\Pi)$). Since $\Delta(\Pi)$ is weak*-compact, Sion's minimax theorem (Sion, 1958) implies that

$$\min_{p\in\Delta(\Pi)} \sup_{\mu\in\mathcal{Q}} g(p,\mu) = \sup_{\mu\in\mathcal{Q}} \min_{p\in\Delta(\Pi)} g(p,\mu).$$

We proceed to verify that the function

$$g(p,\mu) := \mathbb{E}_{\pi\sim p}\,\mathbb{E}_{M\sim\mu}\Big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\Big]$$

linear and continuous in the sense above. Linearity is immediate. For continuity, since we consider the weak*-topology for both spaces, we only need to show that the function

$$(\pi, M) \mapsto f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)$$

is continuous with respect to π and M individually. This is follows because Π and $\mathcal{M}$ are equipped with the discrete topology and the function is bounded (since $\mathcal{R}$ and $D^2_{\mathrm{H}}(\cdot,\cdot)$ are bounded). We conclude that

$$\mathsf{dec}_\gamma(\mathcal{M},\overline{M}) \leq \min_{p\in\Delta(\Pi)}\sup_{\mu\in\mathcal{Q}} g(p,\mu) = \sup_{\mu\in\mathcal{Q}}\min_{p\in\Delta(\Pi)} g(p,\mu) \leq \underline{\mathsf{dec}}_\gamma(\mathcal{M},\overline{M}).$$

$\square$

D.3 Proofs from Section 4.3

Theorem 4.3. *Algorithm 3 with exploration parameter $\gamma>0$ guarantees that*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sup_{\nu\in\Delta(\widehat{\mathcal{M}})} \mathsf{dec}^D_\gamma(\mathcal{M},\nu)\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{D}} \tag{47}$$

almost surely.

Proof of Theorem 4.3. This proof closely follows that of Theorem 4.1. We have

$$\begin{aligned}
\mathbf{Reg}_{\mathsf{DM}} &= \sum_{t=1}^T \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] \\
&= \sum_{t=1}^T \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\mathbb{E}_{\widehat{M}^{(t)}\sim\nu^{(t)}}\left[D\Big(M^\star(\pi^{(t)})\ \|\ \widehat{M}^{(t)}(\pi^{(t)})\Big)\right] + \gamma\cdot\mathbf{Est}_{\mathsf{D}}.
\end{aligned}$$

For each t, since $M^\star\in\mathcal{M}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\right] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\mathbb{E}_{\widehat{M}^{(t)}\sim\nu^{(t)}}\left[D\Big(M^\star(\pi^{(t)})\ \|\ \widehat{M}^{(t)}(\pi^{(t)})\Big)\right] \\
&\leq \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\left[f^{M}(\pi_{M}) - f^{M}(\pi^{(t)})\right] - \gamma\cdot\mathbb{E}_{\pi^{(t)}\sim p^{(t)}}\mathbb{E}_{\widehat{M}^{(t)}\sim\nu^{(t)}}\left[D\Big(M(\pi^{(t)})\ \|\ \widehat{M}^{(t)}(\pi^{(t)})\Big)\right] \\
&= \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot\mathbb{E}_{\widehat{M}^{(t)}\sim\nu^{(t)}} D\Big(M(\pi)\ \|\ \widehat{M}^{(t)}(\pi)\Big)\right] \\
&= \mathsf{dec}^D_\gamma(\mathcal{M},\nu^{(t)}).
\end{aligned}$$

Since $\mathsf{dec}^D_\gamma(\mathcal{M},\nu^{(t)}) \leq \sup_{\nu\in\Delta(\widehat{\mathcal{M}})}\mathsf{dec}^D_\gamma(\mathcal{M},\nu)$, this establishes the result. $\square$

E Proofs from Section 6

E.1 Proofs from Section 6.1

E.1.1 Linear Bandits

Proof of Proposition 6.2. Let $\Delta\in[0,1]$ be a parameter. We construct a hard family of models $\mathcal{M}'=\{M_i\}_{i\in[d]}$ as follows. First, define $\theta_i=\Delta\cdot e_i$, where e_i denotes the standard basis vector, then let $M_i(\pi)=\mathrm{Rad}(\langle\theta_i,\pi\rangle)$; this construction has $f^{M_i}(\pi)=\langle\theta_i,\pi\rangle$. Then, take $\overline{M}(\pi)=\mathrm{Rad}(\langle\mathbf{0},\pi\rangle)=\mathrm{Rad}(0)$ as the reference model.

We now show that $\mathcal{M}'$ is a hard family of models in the sense of Lemma 5.1. Define $u_i(\pi)=\pi_i$ and $v_i(\pi)=\pi_i^2$. Then for any i, we have

$$f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) = \langle\theta_i, e_i-\pi\rangle = \Delta(1-u_i(\pi)).$$

Next, using Lemma A.8, we have

$$D^2_{\mathrm{H}}\big(M_i(\pi),\overline{M}(\pi)\big) = D^2_{\mathrm{H}}(\mathrm{Rad}(\langle\theta_i,\pi\rangle),\mathrm{Rad}(0)) \leq \frac{3}{4}\Delta^2\pi_i^2 = \frac{3}{4}\Delta^2 v_i(\pi).$$

Finally, note that since $\|\pi\|_2 \leq 1$, we have $\sum_{i=1}^d u_i(\pi) = \sum_{i=1}^d \pi_i \leq \sqrt{d} \leq \frac{d}{2}$ (whenever $d \geq 4$) and $\sum_{i=1}^d v_i(\pi) = \sum_{i=1}^d \pi_i^2 \leq 1$. It follows that $\mathcal{M}'$ is a $(\Delta, \frac{3}{4}\Delta^2, 0)$-family, so Lemma 5.1 implies that

$$\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{\Delta}{2} - \gamma\frac{3\Delta^2}{4d}.$$

We choose $\Delta = \frac{d}{3\gamma}$, which gives $\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{d}{12\gamma}$. Finally, note that $\mathcal{M}' \subseteq \mathcal{M}^\infty_\Delta(\overline{M})$, and that we may take $V(\mathcal{M}') = O(1)$ in Theorems 3.1 and 3.2 whenever $\Delta \leq 1/2$; it suffices to restrict to $\gamma \geq \frac{2d}{3}$. ☐

E.1.2 Nonparametric Bandits

Proof of Proposition 6.4. Let $M \in \mathcal{M}$ be fixed. Let Π' be an ε-cover for Π. Since f^M is 1-Lipschitz, for all π there exists a corresponding covering element $\iota(\pi) \in \Pi'$ such that $\rho(\pi, \iota(\pi)) \leq \varepsilon$, and consequently for any distribution p,

$$\begin{aligned}
\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] &\leq \mathbb{E}_{\pi\sim p}[f^M(\iota(\pi_M)) - f^M(\pi)] + |f^M(\pi_M) - f^M(\iota(\pi_M))| \\
&\leq \mathbb{E}_{\pi\sim p}[f^M(\iota(\pi_M)) - f^M(\pi)] + \rho(\pi_M, \iota(\pi_M)) \\
&\leq \mathbb{E}_{\pi\sim p}[f^M(\iota(\pi_M)) - f^M(\pi)] + \varepsilon.
\end{aligned}$$

At this point, since $\iota(\pi_M) \in \Pi$, Proposition 5.2 ensures that if we choose p using inverse gap weighting over Π'

$$\mathbb{E}_{\pi\sim p}[f^M(\iota(\pi_M)) - f^M(\pi)] \leq \frac{|\Pi'|}{\gamma} + \gamma \cdot \mathbb{E}_{\pi\sim p}\big[(f^M(\pi) - f^{\overline{M}}(\pi))^2\big].$$

From our assumption on the growth of $\mathcal{N}_\rho(\Pi, \varepsilon)$, $|\Pi'| \leq \varepsilon^{-d}$, so the value is at most

$$\varepsilon + \frac{\varepsilon^{-d}}{\gamma}.$$

We choose $\varepsilon = \gamma^{-\frac{1}{d+1}}$ to balance the terms, leading to the result. ☐

Proof of Proposition 6.5. Let $\varepsilon \in (0, 1/2)$ be fixed, and let $N = \mathcal{N}_\rho(\Pi, 2\varepsilon)$. By the duality of packing and covering, there exists a packing $\pi_1, \ldots, \pi_N$ such that

$$\rho(\pi_i, \pi_j) > 2\varepsilon \quad \text{for all } i \neq j.$$

We define a class of models $\mathcal{M}' = \{M_1, \ldots, M_N\}$ based on this packing as follows. First, define $h(x) = \max\{1 - x, 0\}$. Then, let

$$f_i(\pi) = \frac{1}{2} + \varepsilon h(\rho(\pi, \pi_i)/\varepsilon).$$

We have $f_i(\pi) \in [1/2, 1)$, and since ρ is a metric,

$$|f_i(\pi) - f_i(\pi')| \leq \varepsilon|h(\rho(\pi, \pi_i)/\varepsilon) - h(\rho(\pi', \pi_i)/\varepsilon)| \leq |\rho(\pi, \pi_i) - \rho(\pi', \pi_i)| \leq \rho(\pi, \pi'),$$

so f_i is 1-Lipschitz. Finally, we define

$$M_i(\pi) = \mathrm{Ber}(f_i(\pi)),$$

and define the reference model $\overline{M}$ via

$$\overline{M}(\pi) = \mathrm{Ber}(1/2).$$

We now prove that $\mathcal{M}'$ is a hard family in the sense of Lemma 5.1. Let $\mathcal{I}_i = \{\pi \in \Pi \mid \rho(\pi, \pi_i) \leq \varepsilon\}$, and let

$$u_i(\pi) = v_i(\pi) = \mathbb{I}\{\pi \in \mathcal{I}_i\}.$$

Since $\rho(\pi_i, \pi_j) > 2\varepsilon$ for $i \neq j$, $\mathcal{I}_i \cap \mathcal{I}_j = \varnothing$, so we have $\sum_i u_i(\pi) \leq 1$ and $\sum_i v_i(\pi) \leq 1$ as required.

Now, observe that for all i and all $\pi \in \Pi$,

$$f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \varepsilon \mathbb{I}\{\pi \notin \mathcal{I}_i\} = \varepsilon(1 - u_i(\pi))$$

and by Lemma A.7

$$D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \leq 3\varepsilon^2 h^2(\rho(\pi, \pi_i)/\varepsilon) \leq 3\varepsilon^2 \mathbb{I}\{\pi \in \mathcal{I}_i\}.$$

Hence, $\mathcal{M}'$ is a $(\varepsilon, 3\varepsilon^2, 0)$-family, and Lemma 5.1 implies that

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \geq \frac{\varepsilon}{2} - \gamma\frac{3\varepsilon^2}{N} \geq \frac{\varepsilon}{2} - \gamma\frac{3\varepsilon^2}{(2\varepsilon)^{-d}} \geq \frac{\varepsilon}{2} - \gamma(6\varepsilon)^{2+d}.$$

We choose $\varepsilon = (36)^{-1}\gamma^{-\frac{1}{d+1}}$, which leads to value

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \geq (108)^{-1}\gamma^{-\frac{1}{d+1}}$$

whenever $d \geq 1$. Note that $\mathcal{M}' \subseteq \mathcal{M}^\infty_\varepsilon(\overline{M})$, and that we may take $V(\mathcal{M}') = O(1)$ in Theorems 3.1 and 3.2 whenever $\varepsilon \leq 1/4$; it suffices to restrict to $\gamma \geq 1$. □

E.1.3 ReLU Bandits

Proof of Proposition 6.6. Let $\varepsilon \in (0, 1)$ be a parameter, and recall that $\Pi = \Theta = \{v \in \mathbb{R}^d \mid \|v\|_2 \leq 1\}$. For each $v \in \Theta$, define a mean reward function

$$f_v(\pi) = \mathsf{relu}(\langle v, \pi\rangle - (1 - \varepsilon)),$$

so that $\max_{\pi\in\Pi} f_v(\pi) = \varepsilon$. Define a corresponding model M_v via

$$M_v(\pi) := \mathrm{Rad}(f_v(\pi)).$$

Finally, define the reference model via $\overline{M}(\pi) := M_{\mathbf{0}}(\pi) = \mathrm{Rad}(0)$. Note that if we choose $\mathcal{M}' = \{M_v \mid \|v\|_2 = 1\}$, we have $\mathcal{M}' \subseteq \mathcal{M}^\infty_\varepsilon(\overline{M})$, since $\max_{\pi\in\Pi} f^{\overline{M}}(\pi) = 0$.

We observe that this family satisfies the following properties:

- For all $\pi \in \Pi$, $\max_{\pi'\in\Pi} f_v(\pi') - f_v(\pi) \geq \varepsilon\mathbb{I}\{\langle v, \pi\rangle \leq 1 - \varepsilon\}$.
- For all $\pi \in \Pi$, $D^2_{\mathrm{H}}\big(M_v(\pi), \overline{M}(\pi)\big) = \frac{3}{4}f_v^2(\pi) \leq \frac{3\varepsilon^2}{4}\mathbb{I}\{\langle v, \pi\rangle > 1 - \varepsilon\}$, where we have used Lemma A.8.

Consequently, we have

$$\begin{aligned}
\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) &\geq \inf_{p\in\Delta(\Pi)} \sup_{\|v\|_2=1} \mathbb{E}_{\pi\sim p}\left[\varepsilon\mathbb{I}\{\langle v, \pi\rangle \leq 1 - \varepsilon\} - \gamma\frac{3\varepsilon^2}{4}\mathbb{I}\{\langle v, \pi\rangle > 1 - \varepsilon\}\right] \\
&= \inf_{p\in\Delta(\Pi)} \sup_{\|v\|_2=1} \mathbb{E}_{\pi\sim p}\left[\varepsilon - \varepsilon\mathbb{I}\{\langle v, \pi\rangle > 1 - \varepsilon\} - \gamma\frac{3\varepsilon^2}{4}\mathbb{I}\{\langle v, \pi\rangle > 1 - \varepsilon\}\right] \\
&\geq \inf_{p\in\Delta(\Pi)} \mathbb{E}_{v\sim\mathrm{unif}(\mathbb{S}^d)} \mathbb{E}_{\pi\sim p}\left[\varepsilon - \varepsilon\mathbb{I}\{\langle v, \pi\rangle > 1 - \varepsilon\} - \gamma\frac{3\varepsilon^2}{4}\mathbb{I}\{\langle v, \pi\rangle > 1 - \varepsilon\}\right].
\end{aligned}$$

We now appeal to the following lemma.

Lemma E.1 (Ball (1997), Lecture 8)**.** *Let v be uniform on $\mathbb{S}^d$. Then for any unit vector x and all $\alpha \in [0, 1]$.*

$$\mathbb{P}(|\langle x, v\rangle| > \alpha) \leq 2\exp\left(-\frac{\alpha^2}{2}d\right).$$

Applying this lemma, we have that for all $\varepsilon \in [0, 1/2]$,

$$\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \varepsilon - 2\varepsilon\exp(-d/8) - \frac{3}{2}\gamma\varepsilon^2\exp(-d/8).$$

Whenever $d \geq 16$, this is lower bounded by

$$\frac{\varepsilon}{2} - \frac{3}{2}\gamma\varepsilon^2 \exp(-d/8).$$

We conclude by choosing $\varepsilon = \frac{e^{d/8}}{6\gamma} \wedge \frac{1}{2}$, which leads to $\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{e^{d/8}}{24\gamma} \wedge \frac{1}{8}$, and has $V(\mathcal{M}') = O(1)$ for Theorems 3.1 and 3.2.

□

E.1.4 Gap-Dependent Lower Bounds

Proof of Proposition 6.7. This proof is almost the same as that of Proposition 5.3. The only difference is that we change the construction so that $\overline{M}$ has a gap, which leads to worse constants.

Fix $\Delta \in (0, 1/8)$ and define family of models $\mathcal{M}' = \{M_1, \ldots, M_A\}$ by setting

$$M_1(\pi) = \mathrm{Ber}(1/2 + \Delta\mathbb{I}\{\pi = 1\})$$

and

$$M_i(\pi) = \mathrm{Ber}(1/2 + \Delta\mathbb{I}\{\pi = 1\} + 2\Delta\mathbb{I}\{\pi = i\})$$

for $i > 1$. Let $\overline{M} = M_1$, and note that $\mathcal{M}' \subseteq \mathcal{M}^\infty_\Delta(\overline{M})$.

We now verify that this is a hard family. Set $u_i(\pi) = v_i(\pi) = \mathbb{I}\{\pi = i\}$. We have

$$f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \Delta(1 - \mathbb{I}\{\pi = i\})$$

since all of the models have gap Δ. Furthermore,

$$D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \leq D^2_{\mathrm{H}}(\mathrm{Ber}(1/2 + 2\Delta), \mathrm{Ber}(1/2))\mathbb{I}\{\pi = i\} \leq 12\Delta^2\mathbb{I}\{\pi = i\}$$

where we have used Lemma A.7. It follows that this is a $(\Delta, 5\Delta^2, 0)$-family, so Lemma 5.1 implies that for any $\gamma > 0$,

$$\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{\Delta}{2} - \gamma\frac{12\Delta^2}{A}.$$

In particular, whenever $\gamma \leq \frac{A}{48\Delta}$, we have $\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{\Delta}{4}$. Furthermore, since $\Delta \leq 1/8$, we may take $V(\mathcal{M}') = O(1)$ in Theorems 3.1 and 3.2. □

Proof of Proposition 6.8. This proof is a simple modification to Proposition 6.2 to ensure that $\overline{M}$ has gap Δ.

Let $\Delta \in (0, 1/4)$ be a given. We set $\Pi = \{e_1, \ldots, e_d\}$ and construct a family $\mathcal{M}' = \{M_i\}_{i \in [d]}$ as follows. Let $\theta_1 = \Delta \cdot e_1$, and let $\theta_i = \Delta \cdot e_1 + 2\Delta \cdot e_i$ for all $i \geq 2$; we have $\|\theta_i\|_2 \leq 1$ whenever $\Delta \leq 1/\sqrt{5}$. Define $M_i(\pi) = \mathrm{Rad}(\langle\theta_i, \pi\rangle)$ so that $f^{M_i}(\pi) = \langle\theta_i, \pi\rangle$, and take $\overline{M} = M_1$ as the reference model. We have $\mathcal{M}' \subseteq \mathcal{M}^\infty_\Delta(\overline{M})$.

We now show that $\mathcal{M}'$ is a hard family of models (cf. Lemma 5.1). Define $u_i(\pi) = v(\pi) = \mathbb{I}\{\pi = e_i\}$. Then for any i, we have

$$f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \Delta(1 - u_i(\pi))$$

and, using Lemma A.8,

$$D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \leq D^2_{\mathrm{H}}(\mathrm{Rad}(2\Delta), \mathrm{Rad}(0))\mathbb{I}\{\pi = e_i\} \leq 3\Delta^2 v_i(\pi).$$

It follows that $\mathcal{M}'$ is a $(\Delta, 3\Delta^2, 0)$-family, so Lemma 5.1 implies that

$$\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{\Delta}{2} - \gamma\frac{3\Delta^2}{2d}.$$

In particular, $\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \geq \frac{\Delta}{4}$ whenever $\gamma \leq \frac{d}{12\Delta}$. Furthermore, since $\Delta \leq 1/4$, we may take $V(\mathcal{M}') = O(1)$ in Theorems 3.1 and 3.2. □

E.2 Proofs from Section 6.2

This section is organized as follows. First, we prove a generic decoupling-type lemma which holds for a slightly more general setting than what is described in Section 6.2. We then prove Theorem 6.1 and Theorem 6.2 as consequences.

E.2.1 Decoupling Lemma

Consider a general setting where we are given a set Π and function class $\mathcal{F} : \Pi \to \mathbb{R}$. Let $\mathcal{Z}$ be an abstract set, and let $\{f_z\}_{z\in\mathcal{Z}}$ and $\{\pi_z\}_{z\in\mathcal{Z}}$ be subsets of $\mathcal{F}$ and Π indexed by $\mathcal{Z}$. Suppose we are given a distribution $\mu \in \Delta(\mathcal{Z})$ from which a random index is drawn. The following lemma allows one to decouple the random variables f_z and π_z under $z \sim \mu$.

Lemma E.2. *For any distribution $\mu \in \Delta(\mathcal{Z})$ and class $\mathcal{F} \subseteq (\Pi \to [-R, R])$, we have*

$$\mathbb{E}_{z\sim\mu}[|f_z(\pi_z)|] \leq \inf_{\Delta>0}\left\{2\Delta + 6\frac{\boldsymbol{\theta}(\mathcal{F}, \Delta, \gamma^{-1}; \rho_\mu)\log^2(R\gamma \vee e)}{\gamma}\right\} + \gamma \cdot \mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big]. \tag{144}$$

for all $\gamma > 0$, where $\rho_\mu(\pi) := \mu(\{\pi_z = \pi\})$.

Proof of Lemma E.2. Assume $R = 1$ without loss of generality. Fix a parameter $\Delta \in (0, 1]$. We begin by defining a clipping operator $[X]_\Delta = \sqrt{X^2 - \Delta^2}\mathbb{I}\{|X| \geq \Delta\}$ and upper bounding

$$\mathbb{E}_{z\sim\mu}[|f_z(\pi_z)|] \leq \Delta + \mathbb{E}_{z\sim\mu}[[f_z(\pi_z)]_\Delta].$$

Next, observe that for any $\varepsilon \in (0, 1]$,

$$\mathbb{E}_{z\sim\mu}[[f_z(\pi_z)]_\Delta] = \mathbb{E}_{z\sim\mu}\left[\frac{[f_z(\pi_z)]_\Delta}{(\mathbb{E}_{z'\sim\mu}[f_z^2(\pi_{z'})] \vee \varepsilon^2)^{1/2}}\big(\mathbb{E}_{z'\sim\mu}\big[f_z^2(\pi_{z'})\big] \vee \varepsilon^2\big)^{1/2}\right].$$

By Cauchy Schwarz and the AM-GM inequality, for all $\eta > 0$ this is bounded by

$$\frac{1}{2\eta}\,\mathbb{E}_{z\sim\mu}\left[\frac{[f_z(\pi_z)]_\Delta^2}{\mathbb{E}_{z'\sim\mu}[f_z^2(\pi_{z'})] \vee \varepsilon^2}\right] + \frac{\eta}{2}\,\mathbb{E}_{z\sim\mu}\,\mathbb{E}_{z'\sim\mu}\big[f_z^2(\pi_{z'})\big] + \frac{\eta}{2}\varepsilon^2.$$

The second term above matches the right-hand side of (144), and the ε^2 term will be shown to contribute a negligible error, so it remains to upper bound the first term. To begin, we have

$$\mathbb{E}_{z\sim\mu}\left[\frac{[f_z(\pi_z)]_\Delta^2}{\mathbb{E}_{z'\sim\mu}[f_z^2(\pi_{z'})] \vee \varepsilon^2}\right] \leq \mathbb{E}_{z\sim\mu}\left[\sup_{f\in\mathcal{F}}\frac{[f(\pi_z)]_\Delta^2}{\mathbb{E}_{z'\sim\mu}[f^2(\pi_{z'})] \vee \varepsilon^2}\right].$$

Observe that for any $\varepsilon \leq X \leq 1$, we have

$$\frac{1}{X^2} = 2\int_X^1 \frac{1}{t^3}dt + 1 = 2\int_\varepsilon^1 \frac{1}{t^3}\mathbb{I}\{t \geq X\}dt + 1.$$

Since $|f| \leq 1$, this means that we can upper bound

$$\mathbb{E}_{z\sim\mu}\left[\sup_{f\in\mathcal{F}}\frac{[f(\pi_z)]_\Delta^2}{\mathbb{E}_{z'\sim\mu}[f^2(\pi_{z'})] \vee \varepsilon^2}\right] \leq 1 + 2\,\mathbb{E}_{z\sim\mu}\left[\sup_{f\in\mathcal{F}}\int_\varepsilon^1 \frac{[f(\pi_z)]_\Delta^2}{\epsilon^3}\mathbb{I}\big\{\mathbb{E}_{z'\sim\mu}\big[f^2(\pi_{z'})\big] \leq \epsilon^2\big\}d\epsilon\right].$$

Applying a similar trick, for any $\Delta \leq X \leq 1$ we have

$$[X]_\Delta^2 = X^2 - \Delta^2 = \int_{\Delta^2}^1 \mathbb{I}\{X^2 > t\}dt = 2\int_\Delta^1 \mathbb{I}\{X > t\}tdt.$$

This leads to the upper bound

$$
\begin{aligned}
&2\,\mathbb{E}_{z\sim\mu}\left[\sup_{f\in\mathcal{F}}\int_{\varepsilon}^{1}\frac{[f(\pi_z)]_{\Delta}^{2}}{\epsilon^3}\mathbb{I}\big\{\mathbb{E}_{z'\sim\mu}\big[f^2(\pi_{z'})\big]\leq\epsilon^2\big\}d\epsilon\right]\\
&\leq 4\,\mathbb{E}_{z\sim\mu}\left[\sup_{f\in\mathcal{F}}\int_{\varepsilon}^{1}\int_{\Delta}^{1}\frac{\delta}{\epsilon^3}\mathbb{I}\big\{|f(\pi_z)|>\delta\wedge\mathbb{E}_{z'\sim\mu}\big[f^2(\pi_{z'})\big]\leq\epsilon^2\big\}d\delta d\epsilon\right]\\
&\leq 4\,\mathbb{E}_{z\sim\mu}\left[\int_{\varepsilon}^{1}\int_{\Delta}^{1}\frac{\delta}{\epsilon^3}\mathbb{I}\big\{\exists f\in\mathcal{F}:|f(\pi_z)|>\delta\wedge\mathbb{E}_{z'\sim\mu}\big[f^2(\pi_{z'})\big]\leq\epsilon^2\big\}d\delta d\epsilon\right]\\
&\leq 4\int_{\varepsilon}^{1}\int_{\Delta}^{1}\frac{\delta}{\epsilon^3}\mathbb{P}_{z\sim\mu}\big(\exists f\in\mathcal{F}:|f(\pi_z)|>\delta\wedge\mathbb{E}_{z'\sim\mu}\big[f^2(\pi_{z'})\big]\leq\epsilon^2\big)d\delta d\epsilon.
\end{aligned}
$$

Now, from the definition of the disagreement coefficient, we have that for all $\epsilon\geq\varepsilon$ and $\delta\geq\Delta$,

$$
\mathbb{P}_{z\sim\mu}\big(\exists f\in\mathcal{F}:|f(\pi_z)|>\delta\wedge\mathbb{E}_{z'\sim\mu}\big[f^2(\pi_{z'})\big]\leq\epsilon^2\big)\leq\frac{\epsilon^2}{\delta^2}\boldsymbol{\theta}(\mathcal{F},\Delta,\varepsilon;\rho_\mu).
$$

It follows that

$$
\begin{aligned}
&\int_{\varepsilon}^{1}\int_{\Delta}^{1}\frac{\delta}{\epsilon^3}\mathbb{P}_{z\sim\mu}\big(\exists f\in\mathcal{F}:|f(\pi_z)|>\delta\wedge\mathbb{E}_{z'\sim\mu}\big[f^2(\pi_{z'})\big]\leq\epsilon^2\big)d\delta d\epsilon\\
&\qquad\leq\boldsymbol{\theta}(\mathcal{F},\Delta,\varepsilon;\rho_\mu)\cdot\int_{\varepsilon}^{1}\int_{\Delta}^{1}\frac{1}{\delta\epsilon}d\delta d\epsilon\\
&\qquad=\boldsymbol{\theta}(\mathcal{F},\Delta,\varepsilon;\rho_\mu)\cdot\log(1/\varepsilon)\log(1/\Delta).
\end{aligned}
$$

Altogether, we have

$$
\mathbb{E}_{z\sim\mu}[|f_z(\pi_z)|]\leq\Delta+\frac{\eta\varepsilon^2}{2}+\frac{1}{2\eta}(4\boldsymbol{\theta}(\mathcal{F},\Delta,\varepsilon;\rho_\mu)\cdot\log(1/\varepsilon)\log(1/\Delta)+1)+\frac{\eta}{2}\,\mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big].
$$

We conclude by tuning the parameters in this bound to derive the theorem statement. First note that for any $R\geq 1$, we can apply the result above to the class $\mathcal{F}/R$ to get

$$
\begin{aligned}
&\mathbb{E}_{z\sim\mu}[|f_z(\pi_z)|]\\
&\leq\inf_{\Delta,\varepsilon\in(0,1],\eta>0}\left\{\Delta R+\frac{\eta\varepsilon^2 R}{2}+\frac{R}{2\eta}(4\boldsymbol{\theta}(\mathcal{F}/R,\Delta,\varepsilon;\rho_\mu)\cdot\log(1/\varepsilon)\log(1/\Delta)+1)+\frac{\eta}{2R}\,\mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big]\right\}\\
&=\inf_{\Delta,\varepsilon\in(0,R],\eta>0}\left\{\Delta+\frac{\eta\varepsilon^2}{2R}+\frac{R}{2\eta}(4\boldsymbol{\theta}(\mathcal{F}/R,\Delta/R,\varepsilon/R;\rho_\mu)\cdot\log(R/\varepsilon)\log(R/\Delta)+1)+\frac{\eta}{2R}\,\mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big]\right\}\\
&=\inf_{\Delta,\varepsilon\in(0,R],\eta>0}\left\{\Delta+\frac{\eta\varepsilon^2}{2R}+\frac{R}{2\eta}(4\boldsymbol{\theta}(\mathcal{F},\Delta,\varepsilon;\rho_\mu)\cdot\log(R/\varepsilon)\log(R/\Delta)+1)+\frac{\eta}{2R}\,\mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big]\right\}.
\end{aligned}
$$

We choose $\eta=2\gamma R$ and $\varepsilon=\gamma^{-1}\wedge R$ to get

$$
\inf_{\Delta\in(0,R]}\left\{\Delta+2\gamma^{-1}+\frac{\boldsymbol{\theta}(\mathcal{F},\Delta,\gamma^{-1}\wedge R;\rho_\mu)\cdot\log(R\gamma\vee 1)\log(R/\Delta)}{\gamma}+\gamma\,\mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big]\right\}.
$$

For $\gamma^{-1}>R$, we have $\boldsymbol{\theta}(\mathcal{F},\Delta,\gamma^{-1};\rho_\mu)=\boldsymbol{\theta}(\mathcal{F},\Delta,R;\rho_\mu)=1$, so we can simplify to $\boldsymbol{\theta}(\mathcal{F},\Delta,\gamma^{-1};\rho_\mu)$. We proceed to use that $\boldsymbol{\theta}(\cdots)\geq 1$ to further simplify to

$$
\begin{aligned}
&\inf_{\Delta\in(0,R]}\left\{\Delta+3\frac{\boldsymbol{\theta}(\mathcal{F},\Delta,\gamma^{-1};\rho_\mu)\cdot\log(R\gamma\vee e)\log(R/\Delta\vee e)}{\gamma}+\gamma\,\mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big]\right\}\\
&\leq\inf_{\Delta\in(\gamma^{-1},R]}\left\{\Delta+3\frac{\boldsymbol{\theta}(\mathcal{F},\Delta,\gamma^{-1};\rho_\mu)\cdot\log^2(R\gamma\vee e)}{\gamma}+\gamma\,\mathbb{E}_{z,z'\sim\mu}\big[f_z^2(\pi_{z'})\big]\right\}.
\end{aligned}
$$

Let $f(\Delta) = 3\frac{\boldsymbol{\theta}(\mathcal{F},\Delta,\gamma^{-1};\rho_\mu)\cdot\log^2(R\gamma\vee e)}{\gamma}$, and let us upper bound by $2\inf_{\Delta\in(\gamma^{-1},R]}\max\{\Delta, f(\Delta)\}$. Since $f(\Delta) \leq 1/\gamma$, we have that $\max\{\Delta, f(\Delta)\} = f(\Delta)$ for $\Delta \leq 1/\gamma$. Since $f(\Delta)$ is decreasing, the unconstrained minimizer must have $\Delta > 1/\gamma$. Similarly, since $f(\Delta)$ is constant for $\Delta \geq R$ and Δ is increasing, the unconstrained minimizer must have $\Delta \leq R$ as well. We conclude that the unconstrained minimizer over Δ is contained in the range $[\gamma^{-1}, R]$, so we can relax the infimum to the range $(0,\infty]$, which gives the final theorem statement. $\square$

E.2.2 Proof of Theorem 6.1 and Theorem 6.2

Proof of Theorem 6.1. By the minimax theorem (cf. Proposition 4.2), it suffices to bound the dual Decision-Estimation Coefficient $\underline{\mathsf{dec}}^{\mathrm{Sq}}_\gamma(\mathcal{M}, \overline{M}) = \sup_{\mu\in\Delta(\mathcal{M})}\underline{\mathsf{dec}}^{\mathrm{Sq}}_\gamma(\mu, \overline{M})$. The bound on this quantity is an immediate consequence of Theorem 6.2 and Lemma 6.1. $\square$

Proof of Theorem 6.2. Let $\mu \in \Delta(\mathcal{M})$ be given, and recall that we define $\rho_\mu(\pi) := \mu(\{\pi_M = \pi\})$ as the posterior sampling distribution. We first observe that by the AM-GM inequality, for all $\gamma > 0$,

$$\mathbb{E}_{\pi\sim\rho_\mu}\mathbb{E}_{M\sim\mu}[f^M(\pi_M) - f^M(\pi)] \leq \mathbb{E}_{\pi\sim\rho_\mu}\mathbb{E}_{M\sim\mu}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] + \frac{\gamma}{2}\mathbb{E}_{\pi\sim\rho_\mu}\mathbb{E}_{M\sim\mu}\big[(f^M(\pi) - f^{\overline{M}}(\pi))^2\big] + (2\gamma)^{-1}.$$

Next, we note that, since $\pi \sim \rho_\mu$ and π_M under $M \sim \mu$ are identical in law,

$$\mathbb{E}_{\pi\sim\rho_\mu}\mathbb{E}_{M\sim\mu}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] = \mathbb{E}_{M\sim\mu}\big[f^M(\pi_M) - f^{\overline{M}}(\pi_M)\big].$$

The result now follows by applying Lemma E.2 with the class $\mathcal{F}_\mathcal{M} - f^{\overline{M}}$ and parameter $\frac{\gamma}{2}$. $\square$

F Proofs and Additional Results from Section 7

F.1 Technical Tools

Lemma F.1 (Bellman residual decomposition)**.** *For any pair of MDPs $M = (P^M, R^M)$ and $\overline{M} = (P^{\overline{M}}, R^{\overline{M}})$ with the same initial state distribution,*

$$f^M(\pi) - f^{\overline{M}}(\pi) = \sum_{h=1}^H \mathbb{E}^{\overline{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1})\big] \tag{145}$$

for all policies $\pi \in \Pi_{\mathrm{RNS}}$.

Proof of Lemma F.1. First, we have

$$\begin{aligned}\sum_{h=1}^H \mathbb{E}^{\overline{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1})\big] &= \sum_{h=1}^H \mathbb{E}^{\overline{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - V_{h+1}^{M,\pi}(s_{h+1})\big] - \mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^H r_h\right] \\ &= \sum_{h=1}^H \mathbb{E}^{\overline{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - V_{h+1}^{M,\pi}(s_{h+1})\big] - f^{\overline{M}}(\pi).\end{aligned}$$

On the other hand, since $V_h^{M,\pi}(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_h^{M,\pi}(s,a)]$, a telescoping argument yields

$$\begin{aligned}\sum_{h=1}^H \mathbb{E}^{\overline{M},\pi}\big[Q_h^{M,\pi}(s_h,a_h) - V_{h+1}^{M,\pi}(s_{h+1})\big] &= \sum_{h=1}^H \mathbb{E}^{\overline{M},\pi}\big[V_h^{M,\pi}(s_h) - V_{h+1}^{M,\pi}(s_{h+1})\big] \\ &= \mathbb{E}^{\overline{M},\pi}\big[V_1^{M,\pi}(s_1)\big] - \mathbb{E}^{\overline{M},\pi}\big[V_{H+1}^{M,\pi}(s_{H+1})\big] \\ &= f^M(\pi),\end{aligned}$$

where we have used that $V_{H+1}^{M,\pi} = 0$, and that both MDPs have the same initial state distribution. $\square$

Lemma F.2 (Global simulation lemma)**.** *Let M and M' be MDPs with $\sum_{h=1}^{H} r_h \in [0,1]$ almost surely, and let $\pi \in \Pi_{\mathrm{RNS}}$. Then we have*

$$\left|f^M(\pi) - f^{M'}(\pi)\right| \leq D_{\mathrm{TV}}(M(\pi), M'(\pi)) \tag{146}$$

$$\leq D_{\mathrm{H}}(M(\pi), M'(\pi)) \leq \frac{1}{2\eta} + \frac{\eta}{2} D^2_{\mathrm{H}}(M(\pi), M'(\pi)) \quad \forall \eta > 0. \tag{147}$$

Proof of Lemma F.2. Let $X = \sum_{h=1}^{H} r_h$. Since $X \in [0,1]$ almost surely, we have

$$\left|f^M(\pi) - f^{M'}(\pi)\right| = \left|\mathbb{E}^{M,\pi}[X] - \mathbb{E}^{M',\pi}[X]\right| \leq D_{\mathrm{TV}}(M(\pi), M'(\pi)) \leq D_{\mathrm{H}}(M(\pi), M'(\pi)),$$

where the second inequality is from Lemma A.5. The final result now follows from the AM-GM inequality. $\square$

Lemma F.3 (Local simulation lemma)**.** *For any pair of MDPs $M = (P^M, R^M)$ and $\overline{M} = (P^{\overline{M}}, R^{\overline{M}})$ with the same initial state distribution and $\sum_{h=1}^{H} r_h \in [0,1]$,*

$$f^M(\pi) - f^{\overline{M}}(\pi) = \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\big[\big[(P^M_h - P^{\overline{M}}_h) V^{M,\pi}_{h+1}\big](s_h, a_h)\big] + \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\Big[\mathbb{E}_{r_h \sim R^M_h(s_h,a_h)}[r_h] - \mathbb{E}_{r_h \sim R^{\overline{M}}_h(s_h,a_h)}[r_h]\Big] \tag{148}$$

$$\leq \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\big[D_{\mathrm{TV}}\big(P^M_h(s_h,a_h), P^{\overline{M}}_h(s_h,a_h)\big) + D_{\mathrm{TV}}\big(R^M_h(s_h,a_h), R^{\overline{M}}_h(s_h,a_h)\big)\big]. \tag{149}$$

Proof of Lemma F.3. From Lemma F.1, we have

$$\begin{aligned}
f^M(\pi) - f^{\overline{M}}(\pi) &= \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\big[Q^{M,\pi}_h(s_h,a_h) - r_h - V^{M,\pi}_{h+1}(s_{h+1})\big] \\
&= \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\Big[\big[P^M_h V^{M,\pi}_{h+1}\big](s_h,a_h) - V^{M,\pi}_{h+1}(s_{h+1}) + \mathbb{E}_{r_h \sim R^M_h(s_h,a_h)}[r_h] - \mathbb{E}_{r_h \sim R^{\overline{M}}_h(s_h,a_h)}[r_h]\Big] \\
&= \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\big[\big[(P^M_h - P^{\overline{M}}_h) V^{M,\pi}_{h+1}\big](s_h,a_h)\big] + \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\Big[\mathbb{E}_{r_h \sim R^M_h(s_h,a_h)}[r_h] - \mathbb{E}_{r_h \sim R^{\overline{M}}_h(s_h,a_h)}[r_h]\Big] \\
&\leq \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi}\big[D_{\mathrm{TV}}\big(P^M_h(s_h,a_h), P^{\overline{M}}_h(s_h,a_h)\big) + D_{\mathrm{TV}}\big(R^M_h(s_h,a_h), R^{\overline{M}}_h(s_h,a_h)\big)\big],
\end{aligned}$$

where we have used that $V^{M,\pi}_{h+1}(s) \in [0,1]$.

$\square$

Lemma 5.2 (Change of measure for reinforcement learning)**.** *Consider any family of MDPs $\mathcal{M}$ and reference MDP $\overline{M}$, all of which satisfy $\sum_{h=1}^{H} r_h \in [0,1]$. Suppose that $\mu \in \Delta(\mathcal{M})$ and $p \in \Delta(\Pi_{\mathrm{RNS}})$ are such that*

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] \tag{52}$$

$$\leq C_1 + C_2 \cdot \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D^2_{\mathrm{H}}\big(P^M(s_h,a_h), P^{\overline{M}}(s_h,a_h)\big) + D^2_{\mathrm{H}}\big(R^M(s_h,a_h), R^{\overline{M}}(s_h,a_h)\big)\right], \tag{53}$$

for parameters $C_1, C_2 > 0$. Then for all $\eta > 0$, it holds that

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \leq C_1 + (4\eta)^{-1} + (40HC_2 + \eta) \cdot \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \tag{54}$$

Proof of Lemma 5.2. We first relate the left-hand side of (53) to that of (54) using Lemma F.2, which implies that for all $\eta > 0$,

$$\begin{aligned}
&\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \\
&\leq \mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi)\big] + \eta\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big)\big] + \frac{1}{4\eta}.
\end{aligned}$$

We now relate the right-hand sides of (53) and (54). By Lemma A.11, since $D^2_{\mathrm{H}}(\cdot,\cdot) \leq 2$, we have that for any fixed draw of M and π,

$$\begin{aligned}
&\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D^2_{\mathrm{H}}\big(P^{M}(s_h,a_h),P^{\overline{M}}(s_h,a_h)\big) + D^2_{\mathrm{H}}\big(R^{M}(s_h,a_h),R^{\overline{M}}(s_h,a_h)\big)\right] \\
&\leq 3\,\mathbb{E}^{M,\pi}\left[\sum_{h=1}^{H} D^2_{\mathrm{H}}\big(P^{M}(s_h,a_h),P^{\overline{M}}(s_h,a_h)\big) + D^2_{\mathrm{H}}\big(R^{M}(s_h,a_h),R^{\overline{M}}(s_h,a_h)\big)\right] + 16H D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big),
\end{aligned}$$

Next, we recall from Lemma A.9 that for all h,

$$\mathbb{E}^{M,\pi}\big[D^2_{\mathrm{H}}\big(P^{M}(s_h,a_h),P^{\overline{M}}(s_h,a_h)\big)\big] + \mathbb{E}^{M,\pi}\big[D^2_{\mathrm{H}}\big(R^{M}(s_h,a_h),R^{\overline{M}}(s_h,a_h)\big)\big] \leq 8 D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big).$$

As a result, we have

$$\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D^2_{\mathrm{H}}\big(P^{M}(s_h,a_h),P^{\overline{M}}(s_h,a_h)\big) + D^2_{\mathrm{H}}\big(R^{M}(s_h,a_h),R^{\overline{M}}(s_h,a_h)\big)\right] \leq 8H D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big).$$

Since this holds uniformly for all M and π, we conclude that

$$\begin{aligned}
&\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\,\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D^2_{\mathrm{TV}}\big(P^{M}(s_h,a_h),P^{\overline{M}}(s_h,a_h)\big) + D^2_{\mathrm{TV}}\big(R^{M}(s_h,a_h),R^{\overline{M}}(s_h,a_h)\big)\right] \\
&\leq 40H\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi),\overline{M}(\pi)\big)\big].
\end{aligned}$$

□

F.2 Proofs from Section 7.1

In this section of the appendix we prove Theorem 7.1 and Theorem 7.2. Before proving the results, we state and prove a helper lemma.

Lemma F.4. *Let $\mathcal{M}$ be a bilinear class, and let π^{α}_{M} be defined as in Section 7.1.1. Then for all h, for any $M, M', \overline{M} \in \mathcal{M}$,*

$$\big\langle X_h(M;\overline{M}), W_h(M';\overline{M})\big\rangle^2 \leq 4C_\alpha L^2_{\mathrm{bi}}(\mathcal{M})\cdot D^2_{\mathrm{H}}\big(M'(\pi^{\alpha}_{M}),\overline{M}(\pi^{\alpha}_{M})\big), \tag{150}$$

where $C_\alpha = 1$ when $\pi^{\mathrm{est}}_{M} = \pi_{M}$ and $C_\alpha \leq \frac{2H}{\alpha}$ in general, as long as $\alpha \in [0, 1/2]$.

Proof of Lemma F.4. Since $W(M';M') = 0$, we have

$$\begin{aligned}
\big|\big\langle X_h(M;\overline{M}), W_h(M';\overline{M})\big\rangle\big| &= \big|\big\langle X_h(M;\overline{M}), W_h(M';\overline{M})\big\rangle - \langle X_h(M;M'), W_h(M';M')\rangle\big| \\
&= \Big|\mathbb{E}^{\overline{M},\pi_M\circ_h\pi^{\mathrm{est}}_{M}}\big[\ell^{\mathrm{est}}_{M}(M';z_h)\big] - \mathbb{E}^{M',\pi_M\circ_h\pi^{\mathrm{est}}_{M}}\big[\ell^{\mathrm{est}}_{M}(M';z_h)\big]\Big| \\
&= \Big|\mathbb{E}^{\overline{M},\pi_M\circ_h\pi^{\mathrm{est}}_{M}\circ_{h+1}\pi_M}\big[\ell^{\mathrm{est}}_{M}(M';z_h)\big] - \mathbb{E}^{M',\pi_M\circ_h\pi^{\mathrm{est}}_{M}\circ_{h+1}\pi_M}\big[\ell^{\mathrm{est}}_{M}(M';z_h)\big]\Big|.
\end{aligned}$$

Hence, since z_h is a measurable function of the learner's trajectory τ_H, and since $|\ell^{\mathrm{est}}_{M}(M';\cdot)| \leq L_{\mathrm{bi}}(\mathcal{M})$ almost surely under both models, we have

$$\begin{aligned}\left|\left\langle X_h(M;\overline{M}), W_h(M';\overline{M})\right\rangle\right| &\leq 2L_{\mathrm{bi}}(\mathcal{M}) D_{\mathrm{TV}}\big(M'(\pi_M \circ_h \pi^{\mathrm{est}}_M \circ_{h+1} \pi_M), \overline{M}(\pi_M \circ_h \pi^{\mathrm{est}}_M \circ_{h+1} \pi_M)\big) \\ &\leq 2L_{\mathrm{bi}}(\mathcal{M}) D_{\mathrm{H}}\big(M'(\pi_M \circ_h \pi^{\mathrm{est}}_M \circ_{h+1} \pi_M), \overline{M}(\pi_M \circ_h \pi^{\mathrm{est}}_M \circ_{h+1} \pi_M)\big).\end{aligned}$$

If $\pi^{\mathrm{est}}_M = \pi_M$ we are done. Otherwise, let $b_1, \ldots, b_H \in \{0,1\}$, and let $\pi^b_{M,h}(s) := b \cdot \pi_{M,h}(s) + (1-b) \cdot \pi^{\mathrm{est}}_{M,h}(s)$. For any MDP, trajectories induced by π^α_M are equivalent in law to those generated by sampling $b_1, \ldots, b_H \sim \mathrm{Ber}((1-\alpha/H))$ independently and playing π^b_M. As a result, since Hellinger distance has $D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X} \otimes \mathbb{P}_X, \mathbb{Q}_{Y|X} \otimes \mathbb{P}_X\big) = \mathbb{E}_{X \sim \mathbb{P}_X}\big[D^2_{\mathrm{H}}\big(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\big)\big]$, we have

$$\begin{aligned}D^2_{\mathrm{H}}\big(M'(\pi^\alpha_M), \overline{M}(\pi^\alpha_M)\big) &= \mathbb{E}_{b_1,\ldots,b_H}\big[D^2_{\mathrm{H}}\big(M'(\pi^b_M), \overline{M}(\pi^b_M)\big)\big] \\ &\geq (\alpha/H)(1-\alpha/H)^{H-1} \cdot D^2_{\mathrm{H}}\big(M'(\pi_M \circ_h \pi^{\mathrm{est}}_M \circ_{h+1} \pi_M), \overline{M}(\pi_M \circ_h \pi^{\mathrm{est}}_M \circ_{h+1} \pi_M)\big).\end{aligned}$$

Finally, we note that $(1-\alpha/H)^{H-1} \geq 1-\alpha \geq 1/2$ whenever $\alpha \leq 1/2$.

□

Theorem 7.1. *Let $\mathcal{M}$ be a bilinear class and let $\overline{M} \in \mathcal{M}$. Let $\mu \in \Delta(\mathcal{M})$ be given, and consider the modified posterior sampling strategy that samples $M \sim \mu$ and plays π^α_M, where $\alpha \in [0,1]$ is a parameter.*

- *If $\pi^{\mathrm{est}}_M = \pi_M$ (i.e., estimation is on-policy), this strategy with $\alpha = 0$ certifies that*

$$\underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}) \leq \frac{4H^2 L^2_{\mathrm{bi}}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}$$

for all $\gamma > 0$.

- *For general estimation policies, this strategy with $\alpha = \big(\frac{8H^3 L^2_{\mathrm{bi}}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M},\overline{M})}{\gamma}\big)^{1/2}$ certifies that*

$$\underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}) \leq \left(\frac{32H^3 L^2_{\mathrm{bi}}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}\right)^{1/2}$$

whenever $\gamma \geq 32H^3 L^2_{\mathrm{bi}}(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$.

Proof of Theorem 7.1. Throughout this proof we abbreviate $d \equiv d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$. Let a prior $\mu \in \Delta(\mathcal{M})$ be fixed. Given Lemma F.2, it suffices to bound the quantity

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big]$$

in terms of the Hellinger error $\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]$. To begin, we write

$$\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] = \mathbb{E}_{M\sim\mu}[f^M(\pi_M)] - \mathbb{E}_{\pi\sim p}\big[f^{\overline{M}}(\pi)\big].$$

From the definition of p, this is equal to

$$\mathbb{E}_{M\sim\mu}[f^M(\pi_M)] - \mathbb{E}_{M\sim\mu}\big[f^{\overline{M}}(\pi^\alpha_M)\big] \leq \mathbb{E}_{M\sim\mu}\big[f^M(\pi_M) - f^{\overline{M}}(\pi_M)\big] + \alpha,$$

where we have used that

$$\mathbb{P}^{\overline{M},\pi^\alpha_M}\big(\forall h : \pi^\alpha_{M,h}(s_h) = \pi_{M,h}(s_h)\big) \geq (1-\alpha/H)^H \geq 1-\alpha,$$

and that $f^{\overline{M}} \in [0,1]$.

To proceed, using the Bellman residual decomposition (Lemma F.1) and applying the first bilinear class property, we have

$$\begin{aligned}\mathbb{E}_{M\sim\mu}\big[f^M(\pi_M) - f^{\overline{M}}(\pi_M)\big] &= \mathbb{E}_{M\sim\mu}\left[\sum_{h=1}^H \mathbb{E}^{\overline{M},\pi_M}\left[Q^{M,\star}_h(s_h,a_h) - r_h - V^{M,\star}_{h+1}(s_{h+1})\right]\right] \\ &\leq \sum_{h=1}^H \mathbb{E}_{M\sim\mu}\big[\big|\big\langle X_h(M;\overline{M}), W_h(M;\overline{M})\big\rangle\big|\big].\end{aligned}$$

Let $h \in [H]$ be fixed and abbreviate $X_h(M) \equiv X_h(M; \overline{M})$ and $W_h(M) \equiv W_h(M; \overline{M})$. Define $\Sigma := \mathbb{E}_{M\sim\mu}\left[X_h(M)X_h(M)^\top\right]$. Then by Cauchy-Schwarz,

$$
\begin{aligned}
\mathbb{E}_{M\sim\mu}[|\langle W_h(M), X_h(M)\rangle|] &= \mathbb{E}_{M\sim\mu}\left[\left|\langle \Sigma^{1/2}W_h(M), (\Sigma^\dagger)^{1/2}X_h(M)\rangle\right|\right] \\
&\leq \sqrt{\mathbb{E}_{M\sim\mu}\left\|\Sigma^{1/2}W_h(M)\right\|_2^2} \cdot \sqrt{\mathbb{E}_{M\sim\mu}\left\|(\Sigma^\dagger)^{1/2}X_h(M)\right\|_2^2},
\end{aligned}
\tag{151}
$$

where the first inequality uses that $X_h(M) \in \operatorname{span}(\Sigma)$ almost surely. For the first term in (151), we have

$$
\begin{aligned}
\mathbb{E}_{M\sim\mu}\left\|\Sigma^{1/2}W_h(M)\right\|_2^2 &= \mathbb{E}_{M\sim\mu}[\langle \Sigma W_h(M), W_h(M)\rangle] \\
&= \mathbb{E}_{M'\sim\mu}\left[\left\langle \mathbb{E}_{M\sim\mu}\left[X_h(M)X_h(M)^\top\right]W_h(M'), W_h(M')\right\rangle\right] \\
&\leq \mathbb{E}_{M,M'\sim\mu}\left[\langle X_h(M), W_h(M')\rangle^2\right].
\end{aligned}
$$

We bound the second term in (151) by writing

$$
\mathbb{E}_{M\sim\mu}\left\|(\Sigma^\dagger)^{1/2}X_h(M)\right\|_2^2 = \left\langle \Sigma^\dagger, \mathbb{E}_{M\sim\mu}\left[X_h(M)X_h(M)^\top\right]\right\rangle = \langle \Sigma^\dagger, \Sigma\rangle \leq d.
$$

By the AM-GM inequality, we conclude that for any $\eta > 0$,

$$
\mathbb{E}_{M\sim\mu}\left[f^M(\pi_M) - f^{\overline{M}}(\pi_M)\right] \leq \frac{dH}{2\eta} + \frac{\eta}{2}\sum_{h=1}^H \mathbb{E}_{M,M'\sim\mu}\left[\left\langle X_h(M;\overline{M}), W_h(M';\overline{M})\right\rangle^2\right].
\tag{152}
$$

Using Lemma F.4, this implies that whenever $\alpha \leq 1/2$,

$$
\begin{aligned}
\mathbb{E}_{M\sim\mu}\left[f^M(\pi_M) - f^{\overline{M}}(\pi_M)\right] &\leq \frac{dH}{2\eta} + 2\eta H C_\alpha L_{\mathrm{bi}}^2(\mathcal{M})\,\mathbb{E}_{M,M'\sim\mu}\left[D_{\mathrm{H}}^2\left(M'(\pi_M^\alpha), \overline{M}(\pi_M^\alpha)\right)\right] \\
&= \frac{dH}{2\eta} + 2\eta H C_\alpha L_{\mathrm{bi}}^2(\mathcal{M})\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\left[D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right)\right],
\end{aligned}
$$

where $C_\alpha = 1$ when $\pi_M^{\mathrm{est}} = \pi_M$ and $C_\alpha = \frac{2H}{\alpha}$ otherwise.

Altogether, we have shown that

$$
\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^{\overline{M}}(\pi)\right] \leq \frac{dH}{2\eta} + 2\eta H C_\alpha L_{\mathrm{bi}}^2(\mathcal{M})\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\left[D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right)\right] + \alpha,
$$

which, via Lemma F.2, implies that for any $\eta' > 0$,

$$
\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \leq \frac{dH}{2\eta} + \frac{1}{2\eta'} + (2\eta H C_\alpha L_{\mathrm{bi}}^2(\mathcal{M}) + \eta'/2)\,\mathbb{E}_{M\sim\mu}\,\mathbb{E}_{\pi\sim p}\left[D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right)\right] + \alpha.
$$

We set $\eta = \eta' = \frac{\gamma}{4HC_\alpha L_{\mathrm{bi}}^2(\mathcal{M})}$ to conclude that $\underline{\mathsf{dec}}_\gamma(\mathcal{M}, \overline{M}) \leq \frac{4H^2 C_\alpha L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M},\overline{M})}{\gamma} + \alpha$. In the on-policy case we take $\alpha = 0$, and in the general case, we set $\alpha = \sqrt{\frac{8H^3 L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M},\overline{M})}{\gamma}}$, which is admissible whenever $\gamma \geq 16H^3 L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$. □

Theorem 7.2. *Let $\mathcal{M}$ be a bilinear class. Let $\gamma > 0$ and $\overline{M} \in \mathcal{M}$ be given, and consider the* **PC-IGW.Bilinear** *strategy in Algorithm 5. Suppose the optimal design solver in Line 5 has approximation factor $C_{\mathrm{opt}} \geq 1$.*

- *If $\pi_M^{\mathrm{est}} = \pi_M$ (i.e., estimation is on-policy), this strategy with $\eta = \frac{\gamma}{3H^3 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M},\overline{M})}$ and $\alpha = 0$ certifies that*

$$
\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \leq \frac{9H^3 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}.
$$

- *For general estimation policies, choosing* $\eta = \frac{\gamma}{6H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}$ *and* $\alpha = \left(\frac{18H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}\right)^{1/2}$ *certifies that*

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \le \left(\frac{72H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}\right)^{1/2},$$

whenever $\gamma \ge 72H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$.

Proof of Theorem 7.2. We first verify that the strategy in (56) is indeed well-defined, in the sense that a normalizing constant $\lambda \in [1, 2HSA]$ always exists.

Proposition F.1. *There is a unique choice for* $\lambda > 0$ *such that* $\sum_{M \in \mathcal{M}} p(\pi^\alpha_M) = 1$, *and its value lies in* $[1/2, 1]$.

Proof. Let $f(\lambda) = \sum_{M \in \mathcal{M}} \frac{q(M)}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))}$. We observe that if $\lambda > 1$, then $f(\lambda) \le \sum_{M \in \mathcal{M}} \frac{q(M)}{\lambda} = \frac{1}{\lambda} < 1$. On the other hand for $\lambda \in (0, 1/2)$, $f(\lambda) \ge \frac{q(\overline{M})}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{\overline{M}}))} = \frac{1}{2\lambda} > 1$. Hence, since $f(\lambda)$ is continuous and strictly decreasing over $(0, \infty)$, there exists unique $\lambda^\star \in [1/2, 1]$ such that $f(\lambda^\star) = 1$. □

We now show that PC-IGW.Bilinear achieves the stated bound on the DEC. Let $M \in \mathcal{M}$ and $\gamma > 0$ be given. We focus on bounding the quantity

$$\mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big].$$

Throughout the proof we will overload notation and write $M \sim p$ and $\pi^\alpha_M \sim p$ interchangeably.

As in the tabular setting (Proposition 5.6), we begin by writing

$$\mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] = \underbrace{\mathbb{E}_{\pi \sim p}\big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\big]}_{\text{(I)}} + \underbrace{f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}})}_{\text{(II)}}. \tag{153}$$

For the first term (I), we have

$$\begin{aligned}
\mathbb{E}_{\pi \sim p}\big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\big] &= \sum_{M \in \mathcal{M}} p(\pi^\alpha_M)\big(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi^\alpha_M)\big) \\
&\le \sum_{M \in \mathcal{M}} p(\pi^\alpha_M)\big(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)\big) + \alpha \\
&= \sum_{M \in \mathcal{M}} q(M) \frac{f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M))} + \alpha \\
&\le (2\eta)^{-1} + \alpha,
\end{aligned} \tag{154}$$

where we have used that $\lambda \ge 0$ and that $\mathbb{P}^{\overline{M}, \pi^\alpha_M}\big(\forall h : \pi^\alpha_{M,h}(s_h) = \pi_{M,h}(s_h)\big) \ge (1 - \alpha/H)^H \ge 1 - \alpha$.

We now bound the second term (II). Using the Bellman residual decomposition (Lemma F.1) and applying the first bilinear class property, we have

$$\begin{aligned}
f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) &= f^M(\pi_M) - f^{\overline{M}}(\pi_M) - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)) \\
&= \sum_{h=1}^H \mathbb{E}^{\overline{M}, \pi_M}\big[Q^{M,\star}_h(s_h, a_h) - r_h - V^{M,\star}_{h+1}(s_{h+1})\big] - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)) \\
&\le \sum_{h=1}^H \big|\big\langle X_h(M; \overline{M}), W_h(M; \overline{M})\big\rangle\big| - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_M)).
\end{aligned} \tag{155}$$

Let $h \in [H]$ be fixed and abbreviate $X_h(M) \equiv X_h(M;\overline{M})$ and $W_h(M) \equiv W_h(M;\overline{M})$. Define $\Sigma := \mathbb{E}_{M\sim p}\big[X_h(M)X_h(M)^\top\big]$. Then by Cauchy-Schwarz and AM-GM, we have that for any $\eta' > 0$,

$$\begin{aligned}
|\langle W_h(M), X_h(M)\rangle| &= \left|\langle \Sigma^{1/2}W_h(M), (\Sigma^\dagger)^{1/2}X_h(M)\rangle\right| \\
&\leq \frac{\eta'}{2}\big\|\Sigma^{1/2}W_h(M)\big\|_2^2 + \frac{1}{2\eta'}\big\|(\Sigma^\dagger)^{1/2}X_h(M)\big\|_2^2,
\end{aligned} \tag{156}$$

where the first inequality uses that $X_h(M) \in \mathrm{span}(\Sigma)$ almost surely. For the first term in (156), we have

$$\begin{aligned}
\big\|\Sigma^{1/2}W_h(M)\big\|_2^2 &= \langle \Sigma W_h(M), W_h(M)\rangle \\
&= \big\langle \mathbb{E}_{M'\sim p}\big[X_h(M')X_h(M')^\top\big]W_h(M), W_h(M)\big\rangle \\
&\leq \mathbb{E}_{M'\sim p}\big[\langle X_h(M'), W_h(M)\rangle^2\big].
\end{aligned}$$

For the second term in (151), we have

$$\big\|(\Sigma^\dagger)^{1/2}X_h(M)\big\|^2 = X_h(M)^\top\big(\mathbb{E}_{M'\sim p}\big[X_h(M')X_h(M')^\top\big]\big)^\dagger X_h(M).$$

We observe that

$$\begin{aligned}
\mathbb{E}_{M'\sim p}\big[X_h(M')X_h(M')^\top\big] &\geq \frac{1}{2H}\sum_{M'\in\mathcal{M}} \frac{q_h^{\mathrm{opt}}(M')}{\lambda + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M}))}X_h(M')X_h(M')^\top \\
&\geq \frac{1}{2H}\sum_{M'\in\mathcal{M}} \frac{q_h^{\mathrm{opt}}(M')}{1 + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M}))}X_h(M')X_h(M')^\top \\
&= \frac{1}{2H}\sum_{M'\in\mathcal{M}} q_h^{\mathrm{opt}}(M')Y_h(M')Y_h(M')^\top,
\end{aligned}$$

where $Y_h(M') \equiv Y_h(M';\overline{M})$. Hence, by the G-optimal design property for q_h^{opt},

$$\begin{aligned}
&X_h(M)^\top\big(\mathbb{E}_{M'\sim p}\big[X_h(M')X_h(M')^\top\big]\big)^\dagger X_h(M) \\
&= Y_h(M)^\top\big(\mathbb{E}_{M'\sim p}\big[X_h(M')X_h(M')^\top\big]\big)^\dagger Y_h(M)\cdot\big(1 + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M}))\big) \\
&\leq 2H\cdot Y_h(M)^\top\Big(\mathbb{E}_{M'\sim q_h^{\mathrm{opt}}}\big[Y_h(M')Y_h(M')^\top\big]\Big)^\dagger Y_h(M)\cdot\big(1 + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M}))\big) \\
&\leq 2dHC_{\mathrm{opt}}\cdot\big(1 + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M}))\big).
\end{aligned}$$

Returning to (155) and summing over all layers, we conclude that

$$\begin{aligned}
&f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) \\
&\leq \frac{\eta'}{2}\sum_{h=1}^{H}\mathbb{E}_{M'\sim p}\big[\langle X_h(M'), W_h(M)\rangle^2\big] + \frac{dH^2C_{\mathrm{opt}}}{\eta'}\cdot\big(1 + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M}))\big) - (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{M})).
\end{aligned}$$

Choosing $\eta' = \eta dH^2C_{\mathrm{opt}}$ yields

$$f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) \leq \frac{\eta dH^2C_{\mathrm{opt}}}{2}\sum_{h=1}^{H}\mathbb{E}_{M'\sim p}\big[\big\langle X_h(M';\overline{M}), W_h(M;\overline{M})\big\rangle^2\big] + \eta^{-1}.$$

By Lemma F.4, this implies that whenever $\alpha \leq 1/2$,

$$\begin{aligned}
f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) &= 2\eta dH^3C_{\mathrm{opt}}C_\alpha L_{\mathrm{bi}}^2(\mathcal{M})\cdot\mathbb{E}_{M'\sim p}\big[D_{\mathrm{H}}^2\big(M(\pi_{M'}^\alpha), \overline{M}(\pi_{M'}^\alpha)\big)\big] + \eta^{-1} \\
&= 2\eta dH^3C_{\mathrm{opt}}C_\alpha L_{\mathrm{bi}}^2(\mathcal{M})\cdot\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{H}}^2\big(M(\pi), \overline{M}(\pi)\big)\big] + \eta^{-1},
\end{aligned}$$

where $C_\alpha = 1$ when $\pi_M^{\mathrm{est}} = \pi_M$ and $C_\alpha = \frac{2H}{\alpha}$ otherwise.

Altogether, we have shown that

$$\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] \leq 2\eta d H^3 C_{\mathrm{opt}} C_\alpha L_{\mathrm{bi}}^2(\mathcal{M}) \cdot \mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] + 2\eta^{-1} + \alpha$$

which, by Lemma F.2, implies that for any $\eta' > 0$,

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \leq (2\eta d H^3 C_{\mathrm{opt}} C_\alpha L_{\mathrm{bi}}^2(\mathcal{M}) + \eta'/2) \cdot \mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] + 2\eta^{-1} + (2\eta')^{-1} + \alpha.$$

Setting $\eta = \eta' = \frac{\gamma}{3dH^3 C_{\mathrm{opt}} C_\alpha L_{\mathrm{bi}}^2(\mathcal{M})}$ gives $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \leq \frac{9H^3 C_\alpha C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma} + \alpha$. In the on-policy case we take $\alpha = 0$, and in the general case, we set $\alpha = \sqrt{\frac{18H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})}{\gamma}}$, which is admissible whenever $\gamma \geq 72H^4 C_{\mathrm{opt}} L_{\mathrm{bi}}^2(\mathcal{M}) d_{\mathrm{bi}}(\mathcal{M}, \overline{M})$.

□

F.3 Proofs from Section 7.2

Proposition 7.2. *Let $\mathcal{M}$ be a bilinear class. Then for all $\gamma > 0$ and $\nu \in \Delta(\mathcal{M})$,*

$$\underline{\mathsf{dec}}^{\mathrm{bi}}_\gamma(\mathcal{M}, \nu) \leq \frac{H \cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma}.$$

Proof of Proposition 7.2. This proof is a slight modification to that of Theorem 7.1. Resuming from (152) in the proof of Theorem 7.1, we have that for any prior $\mu \in \Delta(\mathcal{M})$, the posterior sampling strategy ($p(\pi) = \mu(\{M : \pi_M = \pi\})$) guarantees that for all $\overline{M} \in \mathcal{M}$, and $\gamma > 0$,

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] \leq \frac{H \cdot d_{\mathrm{bi}}(\mathcal{M})}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{bi}}\big(M(\pi) \parallel \overline{M}(\pi)\big)\big].$$

Next, using Lemma F.1, we have that for all $\gamma > 0$,

$$\begin{aligned}
\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big|f^M(\pi) - f^{\overline{M}}(\pi)\big| &\leq \sum_{h=1}^H \mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big|\big\langle X_h(\pi;\overline{M}), W_h(M;\overline{M})\big\rangle\big| \\
&\leq \frac{H}{2\gamma} + \frac{\gamma}{2}\sum_{h=1}^H \mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big\langle X_h(\pi;\overline{M}), W_h(M;\overline{M})\big\rangle^2 \\
&= \frac{H}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{bi}}\big(M(\pi) \parallel \overline{M}(\pi)\big)\big].
\end{aligned}$$

Combining these results, we have that

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \leq \frac{H \cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma} + \gamma\,\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{bi}}\big(M(\pi) \parallel \overline{M}(\pi)\big)\big].$$

Since this result holds uniformly for all $\overline{M}$ and p itself not depend on $\overline{M}$, this further implies that for all $\nu \in \Delta(\mathcal{M})$,

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \leq \frac{H \cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma} + \gamma\,\mathbb{E}_{\overline{M}\sim\nu}\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{bi}}\big(M(\pi) \parallel \overline{M}(\pi)\big)\big].$$

□

Theorem 7.4. *Suppose that Assumption 7.3 holds, and that estimation is on-policy (i.e., $\pi_M^{\mathrm{est}} = \pi_M$). Consider the* **E2D.Bayes** *algorithm with the optimization problem given in (84). This algorithm guarantees that for any prior $\mu \in \Delta(\mathcal{M})$ and $\gamma > 0$,*

$$\mathbb{E}_{M^\star \sim \mu} \mathbb{E}^{M^\star}[\mathbf{Reg}_{\mathsf{DM}}] \le \widetilde{O}\left(\sqrt{H^2 L_{\mathrm{bi}}^2(\mathcal{M}) \cdot d_{\mathrm{bi}}^2(\mathcal{M}) \cdot T \cdot \mathsf{est}(\mathcal{Q}_{\mathcal{M}}, T)}\right).$$

Consequently, we have $\mathfrak{M}(\mathcal{M}, T) \le \widetilde{O}\big(\sqrt{H^2 L_{\mathrm{bi}}^2(\mathcal{M}) \cdot d_{\mathrm{bi}}^2(\mathcal{M}) \cdot T \cdot \mathsf{est}(\mathcal{Q}_{\mathcal{M}}, T)}\big)$.

Proof of Theorem 7.4. We begin as in the main upper bound for E2D.Bayes (Theorem 3.6). Consider a fixed prior distribution $\mu \in \Delta(\mathcal{M})$. Let $\mathbb{E}[\cdot]$ denote expectation with respect to the joint law over $(M^\star, \mathcal{H}^{(T)})$ when $M^\star \sim \mu$. We have

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

which, by conditional independence, implies that

$$\sum_{t=1}^T \mathbb{E}_{t-1}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\big] = \sum_{t=1}^T \mathbb{E}_{M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\big],$$

where $\mu^{(t)}$ denotes the posterior distribution over $\mathcal{M}$ given $\mathcal{H}^{(t-1)}$. In particular, when $p^{(t)}$ solves (84), we have

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_{M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^{(t)})\big] &\le \sum_{t=1}^T \underline{\mathsf{dec}}_\gamma^{\mathrm{bi}}(\mathcal{M}, \mu^{(t)}) + \gamma \cdot \sum_{t=1}^T \mathbb{E}_{\overline{M}, M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\big[D_{\mathrm{bi}}\big(M^\star(\pi) \,\|\, \overline{M}(\pi)\big)\big] \\ &\le \frac{HT \cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma} + \gamma \cdot \sum_{t=1}^T \mathbb{E}_{\overline{M}, M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\big[D_{\mathrm{bi}}\big(M^\star(\pi) \,\|\, \overline{M}(\pi)\big)\big].\end{aligned}$$

We proceed to bound the estimation error term

$$\sum_{t=1}^T \mathbb{E}_{\overline{M}, M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\big[D_{\mathrm{bi}}\big(M^\star(\pi) \,\|\, \overline{M}(\pi)\big)\big], \tag{157}$$

using the method of confidence sets (Russo and Van Roy, 2013) and the elliptic potential. Define

$$\mathcal{L}_h^{(t)}(f, g) = \sum_{i=1}^{t-1} \Big(f(s_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - \max_a g(s_{h+1}^{(i)}, a)\Big)^2,$$

where we recall that $\tau^{(t)} = (s_1^{(t)}, a_1^{(t)}, r_1^{(t)}), \ldots (s_H^{(t)}, a_H^{(t)}, r_H^{(t)})$ is the trajectory observed at episode t. Let $M_1, \ldots, M_N$ be an ε-cover for $\mathcal{Q}_{\mathcal{M}}$, and abbreviate $Q_h^{i,\star} \equiv Q_h^{M_i,\star}$. Fix $R_1 > 0$ and define a confidence set

$$\mathcal{I}^{(t)} = \left\{ i \in [N] \mid \mathcal{L}_h^{(t)}(Q_h^{i,\star}, Q_{h+1}^{i,\star}) \le \min_{j \in [N]} \mathcal{L}_h^{(t)}(Q_h^{j,\star}, Q_{h+1}^{i,\star}) + R_1^2 \;\; \forall h \in [H] \right\}.$$

Let $\iota : \mathcal{M} \to [N]$ be any map from a model M to its corresponding covering element, and let $\mathcal{M}^{(t)} = \{M \mid \iota(M) \in \mathcal{I}^{(t)}\}$. We have the following guarantee.

Lemma F.5. *Let $\delta \in (0,1)$, and suppose we set $R_1 = C \cdot (\varepsilon^2 T + \log(HTN/\delta))$ for a sufficiently large numerical constant C. Define $R_2^2 := O\big(L_{\mathrm{bi}}^2(\mathcal{M})\big(R_1^2 + \varepsilon^2 T + \log(HTN/\delta)\big)\big)$. Then with probability at least $1 - \delta$, for all $t \in [T]$,*

1. *For all M such that $\iota(M) \in \mathcal{I}^{(t)}$,*

$$\sum_{i=1}^{t-1} \langle X_h(\pi^{(i)}; M^\star), W_h(M; M^\star)\rangle^2 \le R_2^2 \quad \forall h \in [H]. \tag{158}$$

2. $\iota(M^\star) \in \mathcal{I}^{(t)}$.

Let $\delta \in (0,1)$ be fixed, and let R_1 and R_2 be set as in Lemma F.5. Let $\mathcal{E}^{(t)}(M^\star)$ denote the event in which both statements in Lemma F.5 hold at round t.[34] Then, using that $\|X_h(\cdot;\cdot)\|_2, \|W_h(\cdot;\cdot)\|_2 \leq 1$, we have

$$
\begin{aligned}
&\mathbb{E}_{\overline{M}, M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\left[D_{\mathrm{bi}}\big(M^\star(\pi) \,\|\, \overline{M}(\pi)\big)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{\overline{M}, M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\left[\big\langle X_h(\pi; \overline{M}), W_h(M^\star; \overline{M})\big\rangle^2\right] \\
&\leq \sum_{h=1}^{H} \mathbb{E}_{\overline{M}, M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\left[\big\langle X_h(\pi; \overline{M}), W_h(M^\star; \overline{M})\big\rangle^2 \mathbb{I}\{\mathcal{E}^{(t)}(M^\star)\}\right] + \mathbb{E}_{M^\star \sim \mu^{(t)}}[\mathbb{I}\{\neg\mathcal{E}^{(t)}(M^\star)\}]H \\
&\leq \sum_{h=1}^{H} \mathbb{E}_{\overline{M} \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\left[\sup_{M \in \mathcal{M}^{(t)}} \big\langle X_h(\pi; \overline{M}), W_h(M; \overline{M})\big\rangle^2\right] + \mathbb{E}_{M^\star \sim \mu^{(t)}}[\mathbb{I}\{\neg\mathcal{E}^{(t)}(M^\star)\}]H \\
&= \sum_{h=1}^{H} \mathbb{E}_{M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\left[\sup_{M \in \mathcal{M}^{(t)}} \langle X_h(\pi; M^\star), W_h(M; M^\star)\rangle^2\right] + \mathbb{E}_{M^\star \sim \mu^{(t)}}[\mathbb{I}\{\neg\mathcal{E}^{(t)}(M^\star)\}]H,
\end{aligned}
$$

where we have used that $\mathcal{M}^{(t)}$ is $\mathcal{H}^{(t-1)}$-measurable. We further bound

$$
\begin{aligned}
&\mathbb{E}_{M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\left[\sup_{M \in \mathcal{M}^{(t)}} \langle X_h(\pi; M^\star), W_h(M; M^\star)\rangle^2\right] \\
&\leq \mathbb{E}_{M^\star \sim \mu^{(t)}} \mathbb{E}_{\pi \sim p^{(t)}}\left[\sup_{M \in \mathcal{M}^{(t)}} \langle X_h(\pi; M^\star), W_h(M; M^\star)\rangle^2 \mathbb{I}\{\mathcal{E}^{(t)}(M^\star)\}\right] + \mathbb{E}_{M^\star \sim \mu^{(t)}}[\mathbb{I}\{\neg\mathcal{E}^{(t)}(M^\star)\}].
\end{aligned}
$$

Next, using the definition of $\mathcal{E}^{(t)}$,

$$
\begin{aligned}
&\sup_{M \in \mathcal{M}^{(t)}} \langle X_h(\pi, M^\star), W_h(M; M^\star)\rangle^2 \mathbb{I}\{\mathcal{E}^{(t)}(M^\star)\} \\
&\leq \sup_{M}\left\{ \langle X_h(\pi; M^\star), W_h(M; M^\star)\rangle^2 \mid \sum_{i=1}^{t-1} \langle X_h(\pi^{(i)}; M^\star), W_h(M; M^\star)\rangle^2 \leq R_2^2 \right\} \\
&\leq \sup_{W}\left\{ \langle X_h(\pi; M^\star), W\rangle^2 \mid \sum_{i=1}^{t-1} \langle X_h(\pi^{(i)}; M^\star), W\rangle^2 \leq R_2^2,\ \|W\|_2^2 \leq 1 \right\} \\
&\leq \sup_{W}\left\{ \langle X_h(\pi; M^\star), W\rangle^2 \mid \sum_{i=1}^{t-1} \langle X_h(\pi^{(i)}; M^\star), W\rangle^2 + \|W\|_2^2 \leq R_2^2 + 1 \right\} \\
&\leq \|X_h(\pi; M^\star)\|^2_{(\Sigma_h^{(t-1)})^{-1}} (R_2^2 + 1),
\end{aligned}
$$

where $\Sigma_h^{(t)} := \sum_{i=1}^{t-1} X_h(\pi^{(i)}; M^\star) X_h(\pi^{(i)}; M^\star)^\top + I$. Summing across all rounds and marginalizing, we have

$$
\begin{aligned}
\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] &\leq \frac{HT \cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma} + \gamma(R_2^2 + 1)\,\mathbb{E}\left[\sum_{t=1}^{T} \sum_{h=1}^{H} \|X_h(\pi^{(t)}; M^\star)\|^2_{(\Sigma_h^{(t-1)})^{-1}}\right] + 2H \sum_{t=1}^{T} \mathbb{P}(\neg\mathcal{E}^{(t)}(M^\star)) \\
&\leq \frac{HT \cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma} + \gamma(R_2^2 + 1)\,\mathbb{E}\left[\sum_{t=1}^{T} \sum_{h=1}^{H} \|X_h(\pi^{(t)}; M^\star)\|^2_{(\Sigma_h^{(t-1)})^{-1}}\right] + 2\delta HT.
\end{aligned}
$$

Finally, using the standard elliptic potential lemma (e.g., Lattimore and Szepesvári (2020), Lemma 19.4), we are guaranteed that for all h, with probability 1,

$$
\sum_{t=1}^{T} \|X_h(\pi^{(t)}; M^\star)\|^2_{(\Sigma_h^{(t-1)})^{-1}} \leq 2 d_{\mathrm{bi}}(\mathcal{M}) \log(1 + T/d_{\mathrm{bi}}(\mathcal{M})) \leq 2 d_{\mathrm{bi}}(\mathcal{M}) \log(2T).
$$

[34] We write $\mathcal{E}^{(t)}(M^\star)$ to make the fact that $\mathcal{E}^{(t)}$ is a measurable function of $\mathcal{H}^{(t-1)}$ and $M^\star$ explicit.

Altogether, after recalling the definition of R_2, we have that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}\bigg(\frac{HT \cdot d_{\mathrm{bi}}(\mathcal{M})}{\gamma} + \gamma \cdot H \cdot d_{\mathrm{bi}}(\mathcal{M}) \cdot L^2_{\mathrm{bi}}(\mathcal{M})(\varepsilon^2 T + \log(\mathcal{N}(\mathcal{Q}_{\mathcal{M}}, \varepsilon)/\delta)) + \delta HT\bigg).$$

The result now follows by setting $\delta = 1/TH$ and tuning γ and ε. $\square$

Proof of Lemma F.5. We first prove Property 1. For each $M \in \mathcal{M}$, let

$$Z^{(t)}_h(M) := \big(Q^{M,\star}_h(s^{(t)}_h, a^{(t)}_h) - r^{(t)}_h - V^{M,\star}_{h+1}(s^{(t)}_{h+1})\big)^2 - \big([\mathcal{T}^{M^\star}_h V^{M,\star}_{h+1}](s^{(t)}_h, a^{(t)}_h) - r^{(t)}_h - V^{M,\star}_{h+1}(s^{(t)}_{h+1})\big)^2.$$

Define a filtration $\mathscr{F}^{(t)} = \sigma(\mathcal{H}^{(t-1)}, \pi^{(t)})$ and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathscr{F}^{(t)}]$. Then we have $|Z^{(t)}_h| \leq 4$,

$$\mathbb{E}_{t-1}\big[Z^{(t)}_h(M)\big] = \mathbb{E}^{M^\star, \pi^{(t)}}\Big[\big(Q^{M,\star}_h(s_h, a_h) - [\mathcal{T}^{M^\star}_h V^{M,\star}_{h+1}](s_h, a_h)\big)^2\Big]$$

and

$$\mathbb{E}_{t-1}\Big[\big(Z^{(t)}_h(M)\big)^2\Big] \leq 16\,\mathbb{E}_{t-1}\big[Z^{(t)}_h(M)\big],$$

where we have used that $\sum_{h=1}^H r_h \in [0, 1]$. As a result, applying Lemma A.2 and taking a union bound, we have that with probability at least $1 - \delta$, for all $t \in [T]$, $h \in [H]$, and $i \in [N]$,

$$\sum_{k=1}^{t-1} \mathbb{E}_{k-1}[Z^{(k)}_h(M_i)] \leq 2\sum_{k=1}^{t-1} Z^{(k)}_h(M_i) + 64\log(HTN/\delta). \tag{159}$$

Consider $i \in \mathcal{I}^{(t)}$, and let j be any covering element that is ε-close to $[\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}]$; such an element is guaranteed to exist by Assumption 7.3. From the definition of $\mathcal{I}^{(t)}$, we have

$$\begin{aligned}
\sum_{k=1}^{t-1} Z^{(k)}_h(M_i) &= \mathcal{L}^{(t)}_h(Q^{i,\star}_h, Q^{i,\star}_{h+1}) - \mathcal{L}^{(t)}_h([\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}], Q^{i,\star}_{h+1}) \\
&= \mathcal{L}^{(t)}_h(Q^{i,\star}_h, Q^{i,\star}_{h+1}) - \mathcal{L}^{(t)}_h(Q^{j,\star}_h, Q^{i,\star}_{h+1}) + \mathcal{L}^{(t)}_h(Q^{j,\star}_h, Q^{i,\star}_{h+1}) - \mathcal{L}^{(t)}_h([\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}], Q^{i,\star}_{h+1}) \\
&\leq \mathcal{L}^{(t)}_h(Q^{i,\star}_h, Q^{i,\star}_{h+1}) - \min_{j'} \mathcal{L}^{(t)}_h(Q^{j',\star}_h, Q^{i,\star}_{h+1}) + \mathcal{L}^{(t)}_h(Q^{j,\star}_h, Q^{i,\star}_{h+1}) - \mathcal{L}^{(t)}_h([\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}], Q^{i,\star}_{h+1}) \\
&\leq R_1^2 + \mathcal{L}^{(t)}_h(Q^{j,\star}_h, Q^{i,\star}_{h+1}) - \mathcal{L}^{(t)}_h([\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}], Q^{i,\star}_{h+1}).
\end{aligned} \tag{160}$$

Next, we observe that

$$\begin{aligned}
&\mathcal{L}^{(t)}_h(Q^{j,\star}_h, Q^{i,\star}_{h+1}) - \mathcal{L}^{(t)}_h([\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}], Q^{i,\star}_{h+1}) \\
&= \sum_{k=1}^{t-1}\Big(Q^{j,\star}_h(s^{(k)}_h, a^{(k)}_h) - [\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}](s^{(k)}_h, a^{(k)}_h)\Big)^2 \\
&\quad + 2\sum_{k=1}^{t-1}\underbrace{\Big(Q^{j,\star}_h(s^{(k)}_h, a^{(k)}_h) - [\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}](s^{(k)}_h, a^{(k)}_h)\Big)\Big([\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}](s^{(k)}_h, a^{(k)}_h) - r^{(k)}_h + V^{i,\star}_{h+1}(s^{(k)}_{h+1})\Big)}_{=:\Delta^{(k)}_h(i,j)} \\
&\leq \varepsilon^2 T + 2\sum_{k=1}^{t-1}\Delta^{(k)}_h(i,j),
\end{aligned}$$

where the inequality uses the covering property for M_j. We now appeal to the following result.

Lemma F.6. *With probability at least $1 - \delta$, for all $i, j \in [N]$, $h \in [H]$, and $t \in [T]$,*

$$\sum_{k=1}^{t-1}\Delta^{(k)}_h(i,j) \leq 4\sum_{k=1}^{t-1}\mathbb{E}_{k-1}\Big[\Big(Q^{j,\star}_h(s^{(k)}_h, a^{(k)}_h) - [\mathcal{T}^{M^\star}_h V^{i,\star}_{h+1}](s^{(k)}_h, a^{(k)}_h)\Big)^2\Big] + 2\log(HTN^2\delta^{-1}).$$

Conditioning on the event in Lemma F.6, the covering property guarantees that

$$\sum_{k=1}^{t-1} \Delta_h^{(k)}(i,j) \leq 4\varepsilon^2 T + 2\log(HTN^2\delta^{-1}).$$

Combining this with (159) and (160), we have that with probability at least $1-2\delta$, for all $h \in [H]$ and $t \in [T]$, all $i \in \mathcal{I}^{(t)}$ satisfy

$$\sum_{k=1}^{t-1} \mathbb{E}^{M^\star,\pi^{(k)}}\left[\left(Q_h^{i,\star}(s_h,a_h) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h,a_h)\right)^2\right] = \sum_{k=1}^{t-1} \mathbb{E}_{k-1}[Z_h^{(k)}(M_i)] \leq 2R_1^2 + 18\varepsilon^2 T + 72\log(HTN/\delta).$$

It follows immediately from the covering property that any M for which $\iota(M) \in \mathcal{I}^{(t)}$ must have

$$\sum_{k=1}^{t-1} \mathbb{E}^{M^\star,\pi^{(k)}}\left[\left(Q_h^{M,\star}(s_h,a_h) - [\mathcal{T}_h^{M^\star} V_{h+1}^{M,\star}](s_h,a_h)\right)^2\right] \leq O(R_1^2 + \varepsilon^2 T + \log(HTN/\delta)).$$

Property 1 now follows from (82).

We now prove Property 2. Let i be an ε-covering element for $M^\star$, and define

$$E_h^{(t)}(M) = \left(Q^{M,\star}(s_h^{(t)},a_h^{(t)}) - r_h^{(t)} - V_{h+1}^{i,\star}(s_{h+1}^{(t)})\right)^2 - \left(Q^{i,\star}(s_h^{(t)},a_h^{(t)}) - r_h^{(t)} - V_{h+1}^{i,\star}(s_{h+1}^{(t)})\right)^2.$$

Note that $|E_h^{(t)}(M)| \leq 4$. Defining $y_h^{(i)} = r_h^{(t)} - V_{h+1}^{i,\star}(s_{h+1}^{(t)})$, we can write

$$\begin{aligned}
E_h^{(t)}(M) = &\left(Q_h^{M,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)^2 - \left(Q_h^{i,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)^2 \\
&+ \left(Q_h^{M,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)\left([\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)}) - y_h^{(t)}\right) \\
&- \left(Q_h^{M,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)\left([\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)}) - y_h^{(t)}\right).
\end{aligned}$$

It follows that

$$\mathbb{E}_{t-1}\big[E_h^{(t)}(M)\big] = \mathbb{E}_{t-1}\left[\left(Q_h^{M,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)^2\right] - \mathbb{E}_{t-1}\left[\left(Q_h^{i,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)^2\right]$$

and

$$\begin{aligned}
&\mathbb{E}_{t-1}\left[\left(E_h^{(t)}(M)\right)^2\right] \\
&\leq 20\,\mathbb{E}_{t-1}\left[\left(Q_h^{M,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)^2\right] + 20\,\mathbb{E}_{t-1}\left[\left(Q_h^{i,\star}(s_h^{(t)},a_h^{(t)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(t)},a_h^{(t)})\right)^2\right].
\end{aligned}$$

Applying Lemma A.2 to the sequence $\mathbb{E}_{t-1}\big[E_h^{(t)}(M_i)\big] - E_h^{(t)}(M_i)$ and taking a union bound, we have that for any $\eta \leq 1/8$, with probability at least $1-\delta$, for all $t \in [H]$, $h \in [H]$, and $j \in [N]$,

$$-\sum_{k=1}^{t-1} E_h^{(k)}(M_j) \leq -\sum_{k=1}^{t-1} \mathbb{E}_{k-1}\big[E_h^{(k)}(M_j)\big] + \eta \sum_{k=1}^{t-1} \mathbb{E}_{k-1}\big[(E_h^{(k)}(M_j))^2\big] + \eta^{-1}\log(HTN/\delta).$$

We choose $\eta \leq 1/20$ and observe that

$$\begin{aligned}
&-\sum_{k=1}^{t-1} \mathbb{E}_{k-1}\big[E_h^{(k)}(M_j)\big] + \eta \sum_{k=1}^{t-1} \mathbb{E}_{k-1}\big[(E_h^{(k)}(M_j))^2\big] \\
&\leq (1+20\eta) \sum_{k=1}^{t-1} \mathbb{E}_{k-1}\left[\left(Q_h^{i,\star}(s_h^{(k)},a_h^{(k)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(k)},a_h^{(k)})\right)^2\right] \\
&\quad - (1-20\eta) \sum_{k=1}^{t-1} \mathbb{E}_{k-1}\left[\left(Q_h^{j,\star}(s_h^{(k)},a_h^{(k)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(k)},a_h^{(k)})\right)^2\right] \\
&\leq 4(1+20\eta)\varepsilon^2 T,
\end{aligned}$$

where the last inequality uses the covering property of M_i. We conclude that for all $j \in [N]$,

$$\mathcal{L}_h^{(t)}(Q_h^{i,\star}, Q_{h+1}^{i,\star}) - \mathcal{L}_h^{(t)}(Q_h^{j,\star}, Q_{h+1}^{i,\star}) = -\sum_{k=1}^{t-1} E_h^{(k)}(M_j) \leq O(\varepsilon^2 T + \log(HTN/\delta)).$$

Since this bound holds uniformly for all $j \in [N]$, it follows that by setting $R_1 = C \cdot (\varepsilon^2 T + \log(HTN/\delta))$ for a sufficiently large numerical constant C, we have $i \in \mathcal{I}^{(t)}$ for all $h \in [H]$ and $t \in [T]$ whenever the concentration event above holds.

□

Proof of Lemma F.6. Let $i, j \in [N]$, $h \in [H]$, and $t \in [T]$ be fixed. Since $|\Delta_h^{(t)}(i,j)| \leq 2$ and $\mathbb{E}_{t-1}\big[\Delta_h^{(t)}(i,j)\big] = 0$, Lemma A.2 implies that with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=1}^{t-1} \Delta_h^{(k)}(i,j) &\leq \frac{1}{2}\sum_{k=1}^{t-1} \mathbb{E}_{k-1}\big[(\Delta_h^{(k)}(i,j))^2\big] + 2\log(\delta^{-1}) \\
&\leq 4\sum_{k=1}^{t-1} \mathbb{E}_{k-1}\left[\left(Q_h^{j,\star}(s_h^{(k)}, a_h^{(k)}) - [\mathcal{T}_h^{M^\star} V_{h+1}^{i,\star}](s_h^{(k)}, a_h^{(k)})\right)^2\right] + 2\log(\delta^{-1}).
\end{aligned}$$

The result now follows from a union bound.

□

F.4 Proofs from Section 7.3

F.4.1 Proof of Proposition 7.4

Proposition 7.4 (Lower bound for linearly realizable MDPs)**.** *For all $d \geq 2^9$ and $H \geq 2$, there exists a family of linearly realizable MDPs $\mathcal{M}$ with $\mathcal{R} = [-1,+1]$ such that for all $\gamma > 0$, there exists $\overline{M} \in \mathcal{M}$ for which*

$$\mathsf{dec}_\gamma(\mathcal{M}_{1/2}(\overline{M}), \overline{M}) \geq \frac{1}{48}\mathbb{I}\{\gamma \leq 2^{-9}\min\{2^H, \exp(2^{-10}d)\}\}. \tag{86}$$

Proof of Proposition 7.4. We follow the construction of Wang et al. (2021). Let $d \in \mathbb{N}$ be given and fix $\Delta \in [0, 1/6]$. By the Johnson-Lindenstrauss theorem, there exist $m := \exp\big(\frac{1}{8}\Delta^2 d\big)$ unit vectors $\{v_1, \ldots, v_m\}$ such that $|\langle v_i, v_j\rangle| \leq \Delta$ for all $i \neq j$. We set $\mathcal{A} = \{1, \ldots, m\} \cup \{\mathsf{term}\}$ and $\mathcal{S} = \{1, \ldots, m\} \cup \{\mathsf{term}\}$, where term denotes both i) a special terminal state, and ii) a corresponding action that always transitions to said state. For each state $s \in \mathcal{S} \setminus \{\mathsf{term}\}$, the available actions are $\mathcal{A} \setminus \{s\}$. For the terminal state, all actions are available.

We begin with a feature map $\phi(s,a) \in \mathbb{R}^d$ defined via

$$\begin{aligned}
&\phi(s,a) = (\langle v_s, v_a\rangle + 2\Delta)v_a \quad \forall s \in [m], a \notin \{s, \mathsf{term}\}, \\
&\phi(\mathsf{term}, \cdot) = \mathbf{0}, \\
&\phi(\cdot, \mathsf{term}) = \mathbf{0}.
\end{aligned}$$

We define a family of MDPs $\{M_{a^\star}\}_{a^\star \in [m]}$ that are linearly realizable with respect to ϕ. Let $a^\star \in [m]$ be fixed. Each MDP $M_{a^\star}$ has a stationary probability transition kernel $P^{(a^\star)} \equiv P^{M_{a^\star}}$ and initial state distribution d_1 defined as follows:

- The initial state distribution d_1 is uniform over $[m]$.
- $P^{(a^\star)}(\mathsf{term} \mid \mathsf{term}, \cdot) = 1$.
- $P^{(a^\star)}(\mathsf{term} \mid \cdot, \mathsf{term}) = 1$.

- $P^{(a^\star)}(\mathsf{term} \mid s, a^\star) = 1$ for all $s \notin \{a^\star, \mathsf{term}\}$.
- For all $s \neq \mathsf{term}$, $a \notin \{s, a^\star, \mathsf{term}\}$:

$$P^{(a^\star)}(\cdot \mid s, a) = \begin{cases} a : \langle v_s, v_a \rangle + 2\Delta \\ \mathsf{term} : 1 - (\langle v_s, v_a \rangle + 2\Delta) \end{cases}.$$

This is a valid transition distribution whenever $\Delta \leq 1/3$, since $\Delta \leq \langle v_s, v_a \rangle + 2\Delta \leq 3\Delta$. Next, we define a nonstationary reward function $R^{(a^\star)} \equiv R^{M_{a^\star}}$. For $1 \leq h < H$, we have:

$$\begin{aligned}
&R_h^{(a^\star)}(\mathsf{term}, \cdot) = \mathrm{Rad}(0), \\
&R_h^{(a^\star)}(\cdot, \mathsf{term}) = \mathrm{Rad}(0), \\
&R_h^{(a^\star)}(s, a^\star) = \mathrm{Rad}(\langle v_s, v_{a^\star} \rangle + 2\Delta) \quad \forall s \notin \{a^\star, \mathsf{term}\}, \\
&R_h^{(a^\star)}(s, a) = \mathrm{Rad}(-2\Delta(\langle v_s, v_a \rangle + 2\Delta)) \quad \forall s \neq \mathsf{term}, a \notin \{s, a^\star, \mathsf{term}\}.
\end{aligned}$$

We define $R_H^{(a^\star)}(s, a) = \mathrm{Rad}(\langle \phi(s, a), v_{a^\star} \rangle)$.

To finish the construction, we define a reference model $\overline{M}$. As before, we take $d_1 = \mathrm{unif}([m])$. The stationary transition kernel $\overline{P} \equiv P^{\overline{M}}$ is given by

$$\begin{aligned}
&\overline{P}(\mathsf{term} \mid \mathsf{term}, \cdot) = 1, \\
&\overline{P}(\mathsf{term} \mid \cdot, \mathsf{term}) = 1, \\
&\overline{P}(\cdot \mid s, a) = \begin{cases} a : \langle v_s, v_a \rangle + 2\Delta \\ \mathsf{term} : 1 - (\langle v_s, v_a \rangle + 2\Delta) \end{cases} \qquad \forall s \neq \mathsf{term}, a \notin \{s, \mathsf{term}\}.
\end{aligned}$$

We define the nonstationary reward function $\overline{R} \equiv R^{\overline{M}}$ for the reference model as follows. For all $h < H$, set

$$\begin{aligned}
&\overline{R}_h(\mathsf{term}, \cdot) = \mathrm{Rad}(0), \\
&\overline{R}_h(\cdot, \mathsf{term}) = \mathrm{Rad}(0), \\
&\overline{R}_h(s, a) = \mathrm{Rad}(-2\Delta(\langle v_s, v_a \rangle + 2\Delta)) \quad \forall s \neq \mathsf{term}, a \notin \{s, \mathsf{term}\},
\end{aligned}$$

and take $\overline{R}_H(\cdot, \cdot) = \mathrm{Rad}(0)$.

Our lower bound is based on the family of models $\mathcal{M}' := \{M_a\}_{a \in [m]} \cup \{\overline{M}\}$. We first show that this class is indeed linearly realizable; proofs for this and all subsequent lemmas are deferred to the end of the main proof. We abbreviate $Q^{(a)} \equiv Q^{M_a, \star}$, $V^{(a)} \equiv V^{M_a, \star}$, $Q^{\overline{M}} \equiv Q^{\overline{M}, \star}$, and $\pi_a \equiv \pi_{M_a}$ throughout

Lemma F.7. *If $\Delta < 1/3$, then for all $a^\star \in [m]$, $Q_h^{(a^\star)}(s, a) = \langle \phi(s, a), v_{a^\star} \rangle$ for all $h \in [H]$.*

This lemma shows that $\{M_a\}_{a \in [m]}$ is linearly realizable with respect to ϕ. To show ensure $\overline{M}$ is linearly realizable, we consider the expanded feature map feature map $\phi'(s, a) \in \mathbb{R}^{d+1}$ given by

$$\phi'(s, a) = (\phi(s, a), -2\Delta(\langle v_s, v_a \rangle + 2\Delta)\mathbb{I}\{s \neq \mathsf{term}, a \notin \{s, \mathsf{term}\}\}).$$

Lemma F.8. *We have $Q_h^{\overline{M}}(s, a) = \langle \phi'(s, a), (\mathbf{0}, 1) \rangle = 0$ for all $h < H$ and $Q_H^{\overline{M}}(s, a) = \langle \phi'(s, a), \mathbf{0} \rangle$.*

This establishes that $\mathcal{M}'$ is linearly realizable with the feature map ϕ'. Note that $\|\phi'(s, a)\|_2 \leq 1$ whenever $\Delta \leq 1/6$, and that all of the weight parameters for the Q-functions defined above have norm at most 1 as well.

Next, we show that $\mathcal{M}'$ is localized around $\overline{M}$.

Lemma F.9 (Local property)**.** *For all $a^\star \in [m]$, $f^{\overline{M}}(\pi_{\overline{M}}) \geq f^{M_{a^\star}}(\pi_{a^\star}) - 3\Delta^2$, so that $\mathcal{M}' \subseteq \mathcal{M}_{3\Delta^2}(\overline{M})$.*

To proceed, we state three lemmas which will be used to prove that $\mathcal{M}'$ is a hard sub-family of models.

Lemma F.10. *For all $M \in \mathcal{M}'$, and $h \in [H]$, $\sup_{\pi \in \Pi_{\mathrm{RNS}}} \mathbb{P}^{M, \pi}(s_h \neq \mathsf{term}) \leq (3\Delta)^{h-1}$.*

Lemma F.11. *Whenever $\Delta \leq 1/6$, we have that for all $\pi \in \Pi_{\mathrm{RNS}}$,*

$$f^{M_{a^\star}}(\pi_{a^\star}) - f^{M_{a^\star}}(\pi) \geq \frac{\Delta}{2}\big(1 - \mathbb{P}^{\overline{M},\pi}(s_1 = a^\star \vee a_1 = a^\star)\big). \tag{161}$$

Lemma F.12. *Let $\Delta \leq 1/6$. Then for all $a^\star \in [m]$ and all $\pi \in \Pi_{\mathrm{RNS}}$,*

$$D^2_{\mathrm{H}}\big(M_{a^\star}(\pi), \overline{M}(\pi)\big) \leq 11\Delta \sum_{h=1}^{H} \mathbb{P}^{\overline{M},\pi}(s_h \neq \mathsf{term}, a_h = a^\star) + (3\Delta)^{H+1}. \tag{162}$$

We now complete the proof. We claim that $\{M_{a^\star}\}_{a^\star \in [m]}$ is a hard family of models in the sense of Lemma 5.1. Define $u_{a^\star}(\pi) = \mathbb{P}^{\overline{M},\pi}(s_1 = a^\star \vee a_1 = a^\star)$. Then whenever $m \geq 4$, for all $\pi \in \Pi_{\mathrm{RNS}}$,

$$\sum_{a^\star \in [m]} u_{a^\star}(\pi) \leq \sum_{a^\star \in [m]} \mathbb{P}^{\overline{M},\pi}(s_1 = a^\star) + \mathbb{P}^{\overline{M},\pi}(a_1 = a^\star) \leq 2 \leq \frac{m}{2}.$$

Next, let $v_{a^\star}(\pi) = \frac{1}{2}\sum_{h=1}^{H} \mathbb{P}^{\overline{M},\pi}(s_h \neq \mathsf{term}, a_h = a^\star)$. Then by Lemma F.10, whenever $\Delta \leq 1/6$, we have that for all $\pi \in \Pi_{\mathrm{RNS}}$,

$$\sum_{a^\star \in [m]} v_{a^\star}(\pi) \leq \frac{1}{2} \sum_{a^\star \in [m]} \sum_{h=1}^{H} (3\Delta)^{h-1} \mathbb{P}^{\overline{M},\pi}(a_h = a^\star \mid s_h \neq \mathsf{term}) \leq \frac{1}{2} \sum_{h=1}^{H} (3\Delta)^{h-1} \leq 1.$$

As a result, Lemma F.11 and Lemma F.12 imply that $\{M_{a^\star}\}_{a^\star \in [m]}$ is a $(\Delta/2, 22\Delta, (3\Delta)^{H+1})$-family, and Lemma 5.1 gives the lower bound

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \geq \frac{\Delta}{4} - \gamma\left(\frac{22\Delta}{m} + (3\Delta)^{H+1}\right).$$

In particular, whenever $\Delta \leq 1/6$, we see that as long as $\gamma \cdot \big(22 \cdot m^{-1} + 3 \cdot 2^{-H}\big) \leq 1/8$,

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \geq \Delta/8.$$

We set $\Delta = 1/6$. Recalling that $m = \exp(\frac{1}{8}\Delta^2 d)$, it suffices to take

$$\gamma \leq 2^{-9} \min\big\{m, 2^H\big\} = 2^{-9} \min\big\{\exp(288^{-1} d), 2^H\big\}.$$

Finally, we recall that the dimension of the construction is $d+1$ (as opposed to d), and simplify further to

$$\gamma \leq 2^{-9} \min\big\{\exp(2^{-10} d), 2^H\big\}.$$

To ensure that $m \geq 4$ as required, it suffices to take $d \geq 2^9$.

□

F.4.2 Proofs for Auxiliary Lemmas (Proposition 7.4)

Proof of Lemma F.7. Let $a^\star$ be fixed, and let us abbreviate $Q \equiv Q^{(a^\star)}$ and $V \equiv V^{(a^\star)}$. We begin by noting that in the terminal state, $Q_h(\mathsf{term}, \cdot) = V_h(\mathsf{term}, \cdot) = 0$, so that $Q_h(\mathsf{term}, a) = \langle \phi(\mathsf{term}, a), v_{a^\star} \rangle$. Similarly, for any state s, $Q_h(s, \mathsf{term}) = 0 = \langle \phi(s, \mathsf{term}), v_{a^\star} \rangle$. We proceed to prove the result for non-terminal states and actions. We prove inductively that for all $s \neq \mathsf{term}$, $a \notin \{s, \mathsf{term}\}$,

$$Q_h(s, a) = (\langle v_s, v_a \rangle + 2\Delta)\langle v_{a^\star}, v_a \rangle \mathbb{I}\{a \neq \mathsf{term}\} = \langle \phi(s, a), v_{a^\star} \rangle, \tag{163}$$

and that for all $s \notin \{a^\star, \mathsf{term}\}$,

$$V_h(s) = Q_h(s, a^\star) = \langle v_s, v_{a^\star} \rangle + 2\Delta. \tag{164}$$

Condition (163) holds at layer H by definition. We now show that (164) follows from (163) for all h. Indeed, for $s \neq \mathsf{term}$, $a \notin \{s, a^\star\}$,

$$Q_h(s,a) = (\langle v_s, v_a \rangle + 2\Delta)\langle v_{a^\star}, v_a\rangle \mathbb{I}\{a \neq \mathsf{term}\} \leq 3\Delta^2,$$

while for $s \neq a^\star$, $Q_h(s,a^\star) = (\langle v_s, v_{a^\star}\rangle + 2\Delta) \geq \Delta$. This implies that the optimal action is $a^\star$ whenever $\Delta < 1/3$, so that $V_h(s) = Q_h(s,a^\star)$.

We now complete the induction. For any $s \notin \{a^\star, \mathsf{term}\}$, we trivially have

$$Q_h(s,a^\star) = (\langle v_s, v_{a^\star}\rangle + 2\Delta) = (\langle v_s, v_{a^\star}\rangle + 2\Delta)\langle v_{a^\star}, v_{a^\star}\rangle,$$

from the definition of the reward distribution. For any $s \neq \mathsf{term}$ and $a \notin \{s, a^\star, \mathsf{term}\}$, by Bellman optimality,

$$\begin{aligned}
Q_h(s,a) &= \mathbb{E}^{M_{a^\star}}[r_h \mid s_h = s, a_h = a] + \mathbb{E}^{M_{a^\star}}[V_{h+1}(s_{h+1}) \mid s_h = s, a_h = a] \\
&= -2\Delta(\langle v_s, v_a\rangle + 2\Delta) + (\langle v_s, v_a\rangle + 2\Delta)V_{h+1}(a) + (1 - (\langle v_s, v_a\rangle + 2\Delta))V_{h+1}(\mathsf{term}) \\
&= -2\Delta(\langle v_s, v_a\rangle + 2\Delta) + (\langle v_s, v_a\rangle + 2\Delta)(\langle v_a, v_{a^\star}\rangle + 2\Delta) \\
&= (\langle v_s, v_a\rangle + 2\Delta)(\langle v_a, v_{a^\star}\rangle.
\end{aligned}$$

This establishes the result. □

Proof of Lemma F.8. Let us abbreviate $Q \equiv Q^{\overline{M},\star}$ and $V \equiv V^{\overline{M},\star}$. The claim that $Q_H(s,a) = 0$ follows immediately from the definition of $\overline{R}_H$. For $h < H$, we prove the result under the inductive hypothesis that $V_{h+1}(s) = 0$, which is clearly satisfied at layer H.

Fix $h < H$. Then by Bellman optimality,

$$\begin{aligned}
Q_h(s,a) &= \mathbb{E}^{\overline{M}}[r_h \mid s_h =, a_h = a] + \mathbb{E}^{\overline{M}}[V_{h+1}(s_{h+1}) \mid s_h = s, a_h = a] \\
&= \mathbb{E}^{\overline{M}}[r_h \mid s_h = s, a_h = a] \\
&= -2\Delta(\langle v_s, v_a\rangle + 2\Delta)\mathbb{I}\{s \neq \mathsf{term}, a \notin \{s, \mathsf{term}\}\} \\
&= \langle \phi'(s,a), (\mathbf{0}, 1)\rangle.
\end{aligned}$$

We now prove that $V_h(s) = 0$ as a consequence. If $s = \mathsf{term}$ this is trivial, so suppose otherwise. Then we have

$$-2\Delta(\langle v_s, v_a\rangle + 2\Delta) \leq \; -2\Delta^2 < 0,$$

so that $Q_h(s,a) < Q_h(s, \mathsf{term}) = 0$ for all $a \notin \{s, \mathsf{term}\}$. It follows that the optimal action is term, and that $V_h(s) = 0$.

□

Proof of Lemma F.10. For all models in the family, $P(\mathsf{term} \mid s,a) \geq 1 - 3\Delta$ for all admissible (s,a) pairs. The result follows immediately. □

Proof of Lemma F.11. Recall that $\pi_{a^\star} \equiv \pi_{M_{a^\star}}$. By the performance difference lemma (Kakade, 2003), we have that for any $\pi \in \Pi_{\mathrm{RNS}}$,

$$f^{M_{a^\star}}(\pi_{a^\star}) - f^{M_{a^\star}}(\pi) = \mathbb{E}^{M_{a^\star},\pi}\left[\sum_{h=1}^{H} Q_h^{(a^\star)}(s_h, \pi_{a^\star}(s_h)) - Q_h^{(a^\star)}(s_h, \pi(s_h))\right]$$

Lemma F.7 shows that for all $s \notin \{a^\star, \mathsf{term}\}$, we have $\pi_{a^\star}(s) = a^\star$, and for $a \neq \pi_{a^\star}(s)$,

$$Q_h^{(a^\star)}(s,a^\star) - Q_h^{(a^\star)}(s,a) = \langle \phi(s,a^\star) - \phi(s,a), v_{a^\star}\rangle \geq \Delta - 3\Delta^2 \geq \frac{\Delta}{2},$$

whenever $\Delta \leq 1/6$. Hence, since $a^\star$ is only reachable at $h = 1$, we have

$$
\begin{aligned}
f^{M_{a^\star}}(\pi_{a^\star}) - f^{M_{a^\star}}(\pi) &\geq \frac{\Delta}{2} \sum_{h=2}^{H} \mathbb{P}^{M_{a^\star},\pi}(s_h \neq \mathsf{term}, a \neq a^\star) + \frac{\Delta}{2} \mathbb{P}^{M_{a^\star},\pi}(s_1 \notin \{a^\star, \mathsf{term}\}, a_1 \neq a^\star) \\
&\geq \frac{\Delta}{2} \mathbb{P}^{M_{a^\star},\pi}(s_1 \notin \{a^\star, \mathsf{term}\}, a_1 \neq a^\star).
\end{aligned}
$$

Since $s_1 \sim \mathrm{unif}([m])$ for all models, and term is not reachable at $h = 1$, this implies

$$
f^{M_{a^\star}}(\pi_{a^\star}) - f^{M_{a^\star}}(\pi) \geq \frac{\Delta}{2} \mathbb{P}^{\overline{M},\pi}(s_1 \neq a^\star, a_1 \neq a^\star) = \frac{\Delta}{2}\big(1 - \mathbb{P}^{\overline{M},\pi}(s_1 = a^\star \vee a_1 = a^\star)\big).
$$

$\square$

Proof of Lemma F.12. First, observe that $P^{(a^\star)}(\cdot \mid \mathsf{term}, \cdot) = \overline{P}(\cdot \mid \mathsf{term}, \cdot)$, and for all $s \neq \mathsf{term}$ and $a \neq a^\star$, we have

$$
P^{(a^\star)}(\cdot \mid s, a) = \overline{P}(\cdot \mid s, a),
$$

while for $s \neq a^\star$, we have

$$
\mathrm{P}^{(a^\star)}(\mathsf{term} \mid s, a^\star) = 1, \quad \text{and} \quad \overline{P}(\cdot \mid s, a^\star) = \begin{cases} a^\star : \langle v_s, v_{a^\star} \rangle + 2\Delta \\ \mathsf{term} : 1 - (\langle v_s, v_{a^\star} \rangle + 2\Delta) \end{cases} \quad \forall s \neq \mathsf{term}, a \neq s.
$$

Similarly, we have $R_h^{(a^\star)}(\mathsf{term}, \cdot) = \overline{R}_h(\mathsf{term}, \cdot)$ for all $h \in [H]$, and for all $h < H$ and $s \neq \mathsf{term}$, $a \neq a^\star$ we have $R_h(s, a) = \overline{R}(s, a)$. Next,

$$
R_h^{(a^\star)}(s, a^\star) = \mathrm{Rad}(\langle v_s, v_{a^\star} \rangle + 2\Delta) \neq \overline{R}_h(s, a^\star) = \mathrm{Rad}(-2\Delta(\langle v_s, v_{a^\star} \rangle + 2\Delta)).
$$

Lastly, for $s \neq \mathsf{term}$, we have $R_H^{(a^\star)}(s, a^\star) = \mathrm{Rad}(\langle \phi(s, a), v_{a^\star} \rangle)$, while $\overline{R}_H(s, a) = 0$.

Using that $\Delta \leq 1/3$, we have $\langle v_s, v_{a^\star} \rangle + 2\Delta \in [\Delta, 3\Delta]$, $-2\Delta(\langle v_s, v_{a^\star} \rangle + 2\Delta) \in \left[-6\Delta^2, 6\Delta^2\right]$, and

$$
|\langle v_s, v_{a^\star} \rangle + 2\Delta - (-2\Delta(\langle v_s, v_{a^\star} \rangle + 2\Delta))| \leq 5\Delta.
$$

Likewise, we have

$$
|\langle \phi(s, a), v_{a^\star} \rangle| \leq 3\Delta.
$$

Using these observations, we appeal to the following result.

Proposition F.2. *For any $\mu_1, \mu_2 \in (-1, +1)$, $D_{\mathrm{KL}}(\mathrm{Rad}(\mu_1) \,\|\, \mathrm{Rad}(\mu_2)) \leq \frac{(\mu_1 - \mu_2)^2}{2 \min\{1+\mu_1, 1+\mu_2, 1-\mu_1, 1-\mu_2\}}$.*

This implies that whenever $\Delta \leq 1/6$, for all $h < H$,

$$
D_{\mathrm{KL}}\big(\overline{R}_h(s, a) \,\|\, R_h^{(a^\star)}(s, a)\big) \leq 25\Delta^2 \mathbb{I}\{s \neq \mathsf{term}, a = a^\star\},
$$

and that

$$
D_{\mathrm{KL}}\big(\overline{R}_H(s, a) \,\|\, R_H^{(a^\star)}(s, a)\big) \leq 9\Delta^2 \mathbb{I}\{s \neq \mathsf{term}\}.
$$

Next, we calculate that whenever $\Delta \leq 1/6$,

$$
\begin{aligned}
D_{\mathrm{KL}}\big(\overline{P}(s, a) \,\|\, P^{(a^\star)}(s, a)\big) &= D_{\mathrm{KL}}(\mathrm{Mult}((0, 1)) \,\|\, \mathrm{Mult}((\langle v_s, v_{a^\star} \rangle + 2\Delta, 1 - \langle v_s, v_{a^\star} \rangle - 2\Delta))) \mathbb{I}\{s \neq \mathsf{term}, a = a^\star\} \\
&\leq \log\left(\frac{1}{1 - 3\Delta}\right) \mathbb{I}\{s \neq \mathsf{term}, a = a^\star\}, \\
&\leq 6\Delta \mathbb{I}\{s \neq \mathsf{term}, a = a^\star\}.
\end{aligned}
$$

Combining these bounds, we have

$$
\begin{aligned}
D^2_{\mathrm{H}}\big(\overline{M}(\pi), M_{a^\star}(\pi)\big) &\le D_{\mathrm{KL}}\big(\overline{M}(\pi) \,\|\, M_{a^\star}(\pi)\big) \\
&= \mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\overline{P}(s_h,a_h) \,\|\, P^{(a^\star)}(s_h,a_h)\big) + D_{\mathrm{KL}}\big(\overline{R}(s_h,a_h) \,\|\, R^{(a^\star)}(s_h,a_h)\big)\right] \\
&\le 9\Delta^2\mathbb{P}^{\overline{M},\pi}(s_H \neq \mathsf{term}) + \sum_{h=1}^{H} 6\Delta\mathbb{P}^{\overline{M},\pi}(s_h \neq \mathsf{term}, a_h = a^\star) + 25\Delta^2\mathbb{P}^{\overline{M},\pi}(s_h \neq \mathsf{term}, a_h = a^\star) \\
&\le 9\Delta^2\mathbb{P}^{\overline{M},\pi}(s_H \neq \mathsf{term}) + 11\Delta\sum_{h=1}^{H}\mathbb{P}^{\overline{M},\pi}(s_h \neq \mathsf{term}, a_h = a^\star),
\end{aligned}
$$

where we have used that $\Delta \le 1/6$ to simplify. Finally, using Lemma F.10, we have $9\Delta^2\mathbb{P}_{\overline{M},\pi}(s_H \neq \mathsf{term}) \le 9\Delta^2(3\Delta)^{H-1} = (3\Delta)^{H+1}$. $\square$

Proof of Lemma F.9. Lemma F.7 proves that for all $a^\star \in [m]$, $Q^{(a^\star)}_h(s,a) = \langle \phi(s,a), v_{a^\star}\rangle \le 3\Delta^2$ for all h. It follows that $\max_\pi f^{M_{a^\star}}(\pi) \le 3\Delta^2$. On the other hand, Lemma F.8 shows that $\max_\pi f^{\overline{M}}(\pi) = 0$. $\square$

F.4.3 Proof of Proposition 7.5

Proposition 7.5 (Lower bound for deterministic linearly realizable MDPs)**.** *There exists a collection $\mathcal{M}$ of linearly realizable MDPs $\mathcal{M}$ with $\mathcal{R} = [0,1]$, $H = 1$, constant suboptimality gap, and deterministic rewards and dynamics, such that for all $\gamma > 0$, there exists $\overline{M} \in \mathcal{M}$ such that*

$$
\mathsf{dec}_\gamma(\mathcal{M}^\infty_{1/3}(\overline{M}), \overline{M}) \ge \frac{1}{12}\mathbb{I}\left\{\gamma \le \frac{d}{48}\right\}.
$$

Proof of Proposition 7.5. As with Proposition 7.3, we consider a bandit problem with no state, which is a special case of the reinforcement learning setting with $H = 1$. The proof is a small modification to Proposition 6.8, with the only difference being that rewards are no longer stochastic.

We set $\Pi = \mathcal{A} = \{e_1, \ldots, e_d\}$ and construct a family $\mathcal{M}' = \{M_i\}_{i\in[d]}$ as follows. Let $\theta_1 = \frac{1}{3}\cdot e_1$, and let $\theta_i = \frac{1}{3}\cdot e_1 + \frac{2}{3}\cdot e_i$ for all $i \ge 2$; we have $\|\theta_i\|_2 \le 1$. Define $M_i(\pi) = \delta_{\langle\theta_i,\pi\rangle}$ so that $f^{M_i}(\pi) = \langle\theta_i,\pi\rangle$. Take $\overline{M} = M_1$ as the reference model. We have $\mathcal{M}' \subseteq \mathcal{M}^\infty_{1/3}(\overline{M})$.

We now show that $\mathcal{M}'$ is a hard family of models in the sense of Lemma 5.1. Define $u_i(\pi) = v(\pi) = \mathbb{I}\{\pi = e_i\}$. Then for any i, we have

$$
f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \ge \tfrac{1}{3}(1 - u_i(\pi)),
$$

and

$$
D^2_{\mathrm{H}}\big(M_i(\pi), \overline{M}(\pi)\big) \le 2\mathbb{I}\{\pi = e_i\} \le 2v_i(\pi).
$$

It follows that $\mathcal{M}'$ is a $(\frac{1}{3}, 2, 0)$-family, so Lemma 5.1 implies that

$$
\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \ge \frac{1}{6} - \gamma\frac{4}{d}.
$$

In particular, $\mathsf{dec}_\gamma(\mathcal{M}', \overline{M}) \ge \frac{1}{12}$ whenever $\gamma \le \frac{d}{48}$. $\square$

F.5 Additional Results

F.5.1 Bellman Representability and Bellman-Eluder Dimension

In this section we provide a nonlinear generalization of the bilinear class property which we refer to as *Bellman representability*, and give a bound on the Decision-Estimation Coefficient based on this property. The results here recover the Bellman-eluder dimension (Jin et al., 2021) as a special case. Throughout the section, we assume that $\sum_{h=1}^{H} r_h \in [0,1]$.

Definition F.1 (Bellman representability)**.** *Let $\mathcal{M}$ and $\overline{M}$ be given. Let $\mathcal{G}^{\overline{M}} = \mathcal{G}_1^{\overline{M}}, \ldots, \mathcal{G}_H^{\overline{M}}$ be a collection of a function classes of the form $\mathcal{G}_h^{\overline{M}} = \left\{ g_h^{M;\overline{M}} : \mathcal{M} \to [-1,+1] \right\}_{M \in \mathcal{M}}$. We say that $\mathcal{G}^{\overline{M}}$ is a Bellman representation for $\mathcal{M}$ (relative to $\overline{M}$) if:*

1. *Fo all h and $M \in \mathcal{M}$:*

$$\left| \mathbb{E}^{\overline{M},\pi_M}\left[Q_h^{M,\star}(s_h,a_h) - r_h - V_h^{M,\star}(s_{h+1}) \right] \right| \le |g_h^{M;\overline{M}}(M)|. \tag{165}$$

   We assume that $g_h^{\overline{M};\overline{M}}(M) = 0$ for all $M \in \mathcal{M}$.

2. *Let $z_h = (s_h, a_h, r_h, s_{h+1})$. There exists a collection of estimation policies $\{\pi_M^{\mathrm{est}}\}_{M \in \mathcal{M}}$ and estimation functions $\{\ell_M^{\mathrm{est}}(\cdot;\cdot)\}_{M \in \mathcal{M}}$ such that for all $M, M' \in \mathcal{M}$, $h \in [H]$,*

$$g_h^{M';\overline{M}}(M) = \mathbb{E}^{\overline{M},\pi_M \circ_h \pi_M^{\mathrm{est}}}\left[\ell_M^{\mathrm{est}}(M';z_h) \right]. \tag{166}$$

   If $\pi_M^{\mathrm{est}} = \pi_M$, we say that estimation is on-policy.

We let $L_{\mathrm{br}}(\mathcal{M};\overline{M}) \ge 1$ denote any almost sure upper bound on $|\ell_M^{\mathrm{est}}(M';z_h)|$ under $\overline{M}$, and let $L_{\mathrm{br}}(\mathcal{M}) = \sup_{\overline{M} \in \mathcal{M}} L_{\mathrm{br}}(\mathcal{M};\overline{M})$.

The following result generalizes Theorem 7.1. Recall that π_M^{α} denotes a randomized policy that, for each h, plays $\pi_{M,h}$ with probability $1 - \alpha/H$ and $\pi_{M,h}^{\mathrm{est}}$ with probability α/H.

Theorem F.1. *Let $\mathcal{M}$ be a bilinear class with $\overline{M} \in \mathcal{M}$. Let $\mu \in \Delta(\mathcal{M})$ be given, and consider the modified posterior sampling strategy that samples $M \sim \mu$ and plays π_M^{α} for $\alpha \in [0,1]$.*

- *If $\pi_M^{\mathrm{est}} = \pi_M$ (i.e., estimation is on-policy), this strategy with $\alpha = 0$ certifies that*

$$\underline{\mathsf{dec}}_\gamma(\mathcal{M},\overline{M}) \le O(H^2 L_{\mathrm{br}}^2(\mathcal{M})) \cdot \inf_{\Delta > 0} \left\{ \Delta + \frac{\max_h \min\{\mathfrak{e}(\mathcal{G}_h^{\overline{M}},\Delta), \mathfrak{s}^2(\mathcal{G}_h^{\overline{M}},\Delta)\} \log^2(\gamma)}{\gamma} \right\}.$$

  for all $\gamma \ge e$.

- *For general estimation policies, this strategy (with an appropriate choice of α) certifies that*

$$\mathsf{dec}_\gamma(\mathcal{M},\overline{M}) \le O(H^{3/2} L_{\mathrm{br}}(\mathcal{M})) \cdot \inf_{\Delta > 0} \left(\Delta + \frac{\max_h \min\{\mathfrak{e}(\mathcal{G}_h^{\overline{M}},\Delta), \mathfrak{s}^2(\mathcal{G}_h^{\overline{M}},\Delta)\} \log^2(\gamma)}{\gamma} \right)^{1/2},$$

  whenever $\gamma \ge e$ is sufficiently large.

Example F.1 (Bellman-eluder dimension)**.** By taking

$$g_h^{M;\overline{M}}(M') = \mathbb{E}^{\overline{M},\pi_{M'}}\left[Q_h^{M,\star}(s_h,a_h) - r_h - V_{h+1}^{M,\star}(s_{h+1}) \right],$$

and $\pi_M^{\mathrm{est}} = \pi_M$, Theorem F.1 recovers the Q-type Bellman-eluder dimension of Jin et al. (2021). Taking

$$g_h^{M;\overline{M}}(M') = \mathbb{E}^{\overline{M},\pi_{M'} \circ_h \pi_M}\left[Q_h^{M,\star}(s_h,a_h) - r_h - V_{h+1}^{M,\star}(s_{h+1}) \right],$$

and $\pi_M^{\mathrm{est}} = \mathrm{unif}(\mathcal{A})$ recovers the V-type Bellman-eluder dimension.

On the other hand, the notion of Bellman representability goes well beyond the Bellman-eluder dimension, since it recovers the usual Bilinear class definition as a special case. Interestingly Theorem F.1 shows that the Bellman-eluder dimension can be replaced by a weaker parameter we call the *Bellman-star number*, which, in general, can be arbitrarily small compared to the former quantity. ◁

Proof of Theorem F.1. We only sketch the proof, as it closely follows Theorem 7.1. Let $\mu \in \Delta(\mathcal{M})$, $\overline{M} \in \mathcal{M}$, and $\gamma \geq 1$ be fixed. Following the same steps as Theorem 7.1 and using the first property of Definition F.1, we have

$$\begin{aligned}\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] &\leq \mathbb{E}_{M\sim\mu}\left[\sum_{h=1}^H \mathbb{E}^{\overline{M},\pi_M}\big[Q_h^{M,\star}(s_h,a_h) - r_h - V_{h+1}^{M,\star}(s_{h+1})\big]\right] + \alpha \\ &\leq \sum_{h=1}^H \mathbb{E}_{M\sim\mu}|g_h^{M;\overline{M}}(M)| + \alpha.\end{aligned}$$

Using Lemma E.2, we have that for each h, for all $\eta > e$,

$$\mathbb{E}_{M\sim\mu}|g_h^{M;\overline{M}}(M)| \leq \inf_{\Delta>0}\left\{2\Delta + 6\frac{\boldsymbol{\theta}(\mathcal{G}_h^{\overline{M}},\Delta,\eta^{-1};\rho_\mu)\log^2(\eta)}{\eta}\right\} + \eta\cdot\mathbb{E}_{M,M'\sim\mu}\big[(g_h^{M;\overline{M}}(M'))^2\big],$$

where $\boldsymbol{\theta}$ is the disagreement coefficient (Definition 6.3) and ρ_μ is the induced distribution over policies. Lemma 6.1 implies that

$$\boldsymbol{\theta}(\mathcal{G}_h^{\overline{M}},\Delta,\eta^{-1};\rho_\mu) \leq 4\min\big\{\mathfrak{e}(\mathcal{G}_h^{\overline{M}},\Delta),\mathfrak{s}^2(\mathcal{G}_h^{\overline{M}},\Delta)\big\},$$

and Lemma F.4 (via the second property of Definition F.1) implies that for all $\alpha \leq 1/2$,

$$\begin{aligned}\mathbb{E}_{M,M'\sim\mu}\big[(g_h^{M;\overline{M}}(M'))^2\big] &\leq 4C_\alpha L_{\mathrm{br}}^2(\mathcal{M})\cdot\mathbb{E}_{M,M'\sim\mu}\big[D_{\mathrm{H}}^2\big(M(\pi_{M'}^\alpha),\overline{M}(\pi_{M'}^\alpha)\big)\big] \\ &= 4C_\alpha L_{\mathrm{br}}^2(\mathcal{M})\cdot\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{H}}^2\big(M(\pi),\overline{M}(\pi)\big)\big],\end{aligned}$$

where $C_\alpha = 1$ in the on-policy case and $C_\alpha \leq 2H/\alpha$ in the general case.

Altogether, we have that

$$\begin{aligned}&\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] \\ &\leq O(H)\cdot\inf_{\Delta>0}\left\{\Delta + \frac{\max_h\min\big\{\mathfrak{e}(\mathcal{G}_h^{\overline{M}},\Delta),\mathfrak{s}^2(\mathcal{G}_h^{\overline{M}},\Delta)\big\}\log^2(\eta)}{\eta}\right\} + \eta 4HC_\alpha L_{\mathrm{br}}^2(\mathcal{M})\cdot\mathbb{E}_{M\sim\mu}\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{H}}^2\big(M(\pi),\overline{M}(\pi)\big)\big] + \alpha.\end{aligned}$$

From here, following the same steps as in Theorem 7.1 and tuning η and α appropriately leads to the result. $\square$

G Additional Proofs

G.1 Proofs from Section 1

Proof of Proposition 2.1. We first state a minimax theorem for *finitely supported* models, which is a straightforward adaptation of Theorem 1 from Lattimore and Szepesvári (2019).

Lemma G.1. *Suppose $\mathcal{M}$ is finitely supported in the sense that $\big|\bigcup_{\pi\in\Pi}\mathrm{supp}(M(\pi))\big| < \infty$ for all $M \in \mathcal{M}$. In addition, assume that Π is finite and $\mathcal{R}$ is bounded. Then we have*

$$\mathfrak{M}(\mathcal{M},T) = \underline{\mathfrak{M}}(\mathcal{M},T). \tag{167}$$

We prove Proposition 2.1 by arguing that any countably supported model class can be approximated by a finitely supported class, then applying Lemma G.1 to the approximating class.

Fix $\varepsilon > 0$. For each $M \in \mathcal{M}$, since $M(\pi)$ has countable support and Π is finite, there exists a model M_ε with finite support such that

$$\max_{\pi\in\Pi} D_{\mathrm{TV}}(M(\pi),M_\varepsilon(\pi)) \leq \varepsilon.$$

Let $\mathcal{M}_\varepsilon := \{M_\varepsilon \mid M \in \mathcal{M}\}$ be a class of approximating models. We use the following lemma to relate the regret under each model in $\mathcal{M}$ to its approximate counterpart.

Lemma G.2. *Let $(\mathcal{X}_1, \mathscr{F}_1), \ldots, (\mathcal{X}_n, \mathscr{F}_n)$ be a sequence of measurable spaces, and let $\mathcal{X}^{(i)} = \prod_{i=t}^{i} \mathcal{X}_t$ and $\mathscr{F}^{(i)} = \bigotimes_{t=1}^{i} \mathscr{F}_t$. For each i, let $\mathbb{P}^{(i)}(\cdot \mid \cdot)$ and $\mathbb{Q}^{(i)}(\cdot \mid \cdot)$ be probability kernels from $(\mathcal{X}^{(i-1)}, \mathscr{F}^{(i-1)})$ to $(\mathcal{X}_i, \mathscr{F}_i)$. Let $\mathbb{P}$ and $\mathbb{Q}$ be the laws of $X_1, \ldots, X_n$ under $X_i \sim \mathbb{P}^{(i)}(\cdot \mid X_{1:i-1})$ and $X_i \sim \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1})$ respectively, and supposed that*

$$D_{\mathrm{TV}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1})) \leq \varepsilon$$

almost surely under $X \sim \mathbb{P}$. Then it holds that

$$D^2_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq O(\varepsilon \cdot n \log n).$$

Using Lemma G.2, we have that for any $M \in \mathcal{M}$ and any algorithm p,

$$|\mathbb{E}^{M,p}[\mathbf{Reg}_{\mathsf{DM}}] - \mathbb{E}^{M_\varepsilon,p}[\mathbf{Reg}_{\mathsf{DM}}]| \leq \widetilde{O}(T^{3/2}\varepsilon^{1/2}).$$

As a result, we have

$$\mathfrak{M}(\mathcal{M}, T) \leq \mathfrak{M}(\mathcal{M}_\varepsilon, T) + \widetilde{O}(T^{3/2}\varepsilon^{1/2}) = \underline{\mathfrak{M}}(\mathcal{M}_\varepsilon, T) + \widetilde{O}(T^{3/2}\varepsilon^{1/2}) \leq \underline{\mathfrak{M}}(\mathcal{M}, T) + \widetilde{O}(T^{3/2}\varepsilon^{1/2}),$$

where the equality holds due to Lemma G.1. Finally, since neither $\mathfrak{M}(\mathcal{M}, T)$ nor $\underline{\mathfrak{M}}(\mathcal{M}, T)$ depends on ε, we can take $\varepsilon \to 0$ to finish the proof.

□

Proof of Lemma G.2. Using Lemma A.13, we have

$$D^2_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leq D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \leq O(\log(n)) \cdot \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))\right].$$

Noting that $D^2_{\mathrm{H}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1})) \leq 2D_{\mathrm{TV}}(\mathbb{P}^{(i)}(\cdot \mid X_{1:i-1}), \mathbb{Q}^{(i)}(\cdot \mid X_{1:i-1}))$ concludes the proof. □

G.2 Proofs from Section 5

G.2.1 Proof of Proposition 5.5

In this section we prove the following quantitative version of Proposition 5.5.

Proposition 5.5a. *Let $\overline{M} \in \mathcal{M}$ and $\eta \geq e$ be given. Suppose there exists $\delta \in (0, 1)$ such that i) the initial distribution has $d_1(s) \geq \delta$ for all s, and ii) for all (s, a, s') and $h \in [H]$, $P_h^{\overline{M}}(s'|s, a) \geq \delta$. Then an approximate version* PC-IGW *algorithm (Algorithm 4) which satisfies*

$$\sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \leq 175 \frac{H^3 SA}{\gamma}.$$

can be implemented in $\mathrm{poly}(H, S, A, \log(\eta/\delta))$ *time with high probability via linear programming.*

Proposition 5.5a requires the additional assumption that the initial state probabilities and transition probabilities are lower bounded by a constant δ. This assumption is fairly mild because the runtime scales with $\log(\delta^{-1})$. In fact, when Algorithm 4 is invoked within $\mathsf{E2D}$, where $\overline{M}$ represents an estimator, one can always modify $\overline{M}$ such that $\delta = 1/\mathrm{poly}(T)$, without worsening the regret by more than a constant factor; we omit the details.

Proof of Proposition 5.5a. It suffices to provide an algorithm that, for any fixed $\bar{s} \in \mathcal{S}$, $\bar{a} \in \mathcal{A}$, and $\bar{h} \in [H]$, solves the optimization problem.

$$\pi_{\bar{h},\bar{s},\bar{a}} = \operatorname*{arg\,max}_{\pi \in \Pi_{\mathrm{RNS}}} \frac{d_{\bar{h}}^{\overline{M},\pi}(\bar{s}, \bar{a})}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))}. \tag{168}$$

The following proposition—proven at the end—shows that an approximate solution suffices.

Proposition G.1. *Consider the setting of Proposition 5.5a, with parameter $\delta > 0$. If we run the* **PC-IGW** *strategy in Algorithm 4 with a collection of policies $\Psi = \{\pi_{h,s,a}\}_{h\in[H],s\in[S],a\in[A]}$ such that*

$$\frac{d_h^{\overline{M},\pi_{h,s,a}}(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{h,s,a}))} \geq \operatorname*{arg\,max}_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d_h^{\overline{M},\pi}(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))} - \frac{\delta}{4HSA + 2\eta}, \tag{169}$$

then by setting $\eta = \frac{\gamma}{41H^2}$, we have

$$\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \leq 175\frac{H^3SA}{\gamma}.$$

To proceed, we formulate (168) as a linear-fractional program and solve it.

Formulating the problem as a linear-fractional program. Let $(\bar{h}, \bar{s}, \bar{a})$ be fixed. We adopt a standard dual approach (e.g., Neu and Pike-Burke (2020)) and—rather than optimizing over policies directly—optimize over occupancy measures, from which an optimal policy $\pi_{\bar{h},\bar{s},\bar{a}}$ can extracted. Let $\boldsymbol{r} := \big(\mathbb{E}_{r_h\sim R_h^{\overline{M}}(s,a)}[r_h]\big)_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}$ be the vector of average rewards for $\overline{M}$. We consider a following linear-fractional program with decision variables $\boldsymbol{d} = (d_{h,s,a})_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}$ representing feasible occupancy measures for $\overline{M}$:

$$\begin{aligned}
\operatorname*{maximize}_{\boldsymbol{d}} \quad & \frac{d_{\bar{h},\bar{s},\bar{a}}}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \boldsymbol{d}, \boldsymbol{r}\rangle)}, \\
\text{subject to} \quad & d_{h,s,a} \geq 0, && \forall h\in[H], s\in\mathcal{S}, a\in\mathcal{A} \\
& \sum_a d_{1,s,a} = d_1(s), && \forall s\in\mathcal{S}, \\
& \sum_{s,a} d_{h,s,a}P_h^{\overline{M}}(s'|s,a) = \sum_a d_{h+1,s',a}, && \forall h\in[H-1], s'\in\mathcal{S}.
\end{aligned} \tag{170}$$

This is a linear-fractional program of HSA decision variables with $HSA + HS$ constraints. We let $\mathrm{OPT}_{\mathrm{LFP}}$ denote the value of the program.

Applying the Charnes-Cooper transformation. Next, we apply the Charnes-Cooper transformation (Charnes and Cooper, 1962) to transform the linear-fractional program above into a linear pogram wth $HSA + 1$ decision variables and $HSA + HS + 2$ constraints. The new program has decision variables $\boldsymbol{w} = (w_{h,s,a})_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}$ and $t \in \mathbb{R}$, with the (implicit) correspondence

$$\begin{aligned}
\boldsymbol{w} &= \frac{\boldsymbol{d}}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \boldsymbol{d}, \boldsymbol{r}\rangle)}, \\
t &= \frac{1}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \boldsymbol{d}, \boldsymbol{r}\rangle)},
\end{aligned}$$

and is defined as follows:

$$\begin{aligned}
\operatorname*{maximize}_{\boldsymbol{w},t} \quad & w_{\bar{h},\bar{s},\bar{a}}, \\
\text{subject to} \quad & 2HSAt + \eta f^{\overline{M}}(\pi_{\overline{M}})t - \eta\langle \boldsymbol{w}, \boldsymbol{r}\rangle = 1, \\
& \sum_{s,a} w_{h,s,a}P_h^{\overline{M}}(s'|s,a) = \sum_a w_{h+1,s',a}, && \forall h\in[H-1], s'\in\mathcal{S}, \\
& \sum_a w_{1,s,a} = td_1(s), && \forall s\in\mathcal{S}, \\
& 1 \geq w_{h,s,a} \geq 0, && \forall h\in[H], s\in\mathcal{S}, a\in\mathcal{A}, \\
& 1 \geq t \geq 0.
\end{aligned} \tag{171}$$

Let $\mathrm{OPT}_{\mathrm{LP}}$ denote the value of this program, which has the following properties:

- Value is preserved: $\mathrm{OPT}_{\mathrm{LFP}} = \mathrm{OPT}_{\mathrm{LP}}$.
- For feasible point $(\boldsymbol{w}, t)$ for (171), the corresponding solution $\boldsymbol{d} = \boldsymbol{w}/t$ is feasible to the linear-fractional program (170).
- If $\boldsymbol{d}$ is feasible for (170), then the variables $(\boldsymbol{w}, t)$ given by

$$
\begin{aligned}
w_{h,s,a} &= \frac{d_{h,s,a}}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \boldsymbol{d}, \boldsymbol{r} \rangle)} \le \frac{1}{HSA} \le 1, \\
t &= \frac{1}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \boldsymbol{d}, \boldsymbol{r} \rangle)} \le \frac{1}{HSA} \le 1.
\end{aligned}
$$

  are feasible for (171).

Solving the linear program. To deduce the final runtime bound, we appeal to a standard linear program solver for (171).

Proposition G.2 (Lee and Sidford (2019), Theorem 1[35])**.** *Let* OPT *denote the value of the linear program*

$$
\begin{aligned}
\underset{x}{\text{maximize}} \quad & \langle c, x \rangle, \\
\text{subject to} \quad & G^\top x = b, \\
& l_i \le x_i \le u_i, \quad \forall i \in [m],
\end{aligned}
\tag{172}
$$

where $G \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$ and $l_i \in \mathbb{R} \cup \{-\infty\}, u_i \in \mathbb{R} \cup \{\infty\}$ are given $x \in \mathbb{R}^m$ is the decision variable. Suppose the following technical conditions hold:

- *$G^\top$ has full row-rank, i.e., no constraints are linearly dependent (this implies that $m \ge n$).*
- *For each $i \in [m]$, $\mathrm{dom}(x_i) := \{x : l_i < x < u_i\}$ is neither the empty set nor the entire real line.*
- *The interior $\Omega := \{x \in R^m : G^\top x = b, l_i < x_i < u_i\}$ is not empty.*

There exists a randomized algorithm LPSolve *which, for any $\varepsilon > 0$, $\alpha > 0$, and initial point $x^0 \in \Omega$, outputs a point $x \in \Omega$ for which $c^\top x \ge \mathrm{OPT} - \varepsilon$ with probability at least $1 - \alpha$, and does so using at most*

$$
O(n^{1.5} m^2 \log^{13} m \cdot \log(mU/\varepsilon) \log(1/\alpha))
$$

steps, where $U := \max\{1/\|u - x^0\|_\infty, 1/\|x^0 - l\|_\infty, \|u - l\|_\infty, \|c\|_\infty\}$.

We now address the technical conditions required to apply LPSolve to (171).

- *No constraints are linearly dependent.* This can be achieved by applying Gaussian elimination to remove dependent constraints, which takes no more than $O(H^3 S^3 A^3)$ steps.
- *The domain of each decision variable is neither the empty set or the entire real line.* This is trivially satisfied, since $1 \ge t, \boldsymbol{w} \ge 0$.
- *The interior of the polytope is not empty.* Below we construct an initial point $x^0 \in \Omega$, certifies that the interior is non-empty.

Finding an initial point. We construct an interior point $x^0 = (\boldsymbol{w}^0, t^0) \in \Omega$ by first finding an interior point $\boldsymbol{d}^0$ of the linear-fractional program (170), and then applying the Charnes-Cooper transformation to obtain $(\boldsymbol{w}^0, t^0)$. First, note that under the assumption in Proposition 5.5a we have that for any policy π, $\sum_a d_h^{\overline{M},\pi}(s,a) \ge \delta$ for all $h \in [H]$, $s \in \mathcal{S}$. Thus, the uniform policy $\pi_{\mathrm{unif}}(s) := \mathrm{unif}(\mathcal{A})$ induces an interior point $\boldsymbol{d}^0 = (d^0_{h,s,a})_{h \in [H], s \in \mathcal{S}, a \in \mathcal{A}}$ for the linear-fractional program (170) by taking

$$
d^0_{h,s,a} := d_h^{\overline{M},\pi_{\mathrm{unif}}}(s,a) = \frac{1}{A} d_h^{\overline{M},\pi_{\mathrm{unif}}}(s) \ge \frac{\delta}{A}.
$$

[35] Theorem 1 in Lee and Sidford (2019) is stated with constant probability, but the high probability result here trivially follows via confidence boosting.

Hence, if we obtain $(\boldsymbol{w}^0, t^0)$ by applying the Charnes-Cooper transformation to $\boldsymbol{d}^0$, we have

$$\frac{1}{HSA} \geq w^0_{h,s,a} := \frac{d^0_{h,s,a}}{2HSA + \eta\big(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \boldsymbol{d}^0, \boldsymbol{r}\rangle\big)} \geq \frac{\delta}{2A(HSA+\eta)},$$

and

$$\frac{1}{HSA} \geq t^0 := \frac{1}{2HSA + \eta\big(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \boldsymbol{d}^0, \boldsymbol{r}\rangle\big)} \geq \frac{1}{2(HSA+\eta)}.$$

It follows that the interior of (171) is non-empty, and that by initializing with $x^0 = (\boldsymbol{w}^0, t^0)$, we have

$$U \leq \max\left\{\left(1 - \tfrac{1}{HSA}\right)^{-1}, \left(\tfrac{\delta}{2A(HSA+\eta)} - 0\right)^{-1}, 1, 1\right\} \leq 2A(HSA+\eta)/\delta.$$

Final guarantee. Applying $\mathsf{LPSolve}$ from $(\boldsymbol{w}^0, t^0)$, in $\widetilde{O}\big((HSA)^{3.5}\log(\eta/(\delta\varepsilon))\log(1\alpha)\big)$ steps we obtain an interior point $(\widetilde{\boldsymbol{w}}, \widetilde{t})$ for which

$$\widetilde{w}_{\bar{h},\bar{s},\bar{a}} \geq \mathrm{OPT}_{\mathrm{LP}} - \varepsilon,$$

with probability at least $1-\alpha$. To obtain a policy, we define an occupancy measure $\widetilde{\boldsymbol{d}} = \widetilde{\boldsymbol{w}}/\widetilde{t}$, then take

$$\widetilde{\pi}_h(s,a) = \widetilde{d}_{h,s,a}/\sum_{a\in\mathcal{A}} \widetilde{d}_{h,s,a}.$$

This policy has the property that $d_h^{\overline{M},\widetilde{\pi}}(s,a) = \widetilde{d}_h(s,a)$. As a result, the policy satisfies

$$\frac{d_{\bar{h}}^{\overline{M},\widetilde{\pi}}(\bar{s},\bar{a})}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\widetilde{\pi}))} = \frac{\widetilde{d}_{\bar{h},\bar{s},\bar{a}}}{2HSA + \eta\big(f^{\overline{M}}(\pi_{\overline{M}}) - \langle \widetilde{\boldsymbol{d}}, \boldsymbol{r}\rangle\big)} = \widetilde{w}_{\bar{h},\bar{s},\bar{a}},$$

and hence

$$\frac{d_{\bar{h}}^{\overline{M},\widetilde{\pi}}(\bar{s},\bar{a})}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\widetilde{\pi}))} \geq \mathrm{OPT}_{\mathrm{LP}} - \varepsilon = \mathrm{OPT}_{\mathrm{LFP}} - \varepsilon = \max_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d_{\bar{h}}^{\overline{M},\pi}(\bar{s},\bar{a})}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))} - \varepsilon.$$

To conclude, we set ε so that the condition of Proposition G.1 holds.

□

Proof sketch for Proposition G.1. We show that the proof of Proposition 5.6 goes through essentially as-is under the condition in (169). Observe that each policy satisfies

$$\begin{aligned}\frac{d_h^{\overline{M},\pi_{h,s,a}}(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi_{h,s,a}))} &\geq \operatorname*{arg\,max}_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d_h^{\overline{M},\pi}(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))} - \frac{\delta}{4HSA+2\eta} \\ &\geq \frac{1}{2}\operatorname*{arg\,max}_{\pi\in\Pi_{\mathrm{RNS}}} \frac{d_h^{\overline{M},\pi}(s,a)}{2HSA + \eta(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi))}.\end{aligned}$$

As a result, Eq. (60) in the proof of Proposition 5.6 continues to hold up to a factor of 2, and changing the parameters η and η' accordingly yields the result. □

G.3 Proofs from Section 8

Theorem 8.1. *Algorithm 8 with exploration parameter $\gamma > 0$ guarantees that*

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sup_{x \in \mathcal{X}} \mathsf{dec}^{D}_{\gamma}(\mathcal{M}|_x, \widehat{\mathcal{M}}|_x) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{D}} \tag{90}$$

almost surely.

Proof of Theorem 8.1. This proof follows the same template as Theorems 4.1 and 4.3. We have

$$\begin{aligned}
\mathbf{Reg}_{\mathsf{DM}} &= \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \left[f^{M^\star}(x^{(t)}, \boldsymbol{\pi}^\star(x^{(t)})) - f^{M^\star}(x^{(t)}, \pi^{(t)}) \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \left[f^{M^\star}(x^{(t)}, \boldsymbol{\pi}^\star(x^{(t)})) - f^{M^\star}(x^{(t)}, \pi^{(t)}) \right] - \gamma \cdot \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \left[D\Big(M^\star(x^{(t)}, \pi^{(t)}) \parallel \widehat{M}^{(t)}(x^{(t)}, \pi^{(t)}) \Big) \right] \\
&\quad + \gamma \cdot \mathbf{Est}_{\mathsf{D}}.
\end{aligned}$$

For each t, since $M^\star \in \mathcal{M}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \left[f^{M^\star}(x^{(t)}, \boldsymbol{\pi}^\star(x^{(t)})) - f^{M^\star}(x^{(t)}, \pi^{(t)}) \right] - \gamma \cdot \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \left[D\Big(M^\star(x^{(t)}, \pi^{(t)}) \parallel \widehat{M}^{(t)}(x^{(t)}, \pi^{(t)}) \Big) \right] \\
&\leq \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} [f^{M}(x^{(t)}, \boldsymbol{\pi}_{M}(x^{(t)})) - f^{M}(x^{(t)}, \pi^{(t)})] - \gamma \cdot \mathbb{E}_{\pi^{(t)} \sim p^{(t)}} \left[D\Big(M(x^{(t)}, \pi^{(t)}) \parallel \widehat{M}^{(t)}(x^{(t)}, \pi^{(t)}) \Big) \right] \\
&= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} [f^{M}(x^{(t)}, \boldsymbol{\pi}_{M}(x^{(t)})) - f^{M}(x^{(t)}, \pi)] - \gamma \cdot \mathbb{E}_{\pi \sim p} \left[D\Big(M(x^{(t)}, \pi) \parallel \widehat{M}^{(t)}(x^{(t)}, \pi) \Big) \right] \\
&= \mathsf{dec}^{D}_{\gamma}(\mathcal{M}|_{x^{(t)}}, \widehat{M}^{(t)}(x^{(t)}, \cdot)).
\end{aligned}$$

We conclude that

$$\mathbf{Reg}_{\mathsf{DM}} \leq \sup_{x \in \mathcal{X}} \sup_{\overline{M} \in \widehat{\mathcal{M}}} \mathsf{dec}^{D}_{\gamma}(\mathcal{M}|_x, \overline{M}(x, \cdot)) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{D}}.$$

$\square$

G.4 Proofs from Section 9

Proposition 9.2. *Consider the Lipschitz bandit problem in which $\Pi = [0,1]^d$, $\mathcal{F}_{\mathcal{M}} = \big\{ f : [0,1]^d \to [0,1] \mid |f(x) - f(y)| \leq \|x - y\|_\infty \big\}$, $\mathcal{O} = \{\varnothing\}$, and $\mathcal{M} = \big\{ \pi \mapsto \mathcal{N}(f(\pi), 1) \mid f \in \mathcal{F}_{\mathcal{M}} \big\}$. For this setting, the information ratio is infinite for all $d \geq 1$:*

$$\mathcal{I}_{\mathrm{B}}(\mathcal{M}) = +\infty.$$

On the other hand, we have $\mathsf{dec}_{\gamma}(\mathcal{M}) \leq \widetilde{O}(\gamma^{-\frac{1}{d+1}})$, and consequently Theorem 3.6 recovers the optimal regret bound $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}] \leq \widetilde{O}(T^{\frac{d+1}{d+2}})$.

Proof of Proposition 9.2. For the upper bound on the DEC, refer to Section 6. We focus on proving the lower bound on the information ratio. We consider the case $d = 1$, since this immediately implies a lower bound for higher dimensions.

We define a subclass of models as follows. Let $\varepsilon \in (0, 1/2)$ be fixed, and let $N = \lfloor \frac{1}{\varepsilon} \rfloor \geq \frac{1}{2\varepsilon}$. For each $i \in [N]$, define $\pi_i = \varepsilon \cdot i - \varepsilon/2$. Let $h(\pi) = \max\{1 - |\pi|, 0\}$, which is 1-Lipschitz. For each i, define M_i via

$$f^{M_i}(\pi) = f_i(\pi) := \frac{1}{2} + \varepsilon \cdot h((\pi - \pi_i)/\varepsilon),$$

which is also 1-Lipschitz. Finally, define $\overline{M}$ via $f^{\overline{M}}(\pi) = \bar{f}(\pi) := \frac{1}{2}$. Let $\mathcal{I}_i = [\pi_i - \varepsilon/2, \pi_i + \varepsilon/2]$, and observe that

$$f_i(\pi_{M_i}) - f_i(\pi) \geq \varepsilon \cdot \mathbb{I}\{\pi \notin \mathcal{I}_i\} \tag{173}$$

and

$$D_{\mathrm{KL}}\big(M_i(\pi) \,\|\, \overline{M}(\pi)\big) = \frac{1}{2}(f_i(\pi) - \bar{f}(\pi))^2 \le \frac{\varepsilon^2}{2}\mathbb{I}\{\pi \in \mathcal{I}_i\}. \tag{174}$$

Observe that for any $x \ge 0, y \ge 0$,

$$\frac{x}{y} = \sup_{\eta>0}\Big\{2\eta x^{1/2} - \eta^2 y\Big\}, \tag{175}$$

with the convention that $0/0 = 0$. As a result, we can lower bound

$$\begin{aligned}
\mathcal{I}_{\mathrm{B}}(\mathcal{M}, \overline{M}) &= \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \frac{(\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^M(\pi) - f^M(\pi_M)])^2}{\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\big]} \\
&\ge \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \sup_{\eta>0}\big\{2\eta\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^M(\pi) - f^M(\pi_M)] - \eta^2\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\big]\big\}.
\end{aligned}$$

Let $\mu = \mathrm{unif}(\{M_i\}_{i=1}^N)$. Then, using the expressions in (173) and (174), we have that for all distributions $p \in \Delta(\Pi)$,

$$\begin{aligned}
&\sup_{\eta>0}\big\{2\eta\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^M(\pi) - f^M(\pi_M)] - \eta^2\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\big]\big\} \\
&\ge \sup_{\eta>0}\bigg\{2\eta\cdot(\varepsilon - 1/N) - \eta^2\frac{\varepsilon^2}{2N}\bigg\} \\
&\ge \sup_{\eta>0}\bigg\{\eta\varepsilon - \eta^2\frac{\varepsilon^2}{2N}\bigg\}.
\end{aligned}$$

This expression is optimized by taking $\eta = \frac{N}{\varepsilon}$, which gives

$$\mathcal{I}_{\mathrm{B}}(\mathcal{M}, \overline{M}) \ge \frac{N}{2} \ge \frac{1}{4\varepsilon}.$$

Since this argument holds uniformly for all $\varepsilon \in (0, 1/2)$, we take $\varepsilon \to 0$ and conclude that $\mathcal{I}_{\mathrm{B}}(\mathcal{M}, \overline{M}) = +\infty$.

□

Proposition 9.3. *Consider the multi-armed bandit setting with $\Pi = [A]$ and $\mathcal{M} = \big\{M(\pi) := \mathcal{N}(f(\pi), 1/2) : f \in [0,1]^A\big\}$. For any $\overline{M} \in \mathcal{M}$ for which $f^{\overline{M}} \in \mathrm{int}([0,1]^A)$ and $\min_{\pi\neq\pi_{\overline{M}}}\{f^{\overline{M}}(\pi) - f^{\overline{M}}(\pi_{\overline{M}})\} > 0$, we have*

$$\mathcal{I}_{\mathrm{F}}(\mathcal{M}, \overline{M}) = +\infty.$$

On the other hand, this example has $\mathsf{dec}_\gamma(\mathcal{M}) \le \mathcal{I}_{\mathrm{B}}(\mathcal{M}) \le O(A/\gamma)$ for all $\gamma > 0$.

Proof of Proposition 9.3. Let $\overline{M}$ be given. Using (175), we can lower bound

$$\begin{aligned}
\mathcal{I}_{\mathrm{F}}(\mathcal{M}, \overline{M}) &= \inf_{p\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M})} \frac{(\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^M(\pi) - f^M(\pi_M)])^2}{\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\big]} \\
&\ge \inf_{p\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M})} \sup_{\eta>0}\big\{2\eta\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^M(\pi) - f^M(\pi_M)] - \eta^2\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\big[D_{\mathrm{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\big]\big\} \\
&\ge \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \sup_{\eta>0}\big\{2\eta\cdot\mathbb{E}_{\pi\sim p}[f^M(\pi) - f^M(\pi_M)] - \eta^2\cdot\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\big]\big\}.
\end{aligned}$$

We further rewrite this as

$$\begin{aligned}
&\inf_{p\in\Delta(\Pi)} \sup_{\eta>0} \sup_{\pi^\star\in\Pi} \sup_{M\in\mathcal{M}} \big\{2\eta\cdot\mathbb{E}_{\pi\sim p}[f^M(\pi) - f^M(\pi^\star)] - \eta^2\cdot\mathbb{E}_{\pi\sim p}\big[D_{\mathrm{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)\big]\big\} \\
&= \inf_{p\in\Delta(\Pi)} \sup_{\eta>0} \sup_{\pi^\star\in\Pi} \sup_{f\in[0,1]^A} \mathbb{E}_{\pi\sim p}\big[2\eta\cdot(f(\pi) - f(\pi^\star)) - \eta^2\cdot(f(\pi) - \bar{f}(\pi))^2\big],
\end{aligned} \tag{176}$$

where $\bar{f} := f_{\overline{M}}$.

Consider any $p \in \Delta(\Pi)$. Recall that $\Delta := \min_{\pi \neq \pi_{\overline{M}}} \{f^{\overline{M}}(\pi) - f^{\overline{M}}(\pi_{\overline{M}})\} > 0$ and $\min \bar{f} > 0$ (since $\bar{f} \in \mathrm{int}([0,1]^A)$). We consider two cases. First, if some $\pi^\star \in \Pi$ has $p_{\pi^\star} = 0$, we can choose the vector f in the supremum in (176) to have $f(\pi) = \bar{f}(\pi)$ for $\pi \neq \pi^\star$ and $f(\pi^\star) = 0$. In this case, for any fixed η, the value in (176) is

$$2\eta \sum_{\pi \neq \pi^\star} p_\pi \bar{f}(\pi) \geq 2\eta \cdot \min_\pi \bar{f}(\pi).$$

Hence, if $\min \bar{f} > 0$, the value is $+\infty$, since we can take η to be arbitrarily large.

For the second case, suppose that $p \in \mathrm{int}(\Delta(\Pi))$ and let η and $\pi^\star$ be fixed. The first-order conditions for optimality imply that the maximizer for f in (176) is given by $f(\pi) = \bar{f}(\pi) + \frac{1}{\eta}$ for $\pi \neq \pi^\star$ and $f(\pi^\star) = \bar{f}(\pi^\star) - \frac{(1-p_{\pi^\star})}{\eta p_{\pi^\star}}$; the feasibility of this choice will be verified shortly. The resulting value is

$$\begin{aligned}
&2\eta \sum_\pi p_\pi(\bar{f}(\pi) - \bar{f}(\pi^\star)) + \frac{2}{p_{\pi^\star}} - (1 - p_{\pi^\star}) - \frac{(1 - p_{\pi^\star})^2}{p_{\pi^\star}} \\
&\geq 2\eta \sum_\pi p_\pi(\bar{f}(\pi) - \bar{f}(\pi^\star)) + \frac{1}{p_{\pi^\star}} - 1.
\end{aligned}$$

If we choose $\pi^\star = \pi_{\overline{M}}$, the expression in (176) is lower bounded by

$$2\eta\Delta(1 - p_{\pi_{\overline{M}}}).$$

Since $p \in \mathrm{int}(\Delta(\Pi))$, we have $(1 - p_{\pi_{\overline{M}}}) > 0$, and hence we can drive the value to $+\infty$ by choosing η arbitrarily large. Furthermore, since $\bar{f} \in \mathrm{int}([0,1]^A)$, we have $f \in [0,1]^A$ as required once η is sufficiently large.

$\square$