Proceedings of Machine Learning Research vol 195:1–75, 2023 36th Annual Conference on Learning Theory

Tight Guarantees for Interactive Decision Making with the Decision-Estimation Coefficient

Dylan J. Foster DYLANFOSTER@MICROSOFT.COM
Microsoft Research

Noah Golowich[*] NZG@MIT.EDU
MIT

Yanjun Han YJHAN@MIT.EDU
MIT

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

A foundational problem in reinforcement learning and interactive decision making is to understand what modeling assumptions lead to sample-efficient learning guarantees, and what algorithm design principles achieve optimal sample complexity. Recently, Foster et al. (2021) introduced the Decision-Estimation Coefficient (DEC), a measure of statistical complexity which leads to upper and lower bounds on the optimal sample complexity for a general class of problems encompassing bandits and reinforcement learning with function approximation. In this paper, we introduce a new variant of the DEC, the *Constrained Decision-Estimation Coefficient*, and use it to derive new lower bounds that improve upon prior work on three fronts:

- they hold in expectation, with no restrictions on the class of algorithms under consideration.
- they hold globally, and do not rely on the notion of *localization* used by Foster et al. (2021).
- most interestingly, they allow the *reference model* with respect to which the DEC is defined to be *improper*, establishing that improper reference models play a fundamental role.

We provide upper bounds on regret that scale with the same quantity, thereby closing all but one of the gaps between upper and lower bounds in Foster et al. (2021). Our results apply to both the regret framework and PAC framework, and make use of several new analysis and algorithm design techniques that we anticipate will find broader use.

1. Introduction

Sample efficiency in reinforcement learning and decision making is a fundamental challenge. State-of-the-art algorithms (Lillicrap et al., 2015; Mnih et al., 2015; Silver et al., 2016) often require millions of rounds of interaction to achieve human-level performance, which is prohibitive in practice and constitutes a barrier to reliable deployment. For continued progress on challenging real-world domains where the agent must navigate high-dimensional state and observation spaces, it is critical to design algorithms that can take advantage of users' domain knowledge (via modeling and function approximation) to enable generalization and improved sample efficiency. As such, a foundational question is to understand what modeling assumptions lead to sample-efficient learning guarantees, and what algorithms achieve optimal sample complexity.

The non-asymptotic theory of reinforcement learning is rich with sufficient conditions under which sample-efficient learning is possible (Dean et al., 2020; Yang and Wang, 2019; Jin et al., 2020;

[*] Work done in part while interning at Microsoft Research.

Modi et al., 2020; Ayoub et al., 2020; Krishnamurthy et al., 2016; Du et al., 2019; Li, 2009; Dong et al., 2019; Zhou et al., 2021), as well as structural properties that aim to unify these conditions (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), but conditions that are *necessary* have generally been elusive. Recently though, Foster et al. (2021) introduced the Decision-Estimation Coefficient (DEC), a unified notion of statistical complexity that leads to both upper and *lower* bounds on the optimal sample complexity in a general decision making framework. The results of Foster et al. (2021) show that the DEC plays a role for interactive decision making analogous to that of the VC dimension and its relatives in statistical learning, but leave room for tighter quantitative guarantees. In this paper, we introduce a new variant of the DEC, the *Constrained Decision-Estimation Coefficient*, and use it to close several gaps between the upper and lower bounds in Foster et al. (2021).

1.1. Interactive Decision Making

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

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

1. The learner selects a *decision* $\pi^t \in \Pi$, where Π is the *decision space*.
2. The learner receives a reward $r^t \in \mathcal{R} \subseteq \mathbb{R}$ and observation $o^t \in \mathcal{O}$ sampled via $(r^t, o^t) \sim M^\star(\pi^t)$, where $M^\star : \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ is the underlying *model*.

We refer to $\mathcal{R}$ as the *reward space* and $\mathcal{O}$ as the *observation space*. The model $M^\star$, which we formalize as a conditional distribution, plays the role of the underlying environment, and is unknown to the learner. However, the learner is assumed to have access to a *model class* $\mathcal{M} \subset (\Pi \to \Delta(\mathcal{R} \times \mathcal{O}))$ that contains $M^\star$.

Assumption 1.1 (Realizability)**.** *The learner has access to a class $\mathcal{M}$ containing the true model $M^\star$.*

The model class $\mathcal{M}$ represents the learner's prior knowledge of the underlying environment. For structured bandit problems, where models correspond to reward distributions, it encodes structure in the reward landscape (smoothness, linearity, convexity, etc.), and for reinforcement learning problems, where models correspond to Markov decision processes (MDPs), it typically encodes structure in the transition probabilities or value functions. In more detail:

- **Bandits.** In bandit problems, $\mathcal{M}$ is a reward distribution, π^t is referred to as an *action* or *arm* and Π is referred to as the *action space*; there are no observations beyond rewards ($\mathcal{O} = \{\varnothing\}$). By choosing $\mathcal{M}$ so that the mean reward function f^M has appropriate structure, one can capture bandit problems with continuous or infinite action spaces and structured rewards, including linear bandits (Dani et al., 2007; Abernethy et al., 2008; Bubeck et al., 2012), bandit convex optimization, and nonparametric bandits (Kleinberg, 2004; Bubeck et al., 2011; Magureanu et al., 2014).
- **Reinforcement learning.** In episodic reinforcement learning, each model $M \in \mathcal{M}$ specifies a non-stationary horizon-H 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. Decisions $\pi = \{\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A})\}_{h=1}^H$ are *policies* (mappings from states to actions). Given a policy π, an episode proceeds as follows (beginning from $s_1 \sim d_1$). For $h = 1, \ldots, H$: $a_h \sim \pi_h(s_h)$, $r_h \sim R_h^M(s_h, a_h)$, and $s_{h+1} \sim P_h^M(\cdot \mid s_h, a_h)$. This process leads to $(r, o) \sim M(\pi)$, where $r = \sum_{h=1}^H r_h$ is the cumulative reward in the episode, and the observation $o = (s_1, a_1, r_1), \ldots (s_H, a_H, r_H)$ is the episode's trajectory (sequence of observed states, actions, and rewards). By choosing $\mathcal{M}$ appropriately, one can encompass standard classes of MDPs (e.g., tabular MDPs or linear systems) Dean et al. (2020); Yang and Wang (2019); Jin et al. (2020); Modi et al. (2020); Ayoub et al. (2020); Krishnamurthy et al. (2016); Du et al. (2019); Li (2009); Dong et al. (2019), as well as more general structural conditions (Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020; Du et al., 2021; Jin et al., 2021).

For a model $M \in \mathcal{M}$, $\mathbb{E}^{M,\pi}[\cdot]$ denotes expectation under the process $(r, o) \sim M(\pi)$, $f^M(\pi) := \mathbb{E}^{M,\pi}[r]$ is the mean reward function, and $\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$ is the optimal decision.

We consider two types of guarantees for interactive decision making: regret guarantees and PAC (Probably Approximately Correct) guarantees. For regret guarantees, we are concerned with the *cumulative suboptimality* given by

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

where p^t is the learner's randomization distribution (conditional distribution over π^t) for round t. For PAC guarantees, we are only concerned with *final suboptimality*. After rounds $t = 1, \ldots, T$ complete, the learner can use all of the data collected to select a final decision $\widehat{\pi}$ (which may be randomized according to a distribution $p \in \Delta(\Pi)$), and we measure performance via

$$\mathbf{Risk}_{\mathsf{DM}}(T) := \mathbb{E}_{\widehat{\pi} \sim p}\big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi})\big]. \tag{2}$$

1.2. Background: Decision-Estimation Coefficient

To motivate our results, let us first recall the Decision-Estimation Coefficient of Foster et al. (2021); for this discussion, we restrict our attention to regret guarantees. Define the squared Hellinger distance for probability measures $\mathbb{P}$ and $\mathbb{Q}$ with a common dominating measure ν by

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

For a model class $\mathcal{M}$, reference model $\overline{M} : \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$, and scale parameter $\gamma > 0$, the Decision-Estimation Coefficient is given by

$$\text{r-dec}^{\text{o}}_\gamma(\mathcal{M}, \overline{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \tag{4}$$

Here, we depart slightly from the notation in Foster et al. (2021) and append the prefix r- (indicating "regret") and the superscript "o" (indicating "offset") to distinguish from other variants that will be introduced shortly.

Foster et al. (2021) (see also Foster et al. (2022b)) use the DEC to provide upper and lower bounds on the optimal regret for interactive decision making:

- On the upper bound side, there exists an algorithm (*Estimation-to-Decisions*) that obtains

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \lesssim \inf_{\gamma>0}\left\{ \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})\cdot T + \gamma\cdot \mathbf{Est}_{\mathsf{H}}(T)\right\}, \tag{5}$$

  where $\mathrm{co}(\mathcal{M})$ denotes the convex hull of the class $\mathcal{M}$. Here, the term $\mathbf{Est}_{\mathsf{H}}(T)$ represents the sample complexity required to perform online statistical estimation with the class, and has $\mathbf{Est}_{\mathsf{H}}(T) \leq \log|\mathcal{M}|$ for the special case of finite classes.

- On the lower bound side, *any algorithm* must have

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \gtrsim \sup_{\gamma>0}\left\{ \sup_{\overline{M}\in\mathcal{M}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}'_{\gamma},\overline{M})\cdot T\right\}, \tag{6}$$

  where $\mathcal{M}'_{\gamma}$ is a certain "localized" subclass of $\mathcal{M}$ (informally, models with $\|f^{M} - f^{\overline{M}}\|_{\infty} \lesssim \frac{\gamma}{T}$).

While the results of Foster et al. (2021) show that the DEC characterizes learnability for various classes of models (for instance, convex classes with low "estimation complexity"), the quantitative rates leave room for improvement. The aim of this paper is to address the following gaps:

- The lower bound (6) restricts to a localized subclass of $\mathcal{M}'_{\gamma} \subseteq \mathcal{M}$, which may have lower complexity than the original class. This yields reasonable results for many special cases, but can be arbitrarily loose in general.[1]

- The lower bound (6) restricts to reference models $\overline{M} \in \mathcal{M}$, yet the upper bound (5) maximizes over reference models $\overline{M} \in \mathrm{co}(\mathcal{M})$, potentially leading to larger regret. For example, Proposition 3.1 of Foster et al. (2021) gives a model class $\mathcal{M}$ for which this distinction leads to an arbitrarily large gap between the upper and lower bounds.

A third gap, which we do not address, is the presence of the estimation complexity $\mathbf{Est}_{\mathsf{H}}(T)$ in Eq. (5), which is not present in the lower bound. See Foster et al. (2021, 2022a) for further discussion around this issue.

1.3. Constrained Decision-Estimation Coefficient and Overview of Results

To address the issues in the prequel, we introduce a new complexity measure, the *Constrained Decision-Estimation Coefficient*. For a reference model $\overline{M} : \Pi \to \Delta(\mathcal{R}\times\mathcal{O})$, define

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M}) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \mid \mathbb{E}_{\pi\sim p}\left[D^{2}_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq \varepsilon^{2}\right\}, \tag{7}$$

with the convention that the value above is zero whenever the set

$$\mathcal{H}_{p,\varepsilon}(\overline{M}) := \left\{M\in\mathcal{M} \mid \mathbb{E}_{\pi\sim p}\left[D^{2}_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq \varepsilon^{2}\right\} \tag{8}$$

[1] We remark that Foster et al. (2021) provides additional lower bounds that hold with low probability (as opposed to in expectation) and allow slightly larger localized model classes.

is empty; the superscript "c" indicates "constrained", and distinguishes from the offset counterpart. Our main quantity of interest is

$$\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}). \tag{9}$$

The constrained and offset DEC differ only in how the quantity $\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big]$, representing information gain, is incorporated. The constrained DEC places a hard constraint on the information gain, while the offset DEC treats the information gain as a penalty for the max-player, amounting to a soft constraint. At first glance, one might expect these quantities to be equivalent via Lagrangian duality. It is indeed the case that the offset DEC upper bounds the constrained DEC: $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M}) \leq \inf_{\gamma}\{\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M}) \vee 0 + \gamma\varepsilon^2\}\ \forall\overline{M},$ but strong duality fails, and the complexity measures are not equivalent in general; detailed discussion is given in Appendix D.

Main results for regret framework. The constrained DEC possesses a number of useful properties not shared by the offset DEC, including *implicitly* enforcing a form of localization in an adaptive fashion (cf. Appendix D). Leveraging these properties, we provide improved lower and upper bounds that close all but one of the gaps between the bounds in Foster et al. (2021). Our main result is as follows.

Theorem (informal version of Theorems C.1 and C.2). *For any model class $\mathcal{M}$:*

- *Lower bound: For a worst-case model in $\mathcal{M}$, any algorithm must have*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(1)\cdot\text{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M})\cdot T$$

  for $\underline{\varepsilon}(T) = \widetilde{\Theta}\big(\sqrt{1/T}\big)$.

- *Upper bound: There exists an algorithm (Estimation-to-Decisions^{+}) that achieves*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq \widetilde{O}(1)\cdot\text{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})\cdot T$$

  for $\bar{\varepsilon}(T) = \widetilde{\Theta}\big(\sqrt{\mathbf{Est}_{\mathsf{H}}(T)/T}\big) \leq \widetilde{\Theta}\big(\sqrt{\log|\mathcal{M}|/T}\big)$.

In the above theorem, $\widetilde{\Omega}(\cdot), \widetilde{\Theta}(\cdot), \widetilde{O}(\cdot)$ hide logarithmic factors. This lower and upper bound are always tighter than the respective bounds in prior work, and the lower bound in particular improves upon Foster et al. (2021) on two fronts:

- It holds globally, and removes the notion of *localization* used by Foster et al. (2021). In addition, it holds in expectation, with no restriction on the class of algorithms under consideration.
- More interestingly, it scales with $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})$ as opposed to, say, $\sup_{\overline{M}\in\mathcal{M}}\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})$, showing (in tandem with the upper bound) that *improper* reference models $\overline{M}\in\mathrm{co}(\mathcal{M})$ play a fundamental role, and are not simply an artifact of the upper bounds in Foster et al. (2021).

Together, our upper and lower bounds form an important step toward building a simple, user-friendly, and unified theory for interactive decision making based on the Decision-Estimation Coefficient. The only gap our results leave open is the role of the estimation error $\mathbf{Est}_{\mathsf{H}}(T)$, a deep issue that necessitates future research.

On the technical side, our lower bounds make use of several new analysis techniques and structural results that depart sharply from previous approaches (Foster et al., 2021, 2022b). Our upper bounds build upon the Estimation-to-Decisions paradigm introduced in Foster et al. (2021), but the design and analysis are substantially more sophisticated, as certain properties of the constrained DEC that aid in proving lower bounds lead to non-trivial challenges in deriving upper bounds. We anticipate that our techniques will find broader use, and to this end we provide user-friendly tools for working with the constrained DEC in Appendix D. In particular, Appendix D.1 establishes relationships between the offset and constrained DEC, Appendix D.2 elaborates on the relationship between the constrained DEC and the notion of *localization* introduced in Foster et al. (2021), and Appendix D.4 investigates the relationship between different choices for improper estimators $\overline{M}$ with respect to the (offset and constrained) DEC.

Remark 1.1. At this point, the reader may wonder why the definition (9) incorporates the set $\mathcal{M} \cup \{\overline{M}\}$, where $\overline{M}$ may lie outside the class $\mathcal{M}$. We show in Appendix D.3 that this modification, despite not being required for the offset DEC, plays a central role for the constrained DEC.

Main results for PAC framework. Moving from the regret framework to PAC, we work with the following PAC counterparts to the offset and constrained DEC:

$$
\begin{aligned}
\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) &= \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi \sim q}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \right\} \\
\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) &= \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim q}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2 \right\},
\end{aligned}
\tag{10}
$$

with the convention that the value in Eq. (10) is zero when $\mathcal{H}_{q,\varepsilon}(\overline{M}) = \varnothing$. These definitions parallel (4) and (7), but allow the min-player to select a separate *exploration distribution* q under which the information gain (Hellinger distance) is evaluated, and *exploitation distribution* p under which regret is evaluated. Since one can always choose $p = q$, it is immediate that $\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \leq \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M})$, and likewise $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$. Defining[2] $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, we show that the PAC DEC leads to the following lower and upper bounds on PAC sample complexity.

Theorem (informal version of Theorems 2.1 and 2.2). *For any model class $\mathcal{M}$:*

- *Lower bound: For a worst-case model in $\mathcal{M}$, any PAC algorithm with T rounds of interaction must have*
  $$\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(1) \cdot \textsf{p-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M})$$
  for $\underline{\varepsilon}(T) = \widetilde{\Theta}\big(\sqrt{1/T}\big)$.

- *Upper bound: There exists an algorithm (Estimation-to-Decisions^{+}) that achieves*
  $$\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \leq \widetilde{O}(1) \cdot \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})$$
  for $\bar{\varepsilon}(T) = \widetilde{\Theta}\big(\sqrt{\mathbf{Est}_{\mathsf{H}}(T)/T}\big) \leq \widetilde{\Theta}\big(\sqrt{\log|\mathcal{M}|/T}\big)$.

[2] Compared the definition of the constrained DEC for regret, the definition $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$ does not apply the DEC to the class $\mathcal{M} \cup \{\overline{M}\}$. See Appendix D.3 for a detailed explanation.

Organization. In what follows (Section 2), we formally present our main lower and upper bounds for the PAC framework. ***Due to space constraints, our guarantees for regret, as well as additional results, are deferred to Part I of the appendix:*** Appendix C presents lower and upper bounds for regret. Appendix D establishes structural results for the constrained DEC, and Appendix E provides a detailed comparison to bounds from prior work. Additional examples are given in Appendix F.

Related work. Concurrent work of Chen et al. (2022) independently discovered the offset variant of the PAC Decision-Estimation Coefficient, and used it to give upper and lower bounds for PAC sample complexity by adapting the techniques of Foster et al. (2021). Our guarantees for both regret and PAC are always tighter than these results, analogous to the improvement we obtain over Foster et al. (2021) (see Appendix D.4), but our techniques are otherwise complementary.

Additional notation. For an integer $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, \ldots, n\}$. For a set $\mathcal{Z}$, we let $\Delta(\mathcal{Z})$ denote the set of all probability distributions over $\mathcal{Z}$, and let $\mathcal{Z}^{\mathrm{c}}$ denote the complement. We adopt standard big-oh notation, and write $f = \widetilde{O}(g)$ to denote that $f = O(g \cdot \max\{1, \mathrm{polylog}(g)\})$. We use $\lesssim$ only in informal statements to emphasize the most notable elements of an inequality.

We assume throughout the paper that $\mathcal{R} = [0, 1]$ unless otherwise stated. The history up to time t is denoted by $\mathfrak{H}^t = (\pi^1, r^1, o^1), \ldots, (\pi^t, r^t, o^t)$. For the class $\mathcal{M}$, we define $V(\mathcal{M}) := \sup_{M,M' \in \mathcal{M}} \sup_{\pi \in \Pi} \sup_{\mathcal{E} \in \mathscr{R} \otimes \mathscr{O}} \left\{ \frac{M(\mathcal{E}|\pi)}{M'(\mathcal{E}|\pi)} \right\} \vee e$, where $\mathscr{R}, \mathscr{O}$ denote sigma-algebras for the spaces $\mathcal{R}, \mathcal{O}$, respectively (see Appendix B). Finiteness of $V(\mathcal{M})$ is not necessary for our any of our results, but improves our lower bounds (Theorems 2.1 and C.1) by a $\log(T)$ factor. We define $\mathcal{M}^{+} = \{\overline{M} : \Pi \to \Delta(\mathcal{R} \times \mathcal{O}) \mid \sup_{M \in \mathcal{M}} \sup_{\pi \in \Pi} \sup_{\mathcal{E} \in \mathscr{R} \otimes \mathscr{O}} \left\{ \frac{\overline{M}(\mathcal{E}|\pi)}{M(\mathcal{E}|\pi)} \right\} \leq V(\mathcal{M})\}$ as the space of all models with rewards in $\mathcal{R}$ that obey the same density ratio bound; note that $\mathrm{co}(\mathcal{M}) \subseteq \mathcal{M}^{+}$.

2. PAC Framework: Upper and Lower Bounds

In this section, we provide upper and lower bounds based on the constrained Decision-Estimation Coefficient for the PAC framework. Our upper and lower bounds for regret (Appendix C) have a nearly identical form, and build on the techniques we introduce here, but are more involved.

2.1. Lower Bounds

We provide a minimax lower bound for interactive decision making, which show for any model class $\mathcal{M}$ and horizon $T \in \mathbb{N}$, the worst-case PAC sample complexity for any algorithm is lower bounded by the constrained DEC for an appropriate choice of the radius parameter $\varepsilon > 0$. To state the results, define $C(T) := \log(T \wedge V(\mathcal{M}))$.

Theorem 2.1 (Main Lower Bound: PAC). *Let $\underline{\varepsilon}(T) := c \cdot \frac{1}{\sqrt{TC(T)}}$, where $c > 0$ is a sufficiently small numerical constant and $C := 48\sqrt{2}$. For all $T \in \mathbb{N}$ such that the condition*

$$\textsf{p-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \geq C \cdot \underline{\varepsilon}(T) \tag{11}$$

is satisfied, it holds that for any PAC algorithm, there exists a model in $\mathcal{M}$ such that

$$\mathbb{E}\left[\mathbf{Risk}_{\textsf{DM}}(T)\right] \geq \Omega(1) \cdot \sup_{\overline{M} \in \mathcal{M}^{+}} \textsf{p-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}, \overline{M}) \geq \Omega(1) \cdot \textsf{p-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}). \tag{12}$$

Theorem 2.1 shows that the constrained DEC is a fundamental limit for interactive decision making in the PAC framework. We will show in a moment (Section 2.2) that the lower bound can be achieved algorithmically, up to a difference in radius that depends on the estimation capacity for $\mathcal{M}$ ($\sqrt{\log|\mathcal{M}|/T}$ versus $1/\sqrt{T}$ for the case of finite classes). We defer a detailed comparison to prior work to Appendix E, and take this time to build intuition as to the behavior of the lower bound.

Remark 2.1. Whenever $V(\mathcal{M}) = O(1)$, we have $\underline{\varepsilon}(T) \propto 1/\sqrt{T}$ in Theorem 2.1. In the general case where $V(\mathcal{M})$ is not bounded, we have $\underline{\varepsilon}(T) \propto 1/\sqrt{T\log(T)}$, and the lower bounds lose a logarithmic factor. For many classes, one has $V(\mathcal{M}) = \infty$, but there exists a subclass $\mathcal{M}' \subseteq \mathcal{M}$ with $V(\mathcal{M}') = O(1)$ and $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}') \gtrsim \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$. In this case, one can derive a tighter lower bound with $\underline{\varepsilon}(T) \propto 1/\sqrt{T}$ by applying Theorem 2.1 to $\mathcal{M}'$. See Foster et al. (2021) for examples.

Remark 2.2. Theorem 2.1 scales with the quantity $\sup_{\overline{M}\in\mathcal{M}^{+}} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \geq \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$, which allows for reference models $\overline{M} \notin \mathrm{co}(\mathcal{M})$. We show in Appendix D.4 that maximizing over reference models $\overline{M} \in \mathcal{M}^{+}$ does not increase the value of the DEC beyond what is attained by $\overline{M} \in \mathrm{co}(\mathcal{M})$, so this result does not contradict our upper bounds, which take the maximum over $\overline{M} \in \mathrm{co}(\mathcal{M})$.[3] We state the lower bounds in this form because 1) our proof works with $\overline{M} \in \mathcal{M}^{+}$ directly and does not use the structure of $\mathrm{co}(\mathcal{M})$, and 2) allowing for $\overline{M} \in \mathcal{M}^{+}$ often simplifies calculations.

Understanding the lower bound. Let us give a sense for how the lower bound in Theorem 2.1 behaves for standard model classes; we refer to Appendix F for further examples and details.

- $\sqrt{T}$*-rates.* For the most well-studied classes found throughout the literature on bandits and reinforcement learning, we have $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}}$, where $C_{\mathrm{prob}} > 0$ is a problem-dependent constant that reflects some notion of intrinsic complexity. In this case, the condition (11) is satisfied whenever C_{prob} is larger than some numerical constant, and Theorem 2.1 gives $\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}\left(\sqrt{\frac{C_{\mathrm{prob}}}{T}}\right)$, which implies that $\widetilde{\Omega}\left(\frac{C_{\mathrm{prob}}}{\varepsilon^2}\right)$ samples are required to learn an ε-optimal policy. Examples (cf. Appendix F) include multi-armed bandits with A actions, where $C_{\mathrm{prob}} \geq A$ (leading to $\widetilde{\Omega}\big(\frac{A}{\varepsilon^2}\big)$ sample complexity), linear bandits in dimension d, where $C_{\mathrm{prob}} \geq d$ (leading to $\widetilde{\Omega}\big(\frac{d}{\varepsilon^2}\big)$ sample complexity), and tabular reinforcement learning with S states, A actions, and horizon H, where $C_{\mathrm{prob}} \geq HSA$ (leading to $\widetilde{\Omega}\big(\frac{HSA}{\varepsilon^2}\big)$ sample complexity).

- *Nonparametric rates.* For nonparametric model classes, where the optimal rate is slower than $\frac{1}{\sqrt{T}}$, one typically has $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$ for some $\rho \in (0,1)$. In this case, the condition in Eq. (11) is satisfied whenever T is a sufficiently large constant, and Theorem 2.1 gives $\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(T^{-\frac{(1-\rho)}{2}})$, which implies that $\widetilde{\Omega}(\varepsilon^{-\frac{2}{1-\rho}})$ samples are required to learn a ε-optimal policy.

- *Fast rates.* For problems with low noise, such as noiseless bandits, the DEC can exhibit threshold behavior, where $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \geq 1/\sqrt{C_{\mathrm{prob}}}\}$ for a problem-dependent parameter C_{prob}. In this case, Theorem 2.1 gives $\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \geq \Omega(1)$ until $T = \widetilde{\Omega}(C_{\mathrm{prob}})$; that is, at least $\widetilde{\Omega}(C_{\mathrm{prob}})$ samples are required to learn beyond constant suboptimality.

[3] In fact, the results of Appendix D.4 hold even if $\mathcal{M}^{+}$ is defined to be the set of *all* models $\overline{M} : \Pi \rightarrow \Delta(\mathcal{R} \times \mathcal{O})$.

Proof techniques. We now highlight some of the key ideas behind the proof of Theorem 2.1; refer to Appendix H.1 for the detailed proof. The high-level structure of the proof is as follows. For any algorithm, one can construct a "hard" pair of models $M_1, M_2 \in \mathcal{M}$ such that:

1. The laws of the history $\mathfrak{H}^T$ induced by the algorithm under M_1 and M_2 are close in total variation (i.e., $D_{\mathsf{TV}}(\mathbb{P}^{M_1}, \mathbb{P}^{M_2}) \leq 1/4$, where $\mathbb{P}^M$ is the law of $\mathfrak{H}^T$ under M).

2. Any algorithm with risk much smaller than the DEC must query substantially different decisions in Π depending on whether the underlying model is M_1 or M_2.

Since $\mathbb{P}^{M_1}$ are $\mathbb{P}^{M_2}$ close in total variation, the algorithm will fail to distinguish the models with constant probability, and since the optimal decisions for the models are (approximately) exclusive, the lower bound follows.

To make this approach concrete, we select the pair of models (M_1, M_2) in an *adversarial* fashion based on the algorithm under consideration, in a way that generalizes the approach taken in Foster et al. (2021, 2022b). We fix an arbitrary model $M_1 \in \mathcal{M}$, then, letting $q_{M_1} := \mathbb{E}^{M_1}\left[\frac{1}{T}\sum_{t=1}^{T} q^t(\cdot \mid \mathfrak{H}^{t-1})\right]$ and $p_{M_1} := \mathbb{E}^{M_1}[p(\cdot \mid \mathfrak{H}^T)]$ denote the learner's average play under this model, choose M_2 as the model that attains the maximum in Eq. (10) with (p_{M_1}, q_{M_1}) plugged in. This approach suffices to prove lower bounds that scale with $\sup_{\overline{M}\in\mathcal{M}} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, but is not sufficient to incorporate improper reference models and prove a lower bound that scales with $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$. Indeed, unless the class $\mathcal{M}$ is convex, there is no reason why an algorithm with low risk for models in $\mathcal{M}$ should have low risk for improper *mixtures* $\overline{M} \in \mathrm{co}(\mathcal{M})$. Thus, naively choosing $M_1 = \overline{M} \in \mathrm{co}(\mathcal{M})$ is problematic, as we have no way to relate the algorithm's risk under M_1 to that for models in the class.

To circumvent this issue, we iterate the process above: Given a potentially improper reference model $\overline{M} \in \mathcal{M}^+$, we first obtain M_1 by finding the model that attains the maximum in Eq. (10) with $(p_{\overline{M}}, q_{\overline{M}})$ plugged in. With this model in hand, we obtain M_2 in a similar fashion, but condition on the event the learner behaves near-optimally for M_1. That is, we find the maximizer for Eq. (10) with the distribution $(p_{\overline{M}}(\cdot \mid \mathcal{E}_1), q_{\overline{M}})$ plugged in, where $\mathcal{E}_1$ is the set of near-optimal decisions for M_1. This argument leads to lower bounds that scale with $\sup_{\overline{M}\in\mathcal{M}^+} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \geq \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$ because the reference model $\overline{M}$ acts only as a midpoint between M_1 and M_2 (whose existence allows us to control the total variation between the models), and as a result is not required to live in $\mathcal{M}$.

Benefits of the constrained DEC. The main technical advantage gained by working with the constrained DEC over the offset DEC is that, by placing a hard constraint on the Hellinger distance between models under consideration, we can appeal to stronger change-of-measure arguments than those considered in prior work; this is key to deriving in-expectation (as opposed to low probability) lower bounds. In particular, the radius $\underline{\varepsilon}(T) \approx 1/\sqrt{T}$ is the largest possible choice such that for any algorithm, one can find a worst-case pair of models for which the total variation distance $D_{\mathsf{TV}}(\mathbb{P}^{M_1}, \mathbb{P}^{M_2})$ is a *small constant* (say, $1/4$). Whenever the total variation distance between the induced laws is constant, the algorithm must fail to distinguish M_1 and M_2 with constant probability, which entails large risk if the optimal decisions for M_1 and M_2 are significantly different.

Foster et al. (2021) emphasize that the Decision-Estimation Coefficient can be thought of as interactive counterpart to the modulus of continuity in statistical estimation (Donoho and Liu, 1987, 1991a,b). We find the constrained DEC to be a more direct analogue than the offset DEC: The modulus of continuity places a hard constraint on Hellinger distance for similar technical reasons, and lower bounds based on the modulus make use of the same $1/\sqrt{T}$ radius.

2.2. Upper Bounds

We now give an algorithm and upper bound on PAC sample complexity that complements the lower bound in Theorem 2.1. Our algorithms—both for PAC and regret—make use of the *Estimation-to-Decisions* paradigm of Foster et al. (2021), but incorporate substantial refinements tailored to the constrained DEC. Our algorithm for regret (Appendix C) is particularly involved, and builds on the ideas we introduce for PAC here.

2.2.1. ONLINE ESTIMATION

Our algorithms and regret bounds use the primitive of an *online estimation oracle*, denoted by $\mathbf{Alg}_{\mathsf{Est}}$, which is an algorithm that, given knowledge of some class $\mathcal{M}$ containing the true model $M^\star$, estimates the underlying model $M^\star$ from data on the fly. At each round t, given the data $\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$ observed so far, the estimation oracle builds an estimate $\widehat{M}^t = \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi^i, r^i, o^i)\}_{i=1}^{t-1}\right)$ for the true model $M^\star$. We measure the oracle's estimation performance in terms of cumulative Hellinger error, which we assume is bounded as follows.

Assumption 2.1 (Estimation oracle for $\mathcal{M}$). *At each time $t \in [T]$, an online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ for $\mathcal{M}$ returns, given $\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$ with $(r^i, o^i) \sim M^\star(\pi^i)$ and $\pi^i \sim p^i$, an estimator $\widehat{M}^t \in \mathrm{co}(\mathcal{M})$ such that whenever $M^\star \in \mathcal{M}$,*

$$\mathbf{Est}_{\mathsf{H}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right] \leq \mathbf{Est}_{\mathsf{H}}(T, \delta), \tag{13}$$

with probability at least $1 - \delta$, where $\mathbf{Est}_{\mathsf{H}}(T, \delta)$ is a known upper bound.

Algorithms satisfying Assumption 2.1 can be obtained via online conditional density estimation (that is, online learning with the logarithmic loss). Typically, the best possible estimation rate $\mathbf{Est}_{\mathsf{H}}(T, \delta)$ will reflect the statistical capacity of the class $\mathcal{M}$. Standard examples include finite classes, where the exponential weights algorithm (also known as Vovk's aggregating algorithm) achieves $\mathbf{Est}_{\mathsf{H}}(T, \delta) \leq O(\log(|\mathcal{M}|/\delta))$, and parametric classes in $\mathbb{R}^d$, where one can achieve $\mathbf{Est}_{\mathsf{H}}(T, \delta) \leq \widetilde{O}(d)$. See Section 4 of Foster et al. (2021) for further background.

2.2.2. ALGORITHM AND UPPER BOUND FOR PAC

Algorithm 1 displays our main algorithm for the PAC framework, $\mathsf{E2D}^+$. The algorithm is built upon the Estimation-to-Decisions paradigm of Foster et al. (2021), which uses the following scheme for each round t: **1)** Obtain an estimator $\widehat{M}^t \in \mathrm{co}(\mathcal{M})$ for $M^\star$ from the online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$, **2)** Sample π^t from a decision distribution obtained by solving the min-max optimization problem that defines the DEC, with $\widehat{M}^t$ plugged in as the reference model. $\mathsf{E2D}^+$ follows this template, but incorporates non-trivial changes that are tailored to i) the constrained (as opposed to offset) DEC and ii) PAC guarantees (as opposed to regret). Briefly, the algorithm consists of two phases, an *exploration phase* and an *exploitation phase*, which we outline below.

Exploration phase. In the *exploration phase*, which consists of rounds $t = 1, \ldots, J$, where $J = \widetilde{\Omega}(T)$, Algorithm 1 repeatedly obtains an estimator $\widehat{M}^t$ by querying the estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ with the current dataset $\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$ (Line 5), then computes

Algorithm 1 Estimation-to-Decisions ($\mathsf{E2D}^{+}$) for PAC

```
1:  parameters:
        Number of rounds $T \in \mathbb{N}$.
        Failure probability $\delta > 0$.
        Online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$.
2:  Define $L := \lceil \log 2/\delta \rceil$, $J := \frac{T}{L+1}$, and $\overline{\mathbf{Est}}_{\mathsf{H}} := \mathbf{Est}_{\mathsf{H}}\left(\frac{2T}{\lceil \log 2/\delta \rceil}, \frac{\delta}{4\lceil \log 2/\delta \rceil}\right)$.
3:  Set $\bar{\varepsilon}(T) := 8\sqrt{\frac{\lceil \log 2/\delta \rceil}{T} \cdot \overline{\mathbf{Est}}_{\mathsf{H}}}$.
    /* Exploration phase */
4:  for $t = 1, 2, \cdots, J$ do
5:      Obtain estimate $\widehat{M}^t = \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi^i, r^i, o^i)\}_{i=1}^{t-1}\right)$.
6:      Compute
            $$(p^t, q^t) := \operatorname*{arg\,min}_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{H}_{q,\bar{\varepsilon}(T)}(\widehat{M}^t)} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)],$$
        with the convention that the value is zero if $\mathcal{H}_{q,\bar{\varepsilon}(T)}(\widehat{M}^t) = \varnothing$.
7:      Sample decision $\pi^t \sim q^t$ and update estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ with $(\pi^t, r^t, o^t)$.
    /* Exploitation phase */
8:  Sample $L$ indices $t_1, \ldots, t_L \sim \mathrm{Unif}([J])$ independently.
9:  For each $\ell \in [L]$, draw $J$ independent samples $\pi_\ell^1, \ldots, \pi_\ell^J \sim q^{t_\ell}$, and observe $(\pi_\ell^j, r_\ell^j, o_\ell^j)$ for
    each $j \in [J]$.
10: For each $\ell \in [L]$ and $j \in [J]$, compute
            $$\widetilde{M}_\ell^j := \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi_\ell^i, r_\ell^i, o_\ell^i)\}_{i=1}^{j-1}\right),$$
    and let $\widetilde{M}_\ell := \frac{1}{J}\sum_{j=1}^{J} \widetilde{M}_\ell^j$.        // $\widetilde{M}_\ell$ is a high-quality estimate for $M^\star$ under $q^{t_\ell}$.
11: output: Set $\widehat{p} := p^{t_{\widehat{\ell}}}$ and output $\widehat{\pi} \sim \widehat{p}$, where $\widehat{\ell} := \operatorname*{arg\,min}_{\ell\in[L]} \mathbb{E}_{\pi\sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{t_\ell}(\pi), \widetilde{M}_\ell(\pi)\right)\right]$.
```

the pair of distributions (p^t, q^t) that solve the min-max problem that defines the PAC DEC (Eq. (10)) with $\widehat{M}^t$ plugged in as the reference model (Line 6):

$$(p^t, q^t) := \operatorname*{arg\,min}_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{H}_{q,\bar{\varepsilon}(T)}(\widehat{M}^t)} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)]. \tag{14}$$

By definition, the value above is always bounded by $\mathsf{p\text{-}dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}, \widehat{M}^t)$. Recall that p^t may be though of as an *exploitation distribution*, and that q^t may be thought of as an *exploration distribution*. With these distributions in hand, Algorithm 1 samples $\pi^t \sim q^t$ from the exploration distribution q^t. The distribution p^t is not used in this phase, but is retained for the exploitation phase that follows.

Exploitation phase. In the *exploitation phase* (or, post-processing phase), we aim to identify an exploitation distribution $p^t \in \{p^1, \dots, p^J\}$ from the collection computed during the exploration phase that is "good" in the sense that it has sufficiently low suboptimality under $M^\star$. The motivation behind this stage is as follows. From the expression (14) and the definition of the constrained DEC, we are guaranteed that p^t has

$$\mathbb{E}_{\pi \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] \le \textsf{p-dec}^{\textsf{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}, \widehat{M}^t)$$

for any round $t \in [J]$ where $M^\star \in \mathcal{H}_{q^t, \bar{\varepsilon}(T)}(\widehat{M}^t)$. The challenge here is that the estimation oracle ensures only that the *cumulative* estimation error for the estimators $\widehat{M}^1, \dots, \widehat{M}^J$ is low; there is no guarantee that the per-round estimation error $\mathbb{E}_{\pi^t \sim q^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right]$ will decrease with time, and for any fixed round t of interest, this quantity might be trivially large. This can lead to a problem we term *false exclusion*, where $M^\star \notin \mathcal{H}_{q^t, \bar{\varepsilon}(T)}(\widehat{M}^t)$. False exclusion is problematic because we have no control over the suboptimality under $M^\star$ for rounds $t \in [J]$ where it occurs.

The good news is that while the online estimation oracle may lead to false exclusion for some rounds, Markov's inequality implies that (for our choice of $\bar{\varepsilon}(T)$, which depends on $\mathbf{Est}_{\mathsf{H}}(T, \delta)$) the true model $M^\star$ will be included in $\mathcal{H}_{q^t, \bar{\varepsilon}(T)}(\widehat{M}^t)$ for *at least half of the rounds* $t \in [J]$. Hence, the exploitation phase proceeds by sampling (on Line 8) a small number of rounds $t_1, \dots, t_L \in [J]$—a logarithmic number suffices to ensure that at least one is good with high probability—and performing a test to identify a good distribution $p^{t_{\widehat{\ell}}}$ from the set $\{p^{t_1}, \dots, p^{t_L}\}$, which is then returned by the algorithm.

In more detail, the exploitation phase (Line 8 through Line 11) proceeds by gathering many (namely, $\Theta(T/\log(2/\delta))$) samples from q^{t_ℓ} for each of the L rounds $\{t_\ell\}_{\ell \in [L]}$ and then, for each $\ell \in [L]$, using the estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ to produce an estimated model $\widetilde{M}_\ell \in \mathcal{M}$ based on these samples. Since many samples are used to produce the estimate $\widetilde{M}_\ell$, it is guaranteed to be close to the true model $M^\star$ under q^{t_ℓ}. This means that by choosing a round $t_{\widehat{\ell}}$ that minimizes the Hellinger distance between $\widehat{M}^{t_{\widehat{\ell}}}$ and $\widetilde{M}_{\widehat{\ell}}$ (Line 11), we have that with high probability, $M^\star \in \mathcal{H}_{q^{t_{\widehat{\ell}}}, \varepsilon}(\widehat{M}^{t_{\widehat{\ell}}})$, thus solving the false exclusion problem, and ensuring that the exploitation distribution $p^{t_{\widehat{\ell}}}$ has low risk.

Main result. We show that $\mathsf{E2D}^{+}$ enjoys the following guarantee for PAC.

Theorem 2.2 (Main Upper Bound: PAC). *Fix $\delta \in \left(0, \frac{1}{10}\right)$ and $T \in \mathbb{N}$. Suppose that Assumptions 1.1 and 2.1 hold, and let $\overline{\mathbf{Est}}_{\mathsf{H}} := \mathbf{Est}_{\mathsf{H}}\left(\frac{2T}{\lceil \log 2/\delta \rceil}, \frac{\delta}{4\lceil \log 2/\delta \rceil}\right)$. With $\bar{\varepsilon}(T) := 8\sqrt{\frac{\lceil \log 2/\delta \rceil}{T} \cdot \overline{\mathbf{Est}}_{\mathsf{H}}}$, Algorithm 1 guarantees that with probability at least $1 - \delta$,*

$$\mathbf{Risk}_{\mathsf{DM}}(T) \le \textsf{p-dec}^{\textsf{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}).$$

Thus, the expected risk achieved by Algorithm 1 is bounded by $\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \le \textsf{p-dec}^{\textsf{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}) + \delta$.

This result matches the lower bound in Theorem 2.1, with the only gap being the choice of radius $\varepsilon > 0$ for the constrained DEC: The lower bound (Theorem 2.1) has $\underline{\varepsilon}(T) \propto \sqrt{\frac{1}{T}}$, while the upper bound (Theorem 2.2) has $\bar{\varepsilon}(T) \propto \sqrt{\frac{\log(2/\delta) \cdot \overline{\mathbf{Est}}_{\mathsf{H}}}{T}}$. To make the result concrete, let us instantiate it for some standard examples, focusing on the special case where $\mathcal{M}$ is finite and $\mathbf{Est}_{\mathsf{H}}(T, \delta) \le \log(|\mathcal{M}|/\delta)$ for simplicity.

- Whenever $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}}$, Theorem 2.2 gives $\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \leq \widetilde{O}\left(\sqrt{\frac{C_{\mathrm{prob}} \log|\mathcal{M}|}{T}}\right)$, which translates into $\widetilde{O}\left(\frac{C_{\mathrm{prob}} \log|\mathcal{M}|}{\varepsilon^2}\right)$ samples to learn an ε-optimal policy.
- Whenever $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$ for $\rho \in (0,1)$, Theorem 2.2 gives $\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \leq \widetilde{O}((\log|\mathcal{M}|/T)^{\frac{(1-\rho)}{2}})$, which translates into $\widetilde{O}(\log|\mathcal{M}| \cdot \varepsilon^{-\frac{2}{1-\rho}})$ samples to learn a ε-optimal policy.
- Whenever $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \geq 1/\sqrt{C_{\mathrm{prob}}}\}$, where C_{prob} is a problem-dependent parameter, Theorem 2.2 gives $\mathbf{Risk}_{\mathsf{DM}}(T) = 0$ with high probability whenever $T \geq \widetilde{\Omega}(C_{\mathrm{prob}} \cdot \log|\mathcal{M}|)$.

See Appendix C for analogous guarantees for regret. As discussed in Foster et al. (2021), understanding when the estimation complexity $\overline{\mathbf{Est}}_{\mathsf{H}} \lesssim \log|\mathcal{M}|$ can be removed or weakened is subtle issue, as there are some classes $\mathcal{M}$ for which this term is necessary, and others for which it is superfluous. This is the main question left open by our research.

Acknowledgments

We thank Jian Qian, Sasha Rakhlin, Rob Schapire, and Andrew Wagenmaker for helpful comments and discussions. NG is supported at MIT by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship.

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proc. of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

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.

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.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

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

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

Varsha Dani, Thomas P Hayes, and Sham Kakade. The price of bandit information for online optimization. 2007.

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.

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, 2019.

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.

Richard Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, fifth edition, 2019. ISBN 0-534-24318-5.

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

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

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

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, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

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.

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

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. *arXiv preprint arXiv:2202.10997*, 2022.

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.

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.

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, 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.

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.

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.

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.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

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.

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.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

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

Contents of Appendix

I Main Results

Part I

Main Results

Appendix A. Organization

Part I of the appendix presents the remainder of our main results, which are omitted from the main body due to space constraints. In particular:

- Appendix C presents our main upper and lower bounds for the regret framework, building on the PAC techniques in Section 2.
- Appendix D establishes structural results concerning the constrained DEC, the offset DEC, and their relationship.
- Appendix E provides a detailed comparison to bounds from prior work.
- Appendix F gives additional examples.

All proofs are given in Part II of the appendix.

Appendix B. Preliminaries

Interactive decision making. We assume throughout the remainder of the paper that $\mathcal{R} = [0,1]$ unless otherwise stated. We let $\mathcal{M}^{+} = \{\overline{M} : \Pi \to \Delta([0,1]\times\mathcal{O}) \mid \sup_{M\in\mathcal{M}} \sup_{\pi\in\Pi} \sup_{\mathcal{E}\in\mathscr{R}\otimes\mathscr{O}} \{ \frac{\overline{M}(\mathcal{E}|\pi)}{M(\mathcal{E}|\pi)} \} \leq V(\mathcal{M})\}$ denote the space of all possible models with rewards in $[0,1]$ that obey the same density ratio bound as $\mathcal{M}$. We adopt the shorthand $g^{M}(\pi) = f^{M}(\pi_{M}) - f^{M}(\pi)$.

We adopt the same formalism for probability spaces as in Foster et al. (2021, 2022b). Decisions are associated with a measurable space $(\Pi, \mathscr{P})$, rewards are associated with the space $(\mathcal{R}, \mathscr{R})$, and observations are associated with the space $(\mathcal{O}, \mathscr{O})$. The history up to time t is denoted by $\mathfrak{H}^{t} = (\pi^{1}, r^{1}, o^{1}), \ldots, (\pi^{t}, r^{t}, o^{t})$. We define

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

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

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

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

Appendix C. Regret Framework: Upper and Lower Bounds

We now prove upper and lower bounds based on the constrained Decision-Estimation Coefficient for the regret framework. Both results build on our techniques for PAC, and have nearly identical statements, but have more involved proofs.

C.1. Lower Bounds

To state the lower bound for regret, recall that we define $C(T) := \log(T \wedge V(\mathcal{M}))$.

Theorem C.1 (Main Lower Bound: Regret)**.** *There exist universal constants $C, C' > 0$ and $c, c' > 0$ such that the following holds. Let $\underline{\varepsilon}(T) := c \cdot \frac{1}{\sqrt{TC(T)}}$. For all $T \in \mathbb{N}$ such that the condition*

$$\textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \geq C \cdot \underline{\varepsilon}(T) \tag{15}$$

is satisfied, it holds that for any regret minimization algorithm, there exists a model in $\mathcal{M}$ such that

$$\begin{aligned}\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] &\geq c' \cdot \sup_{\overline{M} \in \mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \cdot \frac{T}{\log T} - C' \cdot \frac{\sqrt{T}}{\log(T)} \\ &\geq c' \cdot \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot \frac{T}{\log T} - C' \cdot \frac{\sqrt{T}}{\log(T)}.\end{aligned} \tag{16}$$

Theorem C.1 shows that the constrained DEC is a fundamental limit for interactive decision making in the regret framework. Importantly, this lower bound removes the notion of localization required by prior work on regret, and shows that the DEC remains a lower bound even if one allows for improper reference models $\overline{M} \in \mathrm{co}(\mathcal{M})$; it is always tighter than the lower bounds found in Foster et al. (2021, 2022b). As with PAC, we will show (Appendix C.2) that the lower bound can be achieved algorithmically, up to a difference in radius that depends on the estimation capacity for $\mathcal{M}$ ($\sqrt{\log|\mathcal{M}|/T}$ versus $1/\sqrt{T}$ for the case of finite classes). A detail comparison to prior work for regret is given in Appendix E.

Let us remark that, similar to our results for PAC, the lower bound in Theorem C.1 scales with the quantity $\sup_{\overline{M} \in \mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \geq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$, which allows for arbitrary reference models $\overline{M} \notin \mathrm{co}(\mathcal{M})$, while our upper bounds for regret scale with $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \geq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$). We show in Appendix D.4 that maximizing over reference models $\overline{M} \in \mathcal{M}^+$ does not increase the value of the DEC beyond what is attained by $\overline{M} \in \mathrm{co}(\mathcal{M})$, so this result does not contradict our upper bounds.

Understanding the lower bound. Standard examples for Theorem C.1 are as follows.

- $\sqrt{T}$*-rates.* For the most well-studied classes found throughout the literature on bandits and reinforcement learning, we have

  $$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}},$$

  where $C_{\mathrm{prob}} > 0$ is a problem-dependent constant that reflects some notion of intrinsic complexity. In this case, the condition (15) is satisfied whenever C_{prob} is larger than some numerical constant, and Theorem C.1 gives[4]

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

  Examples (cf. Appendix F) include multi-armed bandits with A actions, where $C_{\mathrm{prob}} \geq A$ (leading to $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{AT})$), linear bandits in dimension d, where $C_{\mathrm{prob}} \geq d$ (leading to $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{dT})$), and tabular reinforcement learning with S states, A actions, and horizon H, where $C_{\mathrm{prob}} \geq HSA$ (leading to $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{HSAT})$).

[4] This requires $V(\mathcal{M}) = O(1)$ so that $\bar{\varepsilon}(T) \propto 1/\sqrt{T}$; see Remark 2.1.

- *Nonparametric rates.* For nonparametric model classes, where the optimal regret is of larger order than $\sqrt{T}$, one typically has

  $$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$$

  for some $\rho \in (0, 1)$. In this case, the condition in Eq. (15) is satisfied whenever T is a sufficiently large constant, and Theorem C.1 gives

  $$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(T^{\frac{1+\rho}{2}}).$$

  A standard example is Lipschitz bandits over $[0, 1]^d$ (Auer et al., 2007; Kleinberg et al., 2019), where we have $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\frac{d}{d+2}}$, leading to $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}\big(T^{\frac{d+1}{d+2}}\big)$.

- *Fast rates.* For problems with low noise, such as noiseless bandits, the DEC typically exhibits threshold behavior, with

  $$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \geq 1/\sqrt{C_{\text{prob}}}\},$$

  where C_{prob} is a problem-dependent parameter. For example, if $\mathcal{M}$ consists of multi-armed bandit instances with $\Pi = \{1, \ldots, A\}$ and noiseless, binary rewards, one can take $C_{\text{prob}} \propto A$. For such settings, the condition (15) is satisfied whenever $T = \widetilde{O}(C_{\text{prob}})$, and Theorem C.1 gives

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

We refer to Appendix F for further examples and details.

Proof techniques. The proof of Theorem C.1 follows a similar approach to our lower bound for PAC (Theorem 2.1). However, non-trivial difficulties arise in applying the iterative conditioning scheme in the preceding discussion because there are no longer separate distributions for exploration (q_M) and exploitation (p_M), causing the analysis of regret and information to be coupled. To address this, we adopt a somewhat different two-part scheme.

1. First, we prove a lower bound that is similar to Theorem C.1, but qualitatively weaker in the sense that the quantity $\text{r-dec}^{\text{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) = \sup_{\overline{M} \in \text{co}(\mathcal{M})} \text{r-dec}^{\text{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}, \overline{M})$ in Eq. (16) is replaced by $\sup_{\overline{M} \in \mathcal{M}} \text{r-dec}^{\text{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}, \overline{M})$; that is, the lower bound restricts to proper reference models. This is proven using a similar approach to our lower bound for PAC.

2. Then, we upgrade this weaker result to the full claim of Theorem C.1, using the following algorithmic result: for any model class $\mathcal{M}$ and any $\overline{M} \in \mathcal{M}^+$ (not necessarily in $\mathcal{M}$) satisfying mild technical conditions, if there exists an algorithm that achieves expected regret at most R with respect to the class $\mathcal{M}$, then there exists an algorithm that achieves regret at most $O(R \cdot \log T)$ with respect to the enlarged class $\mathcal{M} \cup \{\overline{M}\}$. By then choosing $\overline{M}$ appropriately and applying the proper lower bound from Part 1 to a slightly modified version of the class $\mathcal{M} \cup \{\overline{M}\}$, we are able to establish Theorem C.1.

See Appendix H for the proof.

Algorithm 2 Estimation-to-Decisions ($\mathsf{E2D}^+$) for Regret

```
parameters:
    Number of rounds $T \in \mathbb{N}$ and failure probability $\delta \in (0,1)$.
    Online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$.
    Constant $C_1 > 1$. // Specified in Appendix I.
Define $N := \lceil \log T \rceil$ and $L := \lceil \log 1/\delta \rceil$.
Set $\varepsilon_N := \sqrt{\frac{C_1 \cdot \mathbf{Est}_{\mathsf{H}}(T,\delta) \cdot L}{T}}$ each $\varepsilon_i := \sqrt{2^{N-i}} \cdot \varepsilon_N$ for $i \in [N]$.
Split $[T]$ into $2N$ contiguous blocks $\mathcal{E}_1 \cup \mathcal{R}_1 \cup \cdots \cup \mathcal{E}_N \cup \mathcal{R}_N$, w/ $|\mathcal{R}_i|, |\mathcal{E}_i| \in \left[\frac{2^i}{4}, \frac{2^i}{2}\right]$ $\forall i \in [N]$.
Set $\mathcal{M}_1 := \mathcal{M}$.
for $i \in [N]$ do
    /* Exploration for epoch $i$ */
    Initialize instance of $\mathbf{Alg}_{\mathsf{Est}}$, with horizon $|\mathcal{E}_i|$, failure probability $\delta$, and model class $\mathcal{M}_i$.
    for $t \in \mathcal{E}_i$ do
        Obtain estimate $\widehat{M}^t := \mathbf{Alg}_{\mathsf{Est}}(\{(\pi^s, r^s, o^s)\}_{s \in \mathcal{E}_i, s<t})$, where $\widehat{M}^t \in \mathrm{co}(\mathcal{M}_i)$.
        Compute $p^t := \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{p,\varepsilon_i}(\widehat{M}^t) \cup \{\widehat{M}^t\}} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)]$.
        Sample decision $\pi^t \sim p^t$ and update estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ with $(\pi^t, r^t, o^t)$.
    /* Refinement for epoch $i$ */
    Sample a subset $\mathcal{S}_i \subset \mathcal{E}_i$ of size $L$ uniformly at random (with replacement).
    Initialize $s_i^{\mathrm{tmp}} = \perp$. // With high probability, $s_i^{\mathrm{tmp}}$ will be updated in the for loop.
    for $s \in \mathcal{S}_i$ do
        Initialize instance of $\mathbf{Alg}_{\mathsf{Est}}$ with horizon $J_i := |\mathcal{R}_i|/L$, failure probability $\delta$, and class $\mathcal{M}_i$.
        for $1 \leq j \leq J_i$ do
            Sample a decision $\pi_s^j \sim p^s$ and update $\mathbf{Alg}_{\mathsf{Est}}$ with $(\pi_s^j, r_s^j, o_s^j)$.
            Obtain estimate $\widetilde{M}_s^j := \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi_s^k, r_s^k, o_s^k)_{k=1}^{j-1}\}\right)$, where $\widetilde{M}_s^j \in \mathrm{co}(\mathcal{M}_i)$.
            if $\sum_{k=1}^{j} \mathbb{E}_{\pi \sim p^s}\left[D_{\mathsf{H}}^2\left(\widehat{M}^s(\pi), \widetilde{M}_s^k(\pi)\right)\right] > \frac{J_i \varepsilon_i^2}{4}$ then
                Define $J_{i,s} := j$, and break out of the loop over $j$ (i.e., move to next value in $\mathcal{S}_i$).
            if $j = J_i$ then
                Set $s_i^{\mathrm{tmp}} = s$ and $J_{i,s} = J_i$.[5]
    Set $s_i = s_i^{\mathrm{tmp}}$ if $s_i^{\mathrm{tmp}} \neq \perp$, o.w. $s_i \in \mathcal{S}_i$ arbitrarily. Set $\widehat{M}_i := \frac{1}{J_{i,s}} \sum_{j=1}^{J_{i,s}} \widetilde{M}_{s_i}^j$, $\widehat{p}_i := p^{s_i}$.
    for any remaining rounds $t$ in $\mathcal{R}_i$ do
        Play $\pi^t \sim p^{s_i}$ (the rewards and observations can be ignored).
    Set
        $$\mathcal{M}_{i+1} := \left\{ M \in \mathcal{M} \mid \mathbb{E}_{\pi \sim \widehat{p}_i}\left[D_{\mathsf{H}}^2\left(M(\pi), \widehat{M}_i(\pi)\right)\right] \leq \frac{\mathbf{Est}_{\mathsf{H}}(J_i, \delta)}{J_i} \right\}.$$
```

C.2. Upper Bounds

We now present an algorithm and upper bound for regret that complements Theorem C.1. Our algorithm, Algorithm 2, adapts $\mathsf{E2D}^{+}$ to the regret framework, and attains a regret bound that scales with $\textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})\cdot T$ for an appropriate radius $\bar{\varepsilon}(T) > 0$.

Online estimation. Algorithm 2 makes use of an online estimation oracle in the same fashion as the PAC algorithm (Algorithm 1), but we require a slightly stronger oracle capable of incorporating *constraints* on the model class, specified via a subset $\mathcal{M}' \subseteq \mathcal{M}$.

Assumption C.1 (Constrained estimation oracle for $\mathcal{M}$)**.** *A constrained estimation oracle for $\mathcal{M}$ takes as input a constraint set $\mathcal{M}' \subseteq \mathcal{M}$. At each time $t \in [T]$, the online estimation oracle $\mathbf{Alg}_{\mathsf{Est}}$ returns, given*

$$\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$$

with $(r^i, o^i) \sim M^\star(\pi^i)$ and $\pi^i \sim p^i$, an estimator $\widehat{M}^t \in \mathrm{co}(\mathcal{M}')$ such that whenever $M^\star \in \mathcal{M}'$,

$$\mathbf{Est}_{\mathsf{H}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right] \le \mathbf{Est}_{\mathsf{H}}(T, \delta), \tag{17}$$

with probability at least $1-\delta$, where $\mathbf{Est}_{\mathsf{H}}(T,\delta)$ is a known upper bound.

This assumption is identical to the assumption made for PAC (Assumption 2.1), except that i) the oracle takes a *constraint set* $\mathcal{M}' \subseteq \mathcal{M}$ as an input before the learning process begins, and ii) the oracle is required to produce $\widehat{M}^t \in \mathrm{co}(\mathcal{M}')$; that is, the estimator is required to lie in the convex hull of the constraint set $\mathcal{M}'$. All estimation algorithms that we are aware of can achieve this guarantee with minor or no modifications, including the exponential weights algorithm, which satisfies Assumption C.1 with $\mathbf{Est}_{\mathsf{H}}(T,\delta) \le O(\log(|\mathcal{M}|/\delta))$.

Overview of algorithm. Algorithm 2 employs the Estimation-to-Decisions principle of Foster et al. (2021) but, like Algorithm 1, incorporates substantial modifications tailored to the constrained (as opposed to offset) DEC. The core of the algorithm is Line 10, which—at each round t—obtains an estimator $\widehat{M}^t \in \mathrm{co}(\mathcal{M})$, then computes an exploratory distribution p^t by solving the min-max problem that defines the regret DEC (Eq. (7)) with $\widehat{M}^t$ plugged in:

$$p^t := \operatorname*{arg\,min}_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{p,\varepsilon}(\widehat{M}^t) \cup \{\widehat{M}^t\}} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)], \tag{18}$$

for an appropriate choice of $\varepsilon > 0$ that depends on t.

As with Algorithm 1, the main challenge Algorithm 2 needs to overcome is *false exclusion*: Whenever $M^\star \in \mathcal{H}_{p^t,\varepsilon}(\widehat{M}^t)$, it is immediate from the definition above that

$$\mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\widehat{M}^t\}, \widehat{M}^t),$$

but what happens if $M^\star \notin \mathcal{H}_{p^t,\varepsilon}(\widehat{M}^t)$? For PAC, the solution employed by Algorithm 1 is simple: The online estimation guarantee for $\mathbf{Alg}_{\mathsf{Est}}$ ensures that $M^\star$ is correctly included for at least half the rounds, and we only need to identify a single round where it is included. For regret, this reasoning no longer suffices: We cannot simply ignore the rounds in which $M^\star$ is excluded, as the regret for these rounds must be controlled.

[5] Algorithm 2 will continue to work if we modify it to break out of the loop over $s \in \mathcal{S}_i$ when this if statement is reached, which is somewhat more natural. We use the version presented here because it slightly simplifies the proof.

Epoch scheme. To address the issue of false exclusion, Algorithm 2 breaks the rounds $1, \ldots, T$ into epochs $1, \ldots, N$ of doubling length, with each epoch i consisting of a contiguous set of *exploration rounds* $\mathcal{E}_i \subseteq [T]$ and *refinement rounds* $\mathcal{R}_i \subseteq [T]$. Each epoch proceeds in a similar fashion to Algorithm 1, but takes advantage of the data collected in previous epochs to explore in a fashion that is robust to false exclusions:

- Each exploration phase gathers data using a sequence of exploratory distributions $\{p^t\}_{t\in\mathcal{E}_i}$ computed by solving the DEC with the estimated models $\{\widehat{M}^t\}_{t\in\mathcal{E}_i}$, following Eq. (18). However, the estimated models $\widehat{M}^t$ are restricted to lie in $\mathrm{co}(\mathcal{M}_i)$, where $\mathcal{M}_i$ is a confidence set computed using data from the previous epoch.
- The purpose of the refinement phase in epoch i is to use the distributions generated in the exploration phase to compute a confidence set $\mathcal{M}_{i+1}$ for the *next epoch* that satisfies a certain invariant that allows us to translate low regret with respect to models in the confidence set to low regret under $M^\star$.

To describe the phases for an epoch $i \in [N]$ in greater detail, we define

$$\alpha_i := C_0 \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) + 64\varepsilon_i,$$

for a constant $C_0 > 1$ whose value is specified in the full proof (Appendix I).

Exploration phase. For each step t within the exploration phase $\mathcal{E}_i$, we obtain an estimator $\widehat{M}^t \in \mathrm{co}(\mathcal{M}_i)$ from the estimation oracle and solve

$$p^t := \operatorname*{arg\,min}_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{H}_{p,\varepsilon_i}(\widehat{M}^t)\cup\{\widehat{M}^t\}} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)]. \tag{19}$$

where $\mathcal{M}_i \subseteq \mathcal{M}$ is the confidence set produced by the refinement phase in epoch $i-1$. The set $\mathcal{M}_i$ is constructed so that $M^\star \in \mathcal{M}_i$ with high probability, but in addition, the following *localization property* can be shown (using that $\widehat{M}^t \in \mathrm{co}(\mathcal{M}_i)$ for all $t \in \mathcal{E}_i$; see the proof of Lemma I.3):

$$f^{M^\star}(\pi_{M^\star}) \le f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) + \frac{\alpha_{i-1}}{2} \quad \forall t \in \mathcal{E}_i. \tag{20}$$

The condition (20) implies that we are always in a favorable position with respect to regret:

- If $M^\star \in \mathcal{H}_{p^t,\varepsilon_i}(\widehat{M}^t)$—that is, the true model is not falsely excluded—we have $\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) \le \alpha_i$ by definition.
- Even if $M^\star$ is falsely excluded, Eq. (20) implies that up to certain nuisance terms, we have (as shown in Lemma I.4):

  $$\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \lesssim \mathbb{E}_{\pi\sim p^t}\left[f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) - f^{\widehat{M}^t}(\pi)\right] \lesssim \alpha_{i-1},$$

  where the latter inequality uses the definition (19), which implies that

  $$\mathbb{E}_{\pi\sim p^t}\left[f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) - f^{\widehat{M}^t}(\pi)\right] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\}, \widehat{M}^t) \le \alpha_i.$$

In both situations, we incur no more than $O(\alpha_i)$ regret per round. Due to the doubling epoch schedule, the total contribution to regret for all exploration rounds is no more than $\widetilde{O}\big(\textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}) \cdot T\big)$.

Refinement phase. For the analysis of the exploration phase in epoch $i + 1$ to succeed, the refinement phase at epoch i must construct a confidence set $\mathcal{M}_{i+1}$ so that the localization property (20) is satisfied with scale α_i. To achieve this, we use a similar approach to the exploitation phase in Algorithm 1. For all the rounds $t \in \mathcal{E}_i$ for which $M^\star$ is not falsely excluded, we are guaranteed that i) $M^\star \in \mathcal{H}_{p^t,\varepsilon_i}(\widehat{M}^t)$, and ii) all $M \in \mathcal{H}_{p^t,\varepsilon_i}(\widehat{M}^t)$ satisfy

$$\mathbb{E}_{\pi\sim p^t}\left[f^M(\pi_M) - f^M(\pi)\right] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) \le \frac{\alpha_i}{4},$$

which can be shown to imply the localization property in Eq. (20) holds at epoch i (the proof of this fact uses that $\widehat{M}^t \in \mathrm{co}(\mathcal{M}_{i+1})$ for all $t \in \mathcal{E}_{i+1}$). In addition, the estimation guarantee for the oracle (Assumption C.1) implies that $M^\star$ is included in $\mathcal{H}_{p^t,\varepsilon_i}(\widehat{M}^t)$ for at least half of the rounds $t \in \mathcal{E}_i$. Hence, to construct $\mathcal{M}_{i+1}$, we sample a small number of rounds $t_1, \ldots, t_L \in \mathcal{E}_i$ (Line 12; a logarithmic number L suffices) and perform a test based on Hellinger distance to identify a "good" distribution p^{t_ℓ} from the set $\{p^{t_1}, \ldots, p^{t_L}\}$ with the property that $M^\star \in \mathcal{H}_{p^{t_\ell},\varepsilon_i}(\widehat{M}^{t_\ell})$. We then set $\mathcal{M}_{i+1} = \mathcal{H}_{p^{t_\ell},\varepsilon_i}(\widehat{M}^{t_\ell})$, ensuring that $M^\star \in \mathcal{M}_{i+1}$ and the localization property is satisfied.

Main result. We now present the main regret guarantee for $\textsf{E2D}^{+}$ (Algorithm 2). To state the result in the simplest form possible, we assume that the regret DEC satisfies a mild growth condition.

Assumption C.2 (Regularity of DEC)**.** *The class $\mathcal{M}$ satisfies the following: for some constant $C_{\mathrm{reg}} > 1$ and all $\varepsilon \in (0, 2)$, we have*

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \le C_{\mathrm{reg}}^2 \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}).$$

This condition asserts that the DEC does not shrink too quickly as a function of the parameter ε. It is automatically satisfied whenever $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{\rho}$ for $\rho \le 2$, with $\rho \le 1$ corresponding to the "$\sqrt{T}$-regret or greater" regime; regret bounds that hold under more general assumptions are given in Appendix I.

Theorem C.2 (Main Upper Bound: Regret)**.** *Fix $T \in \mathbb{N}$ and $\delta \in (0, 1)$. Suppose that Assumptions 1.1, C.1 and C.2 hold, and let $\overline{\mathbf{Est}}_{\mathsf{H}} := \mathbf{Est}_{\mathsf{H}}(T, \delta^2)$. Then Algorithm 2, with $C_1 > 0$ chosen appropriately, ensures that with probability at least $1 - \delta$,*

$$\mathbf{Reg}_{\mathsf{DM}}(T) \le \textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}) \cdot O(T\log(T)) + O\Big(\sqrt{T\log(1/\delta)\cdot \overline{\mathbf{Est}}_{\mathsf{H}}}\Big),$$

where $\bar{\varepsilon}(T) := C \cdot \sqrt{\frac{\overline{\mathbf{Est}}_{\mathsf{H}}\cdot\log(1/\delta)}{T}}$ for a numerical constant $C > 0$.

The proof of this result is deferred to Appendix I. As in the PAC setting, this upper bound matches the corresponding lower bound (Theorem C.1), up to the radius for the constrained DEC ($\underline{\varepsilon}(T) \approx \sqrt{\frac{1}{T}}$ for Theorem C.1 versus $\bar{\varepsilon}(T) \approx \sqrt{\frac{\overline{\mathbf{Est}}_{\mathsf{H}}}{T}}$ for Theorem C.2), and cannot be improved beyond logarithmic factors without further assumptions.

Remark C.1 (Relaxing the regularity condition)**.** It follows immediately from the proof of Theorem C.2 that the following holds. Suppose that in place of Assumption C.2, we assume that there is some function $\mathsf{r}(\varepsilon)$ so that, for all $\varepsilon \in (0, 2)$, we have: (1) $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \le \mathsf{r}(\varepsilon)$ and (2) $\mathsf{r}(\varepsilon) \le C_{\mathrm{reg}}^2 \cdot \mathsf{r}(\varepsilon/C_{\mathrm{reg}})$. Then the regret bound of Theorem C.2 holds with $\textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})$ replaced by $\mathsf{r}(\bar{\varepsilon}(T))$.

Examples of Theorem C.2, under the assumption that $\mathcal{M}$ is finite (in which case, $\mathbf{Est}_{\mathsf{H}}(T,\delta) \le O\big(\log(|\mathcal{M}|/\delta)\big)$) include:

- Whenever $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}}$, Theorem C.2 gives $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \le \widetilde{O}\big(\sqrt{C_{\mathrm{prob}} \cdot T \cdot \log|\mathcal{M}|}\big)$.

- Whenever $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$ for $\rho \in (0,1)$, Theorem C.2 gives $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \le \widetilde{O}\big(T^{\frac{(1+\rho)}{2}} \cdot \log^{\frac{1-\rho}{2}}|\mathcal{M}| + T^{\frac{1}{2}}\log^{\frac{1}{2}}|\mathcal{M}|\big)$.

Appendix D. Decision-Estimation Coefficient: Structural Properties

The lower and upper bounds for PAC and regret in Section 2 and Appendix C, which are stated in terms of the constrained Decision-Estimation Coefficient, are tight up to dependence on the model estimation error $\mathbf{Est}_{\mathsf{H}}(T,\delta)$ (recall that the lower bounds use scale $\underline{\varepsilon}(T) = \widetilde{\Omega}\big(\sqrt{1/T}\big)$, while the upper bounds use scale $\bar{\varepsilon}(T) = \widetilde{O}\big(\sqrt{\mathbf{Est}_{\mathsf{H}}(T,\delta)/T}\big)$. It is natural to ask how these results are related to the lower and upper bounds in Foster et al. (2021, 2022b), which are stated in terms of the offset DEC, and at first glance are not obviously comparable. Toward developing such an understanding, this section establishes a number of structural properties for the DEC.

- In Appendix D.1, we show that the constrained and offset DEC are nearly equivalent for PAC. For regret, we show that it is always possible to bound the constrained DEC by the offset DEC in a tight fashion, but the converse is not true in general.

- In Appendix D.2, we show that the constrained DEC implicitly enforces a form of localization, and uncover a tighter relationship between the constrained and offset variants of the regret DEC for localized classes.

Appendices D.3 and D.4 investigate the role of the reference model $\overline{M} \in \mathcal{M}^{+}$.

- First, in Appendix D.3, we show that the definition $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})$, which incorporates suboptimality under the reference model $\overline{M}$, is critical to obtain tight upper and lower bounds. This is a fundamental difference from PAC, where we show that it suffices to use the definition $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, which does not incorporate suboptimality under $\overline{M}$.

- Then, in Appendix D.4, we show that allowing for arbitrary, potentially improper reference models $\overline{M} \in \mathcal{M}^{+}$ never increases the value of the DEC beyond what is achieved by reference models $\overline{M} \in \mathrm{co}(\mathcal{M})$. This illustrates the fundamental role of convexity. In addition, we show that a similar equivalence holds for variants of the DEC that incorporate randomized mixture estimators.

Remark D.1 (Stronger definition of $\mathcal{M}^{+}$)**.** We remark that all results in this section in fact hold even if $\mathcal{M}^{+}$ is defined to be the set of *all* models $\overline{M} : \Pi \to \Delta([0,1] \times \mathcal{O})$, i.e., without the restriction that $\frac{\overline{M}(\mathcal{E}|\pi)}{M(\mathcal{E}|\pi)} \le V(\mathcal{M})$ for all $M \in \mathcal{M}, \pi \in \Pi, \mathcal{E} \in \mathscr{R} \otimes \mathscr{O}$ (since none of the proofs in this section use this density ratio upper bound). This observation strengthens our results somewhat.

D.1. Relationship Between Constrained and Offset DEC

It is immediate that one can bound the constrained DEC by the offset DEC in a tight fashion. Focusing on regret for concreteness, we can use the method of Lagrange multipliers to show that for any $\overline{M} \in \mathcal{M}$ and $\varepsilon > 0$,

$$\begin{aligned}
\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) &= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] \mid \mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \le \varepsilon^2 \right\} \\
&= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \inf_{\gamma > 0} \left\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] - \gamma\big(\mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] - \varepsilon^2\big) \right\} \\
&\le \inf_{\gamma > 0} \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] - \gamma\big(\mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] - \varepsilon^2\big) \right\} \\
&= \inf_{\gamma > 0} \left\{ \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma\varepsilon^2 \right\}. \qquad (21)
\end{aligned}$$

The same approach yields an analogous inequality for PAC. For general models $\overline{M} \notin \mathcal{M}$, the following slightly looser version of Eq. (21) holds.[6]

Proposition D.1. *For all $\overline{M} \in \mathcal{M}^+$ and $\varepsilon > 0$, we have*

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le 8 \cdot \inf_{\gamma > 0} \left\{ \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2 \right\} + 7\varepsilon. \qquad (22)$$

Examples include:

- Whenever $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) \propto \frac{C_{\mathrm{prob}}}{\gamma}$, Proposition D.1 yields $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \lesssim \varepsilon\sqrt{C_{\mathrm{prob}}}$. In this case, both our results and the bounds in Foster et al. (2021) lead to $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \le \widetilde{O}\big(\sqrt{C_{\mathrm{prob}} \cdot T \cdot \mathbf{Est}_{\mathsf{H}}(T, \delta)}\big)$.
- More generally, whenever $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) \propto \left(\frac{C_{\mathrm{prob}}}{\gamma}\right)^{\rho}$ for some $\rho < 1$, then Proposition D.1 yields $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \lesssim C_{\mathrm{prob}}^{\frac{\rho}{1+\rho}} \cdot \varepsilon^{\frac{2\rho}{1+\rho}}$.

In what follows we investigate when and to what extent the constrained and offset variants of the DEC can be related in the opposite direction to Eq. (21)—both for regret and PAC.

D.1.1. PAC DEC: CONSTRAINED VERSUS OFFSET

For the PAC setting, the following result shows that the constrained and offset DEC are equivalent up to logarithmic factors in most parameter regimes.

Proposition D.2. *For all $\varepsilon > 0$ and $\overline{M} \in \mathcal{M}^+$, we have*

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le \inf_{\gamma \ge 0} \left\{ \textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2 \right\}. \qquad (23)$$

On the other hand, for all $\gamma \ge 1$ and $\overline{M} \in \mathcal{M}^+$, letting $L := 2\lceil \log 2\gamma \rceil$, we have

$$\textsf{p-dec}^{\mathrm{o}}_{\gamma \cdot (4L+1)}(\mathcal{M}, \overline{M}) \le \frac{2}{\gamma} + \sup_{\varepsilon > 0} \left\{ \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) - \frac{\gamma\varepsilon^2}{4} \right\}. \qquad (24)$$

[6] An inequality that replaces the right-hand side of Eq. (22) with $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})$ follows immediately by applying Eq. (21) to the class $\mathcal{M} \cup \{\overline{M}\}$. The inequality (22) is stronger, and the proof is more involved.

Whenever $\text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{\rho}$ for $\rho \leq 1$, one loses only logarithmic factors by passing to the offset DEC using Eq. (23) and using Eq. (24) to pass back to the constrained DEC. Yet, in the case where the constrained DEC has "fast" behavior of the form $\text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto C_{\mathrm{prob}} \cdot \varepsilon^2$ or $\text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \geq 1/\sqrt{C_{\mathrm{prob}}}\}$, this process is lossy (due to the $\frac{1}{\gamma}$ term in Eq. (24)), and spoils the prospect of a faster-than-$\sqrt{T}$ rate. This is why we present our results for PAC in terms of the constrained DEC, and why we use a dedicated algorithm tailored to the constrained DEC (Algorithm 1) as opposed to a direct adaptation of the algorithm based on the offset DEC in Foster et al. (2021); taking the latter approach and combining it with Eq. (24) would not lead to fast rates.

D.1.2. REGRET DEC: CONSTRAINED VERSUS OFFSET

In light of the near-equivalence between constrained and offset DEC for the PAC setting, one might hope that a similar equivalence would hold for regret. However, a naive adaptation of the techniques used to prove Proposition D.2 only leads to the following, quantitatively weaker converse to Proposition D.1.

Proposition D.3. *For all $\gamma > 0$ and $\overline{M} \in \mathcal{M}^{+}$,*

$$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \leq \text{r-dec}^{\mathrm{c}}_{\gamma^{-1/2}}(\mathcal{M}, \overline{M}). \tag{25}$$

The bound on the offset DEC in Eq. (25) is unsatisfying due to the scale $\varepsilon = \gamma^{-1/2}$ on the right-hand side. For example, in the case of the multi-armed bandit with A actions, we have $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon\sqrt{A}$ and $\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) \propto \frac{A}{\gamma}$, yet Eq. (25) only yields the inequality $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \lesssim \sqrt{\frac{A}{\gamma}}$. This leads to a suboptimal $A^{1/3}T^{2/3}$-type regret bound when plugged into the upper bounds from Foster et al. (2021) (cf. Eq. (44)). Unfortunately, the following result shows that Proposition D.3 is tight in general, even when $\overline{M} \in \mathcal{M}$.

Proposition D.4. *For all $\gamma \geq 1$, there exists a model class $\mathcal{M}$ such that for all $\varepsilon \in (0,1)$,*

$$\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq O\left(\varepsilon^2 \gamma^{1/2}\right),$$

so in particular $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \leq O(\varepsilon)$ for $\varepsilon \leq \gamma^{-1/2}$. Yet, there exists $\overline{M} \in \mathcal{M}$ such that

$$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \geq \Omega\left(\gamma^{-1/2}\right).$$

This rules out the possibility of an inequality tighter than Eq. (25), and shows that the constrained and offset DEC have fundamentally different behavior for regret.

D.2. Localization

For a scale parameter $\alpha > 0$ and reference model $\overline{M} \in \mathcal{M}^{+}$, define the following *localized* subclass of $\mathcal{M}$:

$$\mathcal{M}_{\alpha}(\overline{M}) := \{M \in \mathcal{M} \mid f^{M}(\pi_M) \leq f^{\overline{M}}(\pi_{\overline{M}}) + \alpha\}. \tag{26}$$

The tightest upper and lower bounds on regret in Foster et al. (2021) are stated in terms of the offset DEC for the localized class (26), for an appropriate choice of $\alpha > 0$ that depends on T (see

Appendix E for precise statements). Our bounds based on the constrained DEC avoid the explicit use of localization, but in what follows, we show that the constrained DEC *implicitly* enforces a form of localization.

Proposition D.5 (Localization for PAC DEC)**.** *For all* $\varepsilon > 0$ *and* $\overline{M} \in \mathcal{M}^+$*, letting* $\alpha(\varepsilon) := \sqrt{3}\varepsilon + \textsf{p-dec}^{\mathrm{c}}_{\sqrt{6}\varepsilon}(\mathcal{M}, \overline{M})$*, we have*

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \textsf{p-dec}^{\mathrm{c}}_{\sqrt{3}\varepsilon}(\mathcal{M}_{\alpha(\varepsilon)}(\overline{M}), \overline{M}).$$

A similar result holds for regret, but we require that the DEC satisfies a slightly stronger version of the regularity condition in Assumption C.2.

Definition D.1 (Strong regularity of DEC)**.** *For* $\overline{M} \in \mathcal{M}^+$*, the constrained DEC is said to satisfy the strong regularity condition relative to* $\overline{M}$ *if there exist constants* $C_{\mathrm{reg}} \geq \sqrt{2}$ *and* $c_{\mathrm{reg}} < C_{\mathrm{reg}}$ *such that for all* $\varepsilon > 0$*,*

$$\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}, \overline{M}) \leq c^2_{\mathrm{reg}} \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}). \tag{27}$$

The constrained DEC is said to satisfy strong regularity relative to a class $\mathcal{M}' \subseteq \mathcal{M}^+$ *if for all* $\varepsilon > 0$*,*

$$\sup_{\overline{M}\in\mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq c^2_{\mathrm{reg}} \cdot \sup_{\overline{M}\in\mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}). \tag{28}$$

This condition is satisfied with $C_{\mathrm{reg}} = 2$ and $c_{\mathrm{reg}} = 2^{\rho/2}$ whenever $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \propto \varepsilon^{\rho}$ for $\rho < 2$.

Proposition D.6 (Localization for regret DEC)**.** *Let* $\overline{M} \in \mathcal{M}^+$ *be given, and assume that the strong regularity condition (27) is satisfied relative to* $\overline{M}$*. Then, for all* $\varepsilon > 0$*, letting* $\alpha(\varepsilon) := C_{\mathrm{reg}} \cdot \varepsilon + \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}, \overline{M}) \leq C^2_{\mathrm{reg}} \cdot \big(\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})\big)$*, we have*

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq C_{\mathrm{loc}} \cdot \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}_{\alpha(\varepsilon)}(\overline{M}), \overline{M}),$$

where $C_{\mathrm{loc}} := \left(\frac{1}{c^2_{\mathrm{reg}}} - \frac{1}{C^2_{\mathrm{reg}}}\right)^{-1}$.

Note that in the case where $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \propto \varepsilon^{\rho}$ for a constant $\rho < 2$, choosing $C_{\mathrm{reg}} = 2$ and $c_{\mathrm{reg}} = 2^{\rho/2}$ gives $C_{\mathrm{loc}} = O(1)$. These results show that the constrained DEC—both for PAC and regret—is equivalent (up to constants) to the constrained DEC for the localized subclass $\mathcal{M}_{\alpha}(\overline{M})$, for a radius α that depends on the value of the DEC itself. In contrast, the offset DEC does not automatically enforce any form of localization, which explains why it was necessary to explicitly restrict to a localized subclass in prior work.

D.2.1. CONSTRAINED VERSUS OFFSET DEC: TIGHTER EQUIVALENCE FOR LOCALIZED CLASSES

Building on the insights in the prequel, we now show that for localized classes, it is possible to bound the offset DEC for regret by the constrained DEC in a tighter fashion that improves upon Proposition D.3.

Proposition D.7. *Let $\alpha, \gamma > 0$ and $\overline{M} \in \mathcal{M}^{+}$ be given. For all $\varepsilon > 0$, we have*

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \max\left\{0,\ \alpha + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^{2}}{2}\right\}, \tag{29}$$

which in particular yields

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\sqrt{2\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \frac{1}{2\gamma}. \tag{30}$$

The bound in Eq. (30) replaces the term $\textsf{r-dec}^{\mathrm{c}}_{\sqrt{1/\gamma}}(\mathcal{M})$ in Proposition D.3 with $\textsf{r-dec}^{\mathrm{c}}_{\sqrt{\alpha/\gamma}}(\mathcal{M})$, leading to improvement when $\alpha \ll 1$. Notably, the bound is strong enough that, by combining it with Proposition D.6, it is possible to upper bound the constrained DEC by the localized offset DEC, and then pass back to the constrained DEC in a fashion that loses only constant factors—at least whenever $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \gtrsim \varepsilon$. The following result uses this approach to derive a near-equivalence for the constrained DEC and localized offset DEC; we also use this approach within the proof of Proposition D.11.

Proposition D.8. *Whenever the strong regularity condition (27) in Definition D.1 is satisfied for $\overline{M} \in \mathcal{M}^{+}$, it holds that for all $\varepsilon > 0$, letting $\alpha(\varepsilon, \gamma) := \gamma\varepsilon^{2}$,*

$$c_1 \cdot \sup_{\gamma > c_3\varepsilon^{-1}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{c_2\cdot\alpha(\varepsilon,\gamma)}(\overline{M}), \overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq c_1' \cdot \sup_{\gamma > c_3'\varepsilon^{-1}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{c_2'\cdot\alpha(\varepsilon,\gamma)}(\overline{M}), \overline{M}) + c_4'\varepsilon, \tag{31}$$

where $c_1, c_2, c_3 > 0$ are numerical constants and $c_1', c_2', c_3', c_4' > 0$ are constants that depend only on C_{reg} and c_{reg}.

In light of this result, our upper bounds (Theorem C.2) can be thought of as improving prior work by achieving the tightest possible localization radius (roughly, $\alpha = O(\gamma\varepsilon^{2})$ instead of $\alpha = O(\gamma\varepsilon^{2} + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}))$). Appendix E gives examples for which this leads to quantitative improvement in rate.

D.3. Reference Models: Role of Suboptimality

We now turn our attention to understanding the role of the reference model $\overline{M}$ with respect to which the Decision-Estimation Coefficient is defined. Recall that for regret, our upper and lower bounds scale with

$$\begin{aligned} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) &= \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \\ &= \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M} \cup \{\overline{M}\}} \left\{ \mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \mid \mathbb{E}_{\pi\sim p}\left[D^{2}_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^{2} \right\}. \end{aligned} \tag{32}$$

By maximizing over $M \in \mathcal{M} \cup \{\overline{M}\}$, this definition forces the min-player to choose $p \in \Delta(\Pi)$ such that the suboptimality $\mathbb{E}_{\pi\sim p}\big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\big]$ under $\overline{M}$ is small. This is somewhat counterintuitive, since $\overline{M} \in \mathrm{co}(\mathcal{M})$ does not necessarily lie in the class $\mathcal{M}$, yet our results show that $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$ characterizes the minimax regret for $\mathcal{M}$. A-priori, one might expect that the quantity $\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, which does not incorporate suboptimality under $\overline{M}$, would be a more natural complexity measure. In what follows, we show that this quantity has fundamentally different behavior from Eq. (32), and that incorporating suboptimality under $\overline{M}$ is essential to characterize minimax regret.

Proposition D.9. *For any $\varepsilon > 0$ sufficiently small, there exists a model class $\mathcal{M}$ such that*

$$\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M}) \leq c\cdot\varepsilon, \tag{33}$$

yet

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \geq c'\cdot\varepsilon^{2/3}, \tag{34}$$

where $c, c' > 0$ are numerical constants.

It is straightforward to show that for the choice $\varepsilon = \underline{\varepsilon}(T) \propto 1/\sqrt{T}$, the optimal regret for the class in Proposition D.9 is $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] = \widetilde{\Theta}(T^{2/3})$. This result is recovered by Theorem C.1, which scales with the quantity in Eq. (34). However, the quantity in Eq. (33) incorrectly suggests a $\sqrt{T}$-type rate, which is not achievable.

For the offset DEC, the role of suboptimality under $\overline{M}$ is more subtle. It is possible to show that in general, $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \gg \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})$, analogous to Proposition D.9, but Proposition D.1 shows that the latter quantity suffices to upper bound bound $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})$.

While the preceding discussion shows that incorporating suboptimality under $\overline{M} \notin \mathcal{M}$ is necessary to obtain tight guarantees for regret, the following result shows that this distinction is largely inconsequential for PAC, and motivates the definition $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})$.

Proposition D.10. *For all $\overline{M} \in \mathcal{M}^{+}$ and $\varepsilon > 0$,*

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq \textsf{p-dec}^{\mathrm{c}}_{\sqrt{3}\varepsilon}(\mathcal{M},\overline{M}) + 4\varepsilon. \tag{35}$$

D.4. Reference Models: Role of Convexity and Randomization

We now focus on understanding the role of *improper* reference models $\overline{M} \notin \mathcal{M}$. Focusing on regret, our upper bound (Theorem C.2) scales with

$$\textsf{r-dec}^{\mathrm{c}}_{\overline{\varepsilon}(T)}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\overline{\varepsilon}(T)}(\mathcal{M}\cup\{\overline{M}\},\overline{M}), \tag{36}$$

which maximizes over all possible reference models in the convex hull $\mathrm{co}(\mathcal{M})$. On the other hand, our lower bound (Theorem C.1) scales with

$$\sup_{\overline{M}\in\mathcal{M}^{+}} \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \geq \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}). \tag{37}$$

Both results allow for improper models $\overline{M} \notin \mathcal{M}$, but the quantity (37) allows the reference model to be unconstrained, and could be larger than the quantity (36) a-priori. Why is there no contradiction here? In what follows, we show that for both the constrained and offset DEC, allowing for arbitrary, reference models $\overline{M} \in \mathcal{M}^{+}$ as in Eq. (37) can only increase the value beyond that achieved by $\overline{M} \in \mathrm{co}(\mathcal{M})$ by constant factors.

Before stating our results, let us mention a secondary, related goal, which is to understand the role of *randomized reference models*. Foster et al. (2021) introduce a variant of the Decision-Estimation Coefficient tailored to randomized (or, mixture) reference models, in which $\overline{M}$ is drawn from a distribution $\nu \in \Delta(\mathcal{M})$. We define constrained and offset variants of this complexity measure for $\nu \in \Delta(\mathcal{M})$ as follows:

$$\textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M},\nu) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \mid \mathbb{E}_{\overline{M}\sim\nu}\,\mathbb{E}_{\pi\sim p}\left[D^{2}_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq \varepsilon^{2}\right\}, \tag{38}$$

$$\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M},\nu) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot\mathbb{E}_{\overline{M}\sim\nu}\left[D^{2}_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\right]. \tag{39}$$

Recent work of Chen et al. (2022) extends the results of Foster et al. (2021) to provide regret bounds that scale with $\sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\nu)$, which one might hope to be smaller than $\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})$ (it is never larger, due to Jensen's inequality). We show that this is not the case: For both constrained and offset, the randomized DEC is sandwiched between the DEC with $\overline{M}\in\mathcal{M}^{+}$ and the DEC with $\overline{M}\in\mathrm{co}(\mathcal{M})$.

Proposition D.11. *Suppose that Assumption G.1 is satisfied. For all $\gamma > 0$, we have*

$$\sup_{\overline{M}\in\mathcal{M}^{+}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M}) \leq \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M},\nu) \leq \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma/4}(\mathcal{M},\overline{M}). \tag{40}$$

In addition, suppose that the strong regularity condition (Definition D.1, Eq. (28)) is satisfied relative to $\mathcal{M}^{+}$. Then for all $\varepsilon > 0$, we have

$$\sup_{\overline{M}\in\mathcal{M}^{+}} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq c_{1} \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{c_{2}\varepsilon}(\mathcal{M}\cup\{\overline{M}_{\nu}\},\nu) + c_{3}\varepsilon \tag{41}$$

$$\leq c_{1} \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{c_{2}\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) + c_{3}\varepsilon, \tag{42}$$

where $\overline{M}_{\nu} := \mathbb{E}_{M'\sim\nu}[M']$ and $c_1, c_2, c_3 > 0$ are constants that depend only on $C_{\mathrm{reg}}, C_{\mathrm{loc}} > 0$.

A similar equivalence holds for PAC; see Appendix J.2.2. The main consequences of this result are as follows.

- Since allowing for arbitrary reference models $\overline{M}\in\mathcal{M}^{+}$ never increases the value over reference models $\overline{M}\in\mathrm{co}(\mathcal{M})$, one can freely work with whichever version is more convenient, either for upper or lower bounds.
- From a statistical perspective, it is not possible to further tighten our results by working with the DEC with randomized estimators, since this complexity measure is never smaller than the variant with $\overline{M}\in\mathrm{co}(\mathcal{M})$ by more than constant factors.

We mention in passing that the proof of the equivalence (40) for the offset DEC is a simple consequence of the minimax theorem and convexity of squared Hellinger distance, but the proof of the equivalence (41) is quite involved, and uses the tools developed in Appendix D.2 to pass back and forth between the constrained and offset DEC. We are curious as to whether there is a simpler proof.

Appendix E. Improvement over Prior Work

In this section, we use the tools developed in Appendix D to show that the regret bounds in Theorems C.1 and C.2 always improve upon those in prior work (Foster et al., 2021, 2022b). We then highlight some concrete model classes for which our bounds provide meaningful improvement, and discuss additional related work.

Regret bounds from prior work. Recall that for a model class $\mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}$, we define the localized subclass around $\overline{M}$ via

$$\mathcal{M}_{\alpha}(\overline{M}) = \left\{ M \in \mathcal{M} : f^{\overline{M}}(\pi_{\overline{M}}) \geq f^{M}(\pi_{M}) - \alpha \right\}. \tag{43}$$

where $\alpha > 0$ is the radius. Focusing on finite classes for simplicity, the best upper bounds from prior work are those of Foster et al. (2021), which take the form

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq \widetilde{O}(1) \cdot \min_{\gamma>0} \max\left\{ \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\overline{\Delta}(\gamma,T)}(\overline{M}), \overline{M}) \cdot T,\ \gamma \cdot \log|\mathcal{M}| \right\}, \tag{44}$$

for $\overline{\Delta}(\gamma, T) = \widetilde{O}\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) + \frac{\gamma}{T}\log|\mathcal{M}| + \gamma^{-1}\big)$. The best lower bounds from prior work are those of Foster et al. (2022b, Theorem D.1), which apply to all algorithms with "sub-Chebychev" tail behavior,[7] and scale as

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \Omega(1) \cdot \max_{\gamma > \sqrt{C(T)T}} \sup_{\overline{M}\in\mathcal{M}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\underline{\Delta}(\gamma,T)}(\overline{M}), \overline{M}) \cdot T, \tag{45}$$

where $C(T) := O(\log(T \wedge V(\mathcal{M})))$ and $\underline{\Delta}(\gamma, T) := C(T)^{-1} \cdot \frac{\gamma}{T}$.

Our improvement. The following result, which follows immediately from Proposition D.8, implies that the upper and lower bounds in Theorem C.2 and Theorem C.1, are always tighter than the guarantees in Eq. (44) and Eq. (45), respectively, under an appropriate regularity condition.

Corollary E.1. Whenever the strong regularity condition (Definition D.1) is satisfied for $\overline{M} \in \mathcal{M}^{+}$ with $C_{\mathrm{loc}}, C_{\mathrm{reg}} = O(1)$, we have that for all $\varepsilon > 0$ and $\gamma > 0$,

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq O\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\overline{\alpha}(\varepsilon,\gamma)}(\overline{M}), \overline{M}) \vee 0 + \gamma\varepsilon^{2} + \varepsilon\big), \tag{46}$$

where $\overline{\alpha}(\varepsilon, \gamma) = O\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^{2} + \gamma^{-1}\big)$. In addition, for all $\varepsilon > 0$, $\gamma \geq \Omega(\varepsilon^{-1})$, and $\overline{M} \in \mathcal{M}^{+}$,

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \geq \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\underline{\alpha}(\varepsilon,\gamma)}(\overline{M}), \overline{M}), \tag{47}$$

where $\underline{\alpha}(\varepsilon, \gamma) = \Omega\big(\gamma\varepsilon^{2}\big)$.

By applying Eq. (46) with $\bar{\varepsilon}(T) = \widetilde{O}\Big(\sqrt{\frac{\log|\mathcal{M}|}{T}}\Big)$, we conclude that the upper bound in Theorem C.2 is always bounded above by the quantity in Eq. (44) up to logarithmic factors in T and $1/\delta$. Similarly, by applying Eq. (47) with $\underline{\varepsilon}(T) = \widetilde{\Omega}\Big(\sqrt{\frac{1}{T}}\Big)$ we see that the lower bound in Theorem C.1 is always bounded below by the quantity in Eq. (45) up to $\log(T)$ factors and an additive $O(\sqrt{T})$ term. Beyond simply scaling with a larger complexity measure, Theorem C.1 1) holds for arbitrary algorithms, removing the sub-Chebychev assumption used by Foster et al. (2022b), and 2) allows for improper reference models $\overline{M} \notin \mathcal{M}$.

We now provide concrete model classes for which our results lead to quantitative improvements in rates. Our first example is a model class for which our main upper bound (Theorem C.2) improves over Foster et al. (2021) by (implicitly) achieving a tighter localization radius than Eq. (44).

[7] Sub-Chebychev algorithms are those for which the root-mean-squared regret is of the same order as the expected regret. Foster et al. (2021) provide lower bounds that do not require the assumption of sub-Chebychev tail behavior, but these results depend on the DEC for a smaller subclass of the form $\mathcal{M}^{\infty}_{\alpha}(\overline{M}) = \left\{ M \in \mathcal{M} : |g^{M}(\pi) - g^{\overline{M}}(\pi)| \leq \alpha \ \ \forall \pi \in \Pi \right\}$, and can be loose compared to Eq. (45).

Example E.1 (Improvement from upper bound)**.** Consider a model class $\mathcal{M}^{\alpha,\beta}$ parameterized by $\alpha \in (0, 1/2]$, $\beta \in (0, 1)$, and $A \in \mathbb{N}$.

1. $\Pi = [A] \cup \{\pi_\circ\}$, where $\pi_\circ$ is a "revealing" decision.
2. $\mathcal{O} = [A] \cup \{\perp\}$, where $\perp$ is a null symbol.
3. We have $\mathcal{M} = \{M_{\alpha,i}\}_{i\in[A]} \cup \{\widetilde{M}\}$. For each $i \in [A]$, the model $M_{\alpha,i} \in \mathcal{M}^{\alpha,\beta}$ has rewards and observations defined as follows:
   (a) For $\pi \in [A]$, $f^{M_{\alpha,i}}(\pi) = \frac{1}{2} + \alpha \cdot \mathbb{1}\{\pi = i\}$, and $f^{M_{\alpha,i}}(\pi_\circ) = 0$. All $\pi \in \Pi$ have $r = f^{M_{\alpha,i}}(\pi)$ almost surely under $r \sim M_{\alpha,i}(\pi)$.
   (b) For $\pi \in [A]$, we receive the observation $o =\perp$. Selecting $\pi_\circ$ gives the observation $o = i \in [A]$ with probability β and $o =\perp$ with probability $1 - \beta$.
4. The model $\widetilde{M}$ is defined as follows:
   (a) We have $f^{\widetilde{M}}(\pi) = \frac{1}{2}$ for all $\pi \in [A]$ and $f^{\widetilde{M}}(\pi_\circ) = 0$, with $r = f^{\widetilde{M}}(\pi)$ almost surely under $r \sim \widetilde{M}(\pi)$ for all $\pi \in \Pi$.
   (b) All $\pi \in [A]$ have $o =\perp$ almost surely. For $\pi_\circ$, we observe $o =\perp$ with probability $1 - \beta$ and $o \sim \mathrm{Unif}([A])$ with probability β.

Let $\mathcal{M} := \mathcal{M}^{\alpha_1,\beta} \cup \mathcal{M}^{\alpha_2,\beta}$, with $\alpha_1 = 1/2$, $\alpha_2 \propto T^{-1/4}$, $\beta \propto T^{-1/2}$, and $A \propto T^2$. Then:

- The E2D$^+$ algorithm, via Theorem C.2, achieves $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq \widetilde{O}(\sqrt{T})$.
- The regret bound in Eq. (44) scales with $\widetilde{\Omega}(T^{5/8})$.

◁

The next example is a model class for which our main lower bound (Theorem C.1) improves over Foster et al. (2021), as a consequence of allowing for improper reference models $\overline{M} \notin \mathcal{M}$.

Example E.2 (Improvement from lower bound)**.** Let $A \in \mathbb{N}$ and $\Pi = \{1, \ldots, A\}$. Consider the multi-armed bandit model class $\mathcal{M} = \{M_1, \ldots, M_A\}$ consisting models of the form

$$M_i(\pi) = \mathrm{Ber}(f_i(\pi)),$$

where $f_i(\pi) := \frac{1}{2} + \Delta\mathbb{I}\{\pi = i\}$. Foster et al. (2021) show that regardless of how $\Delta > 0$ is chosen, $\textsf{r-dec}^{\mathrm{o}}_\gamma(\mathcal{M}, \overline{M}) \leq \frac{1}{\gamma}$ for all $\gamma > 0$ and $\overline{M} \in \mathcal{M}$, so the lower bound (45) can at most give $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \Omega(\sqrt{T})$. On the other hand, by choosing $\overline{M}(\pi) = \mathrm{Ber}(\frac{1}{2})$, which has $\overline{M} \notin \mathcal{M}$, it is straightforward to see that whenever $\Delta \propto \varepsilon\sqrt{A}$, we have $\textsf{r-dec}^{\mathrm{c}}_\varepsilon(\mathcal{M}) \geq \textsf{r-dec}^{\mathrm{c}}_\varepsilon(\mathcal{M}, \overline{M}) \geq \Omega(\varepsilon\sqrt{A})$. Setting $\Delta \propto \underline{\varepsilon}(T) \cdot \sqrt{A}$, Theorem C.1 gives

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

which is optimal. This shows that in general, allowing for improper reference models $\overline{M} \notin \mathcal{M}$ is necessary to obtain tight lower bounds. ◁

E.1. Additional Related Work

Concurrent work of Chen et al. (2022) independently discovered the offset variant of the PAC Decision-Estimation Coefficient, and used it to give upper and lower bounds for PAC sample complexity by adapting the techniques of Foster et al. (2021). Our guarantees for both regret and PAC are always tighter than these results, analogous to the improvement we obtain over Foster et al. (2021) (see also Appendix D.4), but our techniques are otherwise complementary.

Additionally, recent work of Lattimore (2022) considers the closely related framework of adversarial partial monitoring, and gives upper and lower bounds on regret based on a generalization of the information ratio (Russo and Van Roy, 2014, 2018), which is related to the DEC (Foster et al., 2022b). The upper and lower bounds on regret given by Lattimore (2022) are loose by $\mathrm{poly}(|\Pi|)$ factors, and consequently it appears unlikely that this complexity measure can give tight guarantees in the "large decision-space/model class" regime where $T \ll \min\{|\mathcal{M}|, |\Pi|\}$, which is our focus.

Appendix F. Additional Examples

We close with some brief examples that showcase the behavior of the constrained Decision-Estimation Coefficient, as well as our upper and lower bounds, for standard model classes of interest. For regret, Foster et al. (2021) provide lower bounds on the (localized) offset DEC for a number of canonical models in bandits and reinforcement learning.[8] It is straightforward to derive lower bounds on the constrained DEC by combining these results with Corollary E.1.

Likewise, Foster et al. (2021) give global upper bounds on the offset DEC for the same examples, which immediately lead to upper bounds on the constrained DEC via Proposition D.1. This approach leads to lower and upper bounds on the constrained Decision-Estimation Coefficient for all of the examples considered in Foster et al. (2021). We summarize these results and the implied lower bounds on regret, in Table 1; the upper bounds on the DEC and regret are similar, but depend additionally on $\mathbf{Est}_{\mathsf{H}}(T, \delta)$. See Foster et al. (2021) for further background, including references to papers originally deriving upper and lower bounds for each of the model classes.

Setting	$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$ Lower Bound	Regret Lower Bound (Theorem C.1)
Multi-Armed Bandit	$\varepsilon\sqrt{A}$	$\sqrt{AT}$
Multi-Armed Bandit w/ gap	$\Delta\mathbb{I}\{\varepsilon > \Delta/\sqrt{A}\}$	A/Δ
Linear Bandit	$\varepsilon\sqrt{d}$	$\sqrt{dT}$
Lipschitz Bandit	$\varepsilon^{1-\frac{d}{d+2}}$	$T^{\frac{d+1}{d+2}}$
ReLU Bandit	$\mathbb{I}\{\varepsilon > 2^{-\Omega(d)}\}$	$2^{\Omega(d)}$
Tabular RL	$\varepsilon\sqrt{HSA}$	$\sqrt{HSAT}$
Linear MDP	$\varepsilon\sqrt{d}$	$\sqrt{dT}$
RL w/ linear $Q^{\star}$	$\mathbb{I}\{\varepsilon \geq 2^{-\Omega(d)} \vee 2^{-\Omega(H)}\}$	$2^{\Omega(d)} \wedge 2^{\Omega(H)}$
Deterministic RL w/ linear $Q^{\star}$	$\mathbb{I}\{\varepsilon \geq 1/\sqrt{d}\}$	d

Table 1: Lower bounds for bandits and reinforcement learning recovered by the constrained Decision-Estimation Coefficient, where A = #actions, Δ = gap, d = feature dim., H = episode horizon, and S = #states. Numerical constants and $\log(T)$ factors are suppressed.

[8] While many of their derived lower bounds are tight, some are not, such as the Linear Bandit and Linear MDP lower bounds, which are off by $\mathrm{poly}(d)$, and $\mathrm{poly}(H, d)$ factors, respectively (see Foster et al. (2021) for details).

Example: Multi-armed bandit. We now sketch the approach to lower bounds outlined above in greater detail, focusing on multi-armed bandits for concreteness. Foster et al. (2021) show that for when $\mathcal{M}$ is the class of all multi-armed bandit instances with $\Pi = \{1, \ldots, A\}$ and Bernoulli rewards, there exists $\overline{M} \in \mathcal{M}$ such that for all $\gamma \geq c_1 \cdot A$,

$$\sup_{\overline{M} \in \mathcal{M}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha_\gamma}(\overline{M}), \overline{M}) \geq c_2 \cdot \frac{A}{\gamma},$$

where $\alpha_\gamma := c_3 \cdot \frac{A}{\gamma}$, and $c_1, c_2, c_3 > 0$ are numerical constants. Corollary E.1 implies that for all $\varepsilon > 0$ and $\overline{M} \in \mathcal{M}^+$,

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \geq \sup_{\gamma > 0} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\underline{\alpha}(\varepsilon,\gamma)}(\overline{M}), \overline{M}),$$

where $\underline{\alpha}(\varepsilon, \gamma) = c \cdot \gamma\varepsilon^2$ for a sufficiently small numerical constant c. For any given $\varepsilon > 0$, if we set $\gamma = c' \cdot A^{1/2}/\varepsilon$ for a sufficiently large constant c', we have $\mathcal{M}_{\alpha_\gamma}(\overline{M}) \subseteq \mathcal{M}_{\underline{\alpha}(\varepsilon,\gamma)}(\overline{M})$, and we conclude that

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \geq \Omega\Big(\varepsilon\sqrt{A}\Big)$$

for all $\varepsilon \leq c'' \cdot A^{-1/2}$, where c'' is a sufficiently small constant. Plugging this lower bound on the DEC into Theorem C.1 yields a lower bound on regret of the form $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{AT})$.

Part II
Proofs

Appendix G. Preliminaries

G.1. Minimax Theorem

For certain structural results, we require that the offset Decision-Estimation Coefficient (either the regret or PAC variant) is equal to its Bayesian counterpart. This is a consequence of the minimax theorem whenever mild topological conditions are satisfied; note that our objective can always be made convex-concave by writing

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

so all that is required to invoke the minimax theorem is compactness. We state this as an assumption to avoid committing to a particular set of technical conditions.

Assumption G.1 (Minimax swap)**.** *For the regret DEC, we have*

$$\begin{aligned}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) &= \underline{\textsf{r-dec}}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) &(48)\\ &:= \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p \in \Delta(\Pi)} \mathbb{E}_{\pi \sim p, M \sim \mu}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. &(49)\end{aligned}$$

For the PAC DEC, we have

$$\begin{aligned}\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) &= \underline{\textsf{p-dec}}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) &(50)\\ &:= \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p,q \in \Delta(\Pi)} \mathbb{E}_{M \sim \mu}\big[\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi \sim q}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big].\end{aligned}$$

As the simplest possible example, Assumption G.1 is satisfied whenever $\mathcal{R}$ is bounded and Π is finite (cf. Proposition 4.2 in Foster et al. (2021)), but assumption can be shown to hold under substantially more general conditions.

Appendix H. Proofs for Lower Bounds

In what follows, we prove the PAC lower bound (Theorem 2.1), then prove the lower bound for regret (Theorem C.1); the latter is similar to the former, but carries a number of additional challenges that lead to slightly different techniques.

H.1. Proof of PAC Lower Bound (Theorem 2.1)

Preliminaries. Formally, for $T \in \mathbb{N}$, an *algorithm* for the PAC framework is a collection of mappings $(p, q) = \left(\{q^t(\cdot \mid \cdot)\}_{t=1}^T, p(\cdot \mid \cdot)\right)$ that (adaptively) draws decisions $\pi^t \sim q^t(\cdot \mid \mathfrak{H}^{t-1})$ (for $t \in [T]$), and then outputs the final decision $\widehat{\pi} \sim p(\cdot \mid \mathfrak{H}^T)$ conditioned on the history $\mathfrak{H}^T$. We define $\mathbb{P}^{M,(p,q)}$ as the law of $\mathfrak{H}^T$ when the underlying model is M and the algorithm is (p, q). Throughout the proof, we will use the elementary property $D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) \leq D_{\mathsf{H}}(\mathbb{P}, \mathbb{Q})$.
Proof of Theorem 2.1. For later reference, we define a constant

$$c_0 = 1/16,$$

and define $C(T) = 2^8 \cdot \log(T \wedge V(\mathcal{M}))$. Fix $T \in \mathbb{N}$ and an algorithm $(p, q) = \{q^t(\cdot \mid \cdot), p(\cdot \mid \cdot)\}_{t=1}^T$. For each model $M \in \mathcal{M}^+$, we use the abbreviation $\mathbb{P}^M \equiv \mathbb{P}^{M,(p,q)}$, and write $\mathbb{E}^M$ for the corresponding expectation. In addition, we define

$$p_M = \mathbb{E}^M\left[p(\cdot \mid \mathfrak{H}^T)\right], \quad \text{and} \quad q_M = \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^T q^t(\cdot \mid \mathfrak{H}^{t-1})\right].$$

Choosing a hard pair of models. Fix an arbitrary reference model $\overline{M} \in \mathcal{M}^+$ and define $\varepsilon := \frac{1}{10\sqrt{C(T) \cdot T}}$ and $\underline{\varepsilon}(T) := \varepsilon/\sqrt{2}$. We will prove a lower bound in terms of $\mathsf{p\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}, \overline{M})$. We abbreviate $\delta := \mathsf{p\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}, \overline{M})$, so that the assumption (11) gives $\delta \geq 48\varepsilon$.

To begin, choose any model $M_1 \in \mathcal{M}$ satisfying:

$$M_1 \in \left\{M \in \mathcal{M} \,:\, \mathbb{E}_{\pi \sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2 \,\wedge\, \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2\right\}. \tag{51}$$

We will make use of the fact that, by Lemma J.3, we have that for all $p \in \Delta(\Pi)$,

$$\begin{aligned}&\mathsf{dec}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M})\\ &\leq \sup_{M \in \mathcal{M}}\left\{\mathbb{E}_{\pi \sim p}\left[g^M(\pi)\right] \mid \mathbb{E}_{\pi \sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2 \,\wedge\, \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2\right\},\end{aligned} \tag{52}$$

which implies (along with Eq. (11)) that the set in Eq. (51) is non-empty. For any model $M \in \mathcal{M}^+$, define

$$\mathcal{E}^M := \left\{\pi \in \Pi : g^M(\pi) \geq c_0 \cdot \delta\right\},$$

and define $\mathcal{A}_1 := \mathcal{E}^{M_1}$. Let $p' := p_{\overline{M}}(\cdot \mid \mathcal{A}_1^{\mathrm{c}})$, and set

$$M_2 := \operatorname*{arg\,max}_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p'}\left[g^M(\pi)\right] \mid \mathbb{E}_{\pi \sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2 \ \wedge \ \mathbb{E}_{\pi \sim p'}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2 \right\}; \tag{53}$$

as with M_1, Lemma J.3 implies that this set is non-empty. Finally, define $\mathcal{A}_2 := \mathcal{E}^{M_2} \cap \mathcal{A}_1^{\mathrm{c}}$.

Lower bounding the algorithm's risk. We now recall Lemma A.13 from Foster et al. (2021), which states that for all models M,

$$D^2_{\mathsf{H}}\left(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\right) \le C(T) \cdot T \cdot \mathbb{E}_{\pi \sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]. \tag{54}$$

By the data processing inequality, this further implies that

$$D^2_{\mathsf{TV}}(p_M, p_{\overline{M}}) \le C(T) \cdot T \cdot \mathbb{E}_{\pi \sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]. \tag{55}$$

Since $\mathbb{E}_{\pi \sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M_i(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2$ for $i \in \{1, 2\}$, our choice $\varepsilon \le \frac{1}{10 \cdot \sqrt{T \cdot C(T)}}$ implies that

$$D_{\mathsf{TV}}\big(p_{M_i}, p_{\overline{M}}\big) \le \frac{1}{10}, \quad \text{for} \quad i \in \{1, 2\}.$$

As a result, for each $i \in \{1, 2\}$, we have

$$\begin{aligned} \mathbb{E}_{\pi \sim p_{M_i}}\left[g^{M_i}(\pi)\right] &\ge c_0 \delta \cdot p_{M_i}(\pi \in \mathcal{E}^{M_i}) && (56) \\ &\ge c_0 \delta \cdot \big(p_{\overline{M}}(\pi \in \mathcal{E}^{M_i}) - D_{\mathsf{TV}}\big(p_{\overline{M}}, p_{M_i}\big)\big) \\ &\ge c_0 \delta \cdot (p_{\overline{M}}(\pi \in \mathcal{E}^{M_i}) - 1/10). && (57) \end{aligned}$$

Thus, to prove the theorem, it suffices to lower bound $p_{\overline{M}}(\pi \in \mathcal{E}^{M_i})$ by $1/4$ for at least one of $i \in \{1, 2\}$, which will show that the quantity in Eq. (57) is at least $\frac{3c_0\delta}{20}$ (in fact, any constant lower bound greater than $1/10$ suffices).

We assume henceforth that $p_{\overline{M}}(\mathcal{A}_1^{\mathrm{c}}) \ge 1/2$, as otherwise we have $p_{\overline{M}}(\mathcal{E}^{M_1}) = p_{\overline{M}}(\mathcal{A}_1) \ge 1/2$, in which case the result immediately follows from Eq. (57). Before continuing, we note that since $\mathbb{E}_{\pi \sim p_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M_1(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2$ and

$$\mathbb{E}_{\pi \sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot D^2_{\mathsf{H}}\big(M_2(\pi), \overline{M}(\pi)\big)\right] \le \mathbb{E}_{\pi \sim p'}\left[D^2_{\mathsf{H}}\big(M_2(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2,$$

the triangle inequality for Hellinger distance implies that $\mathbb{E}_{\pi \sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot D^2_{\mathsf{H}}(M_1(\pi), M_2(\pi))\right] \le 4\varepsilon^2$. Hence, using the fact that $|f^{M_1}(\pi) - f^{M_2}(\pi)| \le \sqrt{D^2_{\mathsf{H}}(M_1(\pi), M_2(\pi))}$ for all π and Jensen's inequality, we have

$$\mathbb{E}_{\pi \sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot |f^{M_1}(\pi) - f^{M_2}(\pi)|\right] \le \sqrt{\mathbb{E}_{\pi \sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot D^2_{\mathsf{H}}(M_1(\pi), M_2(\pi))\right]} \le 2\varepsilon. \tag{58}$$

Lower bounding the gap. To proceed, we will first establish that

$$f^{M_2}(\pi_{M_2}) \geq f^{M_1}(\pi_{M_1}) + \frac{\delta}{2} \cdot (1 - 4c_0) - 2\varepsilon. \tag{59}$$

To do this, we note that from the definition of M_2,

$$\mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_1^{\mathrm{c}} \right\} \cdot g^{M_2}(\pi) \right] \geq p_{\overline{M}}(\mathcal{A}_1^{\mathrm{c}}) \cdot \mathbb{E}_{\pi \sim p'} \left[g^{M_2}(\pi) \right] \geq \frac{1}{2} \cdot \textsf{p-dec}^{\mathrm{c}}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M}) = \frac{\delta}{2},$$

where we have used the assumption that $p_{\overline{M}}(\mathcal{A}_1^{\mathrm{c}}) \geq \frac{1}{2}$. Thus,

$$\begin{aligned} \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_2 \right\} \cdot g^{M_2}(\pi) \right] &= \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_1^{\mathrm{c}} \right\} \cdot g^{M_2}(\pi) \right] - \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_1^{\mathrm{c}} \cap (\mathcal{E}^{M_2})^{\mathrm{c}} \right\} \cdot g^{M_2}(\pi) \right] \\ &\geq \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_1^{\mathrm{c}} \right\} \cdot g^{M_2}(\pi) \right] - c_0\delta \geq \frac{\delta}{2} \cdot (1 - 2c_0), \end{aligned} \tag{60}$$

where the first inequality follows because $g^{M_2}(\pi) < c_0\delta$ for $\pi \in (\mathcal{E}^{M_2})^{\mathrm{c}}$.

Next, we compute

$$\begin{aligned} c_0\delta &\geq \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_2 \right\} \cdot \left(f^{M_1}(\pi_{M_1}) - f^{M_1}(\pi) \right) \right] \\ &= \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_2 \right\} \cdot \left(f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi) \right) \right] + p_{\overline{M}}(\mathcal{A}_2) \cdot \left(f^{M_1}(\pi_{M_1}) - f^{M_2}(\pi_{M_2}) \right) \\ &\geq \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ \pi \in \mathcal{A}_2 \right\} \cdot \left(f^{M_2}(\pi_{M_2}) - f^{M_2}(\pi) \right) \right] - 2\varepsilon + p_{\overline{M}}(\mathcal{A}_2) \cdot \left(f^{M_1}(\pi_{M_1}) - f^{M_2}(\pi_{M_2}) \right) \\ &\geq \frac{\delta}{2} \cdot (1 - 2c_0) - 2\varepsilon + p_{\overline{M}}(\mathcal{A}_2) \cdot \left(f^{M_1}(\pi_{M_1}) - f^{M_2}(\pi_{M_2}) \right), \end{aligned}$$

where the first inequality follows because $g^{M_1}(\pi) < c_0\delta$ for $\pi \in \mathcal{A}_2 \subset \mathcal{A}_1^{\mathrm{c}} = (\mathcal{E}^{M_1})^{\mathrm{c}}$, the second-to-last inequality follows from Eq. (58) and the fact that $\mathcal{A}_2 \subset \mathcal{A}_1^{\mathrm{c}}$, and the final inequality follows by Eq. (60). Rearranging and using that $p_{\overline{M}}(\mathcal{A}_2) \in [0, 1]$, we obtain

$$\begin{aligned} f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi_{M_1}) &\geq p_{\overline{M}}(\mathcal{A}_2) \cdot \left(f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi_{M_1}) \right) \\ &\geq \frac{\delta}{2} \cdot (1 - 4c_0) - 2\varepsilon. \end{aligned}$$

Lower bounding the failure probability and concluding. To finish the proof, we use the inequality (59) to bound the probability $p_{\overline{M}}((\mathcal{E}^{M_2})^{\mathrm{c}} \cap \mathcal{A}_1^{\mathrm{c}})$ as follows:

$$\begin{aligned} & p_{\overline{M}}((\mathcal{E}^{M_2})^{\mathrm{c}} \cap \mathcal{A}_1^{\mathrm{c}}) \cdot \left(\frac{\delta}{2} \cdot (1 - 4c_0) - 2\varepsilon \right) \\ &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ (\mathcal{E}^{M_2})^{\mathrm{c}} \cap \mathcal{A}_1^{\mathrm{c}} \right\} \cdot \left(f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi_{M_1}) \right) \right] \\ &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ (\mathcal{E}^{M_2})^{\mathrm{c}} \cap \mathcal{A}_1^{\mathrm{c}} \right\} \cdot \left(\left(f^{M_2}(\pi_{M_2}) - f^{M_2}(\pi) \right) - \left(f^{M_1}(\pi_{M_1}) - f^{M_1}(\pi) \right) \right) \right] + 2\epsilon \\ &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}} \left[\mathbb{1}\left\{ (\mathcal{E}^{M_2})^{\mathrm{c}} \cap \mathcal{A}_1^{\mathrm{c}} \right\} \cdot g^{M_2}(\pi) \right] + 2\varepsilon \\ &\leq c_0\delta + 2\varepsilon, \end{aligned}$$

where the first inequality uses Eq. (59), the second inequality uses Eq. (58) and the fact that $(\mathcal{E}^{M_2})^{\mathrm{c}} \cap \mathcal{A}_1^{\mathrm{c}} \subset \mathcal{A}_1^{\mathrm{c}}$, the third inequality uses that $g^{M_1} \geq 0$, and the final inequality uses the definition of $\mathcal{E}^{M_2}$. Since we have assumed that $\varepsilon \leq \frac{\delta}{48}$ and $c_0 \leq 1/16$, it follows that $p_{\overline{M}}((\mathcal{E}^{M_2})^{\mathrm{c}} \cap \mathcal{A}_1^{\mathrm{c}}) \leq \frac{\delta/8}{\delta/4} = 1/2$. Thus, we have established that

$$p_{\overline{M}}(\mathcal{A}_1) + p_{\overline{M}}(\mathcal{E}^{M_2}) \geq p_{\overline{M}}(\mathcal{A}_1 \cup \mathcal{E}^{M_2}) \geq 1/2,$$

which implies that either $p_{\overline{M}}(\mathcal{E}^{M_1}) = p_{\overline{M}}(\mathcal{A}_1) \geq 1/4$ or $p_{\overline{M}}(\mathcal{E}^{M_2}) \geq 1/4$. As a consequence, by Eq. (57), we have that for some $i \in \{1, 2\}$, $\mathbb{E}_{\pi \sim p_{M_i}}\left[g^{M_i}(\pi)\right] \geq \frac{3c_0\delta}{20}$, thus establishing the desired lower bound.

To wrap up, we note that the lower bound we have established holds for an arbitrary reference model $\overline{M} \in \mathcal{M}^+$, so we are free to choose $\overline{M} \in \mathcal{M}^+$ to maximize $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M})$.

□

Remark H.1. The structure of the proof Theorem 2.1 bears some superficial similarities to that of the classical two-point method (Donoho and Liu, 1987, 1991a,b; Yu, 1997; Tsybakov, 2008) in statistics and information theory, but has a number of fundamental differences.

1. First, in our lower bound, the pair of models (M_1, M_2) is chosen in an *adversarial* fashion based on the algorithm under consideration, while the classical approach selects a pair of models obliviously. When considering only two models, choosing the models adversarially is critical to capture the complexity of classes $\mathcal{M}$ that require distinguishing between many distinct decisions. For example, even in the simple special case of multi-armed bandits, this is necessary to make the number of actions A appear in the lower bound.

2. Second, and perhaps more importantly, the classical two-point argument cannot be directly applied because the function $g^M(\pi)$ does not enjoy the metric structure required by this approach. In particular, the classical "separation condition", which takes the form $g^{M_1}(\pi) + g^{M_2}(\pi) \geq \delta \;\forall \pi$ when applied to our setting, does not hold. Instead, the crux of the proof is to show that, as a consequence of our choice for M_1 and M_2, we have
$$\mathbb{E}_{\pi \sim p_{M_1}}\left[g^{M_1}(\pi)\right] + \mathbb{E}_{\pi \sim p_{M_2}}\left[g^{M_2}(\pi)\right] \gtrsim \delta.$$
To establish this inequality, we take advantage of the fact that 1) $\textsf{p-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \gtrsim \underline{\varepsilon}(T)$, by assumption, and 2) rewards r are observed and lie in the range $[0, 1]$; the latter implies that $|f^M(\pi) - f^{\overline{M}}(\pi)| \leq D_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)$ for all $M, \overline{M} \in \mathcal{M}^+$.

H.2. Proof of Regret Lower Bound (Theorem C.1)

In this section, we prove Theorem C.1. The proof proceeds in two parts:

- In Appendix H.2.1, we state and prove Lemma H.1, a lower bound which is similar to Theorem C.1, but restricts to proper reference models (specifically, the lower bound scales with $\sup_{\overline{M} \in \mathcal{M}} \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}, \overline{M})$).
- In Appendix H.2.2, we prove the following algorithmic result (Lemma H.2): For any class $\mathcal{M}$ and any $\overline{M} \in \mathcal{M}^+$ (not necessarily in $\mathcal{M}$), if there is an algorithm that achieves regret at most R with respect to the model class $\mathcal{M}$, then there is an algorithm that achieves regret at most $O(R \cdot \log T)$ with respect to the model class $\mathcal{M} \cup \{\overline{M}\}$. We then prove Theorem C.1 by combining this result with Lemma H.1.

We mention in passing that the two-part approach in this section can also be applied to derive lower bounds for the PAC framework, but we adopt the alternative approach in Appendix H.1 because it leads to a result with fewer logarithmic factors.

H.2.1. LOWER BOUND FOR PROPER REFERENCE MODELS ($\overline{M} \in \mathcal{M}$)

In this section we prove Lemma H.1, a weaker lower bound analogous to the one stated in Theorem C.1, but with the DEC replaced by a smaller quantity constrained to have $\overline{M} \in \mathcal{M}$. This weaker version is shown below.

Lemma H.1. *Let $\widetilde{\varepsilon}(T) := c_1 \cdot \frac{1}{\sqrt{TC(T)}}$, where $c_1 > 0$ is a sufficiently small numerical constant. For all $T \in \mathbb{N}$ such that the condition*

$$\sup_{\overline{M} \in \mathcal{M}} \textsf{r-dec}^{\mathrm{c}}_{\widetilde{\varepsilon}(T)}(\mathcal{M}, \overline{M}) \geq 8 \cdot \widetilde{\varepsilon}(T) \tag{61}$$

is satisfied, we have that for any regret minimization algorithm, there exists a model in $\mathcal{M}$ such that, under this model,

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \Omega(T) \cdot \sup_{\overline{M} \in \mathcal{M}} \textsf{r-dec}^{\mathrm{c}}_{\widetilde{\varepsilon}(T)}(\mathcal{M}, \overline{M}). \tag{62}$$

We remark that the condition (61) can be relaxed by replacing the constant 8 on the right-hand side with any constant strictly greater than 1.

Proof of Lemma H.1. Let the algorithm under consideration be fixed, and let $\mathbb{P}^M$ denote the induced law of $\mathfrak{H}^T$ when M is the underlying model. Let $\mathbb{E}^M$ denote the corresponding expectation, and let $p_M := \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^T p^t\right]$. Define $\varepsilon := \widetilde{\varepsilon}(T) = \frac{c_1}{\sqrt{TC(T)}}$, where the constant $c_1 > 0$ will be specified below. Let $\overline{M} \in \mathcal{M}$ be chosen to maximize $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, and define $\delta := \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$.

Restricting to models performing poorly on $\overline{M}$. If $\mathbb{E}_{\pi \sim p_{\overline{M}}}[g^{\overline{M}}(\pi)] \geq \delta/10$, then, by the definition of $p_{\overline{M}}$, we have $\mathbb{E}^{\overline{M}}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq T \cdot \delta/10$, completing the proof of the lemma. Hence, we may assume going forward that $\mathbb{E}_{\pi \sim p_{\overline{M}}}[g^{\overline{M}}(\pi)] < \delta/10$.

Choosing an alternative model. Define

$$M = \operatorname*{arg\,max}_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2 \right\}, \tag{63}$$

so that

$$\mathbb{E}_{\pi \sim p_{\overline{M}}}[g^M(\pi)] \geq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) = \delta. \tag{64}$$

Let $c_2 \in (0, 1)$ be fixed and define $\mathcal{E} = \{\pi : g^M(\pi) \geq c_2 \cdot \delta\}$. Recall that by Lemma A.13 of Foster et al. (2021), we have

$$D^2_{\mathsf{H}}\left(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\right) \leq C(T) \cdot T \cdot \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq C(T) \cdot T \cdot \varepsilon^2,$$

where we remind the reader that $C(T) = O(\log(T \wedge V(\mathcal{M})))$. We choose the constant $c_1 > 0$ in the definition of $\varepsilon = \widetilde{\varepsilon}(T)$ to be sufficiently small so that $D^2_{\mathsf{H}}\big(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\big) \leq 1/100$ and thus $D_{\mathsf{TV}}\big(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\big) \leq \frac{1}{10}$.

Observe that from the definition of $\mathcal{E}$, we have

$$\begin{aligned} \mathbb{E}_{\pi \sim p_M}[f^M(\pi_M) - f^M(\pi)] &\geq c_2\delta \cdot p_M(\mathcal{E}) \geq c_2\delta \cdot \left(p_{\overline{M}}(\mathcal{E}) - D_{\mathsf{TV}}\left(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\right)\right) \\ &\geq c_2\delta \cdot (p_{\overline{M}}(\mathcal{E}) - 1/10). \end{aligned} \tag{65}$$

Therefore, it suffices to lower bound $p_{\overline{M}}(\mathcal{E})$ by $1/2$.

Lower bounding the gap. We now compute

$$
\begin{aligned}
f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) &\geq \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^{M}(\pi) - g^{\overline{M}}(\pi)] - \mathbb{E}_{\pi\sim p_{\overline{M}}}[|f^{M}(\pi) - f^{\overline{M}}(\pi)|] \\
&\geq \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^{M}(\pi)] - \frac{\delta}{10} - \varepsilon \\
&\geq \frac{9}{10}\delta - \varepsilon, \qquad (66)
\end{aligned}
$$

where the second inequality uses the assumption that $\mathbb{E}_{p_{\overline{M}}}[g^{\overline{M}}] < \delta/10$ and Lemma J.1, and the final inequality uses Eq. (64).

Finishing up. We conclude by noting that

$$
\begin{aligned}
p_{\overline{M}}(\mathcal{E}^{\mathrm{c}}) \cdot \left(\frac{9}{10}\delta - \varepsilon\right) &\leq \mathbb{E}_{\pi\sim p_{\overline{M}}}[\mathbb{1}\left\{\mathcal{E}^{\mathrm{c}}\right\} \cdot (f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}))] \\
&\leq \mathbb{E}_{\pi\sim p_{\overline{M}}}[\mathbb{1}\left\{\mathcal{E}^{\mathrm{c}}\right\} \cdot (g^{M}(\pi) - g^{\overline{M}}(\pi))] + \varepsilon \\
&\leq c_2\delta + \varepsilon,
\end{aligned}
$$

where the first inequality uses Eq. (66), the second inequality uses Lemma J.1, and the third inequality uses that $g^{\overline{M}}(\pi) \geq 0$ for all $\pi \in \Pi$, as well as the fact that $g^{M}(\pi) < c_2\delta$ for $\pi \in \mathcal{E}^{\mathrm{c}}$. Rearranging, we conclude that $p_{\overline{M}}(\mathcal{E}_1^{\mathrm{c}}) \leq 1/2$ as long as $c_2 \leq 1/8$ and $\varepsilon \leq \delta/8$. Our choice of $\overline{M}$, together with the growth condition (61), ensures that we indeed have $\varepsilon \leq \delta/8$, thus establishing via Eq. (65), that $\mathbb{E}_{\pi\sim p_M}[g^{M}(\pi)] \geq \Omega(\delta)$ as desired. □

H.2.2. REDUCING FROM IMPROPER ($\overline{M} \in \mathcal{M}^{+}$) TO PROPER ($\overline{M} \in \mathcal{M}$)

In this section, we work with several choices for the model class and regret minimization algorithm. To avoid ambiguity, let us introduce some additional notation. Recall that an algorithm for the T-timestep interactive decision making problem (in the regret framework) is specified by a sequence $p = (p^1, \ldots, p^T)$, where for each $t \in [T]$, p^t is a probability kernel from $(\Omega^{t-1}, \mathscr{F}^{t-1})$ to $(\Pi, \mathscr{P})$. Given an algorithm p, we let $\mathbb{P}^{M,p}[\cdot]$ denote the law it induces on $\mathfrak{H}^T$ when $M \in \mathcal{M}^{+}$ is the underlying model, and let $\mathbb{E}^{M,p}[\cdot]$ denote the corresponding expectation. With this notation, the algorithm's expected regret when the underlying model is $M \in \mathcal{M}^{+}$ is $\mathbb{E}^{M,p}[\mathbf{Reg}_{\mathsf{DM}}(T)]$.

The following lemma is the main technical result of this section. It shows that any model $\overline{M} \in \mathcal{M}$ with bounded optimal value can be added to a model class $\mathcal{M}$ without substantially increasing the minimax regret.

Lemma H.2. *Let the time $T \in \mathbb{N}$ and model class $\mathcal{M}$ be fixed. Let $\overline{M} \in \mathcal{M}^{+}$ be any model such that for all $M \in \mathcal{M}$, $f^{\overline{M}}(\pi_{\overline{M}}) \leq f^{M}(\pi_{M}) + \delta$ for some $\delta > 0$. The minimax regret for the model class $\mathcal{M} \cup \{\overline{M}\}$ is bounded above as follows:*

$$
\inf_{(p')^1,\ldots,(p')^T} \sup_{M^\star \in \mathcal{M}\cup\{\overline{M}\}} \mathbb{E}^{M^\star,p'}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq C \log T \cdot \inf_{p^1,\ldots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + C \cdot (\sqrt{T} + \delta T). \qquad (67)
$$

where $C > 0$ denotes a universal constant.

Before proving Lemma H.2, we show how it implies Theorem C.1.

Proof of Theorem C.1. Fix $T \in \mathbb{N}$, and write $\varepsilon = \underline{\varepsilon}(T) = \frac{c_1}{\sqrt{2TC(T)}}$, where the constant $c_1 > 0$ is chosen as in Lemma H.1 (in particular, note that $\sqrt{2}\varepsilon = \widetilde{\varepsilon}(T)$, where $\widetilde{\varepsilon}(T)$ is as defined in Lemma H.1). Let $\overline{M} \in \mathcal{M}^+$ be chosen to maximize $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})$, so that $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})$. Define

$$\widetilde{\mathcal{M}} := \{M \in \mathcal{M} \cup \{\overline{M}\} \mid f^M(\pi_{\overline{M}}) \geq f^{\overline{M}}(\pi_{\overline{M}}) - \sqrt{2}\varepsilon\},$$

so that $\overline{M} \in \widetilde{\mathcal{M}}$. Then by Lemma J.2, we have

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}) + \sqrt{2}\varepsilon, \tag{68}$$

Note that for all $M \in \widetilde{\mathcal{M}}$, we have that $f^{\overline{M}}(\pi_{\overline{M}}) \leq f^M(\pi_{\overline{M}}) + \sqrt{2}\varepsilon \leq f^M(\pi_M) + \sqrt{2}\varepsilon$. Thus, by applying Lemma H.2 with $\delta = \sqrt{2}\varepsilon$ to the class $\widetilde{\mathcal{M}}\backslash\{\overline{M}\}$, we see that

$$\begin{aligned}\inf_{(p')^1,\dots,(p')^T} \sup_{M^\star \in \widetilde{\mathcal{M}}} \mathbb{E}^{M^\star,p'}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq& C\log T \cdot \inf_{p^1,\dots,p^T} \sup_{M^\star \in \widetilde{\mathcal{M}}\backslash\{\overline{M}\}} \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + C\cdot(\sqrt{T} + \sqrt{2}\varepsilon T)\\ \leq& C\log T \cdot \inf_{p^1,\dots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + C\cdot(\sqrt{T} + \sqrt{2}\varepsilon T),\end{aligned} \tag{69}$$

where the second inequality follows since $\widetilde{\mathcal{M}}\backslash\{\overline{M}\} \subset \mathcal{M}$.

To proceed, we will apply Lemma H.1 to the class $\widetilde{\mathcal{M}}$. To verify that the condition Eq. (61) is satisfied for this class, we note that, as remarked above, $\widetilde{\varepsilon}(T) = \sqrt{2}\varepsilon$, so that

$$\sup_{\overline{M}_0 \in \widetilde{\mathcal{M}}} \textsf{r-dec}^{\mathrm{c}}_{\widetilde{\varepsilon}(T)}(\widetilde{\mathcal{M}}, \overline{M}_0) = \sup_{\overline{M}_0 \in \widetilde{\mathcal{M}}} \textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}_0) \geq \textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}) \geq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) - \sqrt{2}\varepsilon \geq 8\varepsilon,$$

where the second-to-last inequality uses Eq. (68) and the final inequality uses the assumption from Eq. (15) that $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \geq 10\varepsilon$. Lemma H.1 gives that, for some universal constant $c_2 > 0$,

$$\begin{aligned}\inf_{(p')^1,\dots,(p')^T} \sup_{M^\star \in \widetilde{\mathcal{M}}} \mathbb{E}^{M^\star,p'}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq& c_2 \cdot T \cdot \sup_{\overline{M}_0 \in \widetilde{\mathcal{M}}} \textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}_0)\\ \geq& c_2 \cdot T \cdot \textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}) \geq c_2 \cdot T \cdot (\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) - \sqrt{2}\varepsilon),\end{aligned} \tag{70}$$

where the final inequality uses Eq. (68). Combining Eq. (70) and Eq. (69) gives that

$$\inf_{p^1,\dots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \frac{1}{C\log T} \cdot \left(c_2 \cdot T \cdot (\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) - \sqrt{2}\varepsilon) - C\cdot(\sqrt{T} + \sqrt{2}\varepsilon T)\right),$$

which implies that for some constants $C', C'', c_2' > 0$, we have

$$\inf_{p^1,\dots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \frac{1}{\log T} \cdot \left(c_2' \cdot T \cdot (\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) - C' \cdot \varepsilon) - C'' \cdot \sqrt{T}\right).$$

As long as $\frac{1}{2} \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) > C' \cdot \varepsilon$, it follows that

$$\inf_{p^1,\dots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \frac{c_2' \cdot T}{2\log T} \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) - \frac{C'' \cdot \sqrt{T}}{\log T},$$

as desired. □

Finally, we prove Lemma H.2.

Proof of Lemma H.2. Fix any algorithm $p = (p^1, \ldots, p^T)$. We define a modified algorithm $p' = ((p')^1, \ldots, (p')^T)$ in Algorithm 3. Roughly speaking, p' runs p multiple times, re-initializing p whenever the average reward for the current run falls too far below $f^{\overline{M}}(\pi_{\overline{M}})$. If the algorithm p' finds that it has re-initialized p more than $\log(T)$ times, it will switch to playing $\pi_{\overline{M}}$ for all remaining rounds. The crux of the proof will be to show that the worst-case regret of p' for models in $\mathcal{M} \cup \{\overline{M}\}$ is not much larger than the worst-case regret of p for models in $\mathcal{M}$.

Algorithm 3 Algorithm p' used in proof of Lemma H.2

```
1:  parameters:
        Number of rounds $T \in \mathbb{N}$.
        Algorithm $p = (p^1, \ldots, p^T)$.
2:  Initialize $I = 1$, $T_1 = 1$, and $R = 4 \cdot \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + \delta T + 8\sqrt{T}$.
3:  for $1 \leq t \leq T$ do
4:      Define $(p')^t(\cdot)$ to be the distribution $p^{t-T_I+1}(\cdot \mid \{(\pi^s, r^s, o^s)\}_{s=T_I}^{t-1})$.
5:      Draw $\pi^t \sim (p')^t$, and observe $(\pi^t, r^t, o^t)$.
6:      if $\sum_{s=T_I}^{t} (f^{\overline{M}}(\pi_{\overline{M}}) - r^s) \geq R$ then
7:          Set $T_{I+1} := t + 1$ and then increment $I$.   // This has the effect of re-initializing $p$.
8:      if $I > \lceil \log T \rceil$ then
9:          break out of loop.
10: For remaining time steps $t$ (if any): play $\pi^t := \pi_{\overline{M}}$ (i.e., set $(p')^t = \mathbb{I}_{\pi_{\overline{M}}}$).
```

As per our convention, in the context of Algorithm 3, we let $\mathscr{F}^t$ denote the sigma-algebra generated by $\{(\pi^s, r^s, o^s)\}_{1 \leq s \leq t}$. We bound the regret of the algorithm p' by considering the following cases for $M^\star \in \mathcal{M} \cup \{\overline{M}\}$.

Case 1: $M^\star = \overline{M}$**.** Let $T_0 \in [T+1]$ be defined to be the smallest value of t for which the decision π^t is chosen using the rule at Line 10, or $T + 1$ if there is no such step. By construction, we have that

$$\sum_{t=1}^{T_0 - 1} (f^{\overline{M}}(\pi_{\overline{M}}) - r^t) = \sum_{t=1}^{T} \mathbb{1}\{t < T_0\} \cdot (f^{\overline{M}}(\pi_{\overline{M}}) - r^t) \leq (R+1) \cdot \lceil \log T \rceil. \tag{71}$$

Note that, for each $t \in [T]$, the variable $\mathbb{1}\{t < T_0\}$ is measurable with respect to $\mathscr{F}^{t-1}$. As a result, we have

$$\begin{aligned}
\mathbb{E}^{\overline{M},p'}[\mathbf{Reg}_{\mathsf{DM}}(T)] &= \mathbb{E}^{\overline{M},p'}\left[\sum_{t=1}^{T}\left(f^{\overline{M}}(\pi_{\overline{M}}) - \mathbb{E}_{\pi^t\sim(p')^t}[f^{\overline{M}}(\pi^t)]\right)\right] \\
&= \mathbb{E}^{\overline{M},p'}\left[\sum_{t=1}^{T}\mathbb{1}\{t < T_0\}\cdot\left(f^{\overline{M}}(\pi_{\overline{M}}) - \mathbb{E}_{\pi^t\sim(p')^t}[f^{\overline{M}}(\pi^t)]\right)\right] \\
&= \mathbb{E}^{\overline{M},p'}\left[\sum_{t=1}^{T}\mathbb{1}\{t < T_0\}\cdot(f^{\overline{M}}(\pi_{\overline{M}}) - r^t)\right] + \mathbb{E}^{\overline{M},p'}\left[\sum_{t=1}^{T}\mathbb{1}\{t < T_0\}\cdot\left(r^t - \mathbb{E}_{\pi^t\sim(p')^t}[f^{\overline{M}}(\pi^t)]\right)\right] \\
&\leq (R+1)\cdot\log T,
\end{aligned}$$

where the final inequality uses Eq. (71) and the fact that for each t, we have

$$\mathbb{E}\left[\mathbb{1}\{t < T_0\}\cdot\left(r^t - \mathbb{E}_{\pi^t\sim(p')^t}[f^{\overline{M}}(\pi^t)]\right) \mid \mathscr{F}^{t-1}\right] = \mathbb{1}\{t < T_0\}\cdot\mathbb{E}\Big[r^t - \mathbb{E}_{\pi^t\sim(p')^t}\left[f^{\overline{M}}(\pi^t)\right] \mid \mathscr{F}^{t-1}\Big] = 0.$$

Thus, in the case $M^\star = \overline{M}$, we have verified that the claimed upper bound in Eq. (67) on the regret of p' holds.

Case 2: $M^\star \in \mathcal{M}$. We first state and prove two technical lemmas.

Lemma H.3. *For the algorithm p, any model $M^\star \in \mathcal{M}^+$, and random variable τ (potentially dependent on $\mathfrak{H}^T$) taking values in $[T]$, it holds that*

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

Proof of Lemma H.3. The result follows by noting that

$$\begin{aligned}
&\mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] - \mathbb{E}^{M^\star,p}\left[\sum_{t=1}^{\tau}\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\
&= \mathbb{E}^{M^\star,p}\left[\sum_{t=1}^{T}\mathbb{1}\{\tau < t\}\cdot\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \geq 0,
\end{aligned}$$

where we have used that the random variable $\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]$ is non-negative a.s. □

The next lemma concerns the probability that a *single run* of the algorithm p violates the condition in Line 6 of Algorithm 3.

Lemma H.4. *For any algorithm $p = (p^1, \ldots, p^T)$ and model $M^\star \in \mathcal{M}$, it holds that*

$$\mathbb{P}^{M^\star,p}\left(\exists t \leq T \ :\ \sum_{s=1}^{t}(f^{\overline{M}}(\pi_{\overline{M}}) - r^s) > R\right) \leq \frac{1}{2},$$

where $R = 4\cdot\sup_{M^\star\in\mathcal{M}}\mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + \delta T + 8\sqrt{T}$.

Proof of Lemma H.4. Let $\mathscr{F}^t$ be the sigma-algebra generated by $\{(\pi^s, r^s, o^s)\}_{s=1}^t$. Fix $M^\star \in \mathcal{M}$ and define $R_0 := \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)]$. By Markov's inequality and the fact that the random variables $\mathbb{E}_{\pi^t \sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]$ are all non-negative, it holds that

$$\mathbb{P}^{M^\star,p}\left(\sup_{t\le T}\sum_{s=1}^{t}\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^s)] > 4R_0\right)$$
$$= \mathbb{P}^{M^\star,p}\left(\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)] > 4R_0\right) \le \frac{1}{4}. \tag{72}$$

Now, define $X_0 = 0$ and $X_t = \sum_{s=1}^t \left(\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi^s)] - r^s\right)$ for $t \in [T]$. Note that $(X_t)_{t\ge 0}$ is a martingale with respect to the filtration $\mathscr{F}^t$. Therefore, by Theorem 4.5.1 of Durrett (2019), it holds that

$$\mathbb{E}^{M^\star,p}\left[\sup_{t\le T}|X_t|^2\right] \le 4\cdot\mathbb{E}^{M^\star,p}\left[\sum_{t=1}^{T}\mathbb{E}\left[(\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi^t)] - r^t)^2 \mid \mathscr{F}^{t-1}\right]\right] \le 4T,$$

where the final inequality uses that $|\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi^t)] - r^t| \le 1$ for all t. By Jensen's inequality and Markov's inequality, it follows that for any $\lambda > 0$,

$$\mathbb{P}^{M^\star,p}\left(\sup_{t\le T}|X_t| > 2\lambda\sqrt{T}\right) \le \frac{1}{\lambda},$$

and by choosing $\lambda = 4$, we see that

$$\mathbb{P}^{M^\star,p}\left(\sup_{t\le T}\sum_{s=1}^{t}\left(\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi^s)] - r^s\right) > 8\sqrt{T}\right) \le \frac{1}{4}. \tag{73}$$

Combining Eq. (72) and Eq. (73), we have

$$\mathbb{P}^{M^\star,p}\left(\exists t\le T\ :\ \sum_{s=1}^{t}\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi_{M^\star}) - r^s] > 4R_0 + 8\sqrt{T}\right) \le \frac{1}{2}.$$

Since $M^\star \in \mathcal{M}$, and so $f^{M^\star}(\pi_{M^\star}) \ge f^{\overline{M}}(\pi_{\overline{M}}) - \delta$, it follows that

$$\mathbb{P}^{M^\star,p}\left(\exists t\le T\ :\ \sum_{s=1}^{t}\mathbb{E}_{\pi^s\sim p^s}[f^{\overline{M}}(\pi_{\overline{M}}) - r^s] > 4R_0 + 8\sqrt{T} + \delta T\right) \le \frac{1}{2},$$

as desired. □

We now continue with the proof of Lemma H.2. Write $L = \lceil \log T \rceil$, and denote the final value of I in Algorithm 3 by $I' \le L+1$. If $I' \le L$, then set $T_{I'+1} = \cdots = T_{L+1} = T+1$. Note that for each $\ell \in [L+1]$, $T_\ell - 1$ is a stopping time (since the event $\{T_\ell - 1 = t\}$ is measurable with respect to $\mathscr{F}^t$ for each $t \in [T]$), and thus T_ℓ is a stopping time as well.

Lemma H.4 together with the definition of $(p')^t$ in Line 4 establishes that for each $\ell \in [L]$,

$$\mathbb{P}^{M^\star,p'}(T_{\ell+1} \le T \mid T_\ell \le T) = \mathbb{P}^{M^\star,p'}\left(\exists t \text{ s.t. } T-1 \ge t \ge T_\ell \text{ and } \sum_{s=T_\ell}^{t}(f^{\overline{M}}(\pi_{\overline{M}}) - r^s) > R \mid T_\ell \le T\right) \le \frac{1}{2}.$$

Therefore, since the event $\{T_\ell \leq T\}$ is equal to the event $\{T_{\ell'} \leq T \,\forall \ell' \leq \ell\}$,

$$\begin{aligned}\mathbb{P}^{M^\star,p'}\left(T_{L+1} \leq T\right) &= \mathbb{P}^{M^\star,p'}\left(\forall \ell \leq L+1,\ T_\ell \leq T\right) \\ &= \prod_{\ell=1}^{L} \mathbb{P}^{M^\star,p'}\left(T_{\ell+1} \leq T \mid T_{\ell'} \leq T \,\forall \ell' \leq \ell\right) \\ &= \prod_{\ell=1}^{L} \mathbb{P}^{M^\star,p'}\left(T_{\ell+1} \leq T \mid T_\ell \leq T\right) \\ &\leq (1/2)^L \leq 1/T.\end{aligned}$$

Furthermore, for each $\ell \in [L]$, we have

$$\begin{aligned}&\mathbb{E}^{M^\star,p'}\left[\sum_{t=T_\ell}^{T_{\ell+1}-1} \mathbb{E}_{\pi^t \sim (p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \qquad (74)\\ &= \mathbb{E}^{M^\star,p'}\left[\sum_{t=1}^{T} \mathbb{1}\left\{T_\ell \leq t < T_{\ell+1}\right\} \cdot \mathbb{E}_{\pi^t \sim (p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\ &\leq \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)],\end{aligned}$$

where the inequality uses Lemma H.3 and the definition of p' in Line 4 for steps $T_\ell \leq t < T_{\ell+1}$. It follows that

$$\begin{aligned}&\mathbb{E}^{M^\star,p'}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim (p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\ &\leq T \cdot \mathbb{P}^{M^\star,p'}(T_{L+1} \leq T) + \sum_{\ell=1}^{L} \mathbb{E}^{M^\star,p'}\left[\sum_{t=T_\ell}^{T_{\ell+1}-1} \mathbb{E}_{\pi^t \sim (p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\ &\leq 1 + L \cdot \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)],\end{aligned}$$

which verifies the claimed upper bound on regret in Eq. (67). □

Appendix I. Proofs for Upper Bounds

In this section, we prove Theorem 2.2 and Theorem C.2. The upper bound for regret builds on the ideas used in the proof of the upper bound for PAC, but is somewhat more involved.

I.1. Proof of PAC Upper Bound (Theorem 2.2)

We now prove Theorem 2.2. Before proceeding with the proof, we give some brief background on the notion of online-to-batch conversion, which is used in the exploitation phase and its analysis.

Background: Online-to-batch conversion. As discussed in the prequel, we assume access to an online oracle $\mathbf{Alg}_{\mathsf{Est}}$. We use the *online* guarantee the oracle the algorithm provides—namely, that it ensures that the cumulative estimation error is bounded for an adaptively chosen sequence of decisions—in a non-trivial fashion during the exploration phase, but our analysis additionally makes use the fact that online oracles can be used to provide guarantees for *offline* estimation.

For offline estimation, we consider a setting in which there is some $p \in \Delta(\Pi)$ so that $p^t = p$ for all t, and the algorithm must output a single model estimate $\widehat{M}$ such that $\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(\widehat{M}(\pi), M^\star(\pi)\Big)\Big] \leq \varepsilon^2$. We can obtain such a guarantee using an online estimation oracle via the following online-to-batch conversion process:

- For each $t = 1, \dots, T$, obtain $\widehat{M}^t$ by running $\mathbf{Alg}_{\mathsf{Est}}$ on samples $\{(\pi^i, r^i, o^i)\}_{t=1}^{t-1}$, where $\pi^t \sim p$ and $(r^t, o^t) \sim M^\star(\pi^t)$.
- Let $\widehat{M} := \frac{1}{T}\sum_{t=1}^{T} \widehat{M}^t \in \mathrm{co}(\mathcal{M})$.

It is evident from Assumption 2.1 and convexity of the squared Hellinger distance that with probability at least $1-\delta$, the estimator $\widehat{M}$ constructed above satisfies

$$\begin{aligned}\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M^\star(\pi), \widehat{M}(\pi)\Big)\Big] &= \mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \frac{1}{T}\sum_{t=1}^{T}\widehat{M}^t(\pi)\right)\right] \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim p}\Big[D^2_{\mathsf{H}}\Big(M^\star(\pi), \widehat{M}^t(\pi)\Big)\Big] \leq \frac{\mathbf{Est}_{\mathsf{H}}(T,\delta)}{T}. \qquad (75)\end{aligned}$$

This is the strategy employed in Line 10 of Algorithm 1. Standard offline estimation algorithms (e.g., MLE) can be used to derive similar guarantees, but we make use of online-to-batch in order to keep notation light, since Algorithm 1 already requires the online estimation algorithm for the exploration phase.

Proof of Theorem 2.2. We begin by analyzing the exploitation phase. e Recall that we set $J := \frac{T}{\lceil \log 2/\delta \rceil + 1} \geq \frac{T}{2L}$. By Assumption 2.1, we have that with probability at least $1 - \frac{\delta}{4L}$,

$$\sum_{t=1}^{J}\mathbb{E}_{\pi^t\sim q^t}\Big[D^2_{\mathsf{H}}\Big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\Big)\Big] \leq \mathbf{Est}_{\mathsf{H}}\left(J, \frac{\delta}{4L}\right) \leq \overline{\mathbf{Est}}_{\mathsf{H}}.$$

We denote this event by $\mathscr{E}_0$, and condition on it going forward. Since we have $\bar{\varepsilon}(T)^2 \geq \frac{32}{J}\cdot\overline{\mathbf{Est}}_{\mathsf{H}}$ by definition, it follows from Markov's inequality that if $s \in [J]$ is chosen uniformly at random, then with probability at least $1/2$,

$$\mathbb{E}_{\pi^s\sim q^s}\Big[D^2_{\mathsf{H}}\Big(M^\star(\pi^s), \widehat{M}^s(\pi^s)\Big)\Big] \leq \frac{\bar{\varepsilon}(T)^2}{16}. \qquad (76)$$

Going forward, our aim is to show that the exploitation phase identifies such an index $s \in [J]$. Indeed, for any $s \in [J]$ such that the inequality (76) holds, we have $M^\star \in \mathcal{H}_{q^s,\bar{\varepsilon}(T)}(\widehat{M}^s)$, and hence

$$\mathbb{E}_{\pi\sim p^s}\Big[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\Big] \leq \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}, \widehat{M}^s) \leq \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})$$

from the expression (14) and the definition of the constrained DEC.

To proceed, first observe that for the uniformly sampled indices $t_1, \dots, t_L \in [J]$, a standard confidence boosting argument implies that with probability at least $1 - 2^{-L} \geq 1 - \frac{\delta}{2}$, there is some $\ell \in [L]$ so that Eq. (76) is satisfied with $s = t_\ell$. We denote this event by $\mathscr{F}$.

Next, recall the definition $\widetilde{M}_\ell = \frac{1}{J}\sum_{j=1}^{J} \widetilde{M}_\ell^j$ in Line 10 of Algorithm 1. Using Assumption 2.1 together with Eq. (75), we have that for each $\ell \in [L]$, there is an event that occurs with probability at least $1 - \frac{\delta}{4L}$, denoted by $\mathscr{E}_\ell$, such that under $\mathscr{E}_\ell$ we have

$$\mathbb{E}_{\pi \sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widetilde{M}_\ell(\pi)\right)\right] \leq \frac{\mathbf{Est}_{\mathsf{H}}(J, \delta/4L)}{J} \leq \frac{\overline{\mathbf{Est}}_{\mathsf{H}}}{J} \leq \frac{\bar{\varepsilon}(T)^2}{32},$$

where the second inequality uses our choice for $\bar{\varepsilon}(T)$. Define $\mathscr{E} := \bigcap_{\ell=0}^{L} \mathscr{E}_\ell$, so that $\mathscr{E}$ occurs with probability at least $1 - \frac{(L+1)\delta}{4L} \geq 1 - \frac{\delta}{2}$. We define $\mathscr{E} = \mathscr{F} \cap \bigcap_{\ell=0}^{L} \mathscr{E}_\ell$, so that $\mathscr{E}$ occurs with probability at least $1 - \frac{(L+1)\delta}{4L} - \frac{\delta}{2} \geq 1 - \delta$.

We now show that the exploitation phase succeeds whenever the event $\mathscr{E}$ holds. By the triangle inequality for Hellinger distance, letting $\ell \in [L]$ be any index such that Eq. (76) is satisfied with $s = t_\ell$, we have

$$\mathbb{E}_{\pi \sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\left(\widetilde{M}_\ell(\pi), \widehat{M}^{t_\ell}(\pi)\right)\right] \leq 2 \cdot \left(\mathbb{E}_{\pi \sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widetilde{M}_\ell(\pi)\right) + D^2_{\mathsf{H}}\left(M^\star(\pi), \widehat{M}^{t_\ell}(\pi)\right)\right]\right) \leq \frac{\bar{\varepsilon}(T)^2}{4}.$$

From the definition on Line 11, the index $\widehat{\ell} \in [L]$ satisfies

$$\mathbb{E}_{\pi \sim q^{t_{\widehat{\ell}}}}\left[D^2_{\mathsf{H}}\left(\widetilde{M}_{\widehat{\ell}}(\pi), \widehat{M}^{t_{\widehat{\ell}}}(\pi)\right)\right] \leq \mathbb{E}_{\pi \sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\left(\widetilde{M}_\ell(\pi), \widehat{M}^{t_\ell}(\pi)\right)\right] \leq \frac{\bar{\varepsilon}(T)^2}{4}.$$

Using the triangle inequality for Hellinger distance once more, we obtain that under the event $\mathscr{E}$,

$$\mathbb{E}_{\pi \sim q^{t_{\widehat{\ell}}}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{t_{\widehat{\ell}}}(\pi), M^\star(\pi)\right)\right] \leq 2 \cdot \left(\frac{\bar{\varepsilon}(T)^2}{4} + \frac{\bar{\varepsilon}(T)^2}{32}\right) < \bar{\varepsilon}(T)^2,$$

which means that $M^\star \in \mathcal{H}_{q^{t_{\widehat{\ell}}}, \bar{\varepsilon}(T)}(\widehat{M}^{t_{\widehat{\ell}}})$. It follows that whenever $\mathscr{E}$ holds,

$$\begin{aligned}
\mathbf{Risk}_{\mathsf{DM}}(T) = \mathbb{E}_{\pi \sim p^{t_{\widehat{\ell}}}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] &\leq \sup_{M \in \mathcal{H}_{q^{t_{\widehat{\ell}}}, \bar{\varepsilon}(T)}(\widehat{M}^{t_{\widehat{\ell}}})} \mathbb{E}_{\pi \sim p^{t_{\widehat{\ell}}}}[f^M(\pi_M) - f^M(\pi)] \\
&= \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{q, \bar{\varepsilon}(T)}(\widehat{M}^{t_{\widehat{\ell}}})} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \\
&= \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}}(\mathcal{M}, \widehat{M}^{t_{\widehat{\ell}}}) \\
&\leq \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}, \overline{M}) = \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}),
\end{aligned}$$

where the first equality follows from the choice of $p^{t_{\widehat{\ell}}}, q^{t_{\widehat{\ell}}}$ in Line 6. The bound on expected risk in the theorem statement follows from the observation that $\mathbf{Risk}_{\mathsf{DM}}(T)$ is bounded above by 1, as we assume that rewards lie in $[0, 1]$. □

I.2. Proof of Regret Upper Bound (Theorem C.2)

In this section we prove Theorem C.2, which shows that Algorithm 2 attains a regret bound based on the constrained DEC. Toward proving the result., we introduce a few *success events* that will be used throughout the analysis:

1. For each $i \in [N]$, $\mathscr{A}_i$ denotes the event that all of the following inequalities hold:
$$\forall s \in \mathcal{S}_i, \qquad \sum_{j=1}^{J_{i,s}} \mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widetilde{M}^j_s(\pi)\right)\right] \leq \mathbf{Est}_{\mathsf{H}}(J_i, \delta), \tag{77}$$
$$\mathbb{E}_{\pi\sim p^{s_i}}\left[D^2_{\mathsf{H}}\left(\widehat{M}_i(\pi), M^\star(\pi)\right)\right] \leq \frac{\mathbf{Est}_{\mathsf{H}}(J_i, \delta)}{J_i}, \tag{78}$$
$$\mathbb{E}_{\pi\sim p^{s_i}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{s_i}(\pi), M^\star(\pi)\right)\right] \leq \varepsilon_i^2, \tag{79}$$
where we recall that s_i is the index defined on Line 23.

2. For each $i \in [N]$, $\mathscr{B}_i$ denotes the event that
$$\sum_{t\in\mathcal{E}_i} \mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widehat{M}^t(\pi)\right)\right] \leq \mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|, \delta). \tag{80}$$

3. For each $i \in [N]$, $\mathscr{C}_i$ denotes the event that $M^\star \in \mathcal{M}_i$.

4. For each $i \in [N]$, $\mathscr{D}_i$ denotes the event that there is some $s \in \mathcal{S}_i$ so that
$$\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widehat{M}^s(\pi)\right)\right] \leq \frac{\varepsilon_i^2}{16}.$$

In addition, we define
$$\mathscr{A} = \bigcap_{i\in[N]} \mathscr{A}_i,\ \mathscr{B} = \bigcap_{i\in[N]} \mathscr{B}_i,\ \mathscr{C} = \bigcap_{i\in[N]} \mathscr{C}_i,\ \text{and}\ \mathscr{D} = \bigcap_{i\in[N]} \mathscr{D}_i.$$

We also recall the following notation, which will be used throughout the proof.

- For $i \in [N]$ we set
$$\alpha_i := C_0 \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) + 64\varepsilon_i, \tag{81}$$
with the convention that $\alpha_0 = 1$. The constant $C_0 > 0$ in Eq. (81), as well as the constant $C_1 > 0$ specified in Algorithm 2, will need to be taken sufficiently large; in what follows, we show that $C_0 \geq 20$ and $C_1 \geq 128$ will suffice.

I.2.1. TECHNICAL LEMMAS

Before proving Theorem C.2, we state and prove several technical lemmas concerning the performance of Algorithm 2. The following lemma shows that the event $\mathscr{A} \cap \mathscr{B} \cap \mathscr{C}$ occurs with high probability.

Lemma I.1. *Suppose that $C_1 \geq 128$. The event $\mathscr{A} \cap \mathscr{B} \cap \mathscr{C} \cap \mathscr{D}$ occurs with probability at least $1 - 3L\delta N$.*

Proof of Lemma I.1. We show that $\mathbb{P}\left(\bigcap_{i'\leq i} \mathscr{A}_{i'} \cap \mathscr{B}_{i'} \cap \mathscr{C}_{i'} \cap \mathscr{D}_{i'}\right) \geq 1 - 3L\delta i$ for each $i \in [N]$ using induction on i. Fix $i \in [N]$ and let us condition on $\bigcap_{i'<i}(\mathscr{A}_{i'} \cap \mathscr{B}_{i'} \cap \mathscr{C}_{i'} \cap \mathscr{D}_{i'})$.

Establishing that $\mathscr{C}_i$ holds. The fact that $\mathscr{A}_{i-1}$ holds implies that $\mathbb{E}_{\pi\sim\widehat{p}_{i-1}}\left[D^2_{\mathsf{H}}\left(M^\star(\pi),\widehat{M}_{i-1}(\pi)\right)\right]\le \frac{\mathbf{Est}_{\mathsf{H}}(J_{i-1},\delta)}{J_{i-1}}$, which implies (by definition) that $M^\star\in\mathcal{M}_i$, i.e., $\mathscr{C}_i$ holds.

Establishing that $\mathscr{B}_i$ holds. Conditioned on $\mathscr{C}_i$, it follows from Assumption C.1 that $\mathscr{B}_i$ holds with probability at least $1-\delta$.

Establishing that $\mathscr{D}_i$ holds. Note that as a consequence of our parameter settings,

$$\varepsilon_i^2 \ge 2^{-i}\cdot T\cdot \varepsilon_N^2 = 2^{-i}\cdot C_1\cdot \mathbf{Est}_{\mathsf{H}}(T,\delta)\cdot L \ge \max\left\{32\cdot\frac{\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}{J_i}, 32\cdot\frac{\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta)}{|\mathcal{E}_i|}\right\}, \tag{82}$$

where we have used that $C_1\ge 128$ and that $\mathbf{Est}_{\mathsf{H}}(T,\delta)\ge\max\{\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta),\mathbf{Est}_{\mathsf{H}}(J_i,\delta)\}$. Then since $\mathscr{B}_i$ (i.e., (80)) holds, at least $|\mathcal{E}_i|/2$ rounds $t\in\mathcal{E}_i$ satisfy $M^\star\in\mathcal{H}_{\varepsilon_i/4,p^t}(\widehat{M}^t)$. Since $|\mathcal{S}_i|=L\ge\log 1/\delta$, it follows that with probability at least $1-\delta$, there is some $s\in\mathcal{S}_i$, which we denote by $s_i^\star$, for which $M^\star\in\mathcal{H}_{\varepsilon_i/4,p^{s_i^\star}}(\widehat{M}^{s_i^\star})$. In particular, conditioned on $\mathscr{B}_i$, the event $\mathscr{D}_i$ holds with probability at least $1-\delta$.

Establishing that $\mathscr{A}_i$ holds. Next, from Assumption C.1 and the fact that Hellinger distance is always non-negative, we have that with probability at least $1-L\delta$, for all $s\in\mathcal{S}_i$,

$$\sum_{j=1}^{J_{i,s}}\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(M^\star(\pi),\widetilde{M}^j_s(\pi)\right)\right]\le\mathbf{Est}_{\mathsf{H}}(J_i,\delta), \tag{83}$$

which verifies that (77) holds. Next, let us condition on the event that $\mathscr{D}_i$ holds. Then applying (83) to $s=s_i^\star$ and using the definition of $s_i^\star$ (recall that $s_i^\star$ is defined in the prequel so that $M^\star\in\mathcal{H}_{\varepsilon_i/4,p^{s_i^\star}}(\widehat{M}^{s_i^\star})$), we see that

$$\begin{aligned}\sum_{j=1}^{J_{i,s_i^\star}}\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{s_i^\star}(\pi),\widetilde{M}^j_{s_i^\star}(\pi)\right)\right]\le&\sum_{j=1}^{J_{i,s_i^\star}}2\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{s_i^\star}(\pi),M^\star(\pi)\right)\right]+2\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D^2_{\mathsf{H}}\left(M^\star(\pi),\widetilde{M}^j_{s_i^\star}(\pi)\right)\right]\\ \le&2J_{i,s_i^\star}\cdot\frac{\varepsilon_i^2}{16}+2\sum_{j=1}^{J_{i,s_i^\star}}\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D^2_{\mathsf{H}}\left(M^\star(\pi),\widetilde{M}^j_{s_i^\star}(\pi)\right)\right]\\ \le&\frac{J_i\varepsilon_i^2}{8}+2\cdot\mathbf{Est}_{\mathsf{H}}(J_i,\delta)\le\frac{3J_i\varepsilon_i^2}{16},\end{aligned}$$

where the final inequality uses Eq. (82). By the definition of $J_{i,s_i^\star}$ on Lines 20 and 22, it must be the case that $J_{i,s_i^\star}=J_i$, and thus s_i^{tmp} is assigned at least once on Line 22. Therefore, the value of s_i set on Line 23 satisfies

$$\mathbb{E}_{\pi\sim p^{s_i}}\left[D^2_{\mathsf{H}}\left(\widehat{M}_i(\pi),M^\star(\pi)\right)\right]\le\frac{1}{J_i}\sum_{j=1}^{J_i}\mathbb{E}_{\pi\sim p^{s_i}}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_{s_i}(\pi),M^\star(\pi)\right)\right]\le\frac{\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}{J_i},$$

where the first inequality uses convexity of the squared Hellinger distance and the second inequality uses (83) together with the fact that $J_{i,s_i}=J_i$. The above display verifies (78); to verify (79), we note that, since s_i^{tmp} is assigned at least once,

$$\sum_{j=1}^{J_i} \mathbb{E}_{\pi\sim p^{s_i}} \left[D^2_{\mathsf{H}}\left(\widehat{M}^{s_i}(\pi), \widetilde{M}^j_{s_i}(\pi) \right) \right] \le \frac{J_i \varepsilon_i^2}{4}. \tag{84}$$

Thus, we may compute

$$\begin{aligned}\mathbb{E}_{\pi\sim p^{s_i}} \left[D^2_{\mathsf{H}}\left(\widehat{M}^{s_i}(\pi), M^\star(\pi) \right) \right] \le & 2 \cdot \mathbb{E}_{\pi\sim p^{s_i}} \left[D^2_{\mathsf{H}}\left(\widehat{M}_i(\pi), \widehat{M}^{s_i}(\pi) \right) \right] + 2 \cdot \mathbb{E}_{\pi\sim p^{s_i}} \left[D^2_{\mathsf{H}}\left(\widehat{M}_i(\pi), M^\star(\pi) \right) \right] \\ \le & \frac{2}{J_i} \sum_{j=1}^{J_i} \mathbb{E}_{\pi\sim p^{s_i}} \left[D^2_{\mathsf{H}}\left(\widehat{M}^{s_i}(\pi), \widetilde{M}^j_{s_i}(\pi) \right) \right] + \frac{2}{J_i} \sum_{j=1}^{J_i} \mathbb{E}_{\pi\sim p^{s_i}} \left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widetilde{M}^j_{s_i}(\pi) \right) \right] \\ \le & \frac{\varepsilon_i^2}{2} + \frac{2 \cdot \mathbf{Est}_{\mathsf{H}}(J_i, \delta)}{J_i} \le \varepsilon_i^2, \end{aligned}$$

where the second inequality uses the convexity of squared Hellinger distance, the third inequality uses (83) for $s = s_i$ and (84), and the final inequality uses (82). As the above display verifies (79), we conclude that conditioned on $\mathscr{D}_i$ holding, $\mathscr{A}_i$ holds with probability at least $1 - L\delta$.

Wrapping up. Summarizing, conditioned on $\bigcap_{i'<i} \mathscr{A}_{i'} \cap \mathscr{B}_{i'} \cap \mathscr{C}_{i'} \cap \mathscr{D}_{i'}$, we have shown that $\mathscr{A}_i \cap \mathscr{B}_i \cap \mathscr{C}_i \cap \mathscr{D}_i$ holds with probability $1 - 2\delta - L\delta \ge 1 - 3L\delta$. Thus, the inductive hypothesis that $\mathbb{P}\left(\bigcap_{i'<i} \mathscr{A}_{i'} \cap \mathscr{B}_{i'} \cap \mathscr{C}_{i'}\right) \ge 1 - 3L\delta(i-1)$ implies that $\mathbb{P}\left(\bigcap_{i'\le i} \mathscr{A}_{i'} \cap \mathscr{B}_{i'} \cap \mathscr{C}_{i'} \cap \mathscr{D}_{i'}\right) \ge (1 - 3L\delta(i-1)) \cdot (1 - 3L\delta) \ge 1 - 3L\delta i$.

Summarizing, we get that $\mathbb{P}(\mathscr{A} \cap \mathscr{B} \cap \mathscr{C} \cap \mathscr{D}) \ge 1 - 3L\delta N$. □

Lemma I.2 shows that the distributions $\widehat{p}_i$ computed in Algorithm 2 enjoy low suboptimality with respect to $M^\star$.

Lemma I.2 (Accuracy of refined policies). *Suppose that $C_0 \ge 4$. Then for each epoch $i \in [N]$, under the event $\mathscr{A}_i$, the distribution $\widehat{p}_i$ satisfies*

$$\mathbb{E}_{\pi\sim\widehat{p}_i} \left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi) \right] \le \frac{\alpha_i}{4}. \tag{85}$$

Proof of Lemma I.2. Conditioning on the event $\mathscr{A}_i$ gives that (79) holds, which can in particular be written as $M^\star \in \mathcal{H}_{\varepsilon_i, p^{s_i}}(\widehat{M}^{s_i})$. Therefore, by the choice of $\widehat{p}_i = p^{s_i}$ in Line 23 and the definition in Line 10, under $\mathscr{A}_i$,

$$\begin{aligned}\mathbb{E}_{\pi\sim p^{s_i}} \left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi) \right] \le & \sup_{M \in \mathcal{H}_{\varepsilon_i, p^{s_i}}(\widehat{M}^{s_i}) \cup \{\widehat{M}^{s_i}\}} \mathbb{E}_{\pi\sim p^{s_i}} \left[f^M(\pi_M) - f^M(\pi) \right] \\ = & \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M} \cup \{\widehat{M}^{s_i}\}, \widehat{M}^{s_i}) \le \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le \frac{\alpha_i}{4}, \end{aligned}$$

where the final inequality follows as long as $C_0 \ge 4$. □

Lemma I.3 relates the suboptimality under $M^\star$ for any distribution $p \in \Delta(\Pi)$ to that of any model $M \in \mathrm{co}(\mathcal{M}_{i+1})$, in terms of the distance between M and $M^\star$. We ultimately apply the lemma with $M = \widehat{M}^t$ for each $t \in \mathcal{E}_{i+1}$ to derive the following lemma, Lemma I.4.

Lemma I.3 (Comparison with models in refined class)**.** *Fix $i \in [N]$. Then for all $M \in \mathrm{co}(\mathcal{M}_{i+1})$ and all $p \in \Delta(\Pi)$, under the event $\mathscr{A}_i$,*

$$\mathbb{E}_{\pi\sim p}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \le \mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi)\right] + \frac{\alpha_i}{2} + \sqrt{\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}(M(\pi), M^\star(\pi))\right]}. \tag{86}$$

Proof of Lemma I.3. We first upper bound the optimal value under $M^\star$ by the optimal value under any $M \in \mathrm{co}(\mathcal{M}_{i+1})$. To do so, first note that, by Lemma I.2 (in particular, the fact that Eq. (85) holds at epoch i), we have that $\mathbb{E}_{\pi\sim\widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \le \frac{\alpha_i}{4}$ under $\mathscr{A}_i$. Then, for any $M \in \mathcal{M}_{i+1}$, we have that, under the event $\mathscr{A}_i$,

$$\begin{aligned}
\mathbb{E}_{\pi\sim\widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M}(\pi)\right] &\le \mathbb{E}_{\pi\sim\widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] + \sqrt{\mathbb{E}_{\pi\sim\widehat{p}_i}\left[D^2_{\mathsf{H}}(M(\pi), M^\star(\pi))\right]} \\
&\le \frac{\alpha_i}{4} + \sqrt{2\mathbb{E}_{\pi\sim\widehat{p}_i}\left[D^2_{\mathsf{H}}\left(M(\pi), \widehat{M}_i(\pi)\right)\right] + 2\mathbb{E}_{\pi\sim\widehat{p}_i}\left[D^2_{\mathsf{H}}\left(\widehat{M}_i(\pi), M^\star(\pi)\right)\right]} \\
&\le \frac{\alpha_i}{4} + 2\sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}{J_i}} \le \frac{\alpha_i}{2},
\end{aligned} \tag{87}$$

where the second-to-last inequality holds since $M \in \mathcal{M}_{i+1}$ and by assumption of the event $\mathscr{A}_i$ (in particular, using (78)), and the final inequality holds since $2\sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}{J_i}} \le \frac{\alpha_i}{4}$ by our choice of $\alpha_i \ge 64\varepsilon_i$ and Eq. (82).

Now fix any $M \in \mathrm{co}(\mathcal{M}_{i+1})$, and note that we can write $M = \mathbb{E}_{M'\sim\nu_M}[M']$ for some $\nu_M \in \Delta(\mathcal{M}_{i+1})$. Then for all $\pi \in \Pi$, $f^M(\pi) = \mathbb{E}_{M'\sim\nu_M}[f^{M'}(\pi)]$, and it follows from Eq. (87) that (again under $\mathscr{A}_i$)

$$f^{M^\star}(\pi_{M^\star}) - f^{M}(\pi_M) \le \mathbb{E}_{\pi\sim\widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M}(\pi)\right] = \mathbb{E}_{M'\sim\nu_M}\mathbb{E}_{\pi\sim\widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M'}(\pi)\right] \le \frac{\alpha_i}{2}.$$

Given any $M \in \mathrm{co}(\mathcal{M}_{i+1})$, we have now that under $\mathscr{A}_i$,

$$\begin{aligned}
\mathbb{E}_{\pi\sim p}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] &\le \frac{\alpha_i}{2} + \mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_M) - f^{M}(\pi)\right] + \mathbb{E}_{\pi\sim p}\left[|f^{M}(\pi) - f^{M^\star}(\pi)|\right] \\
&\le \frac{\alpha_i}{2} + \mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_M) - f^{M}(\pi)\right] + \sqrt{\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}(M(\pi), M^\star(\pi))\right]},
\end{aligned}$$

as desired. □

Our final technical lemma, Lemma I.4, bounds the sub-optimality for all policies played in each epoch $\mathcal{E}_i$. The need to establish a result of this type is a crucial difference between the regret and PAC frameworks, and motivates many of the algorithm design choices behind Algorithm 2.

Lemma I.4 ("Backup" regret guarantee)**.** *Fix any $i \in [N]$. Then for all $t \in \mathcal{E}_i$, we have that under the event $\mathscr{A}_{i-1}$,*

$$\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\}, \widehat{M}^t) + \frac{\alpha_{i-1}}{2} + \sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]}.$$

Proof of Lemma I.4. Fix any $t \in \mathcal{E}_i$. The choice of p^t in Line 10 ensures that

$$\mathbb{E}_{\pi\sim p^t}[f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) - f^{\widehat{M}^t}(\pi)] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\}, \widehat{M}^t). \tag{88}$$

Next, under the event $\mathscr{A}_{i-1}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right]\\
&\le \mathbb{E}_{\pi\sim p^t}\left[f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) - f^{\widehat{M}^t}(\pi)\right] + \frac{\alpha_{i-1}}{2} + \sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]}\\
&\le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\}, \widehat{M}^t) + \frac{\alpha_{i-1}}{2} + \sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]},
\end{aligned}$$

where the first inequality uses Lemma I.3 at epoch $i-1$ with $p = p^t$ and $M = \widehat{M}^t$, together with the fact that $\widehat{M}^t \in \mathrm{co}(\mathcal{M}_i)$ by construction, and the second inequality uses Eq. (88). □

I.2.2. PROOF OF THEOREM C.2

Proof of Theorem C.2. Let us condition on the event $\mathscr{A}\cap\mathscr{B}\cap\mathscr{C}\cap\mathscr{D}$, which, by Lemma I.1, holds with probability $1 - 3L\delta = 1 - 3\lceil\log 1/\delta\rceil\cdot\delta$. Fix $i \in [N]$. We analyze the regret in each epoch i as follows.

- We first analyze the rounds in $t \in \mathcal{E}_i$. By Lemma I.4, under the event $\mathscr{A}_{i-1}$, we have

$$\begin{aligned}
&\sum_{t\in\mathcal{E}_i}\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right]\\
&\le |\mathcal{E}_i|\cdot\left(\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\overline{M}\}, \overline{M}) + \alpha_{i-1}\right) + \sum_{t\in\mathcal{E}_i}\sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]}\\
&\le |\mathcal{E}_i|\cdot(\alpha_i + \alpha_{i-1}) + \sqrt{|\mathcal{E}_i|\cdot\sum_{t\in\mathcal{E}_i}\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]}\\
&\le 2\cdot|\mathcal{E}_i|\cdot\alpha_{i-1} + \sqrt{|\mathcal{E}_i|\cdot\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|, \delta)},
\end{aligned} \tag{89}$$

  where the second-to-last inequality follows by our choice of α_i and the final inequality follows from the fact that $\mathscr{B}_i$ holds and $\alpha_i \le \alpha_{i-1}$.

- We next analyze the rounds in $\mathcal{R}_i$. We first analyze those rounds in which a decision $\pi^j_s \sim p^s$ was sampled on Line 17. To do so, fix any $s \in \mathcal{S}_i$. We first note that, by definition of $J_{i,s}$,

$$\begin{aligned}
\sum_{j=1}^{J_{i,s}}\sqrt{\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi), \widehat{M}^s(\pi)\right)\right]} &\le \sqrt{J_{i,s}\cdot\sum_{j=1}^{J_{i,s}}\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi), \widehat{M}^s(\pi)\right)\right]}\\
&\le \sqrt{J_{i,s}\cdot\left(\frac{J_i\varepsilon_i^2}{4} + 2\right)} \le \sqrt{2J_i} + J_i\varepsilon_i/2.
\end{aligned}$$

Furthermore, since the event $\mathscr{A}_i$ holds (in particular, using (77)), we have

$$\sum_{j=1}^{J_{i,s}} \sqrt{\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi), M^\star(\pi)\right)\right]} \leq \sqrt{J_{i,s}\cdot \sum_{j=1}^{J_{i,s}} \mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi), M^\star(\pi)\right)\right]}$$
$$\leq \sqrt{J_{i,s}\cdot \mathbf{Est}_{\mathsf{H}}(J_i,\delta)} \leq \sqrt{J_i^2\varepsilon_i^2/32} \leq J_i\varepsilon_i,$$

where the second-to-last inequality uses (82). Using the above displays, we have

$$\begin{aligned}
&\sum_{j=1}^{J_{i,s}} \mathbb{E}_{\pi^j_s\sim p^s}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^j_s)\right] \\
\leq & J_{i,s}\cdot\left(\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) + \alpha_{i-1} + \sqrt{\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widehat{M}^s(\pi), M^\star(\pi)\right)\right]}\right) \\
\leq & 2J_{i,s}\cdot\alpha_{i-1} + \sum_{j=1}^{J_{i,s}} \sqrt{2\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widehat{M}^s(\pi), \widetilde{M}^j_s(\pi)\right)\right]} + \sqrt{2\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widetilde{M}^j_s(\pi)\right)\right]} \\
\leq & 2J_i\cdot\alpha_{i-1} + \frac{3J_i\varepsilon_i}{2} + \sqrt{2J_i}.
\end{aligned}$$

Next we analyze the rounds $t\in\mathcal{R}_i$ where $\pi^t\sim p^{s_i}=\widehat{p}_i$ on Line 25. Since $\mathscr{A}_i$ holds, we have from Lemma I.2 that $\mathbb{E}_{\pi^t\sim\widehat{p}_i}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)] \leq \alpha_i/4$, meaning that the total contribution to the regret from such rounds $t\in\mathcal{R}_i$ is at most $|\mathcal{R}_i|\cdot\alpha_i/4$. Thus, the overall contribution to regret from rounds in $\mathcal{R}_i$ is bounded above as follows:

$$\sum_{t\in\mathcal{R}_i} \mathbb{E}_{\pi\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)] \leq \frac{|\mathcal{R}_i|\alpha_i}{4} + L\cdot\left(2J_i\alpha_i + \frac{3J_i\varepsilon_i}{2} + \sqrt{2J_i}\right)$$
$$\leq 4|\mathcal{R}_i|\alpha_i + \sqrt{2L|\mathcal{R}_i|},$$

where in the second inequality we have used that $|\mathcal{R}_i| = J_i\cdot L$ and $3\varepsilon_i/2\leq\alpha_i$.

Summarizing, under the event $\mathscr{A}\cap\mathscr{B}\cap\mathscr{C}\cap\mathscr{D}$, the total regret is bounded above by

$$\sum_{t=1}^{T}\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \leq \sum_{i=1}^{N}\left(4\alpha_{i-1}\cdot(|\mathcal{R}_i| + |\mathcal{E}_i|) + \sqrt{|\mathcal{E}_i|\cdot\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta)} + \sqrt{2L|\mathcal{R}_i|}\right). \tag{90}$$

We now simplify the expression in Eq. (90). Recall that Assumption C.2 gives that for all $\varepsilon>0$,

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \leq C^2_{\mathrm{reg}}\cdot\textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}).$$

Applying this inequality a total of $\left\lceil\frac{\log(\varepsilon_i/\varepsilon_N)}{\log(C_{\mathrm{reg}})}\right\rceil$ times for each $i\in[N]$ gives that

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) \leq C^2_{\mathrm{reg}}\cdot\left(\frac{\varepsilon_i}{\varepsilon_N}\right)^2\cdot\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M}) = C^2_{\mathrm{reg}}\cdot 2^{N-i}\cdot\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M}).$$

Then by the choice $\alpha_i = C_0 \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) + 64\varepsilon_i$ for each $i \in [N]$, we have

$$\begin{aligned}\sum_{i=1}^{N} \alpha_{i-1} \cdot 2^i &\leq 64 \sum_{i=1}^{N} \varepsilon_{i-1} \cdot 2^i + C_0 \sum_{i=1}^{N} 2^i \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_{i-1}}(\mathcal{M}) \\ &\leq 64 \sum_{i=1}^{N} \varepsilon_N \cdot \sqrt{2^{N+1+i}} + O\left(\sum_{i=1}^{N} 2^N \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M})\right) \\ &\leq 128 \cdot \sqrt{C_1 \cdot \mathbf{Est}_{\mathsf{H}}(T,\delta) \cdot L} \cdot \sum_{i=1}^{N} \sqrt{2^{i+1}} + O\left(NT \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M})\right).\end{aligned}$$

Therefore, we may upper bound the expression in Eq. (90) as follows (using that $\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta) \leq \mathbf{Est}_{\mathsf{H}}(T,\delta)$ for each i):

$$\begin{aligned}&\sum_{i=1}^{N} \left(4\alpha_{i-1} \cdot (|\mathcal{R}_i| + |\mathcal{E}_i|) + \sqrt{|\mathcal{E}_i| \cdot \mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta)} + \sqrt{2L|\mathcal{R}_i|}\right) \\ \leq& 4\sum_{i=1}^{N} \alpha_{i-1} \cdot 2^i + \sqrt{\mathbf{Est}_{\mathsf{H}}(T,\delta)} \sum_{i=1}^{N} \sqrt{2^i} + \sqrt{2L} \sum_{i=1}^{N} \sqrt{2^i} \\ \leq& O\left(\sqrt{\mathbf{Est}_{\mathsf{H}}(T,\delta) \cdot L} \sum_{i=1}^{N} \sqrt{2^i} + \sqrt{L} \sum_{i=1}^{N} \sqrt{2^i} + NT \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M})\right) \\ \leq& O\left(\sqrt{T\log(1/\delta) \cdot \mathbf{Est}_{\mathsf{H}}(T,\delta)} + T\log(T) \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M})\right),\end{aligned}$$

where we have used that $L = O(\log 1/\delta)$ in the final inequality. The proof is completed by rescaling from δ to δ^2 and noting that, by construction, we have $\varepsilon_N \leq C \cdot \sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(T,1/\delta)\cdot\log 1/\delta}{T}}$ for a universal constant $C > 0$. □

Appendix J. Proofs and Additional Results from Appendix D

J.1. Technical Lemmas

Lemma J.1. *Let M and $\overline{M}$ have $\mathcal{R} \subseteq [0,1]$. Then for all $\varepsilon \geq 0$ and $p \in \Delta(\Pi)$, if $M \in \mathcal{H}_{p,\varepsilon}(\overline{M})$, then*

$$\mathbb{E}_{\pi\sim p}\left[\left|f^M(\pi) - f^{\overline{M}}(\pi)\right|\right] \leq \varepsilon. \tag{91}$$

Proof of Lemma J.1. Since rewards are in $[0,1]$, we have

$$\mathbb{E}_{\pi\sim p}\left[\left|f^M(\pi) - f^{\overline{M}}(\pi)\right|\right] \leq \mathbb{E}_{\pi\sim p}\left[D_{\mathsf{TV}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq \sqrt{\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]} \leq \varepsilon.$$

□

Lemma J.2. *Fix a model class $\mathcal{M}$. Let $\overline{M} \in \mathcal{M}^+$ and $\varepsilon > 0$ be given, and set*

$$\mathcal{M}' = \{M \in \mathcal{M} \mid f^{\overline{M}}(\pi_{\overline{M}}) \leq f^{M}(\pi_{\overline{M}}) + \varepsilon\}.$$

Then we have

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M}) \leq \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}', \overline{M}) + \varepsilon.$$

Proof of Lemma J.2. Let $p \in \Delta(\Pi)$ achieve the value of $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}', \overline{M})$, and set $p' = \frac{1}{2}p + \frac{1}{2}\mathbb{I}_{\pi_{\overline{M}}}$. Let $M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M}) \subseteq \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{\mathbb{I}_{\pi_{\overline{M}}},\varepsilon}(\overline{M})$. We claim that $M \in \mathcal{M}'$. Indeed,

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{M}(\pi_{\overline{M}}) \leq D_{\mathsf{H}}\big(\overline{M}(\pi_{\overline{M}}), M(\pi_{\overline{M}})\big) \leq \varepsilon$$

by Lemma J.1. It follows that

$$\sup_{M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})} \mathbb{E}_{\pi \sim p}[f^{M}(\pi_M) - f^{M}(\pi)] \leq \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{M}'} \mathbb{E}_{\pi \sim p}[f^{M}(\pi_M) - f^{M}(\pi)] \leq \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}', \overline{M}), \tag{92}$$

so that

$$\sup_{M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})} \mathbb{E}_{\pi \sim p'}[f^{M}(\pi_M) - f^{M}(\pi)] \leq \frac{1}{2}\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}', \overline{M}) + \frac{1}{2}\sup_{M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})} [f^{M}(\pi_M) - f^{M}(\pi_{\overline{M}})].$$

To bound the final term above, we have

$$\begin{aligned}
\sup_{M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})} [f^{M}(\pi_M) - f^{M}(\pi_{\overline{M}})] &\leq \sup_{M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})} \Big[f^{M}(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}})\Big] + \varepsilon \\
&\leq \sup_{M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})} \mathbb{E}_{\pi \sim p}\Big[f^{M}(\pi_M) - f^{\overline{M}}(\pi)\Big] + \varepsilon \\
&\leq \sup_{M \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})} \mathbb{E}_{\pi \sim p}[f^{M}(\pi_M) - f^{M}(\pi)] + 2\varepsilon \\
&\leq \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}', \overline{M}) + 2\varepsilon,
\end{aligned}$$

where the first and third inequalities use Lemma J.1, and the last inequality applies Eq. (92).

□

Lemma J.3. *For a model class $\mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}^+$, define*

$$\widetilde{\mathsf{p\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) = \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{q,\varepsilon}(\overline{M})} \mathbb{E}_{\pi \sim p}[f^{M}(\pi_M) - f^{M}(\pi)], \tag{93}$$

with the convention that the value above is zero when $\mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{q,\varepsilon}(\overline{M}) = \varnothing$. For all $\overline{M} \in \mathcal{M}^+$ and $\varepsilon > 0$, we have

$$\widetilde{\mathsf{p\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \widetilde{\mathsf{p\text{-}dec}}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\mathcal{M}, \overline{M}). \tag{94}$$

Proof of Lemma J.3. The first inequality is immediate. For the second, we have

$$\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{\frac{1}{2}p+\frac{1}{2}q,\varepsilon}(\overline{M})} \mathbb{E}_{\pi \sim p}[f^{M}(\pi_M) - f^{M}(\pi)],$$

by observing that for any minimizer q for $\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}$, we can arrive at an upper bound by substituting $q' = \frac{1}{2}p + \frac{1}{2}q$. The result now follows because $\mathcal{H}_{\frac{1}{2}p+\frac{1}{2}q,\varepsilon}(\overline{M}) \subseteq \mathcal{H}_{p,\sqrt{2}\varepsilon}(\overline{M}) \cap \mathcal{H}_{q,\sqrt{2}\varepsilon}(\overline{M})$. □

J.2. Additional Properties of the Decision-Estimation Coefficient

J.2.1. LOCALIZATION

The following result is an extension of Proposition D.7 which accommodates randomized estimators.

Proposition J.1. *Let $\alpha, \gamma > 0$ and $\nu \in \Delta(\mathcal{M})$ be given. Let $\overline{M}_\nu = \mathbb{E}_{M' \sim \nu}[M']$. For all $\varepsilon > 0$, we have*

$$\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}_\nu) \cup \{\overline{M}_\nu\}, \nu) \leq \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \max\left\{0,\ \alpha + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}\right\}. \tag{95}$$

which in particular yields

$$\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}_\nu) \cup \{\overline{M}_\nu\}, \nu) \leq \textsf{r-dec}^{\mathrm{c,rnd}}_{\sqrt{2\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{1}{2\gamma}. \tag{96}$$

Proof of Proposition J.1. Fix $\nu \in \Delta(\mathcal{M})$ and $\varepsilon > 0$, and let $p \in \Delta(\Pi)$ be a minimizer for $\textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu)$. Fix any $M \in \mathcal{M}_\alpha(\overline{M}_\nu) \cup \{\overline{M}_\nu\}$. We bound the regret under p by considering two cases.

Case 1. If $\mathbb{E}_{\overline{M} \sim \nu} \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2$, it follows from the definition $\textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu)$ of that $\mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi)\right] \leq \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu)$.

Case 2. For the second case, suppose that $\mathbb{E}_{\overline{M} \sim \nu} \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] > \varepsilon^2$. We now compute

$$\begin{aligned}
\mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi)\right] &\leq \alpha + \mathbb{E}_{\pi \sim p}\left[f^{\overline{M}_\nu}(\pi_{\overline{M}_\nu}) - f^M(\pi)\right] \\
&\leq \alpha + \mathbb{E}_{\pi \sim p}\left[f^{\overline{M}_\nu}(\pi_{\overline{M}_\nu}) - f^{\overline{M}_\nu}(\pi)\right] + \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot \mathbb{E}_{\pi \sim p}\left[(f^M(\pi) - f^{\overline{M}_\nu}(\pi))^2\right] \\
&\leq \alpha + \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}_\nu(\pi)\big)\right], \\
&\leq \alpha + \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot \mathbb{E}_{\overline{M} \sim \nu} \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right],
\end{aligned}$$

where the second inequality uses Young's inequality and the final inequality uses convexity of the squared Hellinger distance. Rearranging, we obtain

$$\begin{aligned}
&\mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot \mathbb{E}_{\overline{M} \sim \nu}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]\right] \\
&\leq \alpha + \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \nu) + \frac{1}{2\gamma} - \frac{\gamma}{2} \cdot \mathbb{E}_{\overline{M} \sim \nu} \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \\
&\leq \alpha + \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \overline{M}_\nu) + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}.
\end{aligned}$$

Recalling that M can be any model in $\mathcal{M}_\alpha(\overline{M}_\nu) \cup \{\overline{M}_\nu\}$, we obtain

$$\begin{aligned}
&\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}_\nu) \cup \{\overline{M}_\nu\}, \nu) \\
&\leq \max\left\{\textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu),\ \alpha + \frac{1}{2\gamma} + \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) - \frac{\gamma\varepsilon^2}{2}\right\} \\
&= \textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \max\left\{0,\ \alpha + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}\right\}.
\end{aligned} \tag{97}$$

□

J.2.2. ROLE OF CONVEXITY FOR PAC DEC

For $\nu \in \Delta(\mathcal{M})$, we define "randomized" variants of the PAC DEC, analogous to those introduced in Appendix D, as follows:

$$\textsf{p-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}, \nu) = \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\overline{M}\sim\nu} \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \leq \varepsilon^2 \right\}, \tag{98}$$

$$\textsf{p-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M}, \nu) = \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\overline{M}\sim\nu} \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \right\}. \tag{99}$$

The following result provides a PAC counterpart to Eq. (40) of Proposition D.11.

Proposition J.2. *Suppose that Assumption G.1 holds. For all $\gamma > 0$, we have*

$$\sup_{\overline{M}\in\mathcal{M}^+} \textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \leq \sup_{\nu\in\Delta(\mathcal{M})} \textsf{p-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}, \nu) \leq \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{o}}_{\gamma/4}(\mathcal{M}, \overline{M}). \tag{100}$$

A PAC analogue of Eq. (41) can be proven by adapting the proof of Proposition D.11; we do not include this result.

Proof of Proposition J.2. Let $\overline{M} \in \mathcal{M}^+$ and $\gamma > 0$ be given. We first prove the inequality (100). By Assumption G.1, we have

$$\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) = \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p,q\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu}\left[\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right]\right].$$

Since Hellinger distance satisfies the triangle inequality, we have that for all $\pi \in \Pi$,

$$\begin{aligned} \mathbb{E}_{M,M'\sim\mu}\left[D^2_{\mathsf{H}}\left(M(\pi), M'(\pi)\right)\right] &\leq 2\,\mathbb{E}_{M\sim\mu}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] + 2\,\mathbb{E}_{M'\sim\mu}\left[D^2_{\mathsf{H}}\left(M'(\pi), \overline{M}(\pi)\right)\right] \\ &= 4\,\mathbb{E}_{M\sim\mu}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right]. \end{aligned}$$

It follows that

$$\begin{aligned} &\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \\ &\leq \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p,q\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu}\left[\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \frac{\gamma}{4} \cdot \mathbb{E}_{M'\sim\mu} \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), M'(\pi)\right)\right]\right] \\ &\leq \sup_{\nu\in\Delta(\mathcal{M})} \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p,q\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu}\left[\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \frac{\gamma}{4} \cdot \mathbb{E}_{M'\sim\nu} \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), M'(\pi)\right)\right]\right] \\ &\leq \sup_{\nu\in\Delta(\mathcal{M})} \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \frac{\gamma}{4} \cdot \mathbb{E}_{M'\sim\nu} \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), M'(\pi)\right)\right]\right\} \\ &= \sup_{\nu\in\Delta(\mathcal{M})} \textsf{p-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}, \nu). \end{aligned}$$

Jensen's inequality further implies that $\sup_{\nu\in\Delta(\mathcal{M})} \textsf{p-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}, \nu) \leq \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{o}}_{\gamma/4}(\mathcal{M}, \overline{M})$. □

J.2.3. PAC DEC WITH GREEDY DECISIONS

For a model class $\mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}^{+}$, define

$$\text{p-dec}^{\mathrm{c,greedy}}_{\varepsilon}(\mathcal{M}, \overline{M}) = \inf_{q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ f^{M}(\pi_{M}) - f^{M}(\pi_{\overline{M}}) \mid \mathbb{E}_{\pi \sim q}\left[D^{2}_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^{2} \right\},$$

with the convention that the value above is zero when $\mathcal{H}_{q,\varepsilon}(\overline{M}) = \varnothing$.

Proposition J.3. *For all $\varepsilon > 0$ and $\overline{M} \in \mathcal{M}^{+}$, we have*

$$\text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \text{p-dec}^{\mathrm{c,greedy}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \text{p-dec}^{\mathrm{c}}_{\sqrt{3}\varepsilon}(\mathcal{M}, \overline{M}) + 4\varepsilon. \tag{101}$$

Proof of Proposition J.3. It is immediate that $\text{p-dec}^{\mathrm{c,greedy}}_{\varepsilon}(\mathcal{M}, \overline{M}) \geq \text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, so let us prove the second inequality. Let $\overline{M} \in \mathcal{M}$ and $\varepsilon > 0$ be given, and let (p, q) be minimizers for $\widetilde{\text{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$. Define $q' = \frac{1}{3}q + \frac{1}{3}p + \frac{1}{3}\mathbb{I}_{\pi_{\overline{M}}}$. Note that $\mathcal{H}_{q',\varepsilon/\sqrt{3}}(\overline{M}) \subseteq \mathcal{H}_{q,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{\mathbb{I}_{\pi_{\overline{M}}},\varepsilon}(\overline{M})$. As a result, for all $M \in \mathcal{H}_{q',\varepsilon/\sqrt{3}}(\overline{M})$, we have

$$\begin{aligned}
f^{M}(\pi_{M}) - f^{M}(\pi_{\overline{M}}) &= f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) + \left(f^{\overline{M}}(\pi_{\overline{M}}) - f^{M}(\pi_{\overline{M}})\right) \\
&\leq f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) + \varepsilon \\
&\leq \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi)\right] + \varepsilon \\
&\leq \mathbb{E}_{\pi \sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] + 2\varepsilon \\
&\leq \widetilde{\text{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + 2\varepsilon \\
&\leq \text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + 2\varepsilon,
\end{aligned}$$

where the first and third inequalities use Lemma J.1 and the final inequality uses Lemma J.3. □

J.3. Omitted Proofs

Proof of Proposition D.1. Let $\overline{M} \in \mathcal{M}^{+}$ be given. We first prove a more general result under the assumption that for some $\delta > 0$, $f^{\overline{M}}(\pi_{\overline{M}}) \leq f^{M}(\pi_{M}) + \delta$ for all $M \in \mathcal{M}$:

$$\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \delta + \inf_{\gamma > 0}\left\{\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + 4\gamma\varepsilon^{2} + (4\gamma)^{-1}\right\}. \tag{102}$$

Then, at the end of the proof, we show that it is possible to take $\delta = O(\varepsilon)$ without loss of generality.

Let $\gamma > 0$ be given and let p_0 be the minimizer for $\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M})$, so that

$$\sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p_0}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot D^{2}_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}).$$

Let $M^{\star} := \arg\min_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p_0}\left[D^{2}_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right]$, and let $\Delta^{2} := \mathbb{E}_{\pi \sim p_0}\left[D^{2}_{\mathsf{H}}\big(M^{\star}(\pi), \overline{M}(\pi)\big)\right]$.

We will bound the constrained DEC, $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})$, by playing the distribution

$$p := (1 - q) \cdot \mathbb{I}_{\pi_{\overline{M}}} + q \cdot p_0,$$

where

$$q := \frac{2\varepsilon^{2}}{\Delta^{2}} \wedge 1.$$

Before proceeding, we state a basic technical lemma.

Lemma J.4. *The distribution p_0 satisfies*

$$\mathbb{E}_{\pi\sim p_0}\Big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\Big] \leq \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + (4\gamma)^{-1} + 2\gamma\Delta^2. \tag{103}$$

We bound the value of the constrained DEC for p by considering two cases.
Case 1: $q = 1$. If $q = 1$, then $\Delta^2 \leq 2\varepsilon^2$, and $p = p_0$. For models $M \in \mathcal{H}_{p,\varepsilon}(\overline{M})$, the definition of p_0 implies that

$$\begin{aligned}\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] &\leq \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma\cdot\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \\ &\leq \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma\varepsilon^2.\end{aligned}$$

For the model $\overline{M}$, Lemma J.4 implies that

$$\mathbb{E}_{\pi\sim p}\Big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\Big] \leq \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + (4\gamma)^{-1} + 4\gamma\varepsilon^2.$$

Case 2: $q < 1$. If $q < 1$, then for all $M \in \mathcal{M}$, we have

$$\begin{aligned}\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] &\geq q\cdot\mathbb{E}_{\pi\sim p_0}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \\ &\geq q\cdot\mathbb{E}_{\pi\sim p_0}\big[D^2_{\mathsf{H}}\big(M^\star(\pi), \overline{M}(\pi)\big)\big] = 2\varepsilon^2 > \varepsilon^2,\end{aligned}$$

where the second inequality uses that $M^\star$ minimizes $\mathbb{E}_{\pi\sim p_0}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]$, and the last inequality uses that $q = \frac{2\varepsilon^2}{\Delta^2}$ whenever $q < 1$. It follows that $\mathcal{H}_{p,\varepsilon}(\overline{M}) \cup \{\overline{M}\} = \{\overline{M}\}$, so we only need to bound the regret of the distribution p under $\overline{M}$. To do so, we observe that

$$\begin{aligned}\mathbb{E}_{\pi\sim p}\Big[g^{\overline{M}}(\pi)\Big] = q\cdot\mathbb{E}_{\pi\sim p_0}\Big[g^{\overline{M}}(\pi)\Big] &\leq q\cdot\big(\delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + (4\gamma)^{-1} + 2\gamma\Delta^2\big) \\ &\leq \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + (4\gamma)^{-1} + q\cdot 2\gamma\Delta^2 \\ &= \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + (4\gamma)^{-1} + 4\gamma\varepsilon^2,\end{aligned}$$

where the first inequality uses Lemma J.4, and the final equality uses that $q = \frac{2\varepsilon^2}{\Delta^2}$.
Finishing up. We have established that

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \delta + \inf_{\gamma>0}\big\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + 4\gamma\varepsilon^2 + (4\gamma)^{-1}\big\}$$

whenever $f^{\overline{M}}(\pi_{\overline{M}}) \leq f^M(\pi_M) + \delta$ for all $M \in \mathcal{M}$. To conclude, we appeal to Lemma J.2 applied to the class $\mathcal{M} \cup \{\overline{M}\}$, which implies that

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\mathcal{M}' \cup \{\overline{M}\}, \overline{M}) + \sqrt{2}\varepsilon, \tag{104}$$

where $\mathcal{M}' = \big\{M \in \mathcal{M} \mid f^{\overline{M}}(\pi_{\overline{M}}) \leq f^M(\pi_M) + \sqrt{2}\varepsilon\big\}$. Applying (102) to the quantity $\textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\mathcal{M}' \cup \{\overline{M}\}, \overline{M})$ and combining with (104) yields

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq 2\sqrt{2}\varepsilon + \inf_{\gamma>0}\big\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + 8\gamma\varepsilon^2 + (4\gamma)^{-1}\big\}.$$

To simplify this result slightly, we consider two cases. If $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \leq (4\gamma)^{-1}$, then choosing $\gamma = (4\varepsilon)^{-1}$ gives $\inf_{\gamma>0}\big\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + 4\gamma\varepsilon^2 + (4\gamma)^{-1}\big\} \leq 4\varepsilon$. Otherwise, we have $\inf_{\gamma>0}\big\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + 4\gamma\varepsilon^2 + (4\gamma)^{-1}\big\} \leq \inf_{\gamma>0}\big\{2\cdot\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + 8\gamma\varepsilon^2\big\}$.

□

Proof of Lemma J.4. Observe that

$$\begin{aligned}\mathbb{E}_{\pi\sim p_0}\Big[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\Big] &\leq \delta+\mathbb{E}_{\pi\sim p_0}\Big[f^{M^\star}(\pi_{M^\star})-f^{\overline{M}}(\pi)\Big]\\ &\leq \delta+\mathbb{E}_{\pi\sim p_0}\big[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)\big]+\Delta,\end{aligned}$$

where the first inequality uses the fact that $f^{\overline{M}}(\pi_{\overline{M}})\leq f^{M}(\pi_M)+\delta$ for all $M\in\mathcal{M}$, and the second inequality uses Lemma J.1. In addition, the definition of p_0 implies that $\mathbb{E}_{\pi\sim p_0}\big[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)\big]\leq$ $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})+\gamma\Delta^2$, so we have

$$\begin{aligned}\mathbb{E}_{\pi\sim p_0}\Big[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\Big] &\leq \delta+\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})+\gamma\Delta^2+\Delta\\ &\leq \delta+\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})+2\gamma\Delta^2+(4\gamma)^{-1}.\end{aligned}$$

□

Proof of Proposition D.2. We first prove the inequality (23). Let $\varepsilon>0$ and $\overline{M}\in\mathcal{M}^+$ be fixed. Using the method of Lagrange multipliers, we have

$$\begin{aligned}\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M}) &= \inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\big\{\mathbb{E}_{\pi\sim p}[g^M(\pi)]\mid \mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big]\leq\varepsilon^2\big\}\\ &= \inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\max\left\{\inf_{\gamma\geq 0}\big\{\mathbb{E}_{\pi\sim p}[g^M(\pi)]-\gamma\big(\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big]-\varepsilon^2\big)\big\},0\right\}\\ &\leq \inf_{\gamma\geq 0}\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\max\big\{\mathbb{E}_{\pi\sim p}[g^M(\pi)]-\gamma\big(\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big]-\varepsilon^2\big),0\big\}\\ &\leq \inf_{\gamma\geq 0}\big\{\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})\vee 0+\gamma\varepsilon^2\big\}.\end{aligned}$$

We now prove the inequality (24). Let $\gamma\geq 1$ and $\overline{M}\in\mathcal{M}^+$ be fixed. For integers $i\geq 0$, define $\varepsilon_i=2^{-i/2}$. For each $i\geq 0$, let (p_i,q_i) denote a minimizer to the following expression:

$$\inf_{p_i,q_i\in\Delta(\Pi)}\sup_{M\in\mathcal{H}_{q_i,\varepsilon_i}(\overline{M})}\mathbb{E}_{\pi\sim p_i}[g^M(\pi)]=\textsf{p-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M},\overline{M}).$$

Recalling that $L=2\lceil\log 2\gamma\rceil$, set

$$q=\frac{1}{2}\cdot\mathbb{I}_{\pi_{\overline{M}}}+\frac{q_0+\cdots+q_{L-1}}{4L}+\frac{p_0+\cdots+p_{L-1}}{4L},\quad\text{ and }\quad p=\mathbb{I}_{\pi_{\overline{M}}}.$$

Consider any $M\in\mathcal{M}$. We will upper bound the value

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)]-4\gamma L\cdot\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big].$$

Choose $j\in\{0,\ldots,L-1\}$ as large as possible so that

$$\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\leq\frac{\varepsilon_j^2}{4L}. \tag{105}$$

If such an index j does not exist, we must have $\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] > 1/(4L)$. In this case, we have

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - 4\gamma L\cdot\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq 1-\gamma \leq 0 \leq \textsf{p-dec}^{\mathrm{c}}_{0}(\mathcal{M},\overline{M}).$$

Suppose going forward that $0\leq j\leq L-1$ satisfying (105) exists. If $j<L-1$, since we chose the largest possible value of j, we have $\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \geq \frac{\varepsilon_j^2}{8L}$. In addition, regardless of the value of $j\in\{0,1,\ldots,L-1\}$, by the definition of q, we have

$$\mathbb{E}_{\pi\sim q_j}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq 4L\cdot\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq \varepsilon_j^2,$$

and

$$\mathbb{E}_{\pi\sim p_j}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq 4L\cdot\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq \varepsilon_j^2,$$

that is, $M\in\mathcal{H}_{p_j,\varepsilon_j}(\overline{M})\cap\mathcal{H}_{q_j,\varepsilon_j}(\overline{M})$. It follows that

$$\begin{aligned} f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) &\leq \mathbb{E}_{\pi\sim p_j}\left[f^M(\pi_M) - f^{\overline{M}}(\pi)\right] \\ &\leq \mathbb{E}_{\pi\sim p_j}\left[f^M(\pi_M) - f^M(\pi)\right] + \varepsilon_j \\ &\leq \textsf{p-dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M},\overline{M}) + \varepsilon_j, \end{aligned}$$

where the second inequality uses that $M\in\mathcal{H}_{p_j,\varepsilon_j}(\overline{M})$ and the final inequality uses that $M\in\mathcal{H}_{q_j,\varepsilon_j}(\overline{M})$. As a result, we can compute

$$\begin{aligned} &\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi)\right] - 4\gamma L\cdot\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \\ &\leq f^M(\pi_M) - f^M(\pi_{\overline{M}}) - \frac{4\gamma L}{8L}\cdot\varepsilon_j^2\cdot\mathbb{1}\left\{j<L-1\right\} \\ &\leq f^{\overline{M}}(\pi_{\overline{M}}) - f^M(\pi_{\overline{M}}) + \textsf{p-dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M},\overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2\cdot\mathbb{1}\left\{j<L-1\right\} \\ &\leq \frac{1}{2\gamma} + \frac{\gamma}{2}\cdot\left(f^{\overline{M}}(\pi_{\overline{M}}) - f^M(\pi_{\overline{M}})\right)^2 + \textsf{p-dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M},\overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2\cdot\mathbb{1}\left\{j<L-1\right\} \\ &\leq \frac{1}{2\gamma} + \gamma\cdot\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] + \textsf{p-dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M},\overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2\cdot\mathbb{1}\left\{j<L-1\right\} \\ &\leq \frac{1}{\gamma} + \gamma\cdot\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] + \textsf{p-dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M},\overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2, \end{aligned}$$

where the final inequality uses that $\varepsilon_{L-1}^2\leq 1/\gamma^2$ since $L\geq 2\log(2\gamma)$. Rearranging, we obtain

$$\begin{aligned} &\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma\cdot(4L+1)\cdot\mathbb{E}_{\pi\sim q}[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)] \\ &\leq \frac{1}{\gamma} + \textsf{p-dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M},\overline{M}) + \varepsilon_j - \frac{\gamma}{2}\cdot\varepsilon_j^2 \\ &\leq \frac{2}{\gamma} + \textsf{p-dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M},\overline{M}) - \frac{\gamma}{4}\cdot\varepsilon_j^2, \end{aligned} \tag{106}$$

as desired. □

Proof of Proposition D.3. Let $\gamma > 0$ and $\overline{M} \in \mathcal{M}^+$ be given. Fix $\varepsilon > 0$ to be chosen later, and let p be the minimizer for $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$. Consider the value of the offset DEC for $M \in \mathcal{M}$:

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

We consider two cases. First, if $M \in \mathcal{H}_{p,\varepsilon}(\overline{M})$, it is immediate that

$$\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \leq \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}).$$

On the other hand if $M \notin \mathcal{H}_{p,\varepsilon}(\overline{M})$, we have $\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \geq \varepsilon^2$, and since $g^M \leq 1$, we have

$$\mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \leq 1 - \gamma\varepsilon^2 \leq 0$$

by choosing $\varepsilon = \gamma^{-1/2}$. We conclude that

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\gamma^{-1/2}}(\mathcal{M}, \overline{M}).$$

□

Proof of Proposition D.4. Recall the model classes $\mathcal{M}^{\alpha,\beta}$ defined in Example E.1, parametrized by $\alpha \in (0, 1/2], \beta \in (0, 1), A \in \mathbb{N}$. Consider any choice for α, β, and A; we will specify concrete values below. Lemma K.2 gives that for all $\varepsilon > 0$,

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}^{\alpha,\beta}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M}^{\alpha,\beta})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}^{\alpha,\beta} \cup \{\overline{M}\}, \overline{M}) \leq O\left(\frac{\varepsilon^2}{\beta}\right).$$

On the other hand, Lemma K.1 gives that for the choice of $\overline{M} = \widetilde{M} \in \mathcal{M}^{\alpha,\beta}$, we have, for all $\gamma > 0$,

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}, \widetilde{M}) \geq \frac{\alpha}{2 + 8\gamma\beta} - 4\gamma/A.$$

Given $\gamma > 0$, let us choose $\alpha = 1/2,\ \beta = 1/\sqrt{\gamma}$, and $A = 256\gamma^2/\beta$. Then the resulting model class $\mathcal{M} = \mathcal{M}^{\alpha,\beta}$ satisfies $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \leq O(\varepsilon^2\gamma^{1/2})$ for all $\varepsilon > 0$, yet

$$\sup_{\overline{M}\in\mathcal{M}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \geq \frac{1}{4 + 16\gamma\beta} - \frac{4\gamma}{A} \geq \frac{1}{32} \cdot \left(\frac{1}{\gamma^{1/2}} \wedge 1\right) - \frac{4\gamma}{A} \geq \frac{1}{64} \cdot \left(\frac{1}{\gamma^{1/2}} \wedge 1\right) \geq \Omega(\gamma^{-1/2}).$$

□

Proof of Proposition D.5. Recall the definition of $\widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$ in (93). Let $\varepsilon > 0$ and $\overline{M} \in \mathcal{M}^+$ be given, and let (p', q') be minimizers for $\widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$. Then for all $M \in \mathcal{H}_{p',\varepsilon}(\overline{M}) \cap \mathcal{H}_{q',\varepsilon}(\overline{M})$, we have

$$\begin{aligned}\widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) &\geq \mathbb{E}_{\pi\sim p'}[f^M(\pi_M) - f^M(\pi)] \\ &\geq \mathbb{E}_{\pi\sim p'}\Big[f^M(\pi_M) - f^{\overline{M}}(\pi)\Big] - \varepsilon \\ &\geq f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) - \varepsilon,\end{aligned}$$

where the second inequality uses Lemma J.1. That is, if we define $\alpha = \varepsilon + \widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, we have $\mathcal{H}_{p',\varepsilon}(\overline{M}) \cap \mathcal{H}_{q',\varepsilon}(\overline{M}) \subseteq \mathcal{M}_{\alpha}(\overline{M})$. Now, let (p,q) be minimizers for $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M})$, and set $\bar{q} = \frac{1}{3}q + \frac{1}{3}q' + \frac{1}{3}p'$. We have

$$\begin{aligned}
\textsf{p-dec}^{\mathrm{c}}_{\varepsilon/\sqrt{3}}(\mathcal{M}, \overline{M}) &\leq \sup_{M \in \mathcal{H}_{\bar{q},\varepsilon/\sqrt{3}}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \\
&\leq \sup_{M \in \mathcal{H}_{q,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M}) \cap \mathcal{H}_{q',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \\
&\leq \sup_{M \in \mathcal{H}_{q,\varepsilon}(\overline{M}) \cap \mathcal{M}_{\alpha}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \\
&= \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M}).
\end{aligned}$$

Finally, using Lemma J.3, we have $\alpha \leq \varepsilon + \textsf{p-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\mathcal{M}, \overline{M})$.

□

Proof of Proposition D.6. We first prove the following result, which does not requite any regularity condition.

Lemma J.5. *Fix any $\overline{M} \in \mathcal{M}^+$ and $\varepsilon > 0$, and let $\alpha' := \varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$. Then for any constant $C_{\mathrm{reg}} \geq \sqrt{2}$, it holds that*

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}, \overline{M}) \leq \frac{1}{C_{\mathrm{reg}}^2}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha'}(\overline{M}), \overline{M}).$$

Proof of Lemma J.5. Given $\varepsilon > 0$ and $\overline{M}$, let p be the distribution that achieves the value for $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$. Then for all $M \in \mathcal{H}_{p,\varepsilon}(\overline{M})$, we have

$$\begin{aligned}
\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) &\geq \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \\
&\geq \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^{\overline{M}}(\pi)\Big] - \varepsilon \\
&\geq f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) - \varepsilon,
\end{aligned}$$

where the second inequality uses Lemma J.1. Hence, for $\alpha' := \varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, we have $\mathcal{H}_{p,\varepsilon}(\overline{M}) \subseteq \mathcal{M}_{\alpha'}(\overline{M})$.

Now, let p' be the minimizer for $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha'}(\overline{M}), \overline{M})$. Set $\bar{p} = \frac{1}{C_{\mathrm{reg}}^2}p + \Big(1 - \frac{1}{C_{\mathrm{reg}}^2}\Big)p'$. Using that

$$\frac{1}{C_{\mathrm{reg}}^2} \cdot \left(1 - \frac{1}{C_{\mathrm{reg}}^2}\right)^{-1} \leq 1$$

whenever $C_{\mathrm{reg}} \geq \sqrt{2}$, we have

$$\begin{aligned}
\text{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}, \overline{M}) &\leq \sup_{M \in \mathcal{H}_{\bar{p},\varepsilon/C_{\mathrm{reg}}}(\overline{M})} \mathbb{E}_{\pi\sim\bar{p}}[f^M(\pi_M) - f^M(\pi)] \\
&\leq \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim\bar{p}}[f^M(\pi_M) - f^M(\pi)] \\
&\leq \frac{1}{C_{\mathrm{reg}}^2} \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] + \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p'}[f^M(\pi_M) - f^M(\pi)] \\
&= \frac{1}{C_{\mathrm{reg}}^2} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p'}[f^M(\pi_M) - f^M(\pi)] \\
&\leq \frac{1}{C_{\mathrm{reg}}^2} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \sup_{M \in \mathcal{M}_{\alpha'}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p'}[f^M(\pi_M) - f^M(\pi)] \\
&= \frac{1}{C_{\mathrm{reg}}^2} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha'}(\overline{M}), \overline{M}).
\end{aligned}$$

□

We now complete the proof of Proposition D.6. Under the assumed growth condition, we have $\text{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}, \overline{M}) \geq \frac{1}{c_{\mathrm{reg}}^2}\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})$, so rearranging the result of Lemma J.5 (with $\alpha' = \alpha(\varepsilon)$) yields

$$\text{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}, \overline{M}) \leq \left(\frac{1}{c_{\mathrm{reg}}^2} - \frac{1}{C_{\mathrm{reg}}^2}\right)^{-1} \cdot \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha(\varepsilon)}(\overline{M}), \overline{M}).$$

The result in the proposition follows by rescaling ε to $\varepsilon \cdot C_{\mathrm{reg}}$. □

For use later on, we also prove the following variant of Proposition D.6, which concerns the DEC for the class $\mathcal{M} \cup \{\overline{M}\}$.

Proposition J.4 (Localization for regret DEC; global version). *Consider any set $\mathcal{M}' \subseteq \mathcal{M}^+$, and assume that the strong regularity condition (28) is satisfied relative to $\mathcal{M}'$. Then, for all $\varepsilon > 0$, letting $\alpha(\varepsilon) := C_{\mathrm{reg}} \cdot \varepsilon + \sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq C_{\mathrm{reg}}^2 \cdot \big(\varepsilon + \sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})\big)$, we have*

$$\sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq C_{\mathrm{loc}} \cdot \sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}_{\alpha(\varepsilon)}(\overline{M}) \cup \{\overline{M}\}, \overline{M}),$$

where $C_{\mathrm{loc}} := \left(\frac{1}{c_{\mathrm{reg}}^2} - \frac{1}{C_{\mathrm{reg}}^2}\right)^{-1}$.

Proof of Proposition J.4. Define $\alpha := \varepsilon + \sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})$. Applying Lemma J.5 to the class $\mathcal{M} \cup \{\overline{M}\}$ for each choice of $\overline{M} \in \mathcal{M}'$, we obtain that

$$\sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \frac{1}{C_{\mathrm{reg}}^2} \sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}). \tag{107}$$

The growth condition (28) gives that

$$\sup_{\overline{M}\in\mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \geq \frac{1}{c_{\mathrm{reg}}^2}\sup_{\overline{M}\in\mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}).$$

Then rearranging (107) yields

$$\sup_{\overline{M}\in\mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq \left(\frac{1}{c_{\mathrm{reg}}^2}-\frac{1}{C_{\mathrm{reg}}^2}\right)^{-1}\cdot\sup_{\overline{M}\in\mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\},\overline{M}). \tag{108}$$

The result in the proposition statement follows by replacing ε with $\varepsilon\cdot C_{\mathrm{reg}}$. □

Proof of Proposition D.7. Fix $\overline{M}\in\mathcal{M}^+$ and $\varepsilon>0$, and let $p\in\Delta(\Pi)$ be a minimizer for $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})$. Fix any $M\in\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\}$. We bound the regret under p by considering two cases.

Case 1. If $\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\leq\varepsilon^2$, then $M\in\mathcal{H}_{p,\varepsilon}(\overline{M})$, and it follows that $\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^M(\pi)\right]\leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})$.

Case 2. For the second case, suppose that $\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]>\varepsilon^2$. We now compute

$$\begin{aligned}
\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^M(\pi)\right] &\leq \alpha+\mathbb{E}_{\pi\sim p}\left[f^{\overline{M}}(\pi_{\overline{M}})-f^M(\pi)\right]\\
&\leq \alpha+\mathbb{E}_{\pi\sim p}\left[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\right]+\frac{1}{2\gamma}+\frac{\gamma}{2}\cdot\mathbb{E}_{\pi\sim p}\left[(f^M(\pi)-f^{\overline{M}}(\pi))^2\right]\\
&\leq \alpha+\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})+\frac{1}{2\gamma}+\frac{\gamma}{2}\cdot\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right],
\end{aligned}$$

where the second inequality uses Young's inequality. Rearranging, we obtain

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\\
&\leq \alpha+\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})+\frac{1}{2\gamma}-\frac{\gamma}{2}\cdot\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\\
&\leq \alpha+\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})+\frac{1}{2\gamma}-\frac{\gamma\varepsilon^2}{2}.
\end{aligned}\tag{109}$$

Recalling that M can be any model in $\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\}$, we obtain

$$\begin{aligned}
\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\},\overline{M}) &\leq \max\left\{\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}),\ \alpha+\frac{1}{2\gamma}+\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})-\frac{\gamma\varepsilon^2}{2}\right\}\\
&= \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})+\max\left\{0,\ \alpha+\frac{1}{2\gamma}-\frac{\gamma\varepsilon^2}{2}\right\}.
\end{aligned}$$

□

Proof of Proposition D.8. We begin with the upper bound on the constrained DEC. Let $\varepsilon>0$ be fixed. Using Proposition D.6, we have

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\leq C_{\mathrm{loc}}\cdot\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\},\overline{M}),$$

where $\alpha = C_{\mathrm{reg}}\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq C_{\mathrm{reg}}\varepsilon + c^2_{\mathrm{reg}}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq C^2_{\mathrm{reg}}\big(\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})\big)$ (see Definition D.1). Next, for all $\gamma > 0$, using Proposition D.1, we have

$$\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\},\overline{M}) \leq 8\cdot\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}),\overline{M})\vee 0 + C^2_{\mathrm{reg}}\gamma\varepsilon^2\big) + 7C_{\mathrm{reg}}\varepsilon,$$

so that

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq 8C_{\mathrm{loc}}\cdot\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}),\overline{M})\vee 0 + C^2_{\mathrm{reg}}\gamma\varepsilon^2\big) + 7C_{\mathrm{loc}}C_{\mathrm{reg}}\varepsilon.$$

We now set

$$\gamma^\star = (16C^2_{\mathrm{reg}}C_{\mathrm{loc}})^{-1}\cdot\frac{\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})}{\varepsilon^2},$$

which satisfies $\gamma^\star \geq \frac{1}{16C^2_{\mathrm{reg}}C_{\mathrm{loc}}\cdot\varepsilon}$ and gives

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq 8C_{\mathrm{loc}}\cdot\textsf{r-dec}^{\mathrm{o}}_{\gamma^\star}(\mathcal{M}_\alpha(\overline{M}),\overline{M})\vee 0 + \frac{1}{2}\cdot\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\overline{M},\overline{M}) + (7C_{\mathrm{loc}}C_{\mathrm{reg}} + 1/2)\varepsilon,$$

or after rearranging,

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq 16C_{\mathrm{loc}}\cdot\textsf{r-dec}^{\mathrm{o}}_{\gamma^\star}(\mathcal{M}_\alpha(\overline{M}),\overline{M})\vee 0 + 2(7C_{\mathrm{loc}}C_{\mathrm{reg}} + 1/2)\varepsilon.$$

In addition, we have

$$\alpha \leq C^2_{\mathrm{reg}}(16C^2_{\mathrm{reg}}C_{\mathrm{loc}})\cdot\gamma^\star\varepsilon^2 = C^2_{\mathrm{reg}}(16C^2_{\mathrm{reg}}C_{\mathrm{loc}})\cdot\alpha(\varepsilon,\gamma^\star).$$

To conclude, we over-bound by maximizing over $\gamma^\star \geq \frac{1}{16C^2_{\mathrm{reg}}C_{\mathrm{loc}}\cdot\varepsilon}$.

For the lower bound on the constrained DEC, it is an immediate consequence of Proposition D.7 that for all $\varepsilon > 0$ and $\gamma > \sqrt{2}\cdot\varepsilon^{-1}$, letting $\alpha = \frac{\gamma\varepsilon^2}{4}$,

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}),\overline{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) + \max\left\{\alpha + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}, 0\right\} = \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}),$$

for all $\overline{M}\in\mathcal{M}^+$. Since we are free to maximize over $\gamma \geq \sqrt{2}\varepsilon^{-1}$, this establishes the result.

□

Proof of Proposition D.9. Consider the following model class $\mathcal{M}$, parametrized by $\alpha\in(0,1/2)$:

1. $\Pi = \mathbb{N}\cup\{\pi_\circ\}$.
2. We have $\mathcal{M} = \{M_a\}_{a\in\mathbb{N}}$. For each $a\in\mathbb{N}$, the model $M_a\in\mathcal{M}$ has rewards and observations defined as follows:
   (a) For $\pi\in\mathbb{N}$, $f^{M_a}(\pi) = \frac{1}{2} + \alpha\cdot\mathbb{1}\{\pi = a\}$, while $f^{M_a}(\pi_\circ) = 0$.
   (b) For all $\pi\in\Pi$, we have $r = f^{M_a}(\pi)$ almost surely under $r\sim M_a(\pi)$.
   (c) For $\pi\in\mathbb{N}$, we receive the observation $o = \perp$.
   (d) Selecting $\pi_\circ$ gives the observation $o\in\{0,1\}^{\mathbb{N}}$, where for each $i\in\mathbb{N}$, $o_i\sim\mathrm{Ber}(1/2 + \alpha\cdot\mathbb{1}\{a = i\})$ is drawn independently (thus, we have $\mathcal{O} = \{0,1\}^{\mathbb{N}}\cup\{\perp\}$).

Upper bound. We will show that there are constants $c, C > 0$ so that, for $\varepsilon > 0$,

$$\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \alpha \cdot \mathbb{1}\left\{\varepsilon \geq \sqrt{c} \cdot \alpha\right\} \leq C \cdot \varepsilon.$$

Since $\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \alpha$ for all $\varepsilon \geq 0$ (as the choice of $p = \mathbb{I}_1$ satisfies $\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \leq \alpha$ for all $M \in \mathcal{M}$), it suffices to show that for $\varepsilon < \sqrt{c} \cdot \alpha$ and for any $\overline{M} \in \mathrm{co}(\mathcal{M})$, we have $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) = 0$.

Given $\overline{M} \in \mathrm{co}(\mathcal{M})$ and $\varepsilon \leq 1/2$, we can write $\overline{M}(\pi) = \mathbb{E}_{M' \sim \nu}[M'(\pi)]$ for some $\nu \in \Delta(\mathcal{M})$. We define a distribution $p \in \Delta(\Pi)$ according to the following cases:

- If ν puts mass at most $2/3$ on each model $M \in \mathcal{M}$, we define $p = \mathbb{I}_{\pi_\circ}$.
- Otherwise, there is a unique choice for $a^\star \in [A]$ so that $\nu(M_{a^\star}) \geq 2/3$, and in this case, we define $p = \mathbb{I}_{a^\star}$.

Now consider any model $M_a \in \mathcal{M}$. Consider the first case above, and write $o \sim M_a(\pi_\circ)$ and $\bar{o} \sim \overline{M}(\pi_\circ)$. Note that $o_a \sim \mathrm{Ber}(1/2 + \alpha)$, while $\bar{o}_a \sim \mathrm{Ber}(1/2 + \beta)$ for some $\beta \leq \frac{2}{3} \cdot \alpha$. It follows that

$$\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\right] = D^2_{\mathsf{H}}\big(M_a(\pi_\circ), \overline{M}(\pi_\circ)\big) \geq D^2_{\mathsf{H}}(\mathrm{Ber}(1/2 + \alpha), \mathrm{Ber}(1/2 + \beta)) \geq c \cdot \alpha^2,$$

for a numerical constant $c > 0$. Since $c\alpha^2 > \varepsilon^2$, it follows that $M_a \notin \mathcal{H}_{p,\varepsilon}(\overline{M})$; since the choice of M_a is arbitrary, we conclude that $\mathcal{H}_{p,\varepsilon}(\overline{M}) = \varnothing$ in this case.

Now, consider the second case above. For $a = a^\star$, we have that $\mathbb{E}_{\pi \sim p}\left[f^{M_a}(\pi_{M_a}) - f^{M_a}(\pi)\right] = 0$. For $a \neq a^\star$, we have that $\mathbb{P}_{r \sim M_a(a^\star)}(r \neq 1/2) = 0$, while $\mathbb{P}_{r \sim \overline{M}(a^\star)}(r \neq 1/2) \geq 2/3$. As a result,

$$\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\right] = D^2_{\mathsf{H}}\big(M_a(a^\star), \overline{M}(a^\star)\big) \geq D^2_{\mathsf{H}}(\mathrm{Ber}(0), \mathrm{Ber}(2/3)) \geq 4/9 > \varepsilon^2, \tag{110}$$

meaning that $M_a \notin \mathcal{H}_{p,\varepsilon}(\overline{M})$.

Lower bound. Pick $A \geq 2$, and let $\overline{M} = \mathrm{Unif}(\{M_a\}_{a \in [A]})$. Given $p \in \Delta(\Pi)$, let $a = \arg\min_{a \in [A]} p(a)$, so that $p(a) \leq 1/A$. We observe that

$$\mathbb{E}_{\pi \sim p}[f^{M_a}(\pi_{M_a}) - f^{M_a}(\pi)] \geq \alpha(1 - 1/A) \geq \alpha/2$$

and

$$\mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\big] \leq p(\pi_\circ) \cdot D^2_{\mathsf{H}}\big(M_a(\pi_\circ), \overline{M}(\pi_\circ)\big) + 2p(a) + \sum_{i \in [A], i \neq a} p(i) D^2_{\mathsf{H}}\big(M_a(i), \overline{M}(i)\big).$$

For all $i \neq a$, we have $D^2_{\mathsf{H}}\big(M_a(i), \overline{M}(i)\big) \leq D^2_{\mathsf{H}}(\mathrm{Ber}(0), \mathrm{Ber}(1/A)) \leq 2/A$, so that

$$\sum_{i \in [A], i \neq a} p(i) D^2_{\mathsf{H}}\big(M_a(i), \overline{M}(i)\big) \leq 2/A.$$

As long as α is a sufficiently small numerical constant, we also have, using the tensorization property of the squared Hellinger distance,

$$\begin{aligned} D^2_{\mathsf{H}}\big(M_a(\pi_\circ), \overline{M}(\pi_\circ)\big) &\leq D^2_{\mathsf{H}}(\mathrm{Ber}(1/2+\alpha), \mathrm{Ber}(1/2+\alpha/A)) + (A-1)\cdot D^2_{\mathsf{H}}(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2+\alpha/A)) \\ &\leq c\cdot\left(\alpha^2 + (A-1)\cdot\frac{\alpha^2}{A^2}\right) \\ &\leq C\cdot\alpha^2, \end{aligned}$$

where $C, c > 0$ are numerical constants. Altogether, this gives

$$\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\big] \leq p(\pi_\circ)\cdot C\alpha^2 + 4/A.$$

We choose A large enough such that $4/A \leq \varepsilon^2/2$. There are now two cases to consider.

- If $p(\pi_\circ) \leq \frac{\varepsilon^2}{2C\alpha^2}$, then $M_a \in \mathcal{H}_{p,\varepsilon}(\overline{M})$, and
$$\mathbb{E}_{\pi\sim p}[f^{M_a}(\pi_{M_a}) - f^{M_a}(\pi)] \geq \frac{\alpha}{2}.$$
- If this is not the case, we have
$$\mathbb{E}_{\pi\sim p}\Big[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\Big] \geq \frac{1}{2}p(\pi_\circ) \geq \frac{\varepsilon^2}{4C\alpha^2}\wedge 1.$$

By combining these cases, we conclude that there are numerical constants $C, c > 0$ such that

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\}, \overline{M}) \geq C\cdot\alpha\mathbb{I}\{\varepsilon > c\cdot\alpha^{3/2}\}.$$

In particular, choosing $\alpha \propto \varepsilon^{2/3}$ gives $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\}, \overline{M}) \geq \Omega(\varepsilon^{2/3})$, while $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq O(\varepsilon)$.

□

Proof of Proposition D.10. By Proposition J.3, we have that for all $\varepsilon > 0$ and $\overline{M} \in \mathcal{M}^+$,

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\}, \overline{M}) \leq \textsf{p-dec}^{\mathrm{c,greedy}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\}, \overline{M}) = \textsf{p-dec}^{\mathrm{c,greedy}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \textsf{p-dec}^{\mathrm{c}}_{\sqrt{3}\varepsilon}(\mathcal{M}, \overline{M}) + 4\varepsilon.$$

□

Proof of Proposition D.11. By Assumption G.1, we have

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

Observe that since Hellinger distance satisfies the triangle inequality, we have that for all $\pi \in \Pi$,

$$\begin{aligned} \mathbb{E}_{M,M'\sim\mu}\big[D^2_{\mathsf{H}}\big(M(\pi), M'(\pi)\big)\big] &\leq 2\,\mathbb{E}_{M\sim\mu}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] + 2\,\mathbb{E}_{M'\sim\mu}\big[D^2_{\mathsf{H}}\big(M'(\pi), \overline{M}(\pi)\big)\big] \\ &= 4\,\mathbb{E}_{M\sim\mu}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \end{aligned}$$

As a result, we have

$$\begin{aligned}
\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M}) &\leq \sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim p,M\sim\mu}\Big[f^{M}(\pi_M)-f^{M}(\pi)-\frac{\gamma}{4}\cdot\mathbb{E}_{M'\sim\mu}\,D^2_{\mathsf{H}}\big(M(\pi),M'(\pi)\big)\Big]\\
&\leq \sup_{\nu\in\Delta(\mathcal{M})}\sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim p,M\sim\mu}\Big[f^{M}(\pi_M)-f^{M}(\pi)-\frac{\gamma}{4}\cdot\mathbb{E}_{M'\sim\nu}\,D^2_{\mathsf{H}}\big(M(\pi),M'(\pi)\big)\Big]\\
&\leq \sup_{\nu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\Big[f^{M}(\pi_M)-f^{M}(\pi)-\frac{\gamma}{4}\cdot\mathbb{E}_{M'\sim\nu}\,D^2_{\mathsf{H}}\big(M(\pi),M'(\pi)\big)\Big]\\
&= \sup_{\nu\in\Delta(\mathcal{M})}\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M},\nu).
\end{aligned}$$

For the second inequality in (40), it follows immediately from convexity of squared Hellinger distance that $\sup_{\nu\in\Delta(\mathcal{M})}\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M},\nu)\leq\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})$.

We now prove (41). Let $\varepsilon>0$ be given. Since the strong regularity condition is satisfied relative to $\mathcal{M}^{+}$, Proposition J.4 with $\mathcal{M}'=\mathcal{M}^{+}$ implies that

$$\sup_{\overline{M}\in\mathcal{M}^{+}}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\leq C_{\mathrm{loc}}\cdot\sup_{\overline{M}\in\mathcal{M}^{+}}\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}_{\alpha}(\overline{M})\cup\{\overline{M}\},\overline{M}),$$

where $\alpha:=C^2_{\mathrm{reg}}\cdot\big(\varepsilon+\sup_{\overline{M}\in\mathcal{M}^{+}}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\big)$. Now, let

$$\widetilde{\mathcal{M}}_{\alpha}(\overline{M})=\Big\{M\in\mathcal{M}\mid f^{M}(\pi_M)\leq f^{\overline{M}}(\pi_{\overline{M}})+\alpha,f^{\overline{M}}(\pi_{\overline{M}})\leq f^{M}(\pi_{\overline{M}})+\alpha\Big\}.$$

Using Lemma J.2, along with the fact that $\alpha\geq C^2_{\mathrm{reg}}\cdot\varepsilon\geq\sqrt{2}C_{\mathrm{reg}}\cdot\varepsilon$ since $C_{\mathrm{reg}}\geq\sqrt{2}$, we have that

$$\sup_{\overline{M}\in\mathcal{M}^{+}}\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}_{\alpha}(\overline{M})\cup\{\overline{M}\},\overline{M})\leq\sup_{\overline{M}\in\mathcal{M}^{+}}\textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}C_{\mathrm{reg}}\cdot\varepsilon}(\widetilde{\mathcal{M}}_{\alpha}(\overline{M})\cup\{\overline{M}\},\overline{M})+\sqrt{2}C_{\mathrm{reg}}\varepsilon.$$

Let $\widetilde{M}$ be the model in $\mathcal{M}^{+}$ that attains the maximum in the right-hand side above, and set $\mathcal{M}'=\widetilde{\mathcal{M}}_{\alpha}(\widetilde{M})$. Let $\gamma>0$ be fixed. Using Proposition D.1, we have

$$\textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}C_{\mathrm{reg}}\varepsilon}(\mathcal{M}'\cup\{\widetilde{M}\},\widetilde{M})\leq 8\sup_{\overline{M}\in\mathcal{M}^{+}}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}',\overline{M})\vee 0+16C^2_{\mathrm{reg}}\gamma\varepsilon^2+7\sqrt{2}C_{\mathrm{reg}}\varepsilon.$$

By Eq. (40), we have

$$\sup_{\overline{M}\in\mathcal{M}^{+}}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}',\overline{M})\leq\sup_{\nu\in\Delta(\mathcal{M}')}\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}',\nu).$$

Consider any $\nu\in\Delta(\mathcal{M}')$ and let $\overline{M}_{\nu}:=\mathbb{E}_{M'\sim\nu}[M']\in\mathrm{co}(\mathcal{M}')\subseteq\mathrm{co}(\mathcal{M})$. Observe that if $M\in\mathcal{M}'=\widetilde{\mathcal{M}}_{\alpha}(\widetilde{M})$, then

$$f^{M}(\pi_M)\leq f^{\widetilde{M}}(\pi_{\widetilde{M}})+\alpha\leq\mathbb{E}_{M'\sim\nu}\Big[f^{M'}(\pi_{\widetilde{M}})\Big]+2\alpha\leq\max_{\pi\in\Pi}\mathbb{E}_{M'\sim\nu}\Big[f^{M'}(\pi)\Big]+2\alpha=f^{\overline{M}_{\nu}}(\pi_{\overline{M}_{\nu}})+2\alpha.$$

Hence, $\mathcal{M}'\subseteq\mathcal{M}_{2\alpha}(\overline{M}_{\nu})$, and we have the upper bound

$$\sup_{\nu\in\Delta(\mathcal{M}')}\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}',\nu)\leq\sup_{\nu\in\Delta(\mathcal{M})}\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}_{2\alpha}(\overline{M}_{\nu}),\nu).$$

For any $\nu \in \Delta(\mathcal{M})$, Proposition J.1 implies that

$$\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}_{2\alpha}(\overline{M}_\nu), \nu) \le \textsf{r-dec}^{\mathrm{c,rnd}}_{4\sqrt{\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{2}{\gamma}.$$

Putting everything together, this establishes that for all $\gamma > 0$,

$$\begin{aligned}\sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) &\le C_{\mathrm{loc}} \cdot \left(8 \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{4\sqrt{\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + 8\sqrt{2}C_{\mathrm{reg}}\varepsilon + 16\gamma C_{\mathrm{reg}}^2\varepsilon^2 + \frac{16}{\gamma}\right) \\ &\le C_{\mathrm{loc}} \cdot \left(8 \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{4\sqrt{\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + 24C_{\mathrm{reg}}^2\gamma\varepsilon^2 + \frac{24}{\gamma}\right).\end{aligned}$$

We choose $\gamma = \frac{1}{24C_{\mathrm{loc}}C_{\mathrm{reg}}^4} \cdot \frac{\alpha}{\varepsilon^2}$. Since $\varepsilon^2/\alpha \le \varepsilon$, this gives

$$\begin{aligned}\sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) &\le c_1 \cdot \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{c_2\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{1}{2C_{\mathrm{reg}}^2}\alpha + c_4\varepsilon, \\ &\le c_1 \cdot \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{c_2\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{1}{2} \sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + c_3\varepsilon,\end{aligned}$$

where $c_1, c_2, c_3, c_4 > 0$ are constants that depend only on C_{reg} and C_{loc}. Rearranging yields the first inequality in Eq. (41); the second inequality now follows from Jensen's inequality.

□

Appendix K. Omitted Proofs from Appendix E

Proof of Corollary E.1. The lower bound is an immediate corollary of Proposition D.8, so let us prove the upper bound. Let $\varepsilon > 0$ be fixed. Using Proposition D.6, we have

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le C_{\mathrm{loc}} \cdot \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_\alpha(\overline{M}), \overline{M}),$$

where $\alpha = C_{\mathrm{reg}}^2 \cdot \big(\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})\big)$. Recall that we assume $C_{\mathrm{reg}}, C_{\mathrm{loc}} = O(1)$. Hence, for all $\gamma > 0$ be fixed, using Proposition D.1, we have

$$\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_\alpha(\overline{M}), \overline{M}) \le O\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}), \overline{M}) \vee 0 + \gamma\varepsilon^2 + \varepsilon\big).$$

Proposition D.1 also gives

$$\alpha = O(\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})) \le O(\varepsilon + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2) \le O(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2 + \gamma^{-1}) = \overline{\alpha}(\varepsilon, \gamma),$$

where the second inequality is AM-GM. This establishes the result. □

Proof for Example E.1. We first lower bound the quantity in Eq. (44), then prove an upper bound on the regret of $\textsf{E2D}^+$.

Lower bound on offset DEC and regret bound from Eq. (44). We start with a basic lower bound on the offset DEC for the class $\mathcal{M}^{\alpha,\beta}$.

Lemma K.1. *Let $\alpha \in (0, 1/4)$, $\beta \in (0, 1)$ and $A \geq 2$ be given. For all $\gamma > 0$,*

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}, \widetilde{M}) = \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}_{\alpha}(\widetilde{M}), \widetilde{M}) \geq \frac{\alpha}{2 + 8\gamma\beta} - 4\gamma/A.$$

We now prove a lower bound on the quantity

$$R := \min_{\gamma>0} \max\left\{ \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\overline{\Delta}(\gamma,T)}(\overline{M}), \overline{M}) \cdot T,\ \gamma \cdot \log|\mathcal{M}| \right\}$$

appearing in Eq. (44). We begin by lower bounding the localization radius

$$\overline{\Delta}(\gamma, T) = \Omega\left(\frac{\gamma}{T} + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M})\right).$$

We choose $\alpha_1 = 1/2$ and $A = T^2$, so that whenever T is a sufficiently large constant, Lemma K.1 gives

$$\overline{\Delta}(\gamma, T) \geq \Omega\left(\frac{\gamma}{T} + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha_1,\beta}, \widetilde{M})\right) \geq \Omega\left(\frac{\gamma}{T} + \frac{1}{1+\gamma\beta} - 4\gamma/A\right) \geq \Omega\left(\frac{\gamma}{T} + \frac{1}{1+\gamma\beta}\right) \geq \Omega\left(\sqrt{\frac{1}{\beta T}} \wedge 1\right).$$

It follows that as long as $\beta \geq 1/T$, if we set

$$\alpha_2 = c \cdot \sqrt{\frac{1}{\beta T}},$$

where c is a sufficiently small numerical constant, then regardless of how γ is chosen,

$$\mathcal{M}^{\alpha_2,\beta}(\widetilde{M}) \subseteq \mathcal{M}_{\overline{\Delta}(\gamma,T)}(\widetilde{M}),$$

and

$$R \geq \min_{\gamma>0} \max\left\{ \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha_2,\beta}(\widetilde{M}), \widetilde{M}) \cdot T,\ \gamma \right\}.$$

Applying Lemma K.1 once more, we have

$$R \geq \Omega\left(\min_{\gamma>0}\left\{\frac{\alpha_2 T}{1+\gamma\beta} + \gamma\right\}\right) \geq \Omega\left(\alpha_2 T \wedge \sqrt{\frac{\alpha_2 T}{\beta}}\right) \geq \Omega\left(\beta^{-1/2}T^{1/2} \wedge \beta^{-3/4}T^{1/4}\right).$$

We set $\beta = T^{-1/2}$, which gives

$$R \geq \Omega(T^{5/8}),$$

as desired.

Upper bound on constrained DEC and regret of $\textsf{E2D}^{+}$. We now bound the regret of $\textsf{E2D}^{+}$ via Theorem C.2. We first bound the constrained DEC.

Lemma K.2. *Let $\beta \in (0, 1)$ be given, and let $\mathcal{M}^{\mathrm{all}} := \cup_{\alpha\in(0,1/2]}\mathcal{M}^{\alpha,\beta}$. Then for all $\varepsilon > 0$,*

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}^{\mathrm{all}}) \leq O\left(\frac{\varepsilon^2}{\beta}\right).$$

Let $\beta \propto T^{-1/2}$ and $A \propto T^2$ as in the prequel. Plugging the bound from Lemma K.2 into Theorem C.2 (cf. Eq. (90)) gives

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq \widetilde{O}\left(\frac{\bar{\varepsilon}(T)^2}{\beta} \cdot T + \sqrt{T}\right) = \widetilde{O}\left(\sqrt{T}\right),$$

since, with the usual choice of estimation oracle, we can take $\bar{\varepsilon}(T) \leq \widetilde{O}\left(\sqrt{\frac{\log|\mathcal{M}|}{T}}\right)$ and $\log|\mathcal{M}| \leq O(\log(A)) \leq O(\log(T))$.

□

Proof of Lemma K.1. We first remark that $\mathcal{M}^{\alpha,\beta} = \mathcal{M}^{\alpha,\beta}_{\alpha}(\widetilde{M})$, since $f^{\widetilde{M}}(\pi) = 1/2$ for all $\pi \in [A]$, and all $M \in \mathcal{M}^{\alpha,\beta}$ have $f^M(\pi_M) \leq 1/2 + \alpha$.

We now lower bound the value of the offset DEC. Consider any distribution $p \in \Delta(\Pi)$, and let $i := \arg\min_{i\in[A]} p(i)$, so that $p(i) \leq 1/A$. We have

$$\mathbb{E}_{\pi\sim p}\left[f^{M_{\alpha,i}}(\pi_{M_{\alpha,i}}) - f^{M_{\alpha,i}}(\pi)\right] \geq \alpha \cdot (1 - 1/A - p(\pi_\circ)) + \frac{1}{4}p(\pi_\circ) \geq \frac{\alpha}{2},$$

since $\alpha \leq 1/4$ and $A \geq 2$. We now bound the Hellinger distance via

$$\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi), \widetilde{M}(\pi)\right)\right] \leq p(\pi_\circ) \cdot D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi_\circ), \widetilde{M}(\pi_\circ)\right) + 2p(i).$$

Observe that $D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi_\circ), \widetilde{M}(\pi_\circ)\right) \leq 2\beta$ and $p(i) \leq 1/A$, so that

$$\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi), \widetilde{M}(\pi)\right)\right] \leq 2\beta \cdot p(\pi_\circ) + 2/A.$$

Combining the calculations so far gives

$$\mathbb{E}_{\pi\sim p}\left[f^{M_{\alpha,i}}(\pi_{M_{\alpha,i}}) - f^{M_{\alpha,i}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi), \widetilde{M}(\pi)\right)\right] \geq \frac{\alpha}{2} - 2\gamma\beta p(\pi_\circ) - 4\gamma/A.$$

On the other hand, by choosing $M = \widetilde{M}$, we have

$$\mathbb{E}_{\pi\sim p}\left[f^{\widetilde{M}}(\pi_{\widetilde{M}}) - f^{\widetilde{M}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\left(\widetilde{M}(\pi), \widetilde{M}(\pi)\right)\right] = \frac{1}{2}p(\pi_\circ),$$

so that

$$\begin{aligned}\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}, \widetilde{M}) &\geq \min_{p\in\Delta(\Pi)} \max\left\{\frac{\alpha}{2} - 2\gamma\beta p(\pi_\circ), \frac{1}{2}p(\pi_\circ)\right\} - 4\gamma/A, \\ &\geq \frac{\alpha}{2 + 8\gamma\beta} - 4\gamma/A.\end{aligned}$$

□

Proof of Lemma K.2. Let $\overline{M} \in \mathrm{co}(\mathcal{M}^{\mathrm{all}})$ and $\varepsilon \leq 1/10$ be given. Assume that $25\frac{\varepsilon^2}{\beta} \leq 1/2$, as the result is trivial otherwise.

Let $i = \arg\max_{i\in[A]} \mathbb{P}_{o\sim\overline{M}(\pi_\circ)}(o = i)$ and set $p = (1 - 25\frac{\varepsilon^2}{\beta})q + 25\frac{\varepsilon^2}{\beta}\mathbb{I}_{\pi_\circ}$, where $q \in \Delta([A])$ is another distribution whose value will be chosen shortly. We first observe that if $M_{\alpha,j} \in \mathcal{M}^{\alpha,\beta} \subseteq \mathcal{M}^{\mathrm{all}}$ for some $\alpha > 0$ and $j \neq i$, then since $\mathbb{P}_{o\sim\overline{M}(\pi_\circ)}(o = j) \leq \beta/2$, we have

$$D^2_{\mathsf{H}}\big(M_{\alpha,j}(\pi_\circ), \overline{M}(\pi_\circ)\big) \geq \left(\sqrt{\mathbb{P}_{o\sim M_{\alpha,j}(\pi_\circ)}(o = j)} - \sqrt{\mathbb{P}_{o\sim\overline{M}(\pi_\circ)}(o = j)}\right)^2 \geq (\sqrt{\beta} - \sqrt{\beta/2})^2 \geq \frac{\beta}{20},$$

where we have used that $\mathbb{P}_{o\sim M_{\alpha,j}(\pi_\circ)}(o \neq\perp) = \mathbb{P}_{o\sim\overline{M}(\pi_\circ)}(o \neq\perp) = \beta$, since $\overline{M} \in \mathrm{co}(\mathcal{M}^{\mathrm{all}})$. It follows that regardless of how $q \in \Delta([A])$ is chosen, $M_{\alpha,j} \notin \mathcal{H}_{p,\varepsilon}(\overline{M})$, since

$$\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M_{\alpha,j}(\pi), \overline{M}(\pi)\big)\big] \geq \frac{25\beta}{20\beta}\varepsilon^2 > \varepsilon^2.$$

Hence, if we define $\mathcal{M}'_i = \{M_{\alpha,i} \in \mathcal{M}^{\mathrm{all}} \mid \alpha \in (0, 1/2]\}$, we have $\mathcal{H}_{p,\varepsilon}(\overline{M}) \cup \{\overline{M}\} \subseteq \mathcal{M}'_i \cup \{\widetilde{M}\} \cup \{\overline{M}\}$, and it remains to choose q such that the regret on all of these models is small. We note that $\mathbb{E}_{\pi\sim p}\big[g^{\widetilde{M}}\big] \leq 25\frac{\varepsilon^2}{\beta}$ regardless of how q is chosen, so we restrict our attention to $\overline{M}$ and $\mathcal{M}'_i$ going forward.

Let $M^\star = \arg\min_{M\in\mathcal{M}'_i} D^2_{\mathsf{H}}\big(M(i), \overline{M}(i)\big)$. We will show that

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(i) \leq D^2_{\mathsf{H}}\big(M^\star(i), \overline{M}(i)\big). \tag{111}$$

To establish this fact, first note that if $\pi_{\overline{M}} = i$, then (111) is immediate. Otherwise, let $\nu_{\overline{M}} \in \Delta(\mathcal{M}^{\mathrm{all}})$ be such that $\overline{M}(\pi) = \mathbb{E}_{M'\sim\nu_{\overline{M}}}[M'(\pi)]$ for all $\pi \in \Pi$, and then

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(i) \;= \max_{\pi\in[A]} \mathbb{E}_{M'\sim\nu_{\overline{M}}}\Big[f^{M'}(\pi) - f^{M'}(i)\Big] \leq \frac{1}{2}\cdot\mathbb{P}_{M'\sim\nu_{\overline{M}}}\left(M' \notin \mathcal{M}'_i\right) \leq \frac{1}{2}\cdot\mathbb{P}_{r\sim\overline{M}(i)}\left(r = 1/2\right), \tag{112}$$

where the final inequality follows since all models $M' \in \mathcal{M}^{\mathrm{all}}\backslash\mathcal{M}'_i$ satisfy $r = 1/2$ a.s. when $r \sim M'(i)$. Recall the elementary fact that for all events A and distributions $\mathbb{P}$ and $\mathbb{Q}$.

$$\frac{(\mathbb{P}(A) - \mathbb{Q}(A))^2}{\mathbb{P}(A) + \mathbb{Q}(A)} \leq 2D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}). \tag{113}$$

Since $M^\star \in \mathcal{M}'_i$, we have $\mathbb{P}_{r\sim M^\star(i)}(r = 1/2) = 0$, and so, using (113), it follows that

$$\mathbb{P}_{r\sim\overline{M}(i)}(r = 1/2) \leq 2D^2_{\mathsf{H}}\big(\overline{M}(i), M^\star(i)\big),$$

and combining with (112) establishes (111).

To proceed, we choose $q \in \Delta([A])$ by setting $q(i) = \frac{4\varepsilon^2}{D^2_{\mathsf{H}}\big(M^\star(i),\overline{M}(i)\big)} \wedge 1$, and $q(\pi_{\overline{M}}) = 1 - q(i)$. We consider two cases

- If $q(i) = 1$, it is immediate that for all $M \in \mathcal{M}'_i$, $\mathbb{E}_{\pi\sim p}[g^M(\pi)] \leq 25\frac{\varepsilon^2}{\beta}$. In addition,

  $$\frac{4\varepsilon^2}{D^2_{\mathsf{H}}\big(M^\star(i), \overline{M}(i)\big)} \geq 1,$$

  so (111) implies that

  $$f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(i) \leq 4\varepsilon^2.$$

  It follows that $\mathbb{E}_{\pi\sim p}\big[g^{\overline{M}}(\pi)\big] \leq 25\frac{\varepsilon^2}{\beta} + (f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(i)) \leq 25\frac{\varepsilon^2}{\beta} + 4\varepsilon^2$.

- If $q(i) < 1$, then for all $M \in \mathcal{M}'_i$,

  $$\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \geq \frac{1}{2}q(i)D^2_{\mathsf{H}}\big(M(i), \overline{M}(i)\big) \geq \frac{1}{2}q(i)D^2_{\mathsf{H}}\big(M^\star(i), \overline{M}(i)\big) = 2\varepsilon^2,$$

  so $M \notin \mathcal{H}_{p,\varepsilon}(\overline{M})$. It follows that $\mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{M}'_i = \varnothing$. All that remains is to bound the regret under $\overline{M}$, which we do as follows:

  $$\begin{aligned}\mathbb{E}_{\pi\sim p}[g^{\overline{M}}(\pi)] \leq q(i)(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)) + 25\frac{\varepsilon^2}{\beta} &\leq q(i)\cdot D^2_{\mathsf{H}}\big(M^\star(i), \overline{M}(i)\big) + 25\frac{\varepsilon^2}{\beta} \\ &= 4\varepsilon^2 + 25\frac{\varepsilon^2}{\beta}.\end{aligned}$$

Putting the cases above together, we conclude that

$$\sup_{\overline{M}\in\mathrm{co}(\mathcal{M}^{\mathrm{all}})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}^{\mathrm{all}} \cup \{\overline{M}\}, \overline{M}) \leq O\bigg(\frac{\varepsilon^2}{\beta}\bigg).$$

□