Assouad, Fano, and Le Cam with Interaction: A Unifying Lower Bound Framework and Characterization for Bandit Learnability

Fan Chen
MIT
fanchen@mit.edu

Dylan J. Foster
Microsoft Research
dylanfoster@microsoft.com

Yanjun Han
New York University
yanjunhan@nyu.edu

Jian Qian
MIT
jianqian@mit.edu

Alexander Rakhlin
MIT
rakhlin@mit.edu

Yunbei Xu
National University of Singapore
yunbei@nus.edu.sg

Abstract

We develop a unifying framework for information-theoretic lower bound in statistical estimation and interactive decision making. Classical lower bound techniques—such as Fano's inequality, Le Cam's method, and Assouad's lemma—are central to the study of minimax risk in statistical estimation, yet are insufficient to provide tight lower bounds for *interactive decision making* algorithms that collect data interactively (e.g., algorithms for bandits and reinforcement learning). Recent work of Foster et al. [36, 38] provides minimax lower bounds for interactive decision making using seemingly different analysis techniques from the classical methods. These results—which are proven using a complexity measure known as the *Decision-Estimation Coefficient* (DEC)—capture difficulties unique to interactive learning, yet do not recover the tightest known lower bounds for passive estimation. We propose a unified view of these distinct methodologies through a new lower bound approach called *interactive Fano method*. As an application, we introduce a novel complexity measure, the *Decision Dimension*, which facilitates the new lower bounds for interactive decision making that extend the DEC methodology by incorporating the complexity of estimation. Using the Decision Dimension, we (i) provide a unified characterization of learnability for *any* structured bandit problem, (ii) close the remaining gap between the upper and lower bounds in Foster et al. [36, 38] (up to polynomial factors) for any interactive decision making problem in which the underlying model class is convex.

1 Introduction

The minimax criterion is a standard approach to studying the intrinsic difficulty of problems in statistics and machine learning. Stated (somewhat informally) as

$$\min_{\mathtt{ALG}} \max_{M \in \mathcal{M}} \mathsf{Cost}(\mathtt{ALG}, M), \tag{1}$$

where the algorithm ALG collects data (either passively or interactively) from the model M and incurs a cost, and the expression reflects the best cost that can be achieved by an algorithm ALG for a worst-case problem instance in a collection $\mathcal{M}$, measured according to an appropriate cost function Cost. In statistics, the minimax approach was pioneered by A. Wald [82], who made the connection to von Neumann's theory of games [76] and unified statistical estimation and hypothesis testing under the umbrella of *statistical decision theory*. Minimax optimality and minimax rates of convergence of

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

estimators have since become a central object in the modern of non-asymptotic statistics [74, 81]; here, for instance, ALG is an estimator of an unknown parameter based on noisy observations.

Upper bounds on the minimax value (1) are typically achieved by choosing a particular algorithm, while lower bounds often require specialized techniques. In statistics, three such techniques are widely used: Le Cam's two-point method, Fano's inequality, and Assouad's lemma. These techniques entail constructing "difficult" choices of subsets of the class $\mathcal{M}$. Le Cam's method focuses on two hypotheses, while Assouad's lemma and Fano's inequality involve multiple hypotheses indexed by the vertices of a hypercube and a simplex, respectively. The relationships between these methods are explored in Yu [87].

Classical statistical estimation is a purely passive task. A parallel line of research considers the task of *interactive decision making*, where ALG is a multi-round procedure that directly interacts with the data generating process and iteratively makes decisions with the (often contradictory) aims of minimizing cost and collecting information. Proving minimax lower bounds for interactive decision making problems presents unique challenges. The aforementioned lower bound techniques for estimation require quantifying the amount of information that can be gained from passively acquired data from a hard problem instance, but the amount information acquired by an *interactive* algorithm is harder to quantify [3, 61, 62], since it depends on the decisions made by the algorithm itself over multiple rounds.

In spite of the challenges, recent work of Foster et al. [36, 38] provides lower and upper bounds which show that a complexity measure known as the *Decision-Estimation Coefficient* (DEC) characterizes the minimax rates for a general class of interactive decision making problems (up to a gap which is related to the complexity of a certain induced estimation problem). Interestingly, the proof techniques in Foster et al. [36] proceed in a seemingly different fashion from classical lower bounds for statistical estimation; most notably, their techniques involve an *algorithm-dependent* (as opposed to oblivious) choice of a hard-to-distinguish alternative problem instance.

Given the differences between the classical Assouad, Fano, and Le Cam methods, and the even larger disparity between these methods and the interactive decision making techniques of Foster et al. [36, 38], it is natural to ask whether there is a hope of unifying these lower bounds techniques. Beyond the fundamental nature of this question, there is hope that a unified understanding might lead to tighter lower bounds, or even inspire new algorithms and upper bounds; of particular interest is to close the remaining gaps between the upper and lower bounds on the minimax rates for interactive decision making left open by Foster et al. [38], which are closely related to estimation.

Contributions. We present a new framework for information-theoretic lower bounds which allows for a unifying presentation of classical lower bounds in statistical estimation (Assouad, Fano, and Le Cam) and recent DEC-based lower bounds for interactive decision making [36, 38].

- **Interactive lower bound framework (Section 3).** Our main result is to introduce a new lower bound technique, the *interactive Fano method*. The interactive Fano method generalizes the stringent separation condition in the classical Fano inequality to a novel algorithm-dependent condition by introducing the concept of "ghost data" generated from a reference distribution. This technique recovers the Le Cam two-point method (and convex hull method), Assouad method, and Fano method as special cases. By virtue of being algorithm-dependent in nature, the interactive Fano method seamlessly recovers DEC-based lower bounds for interactive decision making as a special case, and leads to refined quantile-based variants.
- **Decision dimension and bandit learnability (Section 4).** As an application of the interactive Fano method, we derive lower bounds for interactive decision making based on a new complexity measure, the *decision dimension*, which quantifies the difficulty of *estimating* a near-optimal policy/decision, and complements the original DEC lower bounds (which reflect difficulty of exploration as opposed to difficulty of estimation). As an application, the decision dimension provides both lower and upper bound for learning any structured bandit problem, up to an exponential gap. In particular, finiteness of the decision dimension is the first necessary and sufficient condition for finite-time learnability of any structured bandit problem. As a secondary result, we use the decision dimension to close the remaining gap between the upper and lower bounds in Foster et al. [36, 38] (up to polynomial factors) for any interactive decision making problem in which the underlying model class is convex.

Related work. Due to space limitations, we discuss the related work in Appendix A.

1.1 Preliminaries

Let P and Q be two distributions over a space Ω such that P is absolutely continuous with respect to Q. Then, for a convex function $f : [0, +\infty) \to (-\infty, +\infty]$ such that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{x\to 0^+} f(x)$, the f-divergence of between P and Q is defined as

$$D_f(P, Q) := \int_\Omega f\left(\frac{dP}{dQ}\right) dQ.$$

Concretely, we make use of three well-known f-divergences: the KL-divergence D_{KL}, the squared Hellinger distance D^2_{H}, and the total variation distance D_{TV}, for which the function $f(x)$ is chosen to be $x \log x$, $\frac{1}{2}(\sqrt{x} - 1)^2$, and $\frac{1}{2}|x - 1|$ respectively. For a pair of random variables (X, Y) with joint distribution $P_{X,Y}$, the mutual information is defined as

$$I(X; Y) = \mathbb{E}_X\big[D_{\mathrm{KL}}\big(P_{Y|X} \,\|\, P_Y\big)\big],$$

where $P_{Y|X}$ is the conditional distribution of $Y|X$ and P_Y is the marginal distribution of Y.

2 Statistical Estimation and Interactive Decision Making

We work in a general framework we refer to as *Interactive Statistical Decision Making* (ISDM). We adopt this framework as a convenient formalism which encompasses statistical estimation and interactive decision making in a unified fashion.

Interactive Statistical Decision Making. An ISDM problem is specified by $(\mathcal{X}, \mathcal{M}, \mathcal{D}, L)$, where $\mathcal{X}$ is the space of outcomes, $\mathcal{M}$ is a model class (parameter space), $\mathcal{D}$ is the space of algorithms, and L is a non-negative risk function. For an algorithm $\mathtt{ALG} \in \mathcal{D}$ chosen by the learner and a model $M \in \mathcal{M}$ specified by the environment, an observation X is generated from a distribution induced by M and $\mathtt{ALG}$: $X \sim \mathbb{P}^{M,\mathtt{ALG}}$. The performance of the algorithm $\mathtt{ALG}$ on the model M is then measured by the risk function $L(M, X)$. The learner's goal is to minimize the risk by choosing the algorithm $\mathtt{ALG}$. As described in the Introduction, the best possible expected risk the learner may achieve is the following *minimax risk*:

$$\inf_{\mathtt{ALG}\in\mathcal{D}} \sup_{M\in\mathcal{M}} \mathbb{E}^{M,\mathtt{ALG}}[L(M, X)]. \tag{2}$$

While main our results concern the general problem formulation in (2), we focus on applications to statistical estimation and interactive decision making throughout. Below, we give additional background on these settings and show how to view them as special cases.

2.1 Statistical estimation

For a general statistical estimation framework known as statistical decision theory [11, 82], the learner is given the parameter space Θ, observation space $\mathcal{Y}$, decision space $\mathcal{A}$, and a loss function L. For an underlying parameter $\theta^\star \in \Theta$, n i.i.d. samples $Y_1, ..., Y_n \sim P_{\theta^\star}$ are drawn and observed by the learner. The learner then chooses a decision $A = A(Y_1, \cdots, Y_n) \in \mathcal{A}$ based on the observations, and then incurs the loss $L(\theta^\star, A)$. This framework subsumes most statistical estimation problems.

Any general statistical estimation problem can be viewed as a ISDM instance, by choosing the model class as $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ and the algorithm space as $\mathcal{D} = \{\mathtt{ALG} : \mathcal{Y}^{\otimes n} \to \mathcal{A}\}$. For model $M = P_\theta$ and algorithm $\mathtt{ALG}$, the distribution of the whole observation $X \sim \mathbb{P}^{M,\mathtt{ALG}}$ is given by

$$X = (Y_1, \cdots, Y_n, A), \qquad Y_1, ..., Y_n \overset{\text{i.i.d}}{\sim} P_\theta,\ A = \mathtt{ALG}(Y_1, \cdots, Y_n).$$

The loss under model M is then measured by the loss of the decision A, i.e., $L(M, X) := L(\theta, A)$.

2.2 Interactive decision making

For interactive decision making, we consider the following variant of the Decision Making with Structured Observations (DMSO) framework [36], which subsumes bandits and reinforcement learning. The learner interacts with the environment (described by an underlying model $M^\star : \Pi \to \Delta(\mathcal{O})$, unknown to the learner) for T rounds. For each round $t = 1, ..., T$:

- The learner selects a decision $\pi^t \in \Pi$, where Π is the decision space.
- The learner receives an observation $o^t \in \mathcal{O}$ via $o^t \sim M^\star(\pi^t)$, where $\mathcal{O}$ is the observation space.

The underlying model $M^\star$ is formally a conditional distribution, and the learner is assumed to have access to a known model class $\mathcal{M} \subseteq (\Pi \to \Delta(\mathcal{O}))$ with the following property.

Assumption 1 (Realizability)**.** *The model class $\mathcal{M}$ contains $M^\star$.*

The model class $\mathcal{M}$ represents the learner's prior knowledge of the structure of the underlying environment. For example, for structured bandit problems, the models specify the reward distributions and hence encode the structural assumptions on the mean reward function (e.g. linearity, smoothness, or concavity). For a more detailed discussion, see Appendix B.

To each model $M \in \mathcal{M}$, we associate a *risk* function $g^M : \Pi \to \mathbb{R}_{\geq 0}$, which measures the performance of a decision in Π. We consider two types of learning goals under the DMSO framework:

- Generalized no-regret learning: The goal of the agent is to minimize the *cumulative* sub-optimality during the course of the interaction, given by

$$\mathbf{Reg}_{\mathsf{DM}}(T) := \sum_{t=1}^{T} g^{M^\star}(\pi^t), \tag{3}$$

  where π^t can be randomly drawn from a distribution $p^t \in \Delta(\Pi)$ chosen by the learner at step t.

- Generalized PAC (Probably Approximately Correct) learning: the goal of the agent is to minimize the sub-optimality of a final output decision $\widehat{\pi}$ (possibly randomized), which is selected by the learner once all T rounds of interaction conclude. We measure performance via

$$\mathbf{Risk}_{\mathsf{DM}}(T) := g^{M^\star}(\widehat{\pi}). \tag{4}$$

With an appropriate choice for g^M, the setting captures reward maximization (regret minimization) [36, 38], model estimation and preference-based learning [19], multi-agent decision making and partial monitoring [33], and various other tasks. In the main text, we focus on reward maximization, and defer the results for more general choices g^M to the appendices (cf. Appendix B).

Example 1 (Reward maximization)**.** Let $R : \mathcal{O} \to [0, 1]$ be a known reward function.[1] For a model $M \in \mathcal{M}$, $\mathbb{E}^{M,\pi}[\cdot]$ denotes expectation under the process $o \sim M(\pi)$, and $f^M(\pi) := \mathbb{E}^{M,\pi}[R(o)]$ denotes the expected value function. An optimal decision is denoted by $\pi_M \in \arg\max_{\pi\in\Pi} f^M(\pi)$, and the sub-optimality measure is defined by $g^M(\pi) = f^M(\pi_M) - f^M(\pi)$.

DMSO as an instance of ISDM. Any DMSO class $(\mathcal{M}, \Pi)$ induces an ISDM as follows. For any $t \in [T]$, denote the full history of decisions and observations up to time t by $\mathcal{H}^{t-1} = (\pi^s, o^s)_{s=1}^{t-1}$. The space of observations $\mathcal{X}$ consists of all such X that $X = \mathcal{H}^T \cup \{\widehat{\pi}\}$, where $\widehat{\pi}$ is a final decision. An algorithm $\mathtt{ALG} = \{q^t\}_{t\in[T]} \cup \{p\}$ is specified by a sequence of mappings, where the t-th mapping $q^t(\cdot \mid \mathcal{H}^{t-1})$ specifies the distribution of π^t based on $\mathcal{H}^{t-1}$, and the final map $p(\cdot \mid \mathcal{H}^T)$ specifies the distribution of the *output decision* $\widehat{\pi}$ based on $\mathcal{H}^T$. The algorithm space $\mathcal{D}$ consists of all such algorithms. The loss function is chosen to be $L(M^\star, X) = \mathbf{Reg}_{\mathsf{DM}}(T)$ for no-regret learning (3), and $L(M^\star, X) = \mathbf{Risk}_{\mathsf{DM}}(T)$ for PAC learning (4). For any algorithm $\mathtt{ALG}$ and model M, $\mathbb{P}^{M,\mathtt{ALG}}(\cdot)$ is the distribution of $X = (\mathcal{H}^T, \widehat{\pi})$ generated by the algorithm $\mathtt{ALG}$ under the model M, and we let $\mathbb{E}^{M,\mathtt{ALG}}[\cdot]$ to be the corresponding expectation.

3 A General Lower Bound

In this section, we introduce our general lower bound technique, the interactive Fano method, and use it to provide minimax lower bounds for the ISDM framework.

Theorem 1 (Interactive Fano method)**.** *Fix an f-divergence D_f. Let $\mathtt{ALG}$ be a given algorithm, $\delta \in (0, 1)$ be a quantile parameter, and $\mu \in \Delta(\mathcal{M})$ be a prior distribution over models. For reference distribution $\mathbb{Q}$ on $\mathcal{X}$ and parameter $\Delta > 0$, we define*

$$\rho_{\Delta,\mathbb{Q}} = \mathbb{P}_{M\sim\mu, X\sim\mathbb{Q}}(L(M, X) < \Delta). \tag{5}$$

Then, the following lower bound holds:

$$\begin{aligned}\sup_{M\in\mathcal{M}} \mathbb{E}_{X\sim\mathbb{P}^{M,\mathtt{ALG}}}[L(M, X)] &\geq \mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathtt{ALG}}}[L(M, X)] \\ &\geq \delta \cdot \sup_{\mathbb{Q}\in\Delta(\mathcal{X}),\Delta>0} \left\{\Delta : \mathbb{E}_{M\sim\mu}\left[D_f\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{Q}\right)\right] < \mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}})\right\},\end{aligned}$$

where we denote $\mathsf{d}_{f,\delta}(p) = D_f(\mathrm{Bern}(1-\delta), \mathrm{Bern}(p))$ if $p \leq 1-\delta$, and $\mathsf{d}_{f,\delta}(p) = 0$ otherwise.

[1] We assume R is known without loss of generality, since the observation o may have a component containing the random reward.

This result generalizes existing statements of Fano's method in multiple ways:

- It encompasses general interactive learning/estimation problems in the ISDM framework, as opposed to purely passive estimation. This is reflected in the fact that the distribution over the outcome X is allowed to depend on ALG itself.
- The most important and novel change is that Theorem 1 generalizes the "hard" separation condition required in the classical Fano method to a "soft" notion of separation captured by the quantile $\rho_{\Delta,\mathbb{Q}}$ in (5). The quantile $\rho_{\Delta,\mathbb{Q}}$ reflects the average separation under "ghost data" X generated from an arbitrary reference distribution $\mathbb{Q}$, which is independent of the true model $M \sim \mu$.
- In addition, instead of relying on mutual information, which is can difficult to quantify for interactive problems, we use divergence with respect to the reference distribution $\mathbb{Q}$, generalizing a central idea in Foster et al. [36, 38].

In what follows, we will show that these generalizations allow the Interactive Fano method to achieve two important desiderata: (1) unifying the methods of Fano, Le Cam, and Assouad (Section 3.1), and (2) integrating these traditional lower bound techniques with contemporary interactive decision making lower bounds to derive new lower bound (see Section 3.2).

3.1 Recovering non-interactive lower bounds

We begin by applying Theorem 1 to recover classical non-interactive lower bounds for statistical estimation. Since a goal of our paper is to integrate the Fano and Assouad methods with the DEC framework, this serves as an important sanity check to demonstrate that our framework can recover the non-interactive versions of these methods.

Fano's method. To recover the Fano method, we specialize Theorem 1 to the KL divergence. Observe that for any reference distribution $\mathbb{Q}$,

$$\mathbb{P}_{M\sim\mu, X\sim\mathbb{Q}}(L(M,X) < \Delta) \le \sup_x \mu(M \in \mathcal{M} \mid L(M,x) < \Delta).$$

By choosing $\mathbb{Q} = \mathbb{E}_{M\sim\mu}\mathbb{P}^{M,\mathsf{ALG}}$ in Theorem 1, we obtain the following proposition, which encompasses prior generalizations of Fano's inequality [88, 32, 23] developed in statistical estimation.

Proposition 2 (Recovering the generalized Fano method)**.** *Fix an algorithm* ALG *and prior distribution $\mu \in \Delta(\mathcal{M})$, and let $I_{\mu,\mathsf{ALG}}(M;X)$ be the mutual information between M and X under $M \sim \mu$ and $X \sim \mathbb{P}^{M,\mathsf{ALG}}$. The following Bayes risk lower bound holds for all $\Delta \ge 0$:*

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}\left[L(M,X)\right] \ge \Delta\left(1 + \frac{I_{\mu,\mathsf{ALG}}(M;X)+\log 2}{\log \sup_x \mu(M\in\mathcal{M}|L(M,x)<\Delta)}\right). \tag{6}$$

When applied to the statistical estimation setting (Section 2.1), the classical Fano inequality corresponds to the special case of Proposition 2 where $\Theta = \mathcal{A} = \{1, 2, \dots, m\}$, $L(\theta, a) = \mathbb{1}(\theta \neq a)$ is the indicator loss, $\mu = \mathrm{Unif}(\Theta)$ is the uniform prior, and $\Delta = 1$.

Note that in Proposition 2, the term $\log \sup_x \mu(M \in \mathcal{M} : L(M,x) < \Delta)$ in the denominator of (6) takes the supremum over the outcome x, resulting in a simplified expression that removes the role of the algorithm ALG. This simplification is often sufficient to derive tight guarantees for estimation, but is insufficient for interactive decision making in general. The DEC, which we define in Section 3.2, more precisely accounts for the role of decisions selected by the algorithm.

Le Cam's method and Assouad's method. To recover Le Cam's two-point method and Assouad's method from Theorem 1, we appeal to the following result, which recovers a more general lower bound known as the Le Cam convex hull method [54, 87].

Proposition 3 (Recovering Le Cam's convex hull method)**.** *For a parameter space Θ and observation space $\mathcal{Y}$, consider a class of distributions $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ indexed by Θ, $P_\theta \in \Delta(\mathcal{Y}^{\otimes n})$. Let $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$ be a loss function. Suppose $\Theta_0 \subseteq \Theta$ and $\Theta_1 \subseteq \Theta$ satisfy the* separation condition

$$L(\theta_0, a) + L(\theta_1, a) \ge 2\Delta, \quad \forall a \in \mathcal{A}, \theta_0 \in \Theta_0, \theta_1 \in \Theta_1.$$

Suppose that there exist probability measures $\nu_0 \in \Delta(\Theta_0)$ and $\nu_1 \in \Delta(\Theta_1)$ such that

$$D_{\mathrm{TV}}\left(\nu_0 \otimes P_\theta, \nu_1 \otimes P_\theta\right) \le 1/2,$$

where $\nu_i \otimes P_\theta$ is the distribution on $\mathcal{Y}^{\otimes n}$ induced by $\theta \sim \nu_i, Y_1, \dots, Y_n \sim P_\theta$ for $i \in \{0,1\}$. Then

$$\inf_{\mathsf{ALG}} \sup_{\theta\in\Theta} \mathbb{E}_{Y_1,\dots,Y_n\sim P_\theta} L(\theta, \mathsf{ALG}(Y_1, \dots, Y_n)) \ge \Delta/4,$$

where the infimum is taken over all algorithms $\mathsf{ALG} : \mathcal{Y}^{\otimes n} \to \mathcal{A}$.

Le Cam's convex hull method is the most general formulation of the Le Cam two-point method, which—in its most basic form—corresponds to the case in which ν_0 and ν_1 are singletons. The convex hull method is also capable of recovering Assouad's method [87]. It is important to note that the classical Fano inequality, e.g. in the form of Proposition 2, cannot recover Proposition 3. This is because of fundamental differences between the divergences (KL versus TV) used in the traditional Fano method and convex hull method.

3.2 Recovering DEC-based lower bounds for interactive decision making

Within the DMSO framework (Section 2.2), Foster et al. [36, 38] introduced the *Decision-Estimation Coefficient* (DEC) as a complexity measure, providing both upper and lower bounds for any model class $\mathcal{M}$. We now show how to recover the lower bounds of Foster et al. [36, 38] through Theorem 1. We focus on the lower bounds from Foster et al. [38], which are based on a variant of the DEC called the *constrained DEC*, and provide the tightest guarantees from prior work.

Background on the Decision-Estimation Coefficient. Consider the reward maximization setting (Example 1) under DMSO. For a model class $\mathcal{M}$ and a reference model $\overline{M} : \Pi \to \Delta(\mathcal{O})$ (not necessarily in $\mathcal{M}$), we define the constrained regret-DEC via

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) := \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] \,|\, \mathbb{E}_{\pi \sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\}, \tag{7}$$

and define the constrained PAC-DEC via

$$\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) := \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] \,|\, \mathbb{E}_{\pi \sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\}. \tag{8}$$

Here, the superscript "c" indicates "constrained", and the superscript "r" (resp. "p") indicates "regret" (resp. "PAC"). We further define

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

where $\mathrm{co}(\mathcal{M})$ denotes the convex hull of the model class $\mathcal{M}$.

Based on these complexity measures, Foster et al. [38] (see also Glasgow and Rakhlin [40]) provide the following lower and upper bounds on optimal risk and regret, under mild growth conditions on the DECs.

Theorem 4 (Informal; Foster et al. [38])**.** *Consider the reward maximization variant of the DMSO setting (Example 1). For any model class $\mathcal{M}$ and $T \in \mathbb{N}$, the following lower and upper bounds hold:*
(1) For PAC learning,

$$\mathsf{p\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \lesssim \inf_{\mathsf{ALG}} \sup_{M \in \mathcal{M}} \mathbb{E}^{M, \mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \lesssim \mathsf{p\text{-}dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}),$$

where $\underline{\varepsilon}(T) \asymp \sqrt{1/T}$ and $\bar{\varepsilon}(T) \asymp \sqrt{\log|\mathcal{M}|/T}$ (up to logarithmic factors).

(2) For no-regret learning,

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot T \lesssim \inf_{\mathsf{ALG}} \sup_{M \in \mathcal{M}} \mathbb{E}^{M, \mathsf{ALG}}[\mathbf{Reg}_{\mathsf{DM}}(T)] \lesssim \mathsf{r\text{-}dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}) \cdot T + T \cdot \bar{\varepsilon}(T).$$

Therefore, up to the $\log|\mathcal{M}|$-gap between the parameters $\underline{\varepsilon}(T)$ and $\bar{\varepsilon}(T)$ appearing in the lower and upper bounds, the constrained PAC-DEC tightly captures the minimax risk of PAC learning, and the constrained regret-DEC captures the minimax regret of no-regret learning.

A new complexity measure: The quantile Decision-Estimation Coefficient. We recover the DEC-based lower bounds from Foster et al. [38] through a new variant we refer to as the *quantile DEC*. To do so, we briefly recount the proof technique used by Foster et al. [38].

Given an algorithm $\mathtt{ALG}$, the proof strategy is to first fix an arbitrary *reference model* $\overline{M}$, then adversarially choose a hard *alternative model* $M \in \mathcal{M}$ (in a way that is guided by the DEC and the algorithm $\mathtt{ALG}$ itself) such that $D_{\mathrm{TV}}(\mathbb{P}^{M, \mathtt{ALG}}, \mathbb{P}^{\overline{M}, \mathtt{ALG}})$ is small, yet $\mathtt{ALG}$ cannot achieve low risk on model M. This lower bound technique does not explicitly require a separation condition between M and $\overline{M}$, which is a departure from the classical Fano and two-point methods. Thus to recover it, the lack of an explicit separation condition in Theorem 1 will be critical. More precisely, for any model M, we consider the following distributions over decisions:

$$q_{M, \mathtt{ALG}} = \mathbb{E}^{M, \mathtt{ALG}}\left[\tfrac{1}{T}\textstyle\sum_{t=1}^{T} q^t(\cdot \mid \mathcal{H}^{t-1})\right] \in \Delta(\Pi), \quad p_{M, \mathtt{ALG}} = \mathbb{E}^{M, \mathtt{ALG}}\left[p(\mathcal{H}^T)\right] \in \Delta(\Pi). \tag{9}$$

That is, $q_{M,\mathsf{ALG}}$ is the expected empirical distribution over the decisions $(\pi^1, \cdots, \pi^T)$ played by the algorithm under M, and $p_{M,\mathsf{ALG}}$ is the expected distribution of the final decision $\widehat{\pi}$.

With these definitions, we instantiate Theorem 1 with the Hellinger distance. We will use the sub-additivity of Hellinger distance (Lemma C.1; Foster et al. [36, 39]), which allows us to bound

$$\tfrac{1}{2} D^2_{\mathrm{TV}}(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{P}^{\overline{M},\mathsf{ALG}}) \leq D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{P}^{\overline{M},\mathsf{ALG}}\right) \leq 7T \cdot \mathbb{E}_{\pi \sim p_{\overline{M},\mathsf{ALG}}}\left[D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)\right]. \tag{10}$$

Theorem 1 then yields the following intermediate result.

Lemma 5 (Interactive Fano method lower bound for interactive decision making)**.** *Let $\delta \in [0,1]$ be given, and consider an algorithm* ALG*. Define*

$$\Delta^\star_{\mathsf{ALG},\delta} := \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \sup_{M \in \mathcal{M}} \sup_{\Delta \geq 0} \left\{ \Delta : p_{\overline{M},\mathsf{ALG}}(\pi : g^M(\pi) \geq \Delta) > \delta + \sqrt{14T\, \mathbb{E}_{\pi \sim q_{\overline{M},\mathsf{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)} \right\}.$$

Then there exists $M \in \mathcal{M}$ such that $\mathbb{P}^{M,\mathsf{ALG}}\left(g^M(\widehat{\pi}) \geq \Delta^\star_{\mathsf{ALG},\delta}\right) \geq \delta$.

Using Lemma 5, as a starting point, we derive a new quantile-based variant of the DEC, which we will show can be viewed as a slight generalization of the original PAC DEC of Foster et al. [38].

For any model $M \in \mathcal{M}$ and any parameter $\delta \in [0,1]$, we define the δ-quantile risk as follows:

$$\widehat{g}^M_\delta(p) = \sup_{\Delta \geq 0}\{\Delta : \mathbb{P}_{\pi \sim p}(g^M(\pi) \geq \Delta) \geq \delta\};$$

this serves as a measure of the sub-optimality of the distribution $p \in \Delta(\Pi)$ in terms of δ-quantile. We now define the quantile PAC DEC as follows:

$$\mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M}) := \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \widehat{g}^M_\delta(p) \mid \mathbb{E}_{\pi \sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \leq \varepsilon^2 \right\}, \tag{11}$$

and define $\mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}) := \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$. Applying Lemma 5, it is immediate to see that the quantile PAC-DEC provides a lower bound on the PAC risk.

Theorem 6 (Quantile DEC lower bound)**.** *Let any $T \geq 1$ and $\delta \in [0,1)$ be given, and define $\underline{\varepsilon}_\delta(T) := \frac{1}{14}\sqrt{\frac{\delta}{T}}$. Then, for any algorithm* ALG*, there exists $M^\star \in \mathcal{M}$ such that*

$$\mathbb{P}^{M^\star,\mathsf{ALG}}\left(\mathbf{Risk}_{\mathsf{DM}}(T) \geq \mathsf{p\text{-}dec}^{\mathrm{q}}_{\underline{\varepsilon}_\delta(T),\delta}(\mathcal{M})\right) \geq \tfrac{\delta}{2}.$$

Unlike the original constrained DEC lower bounds (Theorem 4), which are restricted to the reward maximization variant of the DMSO setting (Example 1), the quantile DEC lower bound in Theorem 6 holds for *any suboptimal measure* g^M. As a result, the lower bound applies to a broader range of generalized PAC learning tasks, including model estimation [19] and multi-agent decision making [33], where DEC-based lower bounds from prior work are loose in general; as a concrete application, we derive a new lower bound for *interactive estimation* (Example 3) in Appendix E.2.

Recovering DEC-based lower bounds using the quantile DEC. At first glance, Theorem 6 might appear to be weaker than the constrained PAC-DEC lower bound in Theorem 4 due to the loose conversion from quantile risk to expected risk. However, by specializing to reward maximization (Example 3) and leveraging the structure of the sub-optimality measure g^M, we show that quantile PAC-DEC is equivalent to its constrained counterpart for this setting, leading to a tight lower bound.

Proposition 7 (Recovering the PAC DEC lower bound)**.** *Under the reward maximization setting (Example 1), for any $\varepsilon > 0$ and $\delta \in [0,1)$ it holds that*

$$\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \leq \mathsf{p\text{-}dec}^{\mathrm{q}}_{\sqrt{2}\varepsilon,\delta}(\mathcal{M}) + \tfrac{4\varepsilon}{1-\delta}.$$

As a corollary, we may choose $\delta = \frac{1}{2}$ and $\underline{\varepsilon}(T) = \frac{1}{20\sqrt{T}}$ in Theorem 6, so that

$$\sup_{M \in \mathcal{M}} \mathbb{E}^{M,\mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \geq \frac{1}{4}\mathsf{p\text{-}dec}^{\mathrm{q}}_{\sqrt{2}\underline{\varepsilon}(T),1/2}(\mathcal{M}) \geq \frac{1}{4}\left(\mathsf{p\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) - 8\underline{\varepsilon}(T)\right).$$

Thus, the quantile PAC-DEC lower bound indeed recovers the constrained PAC-DEC lower bound in Theorem 4.

Our quantile DEC lower bound extends to regret with minor modifications, allowing us to recover the regret lower bounds in Theorem 4. We defer the details to the Appendix E.1 (Theorem E.1).

3.3 Recovering mutual information-based lower bounds for interactive decision making

The following result uses Theorem 1 to extend classical Fano method to interactive decision making and achieves tight dependence on the problem dimension that is not recovered by the standard DEC lower bound in Foster et al. [36, 38].

Proposition 8 (Mutual information-based lower bound)**.** *Consider the DMSO setting. For any $T \geq 1$ and prior $\mu \in \Delta(\mathcal{M})$, we define the maximum T-round mutual information as*

$$I_\mu(T) := \sup_{\mathsf{ALG}} I_{\mu,\mathsf{ALG}}(M; \mathcal{H}^T),$$

where the supremum is taken over all possible DMSO algorithms ALG*. Then for any algorithm* ALG*,*

$$\sup_{M\in\mathcal{M}} \mathbb{E}^{M,\mathsf{ALG}}[g^M(\widehat{\pi})] \geq \frac{1}{2} \sup_{\mu\in\Delta(\mathcal{M})} \sup_{\Delta>0} \left\{ \Delta \mid \sup_{\pi} \mu(M : g^M(\pi) \leq \Delta) \leq \frac{1}{4}\exp(-2I_\mu(T)) \right\}.$$

Using Proposition 8, along with mutual information bounds from Rajaraman et al. [63], we recover a $\Omega(d/\sqrt{T})$ PAC lower bound for linear bandits in d dimensions, which in turn recovers the $\Omega(d\sqrt{T})$ regret lower bound [26, 66, 51, etc.].

Corollary 9. *For $d \geq 2$, consider the d-dimensional linear bandit problem with decision space $\Pi = \{\pi \in \mathbb{R}^d : \|\pi\|_2 \leq 1\}$, parameter space $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$, and model class $\mathcal{M} = \{M_\theta\}_{\theta\in\Theta}$; for each $\theta \in \Theta$ the model M_θ is given by $M_\theta(\pi) = \mathsf{N}(\langle\pi, \theta\rangle, 1)$. Then Proposition 8 implies a minimax risk lower bound:*

$$\inf_{\mathsf{ALG}} \sup_{M\in\mathcal{M}} \mathbb{E}^{M,\mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \geq \Omega\left(\min\{d/\sqrt{T}, 1\}\right). \tag{12}$$

In Section 4, we also instantiate Proposition 8 to derive a new complexity measure for DMSO.

4 Application to Interactive Decision Making: Bandit Learnability and Beyond

In this section, we focus on the DMSO setting and apply our general results (Theorem 1) to derive new lower and upper bounds for interactive decision making that go beyond the previous results based on the Decision-Estimation Coefficient [36, 38] by incorporating hardness of estimation.

Background: Gaps between DEC-based and upper and lower bounds. A fundamental open question of the DEC framework is whether the $\log|\mathcal{M}|$-gap between DEC lower and upper bounds in Theorem 4 can be closed. To highlight this gap in a more interpretable fashion, we re-state Theorem 4 in terms of a quantity we refer to as the *minimax sample complexity*. Let us focus on regret. Recall that for a fixed model class $\mathcal{M}$, the following notion of minimax regret (2) is the central objective of interest:

$$\mathbf{Reg}^\star_T := \inf_{\mathsf{ALG}} \sup_{M\in\mathcal{M}} \mathbb{E}^{M,\mathsf{ALG}}[\mathbf{Reg}_{\mathsf{DM}}(T)].$$

Given a parameter $\Delta > 0$, we define the *minimax sample complexity*

$$T^\star(\mathcal{M}, \Delta) := \inf_{T\geq 1}\{T : \mathbf{Reg}^\star_T \leq T\Delta\} \tag{13}$$

as the least value T for which there exists an algorithm that achieves ΔT regret. Clearly, characterizing $T^\star(\mathcal{M}, \Delta)$ is equivalent to characterizing the minimax regret $\mathbf{Reg}^\star_T$.

Consider the following quantity induced by DEC for a class $\mathcal{M}$ and parameter $\Delta > 0$:

$$T^{\mathsf{DEC}}(\mathcal{M}, \Delta) = \inf_{\varepsilon\in(0,1)}\{\varepsilon^{-2} : \mathsf{r\text{-}dec}^{\mathrm{c}}_\varepsilon(\mathcal{M}) \leq \Delta\}. \tag{14}$$

With this definition, Theorem 4 is equivalent to the following characterization of the minimax sample complexity $T^\star(\mathcal{M}, \Delta)$:

$$T^{\mathsf{DEC}}(\mathcal{M}, \Delta) \lesssim T^\star(\mathcal{M}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathcal{M}, \Delta) \cdot \log|\mathcal{M}|. \tag{15}$$

That is, Theorem 4 characterizes the minimax sample complexity up to a multiplicative $\log|\mathcal{M}|$ factor. Our main result in this section will be to use the decision dimension and interactive Fano method (Theorem 1), to (i) tighten (15) for various special cases of interest, and (ii) give a new characterization for $T^\star(\mathcal{M}, \Delta)$ in structured bandit problems which avoids spurious parameters such as $\log|\mathcal{M}|$ altogether.

4.1 New upper and lower bounds through the decision dimension

For the a model class $\mathcal{M}$ and parameter $\Delta > 0$, we define the *decision dimension* as follows:

$$\mathsf{Ddim}_{\Delta}(\mathcal{M}) := \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \frac{1}{p(\pi : g^M(\pi) \leq \Delta)}. \tag{16}$$

Informal, the decision dimension value $\mathsf{Ddim}_{\Delta}(\mathcal{M})$ represents the best possible coverage over Δ-optimal decisions that can be achieved through a single exploratory distribution in the face of an unknown model $M \in \mathcal{M}$. As will now show, this quantity naturally arises as a lower bound on optimal risk through the interactive Fano method. We begin with the following regularity assumption.

Assumption 2 (Regular model class)**.** *There exists a constant $C_{\mathrm{KL}} > 0$ and a reference model $\overline{M}$ such that $D_{\mathrm{KL}}(M(\pi) \,\|\, \overline{M}(\pi)) \leq C_{\mathrm{KL}}$ for all $M \in \mathcal{M}$ and $\pi \in \Pi$.*

Assumption 2 is a mild assumption on the boundedness of KL divergence. As an example, for structured bandits with means in $[0, 1]$ and Gaussian rewards, Assumption 2 holds with $C_{\mathrm{KL}} = \frac{1}{2}$. Details and more examples are provided in Appendix B.1. Our main lower bound based on the decision dimension follows by specializing Theorem 1 to KL divergence.

Theorem 10 (Decision dimension lower bound)**.** *Suppose that $\mathcal{M}$ satisfies Assumption 2 with parameter $C_{\mathrm{KL}} > 0$. Then for any algorithm ALG and $\Delta > 0$, unless $T < \frac{\log \mathsf{Ddim}_{2\Delta}(\mathcal{M}) - 2}{2C_{\mathrm{KL}}}$, there exists $M^\star \in \mathcal{M}$ such that $\mathbb{P}^{M^\star, \mathsf{ALG}}[g^{M^\star}(\widehat{\pi}) \geq \Delta] \geq \frac{1}{2}$.*

In particular, decision dimension also implies a regret lower bound through Theorem 10: That is, for any algorithm to achieve ΔT-regret, it is necessary to have $T = \Omega(\log \mathsf{Ddim}_{2\Delta}(\mathcal{M}))$. Combining this with Theorem 4, we conclude that boundedness of both the DEC and the decision dimension is necessary for learning with any model class $\mathcal{M}$.

Upper bounds based on the decision dimension. We now complement Theorem 10 by showing that for any reward maximization instance of the DMSO setting (Example 1), boundedness of the decision dimension alone is also sufficient to derive *upper bounds* on the sample complexity of learning. The caveat is that while the lower bound is logarithmic in $\mathsf{Ddim}_{\Delta}(\mathcal{M})$, the upper bound will be polynomial.

Theorem 11 (Decision dimension upper bound)**.** *Consider the reward maximization variant of the DMSO setting (Example 1). There exists an algorithm that for any class $\mathcal{M}$ and $\Delta > 0$, ensures that with probability at least $1 - \delta$,*

$$\mathbf{Reg}_{\mathsf{DM}}(T) \leq T \cdot \Delta + O(\log(1/\delta)) \cdot \sqrt{T \cdot \mathsf{Ddim}_{\Delta}(\mathcal{M})}.$$

Combining Theorem 10 and Theorem 11 yields the following bounds on $T^\star(\mathcal{M}, \Delta)$ (omitting polylogarithmic factors):

$$\frac{\log \mathsf{Ddim}_{2\Delta}(\mathcal{M})}{C_{\mathrm{KL}}} \lesssim T^\star(\mathcal{M}, \Delta) \lesssim \frac{\mathsf{Ddim}_{\Delta/2}(\mathcal{M})}{\Delta^2}. \tag{17}$$

The gap between the lower and upper bounds of (17) is exponential; however, for model classes with $C_{\mathrm{KL}} = O(1)$, (17) suffices to characterize *finite-time learnability*. As a special case, we now show that decision dimension characterizes the learnability of any structured bandit problem.

4.2 Application: Bandit learnability

We consider a structured bandit setting given by a reward function class $\mathcal{H} \subseteq (\Pi \to [0, 1])$. The protocol is as follows: For each round $t \in [T]$, the learner chooses a decision $\pi^t \in \Pi$, then receives a reward $r^t \sim \mathsf{N}(h_\star(\pi^t), 1)$ in response, where the mean reward function $h_\star \in \mathcal{H}$. This corresponds to an instance of the DMSO framework with induced model class $\mathcal{M}_{\mathcal{H}} = \{\pi \mapsto \mathsf{N}(h(\pi), 1) \mid h \in \mathcal{H}\}$. We define the decision dimension for $\mathcal{H}$ via

$$\mathsf{Ddim}_{\Delta}(\mathcal{H}) := \mathsf{Ddim}_{\Delta}(\mathcal{M}_{\mathcal{H}}) = \inf_{p \in \Delta(\Pi)} \sup_{h \in \mathcal{H}} \frac{1}{p(\pi : h(\pi_h) - h(\pi) \leq \Delta)}, \tag{18}$$

where we denote $\pi_h := \arg\max_{\pi \in \Pi} h(\pi)$. This exactly coincides with the notion of *maximin volume* of Hanneke and Yang [41], which was shown to give a tight characterization of learnability for the special case of *noiseless binary-valued* structured bandits. We discuss the connection to Hanneke and Yang [41] in more detail in Appendix H.1.

It is straightforward to show that for any structured bandit problem, the induced class $\mathcal{M}_{\mathcal{H}}$ satisfies Assumption 2 with $C_{\mathrm{KL}} = \frac{1}{2}$ (Example 4). This leads to the following lower bound.

Corollary 12 (Lower bound for stochastic bandits)**.** *For the bandit model class $\mathcal{M}_{\mathcal{H}}$ defined as above, it holds that $T^\star(\mathcal{M}_{\mathcal{H}}, \Delta) \geq 2\mathsf{Ddim}_\Delta(\mathcal{H}) - 2$.*

Combining this result with the upper bound in Theorem 11, we obtain the following bounds on the minimax-optimal sample complexity for the structure bandit problem with class $\mathcal{H}$:

$$\log \mathsf{Ddim}_{2\Delta}(\mathcal{H}) \lesssim T^\star(\mathcal{M}_{\mathcal{H}}, \Delta) \lesssim \frac{\mathsf{Ddim}_{\Delta/2}(\mathcal{H})}{\Delta^2}. \tag{19}$$

This implies that $\mathsf{Ddim}_\Delta(\mathcal{H})$ characterizes learnability for structured bandits.

Theorem 13 (Structured bandit learnability)**.** *For a given reward function class $\mathcal{H}$, the bandit problem class $\mathcal{M}_{\mathcal{H}}$ is learnable for finite T if and only if $\mathsf{Ddim}_\Delta(\mathcal{H}) < +\infty$ for all $\Delta > 0$.*

We remark that the lower and upper bound in (19) cannot be improved in terms of the decision dimension alone: (1) For K-armed bandits, we have $\mathsf{Ddim}_\Delta(\mathcal{H}) \leq K$, meaning the upper bound is tight. (2) For d-dimensional linear bandits, we have $\log \mathsf{Ddim}_\Delta(\mathcal{H}) = \Omega(d)$, meaning the lower bound is nearly tight. Nevertheless, the exponential gap in (19) can be partly mitigated by combining the decision dimension with the DEC, as we will show in Section 4.3.

Our characterization bypasses impossibility results of Hanneke and Yang [41], who show that for *noiseless* structured bandit problems, there exist classes $\mathcal{H}$ for which bandit learnability is independent of the axioms of ZFC. Therefore, their results rule out the possibility of a characterization of noiseless bandit learnability through any *combinatorial dimension* [10] for the problem class. Our characterization is compatible with this result because the argument of Hanneke and Yang [41] relies on the *noiseless* nature of the bandit problem, and hence does not preclude a characterization for the noisy setting. In addition, the decision dimension is not a *combinatorial dimension* under the definition of Ben-David et al. [10]. Additional discussion is deferred to Appendix H.1.

4.3 Improved upper bounds for general decision making

In this section, we derive tighter upper bounds that scale with $\log \mathsf{Ddim}_\Delta(\mathcal{M})$ by combining the decision dimension with the Decision-Estimation Coefficient. For simplicity of presentation, we focus on regret minimization under the setting of Example 1, and we assume the following condition to simplify our bounds (the fully general upper bound is detailed in Appendix G).

Assumption 3 (Regularity of constrained DEC)**.** *A function $\mathsf{d} : [0,1] \to \mathbb{R}$ is said to have* moderate decay *if $\mathsf{d}(\varepsilon) \geq 10\varepsilon \ \forall \varepsilon \in [0,1]$, and there exists a constant $c \geq 1$ such that $c\frac{\mathsf{d}(\varepsilon)}{\varepsilon} \geq \frac{\mathsf{d}(\varepsilon')}{\varepsilon'}$ for all $\varepsilon' \geq \varepsilon$. We assume the function $\varepsilon \mapsto \mathsf{r\text{-}dec}^{\mathsf{c}}_\varepsilon(\mathrm{co}(\mathcal{M}))$, as a function of ε, satisfies moderate decay for a constant $c_{\mathrm{reg}} \geq 1$.*

This condition essentially requires that the DEC for $\mathrm{co}(\mathcal{M})$ exhibits moderate growth, which means that learning with $\mathrm{co}(\mathcal{M})$ is not "too easy".We now state our upper bound, which tightens Theorem 4 by replacing the $\log|\mathcal{M}|$ dependence in the upper bound with $\log \mathsf{Ddim}_\Delta(\mathcal{M})$ (with the caveat that the upper bound scales with the DEC for the *convexified* model class $\mathrm{co}(\mathcal{M})$).

Theorem 14 (Upper bound with DEC and decision dimension)**.** *Consider the reward maximization variant of the DMSO setting. Let $\mathcal{M}$ be any class for which Assumption 3 holds, and assume that Π is finite. Let $\bar{\varepsilon}(T) \asymp \sqrt{\log \mathsf{Ddim}_\Delta(\mathcal{M})/T}$. Then for any $\Delta > 0$, Algorithm 1 (see Appendix G) ensures that with high probability,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq T \cdot \Delta + O\big(c_{\mathrm{reg}} T\sqrt{\log T}\big) \cdot \mathsf{r\text{-}dec}^{\mathsf{c}}_{\bar{\varepsilon}(T)}(\mathrm{co}(\mathcal{M})).$$

Restating this upper bound in terms of minimax sample complexity and combining it with the preceding lower bounds yields the following result.

Theorem 15. *For any class $\mathcal{M}$ that satisfies Assumption 2 and 3, we have*

$$\max\left\{T^{\mathsf{DEC}}(\mathcal{M}, \Delta), \frac{\log \mathsf{Ddim}_{2\Delta}(\mathcal{M})}{C_{\mathrm{KL}}}\right\} \lesssim T^\star(\mathcal{M}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{M}), \Delta) \cdot \log \mathsf{Ddim}_{\Delta/2}(\mathcal{M}), \tag{20}$$

up to dependence on c_{reg} and logarithmic factors.

In particular, when the model class $\mathcal{M}$ is convex (i.e. $\mathrm{co}(\mathcal{M}) = \mathcal{M}$) and $C_{\mathrm{KL}} = O(1)$, Theorem 15 provides lower and upper bounds for learning with $\mathcal{M}$ that match up to a quadratic factor. Indeed, for convex model classes, the upper bound of (20) is always tighter than (15) (and also tighter than the result in Foster et al. [36, 37]), as by definition we have

$$\log \mathsf{Ddim}_\Delta(\mathcal{M}) \leq \log \mathsf{Ddim}_0(\mathcal{M}) \leq \min\{\log|\mathcal{M}|, \log|\Pi|\}, \qquad \forall \Delta > 0.$$

As applications, we apply Theorem 15 to structured bandits and contextual bandits (Appendix H).

Acknowledgements We acknowledge support from ARO through award W911NF-21-1-0328, the Simons Foundation and the NSF through award DMS-2031883, as well as NSF PHY-2019786.

References

[1] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

[2] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

[3] Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[4] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

[5] Noga Alon, Mark Bun, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private and online learnability are equivalent. *ACM Journal of the ACM (JACM)*, 69(4):1–34, 2022.

[6] Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

[7] Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

[8] Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310, 1994.

[9] Shai Ben-David, David Pal, and Shai Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22th Annual Conference on Learning Theory*, 2009.

[10] Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44–48, 2019.

[11] James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 1985.

[12] Lucien Birgé. On estimating a density using hellinger distance and some other strange facts. *Probability theory and related fields*, 71(2):271–291, 1986.

[13] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[14] Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137, 1979.

[15] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

[16] Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.

[17] Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

[18] Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, 2017.

[19] Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: pac, reward-free, preference-based learning, and beyond. *arXiv preprint arXiv:2209.11745*, 2022.

[20] Fan Chen, Junyu Zhang, and Zaiwen Wen. A near-optimal primal-dual method for off-policy learning in cmdp. *Advances in Neural Information Processing Systems*, 35:10521–10532, 2022.

[21] Fan Chen, Huan Wang, Caiming Xiong, Song Mei, and Yu Bai. Lower bounds for learning in revealing pomdps. *arXiv preprint arXiv:2302.01333*, 2023.

[22] Fan Chen, Constantinos Daskalakis, Noah Golowich, and Alexander Rakhlin. Near-optimal learning and planning in separated latent mdps. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 995–1067. PMLR, 30 Jun–03 Jul 2024.

[23] Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On bayes risk lower bounds. *The Journal of Machine Learning Research*, 17(1):7687–7744, 2016.

[24] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[25] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 1999.

[26] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

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

[28] David L Donoho and Richard C Liu. Geometrizing rates of convergence, I. Technical Report 137a, Dept. Statistics, Univ. California, Berkeley, 1987.

[29] David L Donoho and Richard C Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, pages 633–667, 1991.

[30] David L Donoho and Richard C Liu. Geometrizing rates of convergence, III. *The Annals of Statistics*, pages 668–701, 1991.

[31] John C Duchi. Lecture notes on statistics and information theory. 2023.

[32] John C Duchi and Martin J Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

[33] Dean Foster, Dylan J Foster, Noah Golowich, and Alexander Rakhlin. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2678–2792. PMLR, 2023.

[34] Dylan J Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*, pages 3199–3210, 2020.

[35] Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

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

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

[38] Dylan J Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3969–4043. PMLR, 2023.

[39] Dylan J Foster, Yanjun Han, Jian Qian, and Alexander Rakhlin. Online estimation via offline estimation: An information-theoretic framework. *arXiv preprint arXiv:2404.10122*, 2024.

[40] Margalit Glasgow and Alexander Rakhlin. Tight bounds for γ-regret via the decision-estimation coefficient. *arXiv preprint arXiv:2303.03327*, 2023.

[41] Steve Hanneke and Liu Yang. Bandit learnability can be undecidable. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5813–5849. PMLR, 2023.

[42] Rafail Z Hasminskii and Ildar A Ibragimov. On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, volume 473, pages 474–482. North-Holland Amsterdam, 1979.

[43] Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has'Minskii. *Statistical estimation: asymptotic theory*. Springer Science & Business Media, 1981.

[44] Yassir Jedra and Alexandre Proutiere. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2676–2681. IEEE, 2019.

[45] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

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

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

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

[49] Tor Lattimore and Andras Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

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

[51] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

[53] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

[54] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[55] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

[56] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[57] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022.

[58] Guanyu Nie, Mridul Agarwal, Abhishek Kumar Umrawal, Vaneet Aggarwal, and Christopher John Quinn. An explore-then-commit algorithm for submodular maximization under full-bandit feedback. In *Uncertainty in Artificial Intelligence*, pages 1541–1551. PMLR, 2022.

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

[60] Yury Polyanskiy and Yihong Wu. Dualizing le cam's method for functional estimation, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616*, 2019.

[61] Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.

[62] Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems*, pages 1026–1034, 2011.

[63] Nived Rajaraman, Yanjun Han, Jiantao Jiao, and Kannan Ramchandran. Statistical complexity and optimal algorithms for non-linear ridge bandits. *arXiv preprint arXiv:2302.06025*, 2023.

[64] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[65] Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, page 54, 2010.

[66] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[67] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

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

[69] David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *arXiv preprint arXiv:2003.12699*, 2020.

[70] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

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

[72] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

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

[74] Sara A van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[75] Vladimir N. Vapnik and Alexey A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Measures of Complexity*, 16(2):11, 1971.

[76] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior. 1944.

[77] Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

[78] Andrew Wagenmaker, Guanya Shi, and Kevin G Jamieson. Optimal exploration for model-based rl in nonlinear systems. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Andrew J Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR, 2023.

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

[81] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[82] Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, pages 265–280, 1945.

[83] Yuanhao Wang, Ruosong Wang, and Sham M Kakade. An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. *Neural Information Processing Systems (NeurIPS)*, 2021.

[84] Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

[85] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[86] Yunbei Xu and Assaf Zeevi. Bayesian design principles for frequentist sequential learning. In *International Conference on Machine Learning*, pages 38768–38800. PMLR, 2023.

[87] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

[88] Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

[89] 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.

[90] Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning linear quadratic gaussian systems. *IEEE Transactions on Automatic Control*, 2024.

Contents of Appendix

A Related work

In what follows, we briefly survey the most relevant related work.

Minimax bounds for statistical estimation. There is a vast body of literature on minimax risk bounds for statistical estimation, including Hasminskii and Ibragimov [42], Bretagnolle and Huber [14], Birgé [12], Donoho and Liu [29], Cover and Thomas [25], Ibragimov and Has'Minskii [43], Tsybakov [73] as well as references therein. For minimax lower bounds, the most widely applied techniques are Le Cam's two-point method [54], Assouad's lemma [7], and Fano's inequality [25]. Variants and applications of these three methods abound [1, 23, 60, 32]; Fano's inequality in particular has perhaps the largest number of variants, of which the most general version we are aware of is due to Chen et al. [23], which is recovered by our results (Proposition 2). Another celebrated thread, starting from the seminal work of Donoho and Liu [28], provides upper and lower bounds for a large class of non-parametric estimation problems based on Le Cam's two-point method through the study of a complexity measure known as the modulus of continuity [29, 30, 53, 60]. Specifically, for a functional $T : \mathcal{M} \to \mathbb{R}$ on the space of probability models $\mathcal{M}$, the modulus of continuity is

defined as

$$w_\varepsilon(\mathcal{M}, \overline{M}) := \sup_{M \in \mathcal{M}} \left\{ |T(M) - T(\overline{M})| : D^2_{\mathrm{H}}\big(M, \overline{M}\big) \le \varepsilon^2 \right\}. \tag{21}$$

Decision-Estimation Coefficient and Information Ratio. For interactive decision making problems Foster et al. [36, 38] introduce Decision-Estimation Coefficient (DEC) as a complexity measure and show that it characterizes the minimax-optimal regret up to a $\log|\mathcal{M}|$ factor. The DEC can be viewed as an interactive counterpart of the modulus of continuity in (21), and captures hardness of interactive decision making related to exploration, but not necessarily estimation. An active line of research has built on the DEC to encompass a variety of more increasing general decision making settings [36, 37, 19, 38, 33, 40], including adversarial decision making [37], PAC decision making [19, 38], reward-free learning and preference-based learning [19], and multi-agent decision making and partial monitoring [33].

The DEC is closely related to a Bayesian complexity measure known as the information ratio [67, 68, 50, 49, etc.],which was originally introduced to analyze Bayesian algorithms such as posterior sampling. It is also related to a more recent generalization known as the *algorithmic information ratio* (AIR) [86], developed for frequentist algorithms. Additionally, the DEC is connected to asymptotic instance-dependent complexity, as explored by [79].

Lower bounds for interactive learning. There is a long line of work studying the fundamental limits of online learning and reinforcement learning (RL), including lower bounds for structured bandits [26, 66, 71, 51, 46, etc.], contextual bandits [65, 35, etc.], Markov Decision Processes (MDPs) [59, 27, 89, 84, 83, etc.], partially observable RL [47, 57, 21, 22, etc.], dynamical systems and control [72, 44, 70, 77, 90, etc.], and offline RL [64, 85, 80, 45, 20, 55, 78, etc.]. A large portion of these lower bounds are proven through (variants of) the two-point method, and can be recovered by the DEC lower bound approach [36, 38]. Beyond two-point method, there are also (relatively fewer) papers that have used Assouad's lemma or Fano's inequality for proving lower bounds in interactive learning [17, 65, 2, 62, 35, 70, 63]. However, the approachs in these papers are specialized to the specific setting (hypercube structure in particular), and there is not a general principle of Fano's method or Assouad' lemma in interactive setting. The challenge of applying Fano's inequality is also highlighted in various prior works, e.g., Arias-Castro et al. [6, Section 1.3] and Rajaraman et al. [63, Section 1.5.4].

Learnability in statistical learning and decision making. In the literature on statistical learning, there is a long line of work which characterizes *learnability* (i.e., asymptotic achievability of non-trivial sample complexity) of hypothesis classes in terms of abstract complexity measures. Examples include the Vapnik-Chervonenkis dimension for binary classification [75, 13], the Littlestone dimension [56] for online classification [9] and differentially private classification [16, 5], and their real-valued counterparts (e.g. scale-sensitive dimensions) for regression [8, 4].

Beyond the settings above—and in particular for interactive settings—learnability is less well understood. The question of what complexity measure characterizes bandit learnability has been raised in e.g. Simchowitz et al. [71]. Remarkably, Ben-David et al. [10] demonstrate that there exist simple and natural learning for which it is impossible to characterize learnability through any *combinatorial* dimension. More recently, Hanneke and Yang [41] provide similar impossibility results for characterizing learnability of structured bandits in a noiseless setting with real-valued rewards. Hanneke and Yang [41] complement this with a positive result for the case of binary-valued rewards, characterizing learnability through a complexity measure called the *maximin volume*. Our learnability characterization for noisy structured bandits generalizes this complexity measure.

B Additional Background on DMSO Framework

The DMSO framework (Section 1.1) encompasses a wide range of learning goals beyond the reward maximization setting [36, 38], including reward-free learning, model estimation, and preference-based learning [19], and also multi-agent decision making and partial monitoring [33]. We provide two examples below for illustration.

Example 2 (Preference-based learning)**.** In preference-based learning, each model $M \in \mathcal{M}$ is assigned with a comparison function $\mathbb{C}^M : \Pi \times \Pi \to \mathbb{R}$ (where $\mathbb{C}^M(\pi_1, \pi_2)$ typically the probability of $\tau_1 \succ \tau_2$ for $\tau_1 \sim (M, \pi_1)$, $\tau_2 \sim (M, \pi_2)$), and the risk function is specified by $g^M(\pi) =$

$\max_{\pi^\star} \mathbb{C}^M(\pi^\star, \pi)$. Chen et al. [19] provide lower and upper bounds for this setting in terms of Preference-based DEC (PBDEC).

Example 3 (Interactive estimation)**.** In the setting of interactive estimation (a generalized PAC learning goal), each model $M \in \mathcal{M}$ is assigned with a parameter $\theta_M \in \Theta$, which is the parameter that the agent aims to estimate. The decision space $\Pi = \Pi_0 \times \Theta$, where each decision $\pi \in \Pi$ consists of $\pi = (\pi_0, \theta)$, where π_0 is the *explorative* policy to interact with the model[2], and θ is the estimator of the model parameter. In this setting, we define $g^M(\pi) = \mathrm{Dist}(\theta_M, \theta)$ for certain distance $\mathrm{Dist}(\cdot, \cdot)$.

This setting is an interactive version of the statistical estimation task, and it is also a generalization of the model estimation task studied in Chen et al. [19]. Natural examples include estimating some coordinates of the parameter θ in linear bandits. We provide nearly tight guarantee for this setting in Appendix E.2.

Applicability of our results Our general interactive Fano method Lemma 5 applies to any generalized no-regret / PAC learning goal (Section 1.1). Therefore, our risk lower bound in terms of quantile PAC-DEC Theorem 6 and decision dimension lower bound Theorem 10 both apply to any generalized learning goal. For a concrete example, see Appendix E.2 for the application to interactive estimation.

B.1 Examples of Assumption 2

In this section, we provide three general types of model classes where Assumption 2 holds with mild C_{KL}. It is worth noting that in Assumption 2, the reference model $\overline{M}$ does *not* necessarily belong to $\mathrm{co}(\mathcal{M})$.

Example 4 (Gaussian bandits)**.** Suppose that $\mathcal{H} \subseteq (\mathcal{A} \to [0, 1])$ is a class of mean value function, and $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ is the class of the model M associated with a $h^M \in \mathcal{H}$:

$$M(\pi) = \mathsf{N}(h^M(\pi), 1), \qquad \pi \in \mathcal{A}.$$

Then, consider the reference model $\overline{M}$ given by $\overline{M}(\pi) = \mathsf{N}(0, 1) \forall \pi \in \mathcal{A}$. It is clear that for any π, and model $M \in \mathcal{M}_{\mathcal{H},\mathbb{V}}$,

$$D_{\mathrm{KL}}(M(\pi) \,\|\, \overline{M}(\pi)) = \frac{1}{2} h^M(\pi)^2 \le \frac{1}{2},$$

and hence Assumption 2 holds with $C_{\mathrm{KL}} = \frac{1}{2}$.

Example 5 (Problems with finite observations)**.** Suppose that the observation space $\mathcal{O}$ is finite. Then, consider the reference model $\overline{M}$ given by $\overline{M}(\pi) = \mathrm{Unif}(\mathcal{O}) \forall \pi \in \Pi$. It holds that

$$D_{\mathrm{KL}}(M(\pi) \,\|\, \overline{M}(\pi)) \le \log |\mathcal{O}|, \qquad \forall \pi \in \Pi,$$

and hence Assumption 2 holds with $C_{\mathrm{KL}} = \log |\mathcal{O}|$.

Example 5 can further be generalized to infinite observation space, as long as every model in $\mathcal{M}$ admits a bounded density function with respect to the same base measure.

Example 6 (Contextual bandits)**.** Suppose that $\mathcal{H} \subseteq (\mathcal{C} \times \mathcal{A} \to [0, 1])$ is a class of mean value function, and $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ is the class of the model M specified by a value function $h^M \in \mathcal{H}$ and a context distribution $\nu_M \in \Delta(\mathcal{C})$. More specifically, for any $\pi \in \Pi = (\mathcal{C} \to \mathcal{A})$, $M(\pi)$ is the distribution of (c, a, r), generated by $c \sim \nu_M$, $a = \pi(c)$, and $r \sim \mathsf{N}(h^M(c, a), 1)$.

Then, consider the reference model $\overline{M}$ specified by $\nu_{\overline{M}} = \mathrm{Unif}(\mathcal{C})$ and $h^{\overline{M}} \equiv 0$. It is clear that for any π, and model $M \in \mathcal{M}_{\mathcal{H},\mathbb{V}}$,

$$D_{\mathrm{KL}}(M(\pi) \,\|\, \overline{M}(\pi)) \le \log |\mathcal{C}| + 1$$

and hence Assumption 2 holds with $C_{\mathrm{KL}} = \log |\mathcal{C}| + 1$.

The factor of $\log |\mathcal{C}|$ in Example 6 is due to the definition (39) of $\log \mathsf{Ddim}_\Delta(\mathcal{H})$, where we take supremum over all context distribution μ. This factor can be removed if we instead restrict the model class to have a common context distribution (i.e., the setting where context distribution is known or can be estimated from samples).

[2] In other words, $M(\pi)$ only depends on π through π_0.

C Technical Tools

Lemma C.1 (Sub-additivity for squared Hellinger distance, see e.g. [31, Lemma 9.5.3] [39, Lemma D.2]). *Let $(\mathcal{X}_1, \mathcal{F}_1), \ldots, (\mathcal{X}_T, \mathcal{F}_T)$ be a sequence of measurable spaces, and let $\mathcal{X}^t = \prod_{i=1}^t \mathcal{X}_i$ and $\mathcal{F}^t = \bigotimes_{i=1}^t \mathcal{F}_i$. For each t, let $\mathbb{P}^t(\cdot \mid \cdot)$ and $\mathbb{Q}^t(\cdot \mid \cdot)$ be probability kernels from $(\mathcal{X}^{t-1}, \mathcal{F}^{t-1})$ to $(\mathcal{X}_t, \mathcal{F}_t)$.*

Let $\mathbb{P}$ and $\mathbb{Q}$ be the laws of $X_1, \ldots, X_T$ under $X_t \sim \mathbb{P}^t(\cdot \mid X_{1:t-1})$ and $X_t \sim \mathbb{Q}^t(\cdot \mid X_{1:t-1})$ respectively. Then it holds that

$$D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \le 7\, \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^T D^2_{\mathrm{H}}\big(\mathbb{P}^t(\cdot \mid X_{1:t-1}), \mathbb{Q}^t(\cdot \mid X_{1:t-1})\big)\right]. \tag{22}$$

In particular, given a T-round algorithm ALG and a model M, we can consider random variables $X_1 = (\pi^1, o^1), \cdots, X_T = (\pi^T, o^T)$. Then, $\mathbb{P}^{M,\mathtt{ALG}}(X_t = \cdot \mid X_{1:t-1})$ is the distribution of (π^t, o^t), where $\pi^t \sim p^t(\cdot \mid \pi^1, o^1, \cdots, \pi^{t-1}, o^{t-1})$, and $o^t \sim M(\pi^t)$. Therefore, applying Lemma C.1 to $D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{P}^{\overline{M},\mathtt{ALG}}\right)$ gives the following corollary.

Corollary C.2. *For any T-round algorithm ALG, it holds that*

$$\frac{1}{2} D_{\mathrm{TV}}\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{P}^{\overline{M},\mathtt{ALG}}\right)^2 \le D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{P}^{\overline{M},\mathtt{ALG}}\right) \le 7T \cdot \mathbb{E}_{\pi \sim p_{\overline{M},\mathtt{ALG}}}\left[D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)\right].$$

Lemma C.3 (Foster et al. [36, Lemma A.4])**.** *For any sequence of real-valued random variables $(X_t)_{t \le T}$ adapted to a filtration $(\mathcal{F}_t)_{t \le T}$, it holds that with probability at least $1 - \delta$, for all $t \le T$,*

$$\sum_{s=1}^t -\log \mathbb{E}\left[\exp(-X_s) \middle| \mathcal{F}_{s-1}\right] \le \sum_{s=1}^t X_s + \log\left(1/\delta\right).$$

Lemma C.4. *For any pair of random variable (X, Y), it holds that*

$$\mathbb{E}_{X \sim \mathbb{P}_X}\left[D^2_{\mathrm{H}}\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right] \le 2 D^2_{\mathrm{H}}\left(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}\right).$$

Lemma C.5. *Suppose that for a random variable X, its mean and variance under $\mathbb{P}$ is $\mu_{\mathbb{P}}$ and $\sigma^2_{\mathbb{P}}$, and its mean and variance under $\mathbb{Q}$ is $\mu_{\mathbb{Q}}$ and $\sigma^2_{\mathbb{Q}}$. Then it holds that*

$$|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2 \le 4\left(\sigma^2_{\mathbb{P}} + \sigma^2_{\mathbb{Q}} + \frac{1}{2}|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2\right) D^2_{\mathrm{H}}\left(\mathbb{P}, \mathbb{Q}\right).$$

In particular, when $\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}, \sigma_{\mathbb{P}}, \sigma_{\mathbb{Q}} \in [0, 1]$, we have $D^2_{\mathrm{H}}\left(\mathbb{P}, \mathbb{Q}\right) \ge \frac{1}{10}|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2$.

On the other hand, when $\mathbb{P} = \mathsf{N}(\mu_{\mathbb{P}}, 1), \mathbb{Q} = \mathsf{N}(\mu_{\mathbb{Q}}, 1)$, then $D^2_{\mathrm{H}}\left(\mathbb{P}, \mathbb{Q}\right) \le \frac{1}{8}|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2$.

Proof. Let $\nu = \frac{\mathbb{P}+\mathbb{Q}}{2}$ be the common base measure and set $\mu = \frac{\mu_{\mathbb{P}} + \mu_{\mathbb{Q}}}{2}$. Then

$$\begin{aligned}
|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2 &= |\mathbb{E}_{\mathbb{P}}[X - \mu] - \mathbb{E}_{\mathbb{Q}}[X - \mu]|^2 \\
&= \left|\mathbb{E}_\nu\left[\left(\frac{d\mathbb{P}}{d\nu} - \frac{d\mathbb{P}}{d\nu}\right)(X - \mu)\right]\right|^2 \\
&\le \mathbb{E}_\nu\left[\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{P}}{d\nu}}\right)^2\right] \mathbb{E}_\nu\left[\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} + \sqrt{\frac{d\mathbb{P}}{d\nu}}\right)^2 (X - \mu)^2\right] \\
&\le 2 D^2_{\mathrm{H}}\left(\mathbb{P}, \mathbb{Q}\right) \cdot 2\big(\mathbb{E}_{\mathbb{P}}(X - \mu)^2 + \mathbb{E}_{\mathbb{Q}}(X - \mu)^2\big) \\
&= 4\left(\sigma^2_{\mathbb{P}} + \sigma^2_{\mathbb{Q}} + \frac{1}{2}|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2\right) D^2_{\mathrm{H}}\left(\mathbb{P}, \mathbb{Q}\right).
\end{aligned}$$

□

D Proofs from Section 3

In this section, we present proofs for the results in Section 3, except Section 3.2.

D.1 Proof of Theorem 1

In the following, we fix a prior $\mu \in \Delta(\mathcal{M})$, quantile $\delta > 0$, f-divergence D_f, and an algorithm ALG.

We first note that the risk of ALG under prior μ is lower bounded by

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}[L(M,X)] \geq \Delta \cdot \mathbb{P}_{M\sim\mu,X\sim\mathbb{P}^{M,\mathsf{ALG}}}(L(M,X)\geq\Delta).$$

It remains to show the following claim.

Claim. Suppose that there exists a reference distribution $\mathbb{Q}$ such that

$$\mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}}) > \mathbb{E}_{M\sim\mu} D_f\big(\mathbb{P}^{M,\mathsf{ALG}},\mathbb{Q}\big),$$

then $\mathbb{P}_{M\sim\mu,X\sim\mathbb{P}^{M,\mathsf{ALG}}}(L(M,X)\geq\Delta)\geq\delta$.

We denote $\bar{\rho}_\Delta = \mathbb{P}_{M\sim\mu,X\sim\mathbb{P}^{M,\mathsf{ALG}}}(L(M,X)<\Delta)$, and recall that we define $\rho_{\Delta,\mathbb{Q}} = \mathbb{P}_{M\sim\mu,X\sim\mathbb{Q}}(L(M,X)<\Delta)$. We then consider the following two distributions over $\mathcal{M}\times\mathcal{X}$:

$$P_0 : M\sim\mu, X\sim\mathbb{P}^{M,\mathsf{ALG}}, \qquad P_1 : M\sim\mu, X\sim\mathbb{Q}.$$

By the data processing inequality of f-divergence, we have

$$D_f(\bar{\rho}_\Delta,\rho_{\Delta,\mathbb{Q}}) \leq D_f(P_0,P_1) = \mathbb{E}_{M\sim\mu} D_f\big(\mathbb{P}^{M,\mathsf{ALG}},\mathbb{Q}\big).$$

Therefore, using $\mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}}) > \mathbb{E}_{M\sim\mu} D_f\big(\mathbb{P}^{M,\mathsf{ALG}},\mathbb{Q}\big)$, we know that $\mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}}) > D_f(\bar{\rho}_\Delta,\rho_{\Delta,\mathbb{Q}})$. In particular, we know $\rho_{\Delta,\mathbb{Q}} < 1-\delta$. Hence, there are two cases: (1) $\bar{\rho}_\Delta \leq \rho_{\Delta,\mathbb{Q}} < 1-\delta$, or (2) $\bar{\rho}_\Delta > \rho_{\Delta,\mathbb{Q}}$, and we can use the monotone property of D_f (Lemma D.1), which also implies $\bar{\rho}_\Delta \leq 1-\delta$. This immediately gives

$$\mathbb{P}_{M\sim\mu,X\sim\mathbb{P}^{M,\mathsf{ALG}}}(L(M,X)\geq\Delta) = 1-\bar{\rho}_\Delta \geq \delta.$$

Hence, the proof is completed. □

Lemma D.1. *For $x,y\in(0,1)$, the quantity $D_f(x,y)$ is increasing with respect to x when $x\geq y$.*

Proof of Lemma D.1 By definition, we know that

$$D_f(x,y) = yf\left(\frac{x}{y}\right) + (1-y)f\left(\frac{1-x}{1-y}\right).$$

For any $x > z \geq y$, we denote

$$a_x = \frac{x}{y}, \quad b_x = \frac{1-x}{1-y}, \quad a_z = \frac{z}{y}, \quad b_z = \frac{1-z}{1-y},$$

and then because f is convex, we know

$$\frac{a_z-b_x}{a_x-b_x}f(a_x) + \frac{a_x-a_z}{a_x-b_x}f(b_x) \geq f(a_z),$$
$$\frac{b_z-b_x}{a_x-b_x}f(a_x) + \frac{a_x-a_z}{a_x-b_z}f(b_x) \geq f(b_z).$$

Notice that $ya_z + (1-y)b_z = 1$, and hence

$$yf(a_z) + (1-y)f(b_z) \leq yf(a_x) + (1-y)f(b_x).$$

This gives $D_f(x,y) \geq D_f(z,y)$. □

D.2 Proof of Proposition 2

Consider

$$\mathbb{Q} = \mathbb{E}_{M\sim\mu}\mathbb{P}^{M,\mathsf{ALG}}.$$

Then, by the choice of $\mathbb{Q}$ and definition of KL-divergence, we have

$$\mathbb{E}_{M\sim\mu} D_f\big(\mathbb{P}^{M,\mathsf{ALG}},\mathbb{Q}\big) = I_{\mu,\mathsf{ALG}}(M;X),$$

and by definition, we have

$$\rho_{\Delta,\mathbb{Q}} = \mathbb{P}_{M\sim\mu,X'\sim\mathbb{Q}}(L(M,X')<\Delta) \leq \sup_x \mu(M\in\mathcal{M} : L(M,x)<\Delta), \tag{23}$$

For any $\delta \in (0,1)$ and $\Delta > 0$, we apply Theorem 1 to obtain that when

$$I_{\mu,\mathtt{ALG}}(M;X) < D_{\mathrm{KL}}(1-\delta \,\|\, \rho_{\Delta,\mathbb{Q}}), \tag{24}$$

we have

$$\sup_{M\in\mathcal{M}} \mathbb{E}_{X\sim\mathbb{P}^{M,\mathtt{ALG}}}[L(M,X)] \geq \delta\Delta. \tag{25}$$

Note that the KL-divergence between $\mathrm{Bern}(1-\delta)$ and $\mathrm{Bern}(\rho_{\Delta,\mathbb{Q}})$ is lower bounded by

$$\begin{aligned} D_{\mathrm{KL}}(1-\delta \,\|\, \rho_{\Delta,\mathbb{Q}}) &= (1-\delta)\log\frac{1-\delta}{\rho_{\Delta,\mathbb{Q}}} + \delta\log\frac{\delta}{(1-\rho_{\Delta,\mathbb{Q}})} \\ &> (1-\delta)\log\frac{1}{\rho_{\Delta,\mathbb{Q}}} + (1-\delta)\log(1-\delta) + \delta\log\delta \\ &\geq (1-\delta)\log\frac{1}{\rho_{\Delta,\mathbb{Q}}} - \log 2 \\ &\geq (1-\delta)\log\frac{1}{\sup_x \mu(M\in\mathcal{M} : L(M,x) < \Delta)} - \log 2 \end{aligned} \tag{26}$$

where the third inequality is by Jensen's inequality, and the last inequality is by (23).

Taking

$$\delta = 1 + \frac{I_{\mu,\mathtt{ALG}}(M;X) + \log 2}{\log\sup_x \mu(M\in\mathcal{M} : L(M,x) < \Delta)},$$

we know from (26) that (24) is true so that the result (25) holds, which proves Corollary 2. □

D.3 Proof of Proposition 3

We frame the problem in the ISDM framework, where each algorithm corresponds to an estimator $\widehat{\theta} : \mathcal{Y}^{\otimes n} \to \Theta$. Let the model class $\mathcal{M} = \{M_0, M_1\}$, where for each estimator ALG (regarded as an algorithm), $\mathbb{P}^{M_0,\mathtt{ALG}}$ is the distribution of $X \in \mathcal{A}$ generated by

$$X \sim M_0 : \ \theta \sim \nu_0, Y_1, \ldots, Y_n \sim P_\theta, X = \mathtt{ALG}(Y_1, \ldots, Y_n),$$

and $\mathbb{P}^{M_1,\mathtt{ALG}}$ is the distribution of $X \in \mathcal{A}$ generated by

$$X \sim M_1 : \ \theta \sim \nu_1, Y_1, \ldots, Y_n \sim P_\theta, X = \mathtt{ALG}(Y_1, \ldots, Y_n).$$

We further define the new loss for M_i, $i \in \{0,1\}$:

$$\ell(M_i, X) := \inf_{\theta\in\mathrm{supp}(\nu_i)} L(\theta, X), \qquad \forall X \in \mathcal{A}.$$

By the separation condition on Θ_0 and Θ_1, we have for any $X \in \mathcal{A}$,

$$\ell(M_0, X) + \ell(M_1, X) \geq 2\Delta.$$

This implies that

$$\mathbb{P}_{M\sim\mu}(\ell(M,X) \geq \Delta) \geq 1/2, \qquad \forall X \in \Theta.$$

Therefore, choosing prior $\mu = \mathrm{Unif}(\{0,1\})$ and reference $\mathbb{Q} = \mathbb{E}_{M\sim\mu}\mathbb{P}^{M,\mathtt{ALG}}$ gives

$$\rho_{\Delta,\mathbb{Q}} = \mathbb{P}_{M\sim\mu, X\sim\mathbb{Q}}(\ell(M,X) < \Delta) \leq 1/2,$$

and

$$\begin{aligned} \mathbb{E}_{M\sim\mu}[D_{\mathrm{TV}}\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{Q}\right)] &= \frac{1}{2}\left(D_{\mathrm{TV}}\left(\mathbb{P}^{M_0,\mathtt{ALG}}, \mathbb{Q}\right) + D_{\mathrm{TV}}\left(\mathbb{P}^{M_1,\mathtt{ALG}}, \mathbb{Q}\right)\right) \\ &\leq \frac{1}{2} D_{\mathrm{TV}}\left(\mathbb{P}^{M_0,\mathtt{ALG}}, \mathbb{P}^{M_1,\mathtt{ALG}}\right) \\ &\leq \frac{1}{2} D_{\mathrm{TV}}\left(\nu_0 \otimes P_\theta, \nu_1 \otimes P_\theta\right) \leq \frac{1}{4}, \end{aligned}$$

where the first inequality is by the convexity of the TV distance and the second inequality is by the data-processing inequality. This shows that

$$\mathbb{E}_{M\sim\mu}[D_{\mathrm{TV}}(\mathbb{P}^{M,\mathsf{ALG}},\mathbb{Q})] \le 1/4 \le \mathsf{d}_{|\cdot|,1/4}(\rho_{\Delta,\mathbb{Q}}).$$

Therefore, applying Theorem 1 gives

$$\begin{aligned}\sup_{\theta\in\Theta}\mathbb{E}_{Y_1,\dots,Y_n\sim P_\theta}[L(\theta,\mathsf{ALG}(Y_1,\dots,Y_n))] &\ge \mathbb{E}_{\theta\sim\frac{\nu_0+\nu_1}{2}}\mathbb{E}_{Y_1,\dots,Y_n\sim P_\theta}[L(\theta,\mathsf{ALG}(Y_1,\dots,Y_n))]\\ &\ge \mathbb{E}_{M\sim\mu}\mathbb{E}_{Y_1,\dots,Y_n\sim\mathbb{P}^{M,\mathsf{ALG}}}[\ell(M,\mathsf{ALG}(Y_1,\dots,Y_n))]\\ &\ge \mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}[\ell(M,X)] \ge \frac{\Delta}{4},\end{aligned}$$

where the second inequality follows from the fact that

$$\mathbb{E}_{Y_1,\dots,Y_n\sim P_\theta}[L(\theta,\mathsf{ALG}(Y_1,\dots,Y_n))] \ge \mathbb{E}_{Y_1,\dots,Y_n\sim P_\theta}[\ell(M_i,\mathsf{ALG}(Y_1,\dots,Y_n))], \qquad \theta\in\mathrm{supp}(\nu_i),$$

and the last inequality follows from Theorem 1 with $\delta=\frac{1}{4}$. This gives the desired result. □

D.4 Proof of Corollary 9

Consider the following setup of linear bandits: let $\theta^\star\in\mathbb{R}^d$ be an unknown parameter. At time t, the learner chooses an action $\pi^t\in\{\pi\in\mathbb{R}^d : \|\pi\|_2\le 1\}$ and receives a Gaussian reward $r^t\sim\mathsf{N}(\langle\theta^\star,\pi^t\rangle,1)$. For $T\in\mathbb{N}$, let $\mathcal{H}^T=(\pi^1,r^1,\cdots,\pi^T,r^T)$ be the observed history up to time T. The central claim of this section is the following upper bound on the mutual information.

Theorem D.2. *For any $r>0$, we define the prior μ_r over $\mathbb{B}^d(r)$ by*

$$\mu_r : \theta^\star\sim\mathsf{N}\left(0,\frac{r^2}{4d}I_d\right) \mid \|\theta^\star\|\le r.$$

Then for any algorithm ALG*, we have*

$$I_{\mu_r,\mathsf{ALG}}(\theta^\star;\mathcal{H}^T) \le d\log\left(1+\frac{r^2T}{4d^2}\right).$$

Proof. Denote $\lambda=\frac{r^2}{4}$. We first prove that if $\theta^\star\sim\mu=\mathsf{N}(0,\lambda I_d/d)$, then

$$I_{\mu,\mathsf{ALG}}(\theta^\star;\mathcal{H}^T) \le \frac{d}{2}\log\left(1+\frac{\lambda T}{d^2}\right). \tag{27}$$

By the Bayes rule, the posterior distribution of $\theta^\star$ conditioned on $(\mathcal{H}^{t-1},\pi^t)$ is

$$p(\theta^\star\mid\mathcal{H}^{t-1},\pi^t)\propto\exp\left(-\frac{d\|\theta^\star\|_2^2}{2\lambda}-\frac{1}{2}\sum_{s<t}(r^s-\langle\theta^\star,\pi^s\rangle)^2\right),$$

which is a Gaussian distribution with covariance $(\Sigma^{t-1})^{-1}$, where

$$\Sigma^{t-1}=\frac{d}{\lambda}I_d+\sum_{s<t}\pi^s(\pi^s)^\top.$$

Therefore, by the chain rule of mutual information, we have

$$\begin{aligned}I_{\mu,\mathsf{ALG}}(\theta^\star;\mathcal{H}^T) &= \sum_{t=1}^T I_{\mu,\mathsf{ALG}}(\theta^\star;r^t\mid H^{t-1},\pi^t)\\ &= \sum_{t=1}^T\mathbb{E}^{\mu,\mathsf{ALG}}\left[\frac{1}{2}\log\left(1+(\pi^t)^\top(\Sigma^{t-1})^{-1}\pi^t\right)\right]\\ &= \mathbb{E}^{\mu,\mathsf{ALG}}\left[\frac{1}{2}\sum_{t=1}^T\log\frac{\det(\Sigma^t)}{\det(\Sigma^{t-1})}\right]\\ &= \mathbb{E}^{\mu,\mathsf{ALG}}\left[\frac{1}{2}\log\frac{\det(\Sigma^T)}{(d/\lambda)^d}\right]\end{aligned}$$

$$\le \mathbb{E}^{\mu,\mathsf{ALG}}\left[\frac{d}{2}\log\frac{\mathrm{Tr}(\Sigma^T)/d}{d/\lambda}\right]$$
$$\le \frac{d}{2}\log\left(1+\frac{\lambda T}{d^2}\right),$$

which is exactly (27).

Next we deduce the claimed result from (27). Consider the random variable $Z = \mathbf{1}\{\|\theta^\star\|_2 \le r\} \in \{0,1\}$, and then

$$\begin{aligned}\frac{d}{2}\log\left(1+\frac{\lambda T}{d^2}\right) &\ge I_{\mu,\mathsf{ALG}}(\theta^\star;\mathcal{H}^T)\\ &\ge I_{\mu,\mathsf{ALG}}(\theta^\star;\mathcal{H}^T \mid Z)\\ &\ge \mathbb{P}(Z=1)\cdot I_{\mu_r,\mathsf{ALG}}(\theta^\star;\mathcal{H}^T|Z=1)\\ &= \mathbb{P}_\mu(\|\theta^\star\|_2\le r)\cdot I_{\mu_r,\mathsf{ALG}}(\theta^\star;\mathcal{H}^T).\end{aligned}$$

Here the first inequality is (27), the second inequality follows from $I(X;Y) - I(X;Y \mid f(X)) = I(f(X);Y) - I(f(X);Y \mid X) \ge 0$, the third identity follows from the definition of conditional mutual information. Finally, noticing that $\mathbb{P}_\mu(\|\theta^\star\|_2 \le r) \ge \frac{1}{2}$ by concentration of χ^2_d random variable, we arrive at the desired statement. □

Next we show how to translate the mutual information upper bound in Theorem D.2 to lower bounds of estimation and regret.

Theorem D.3. *Let $T \ge 1$, $r = \min\left\{\frac{c_0 d}{\sqrt{T}}, 1\right\}$ for a small absolute constant c_0, and consider the prior $\mu = \mu_r$. For any T-round algorithm with output $\widehat{\pi}$, Proposition 8 implies that*

$$\mathbb{E}^{\mu,\mathsf{ALG}}\left[\left\|\widehat{\pi} - \frac{\theta^\star}{\|\theta^\star\|}\right\|^2\right] \ge \frac{1}{10}.$$

Therefore, we may deduce that

$$\sup_{M^\star\in\mathcal{M}} \mathbb{E}^{M^\star,\mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \gtrsim \min\left\{\frac{d}{\sqrt{T}}, 1\right\}.$$

Proof. We first prove the first inequality by applying Proposition 8 to the following sub-optimality measure

$$\tilde{g}^{M_\theta}(\pi) = \|\pi - \mathsf{normalize}(\theta)\|_2^2,$$

where we denote $\mathsf{normalize}(\theta) = \frac{\theta}{\|\theta\|} \in \mathbb{B}^d(1)$. Notice that for $\theta \in \Theta$, we have

$$g^{M_\theta}(\pi) = \|\theta\| - \langle\theta,\pi\rangle \ge \|\theta\|\cdot\left\|\pi - \frac{\theta}{\|\theta\|}\right\|^2 = \|\theta\|\cdot\tilde{g}^{M_\theta}(\pi).$$

For $\Delta \in (0,1)$, we first claim that

$$\rho_\Delta := \sup_\pi \mu(\theta : \tilde{g}^{M_\theta}(\pi) \le \Delta) = O\left(\sqrt{d}\Delta^{(d-1)/2}\right). \tag{28}$$

To see so, by symmetry of Gaussian distribution, we know for fixed any π,

$$\mu(\theta : \tilde{g}^{M_\theta}(\pi)\le\Delta) = \mathbb{P}_{\theta\sim\mathrm{Unif}(\mathbb{B}^d(1))}\left(\theta : \|\theta-\pi\|^2 \le \Delta\right),$$

and hence we can instead consider the uniform distribution over $\mathbb{B}^d(1)$. By rotational invariance, we may assume that $\pi = (x,0,\cdots,0)$, with $x \ge 0$. Then

$$\left\{\theta\in\mathbb{B}^d(1) : \|\theta-\pi\|_2^2 \le \Delta\right\} = \left\{\theta\in\mathbb{B}^d(1) : \theta_1 \ge \frac{x^2+1-\Delta}{2x}\right\} \subseteq \left\{\theta\in\mathbb{B}^d(1) : \theta_1 \ge \sqrt{1-\Delta}\right\}.$$

By Bubeck et al. [15, Section 2], the density of $\theta_1 \in [-1, 1]$ is given by

$$f(\theta_1) = \frac{\Gamma(d/2)}{\Gamma((d-1)/2)\sqrt{\pi}}(1-\theta_1^2)^{(d-3)/2}.$$

Therefore,

$$\rho_\Delta \le \int_{\sqrt{1-\Delta}}^{1} f(\theta_1)d\theta_1 = O(\sqrt{d}) \cdot (1-\sqrt{1-\Delta})\Delta^{(d-3)/2} = O\Big(\sqrt{d}\Delta^{(d-1)/2}\Big).$$

With the upper bound (28) of ρ_Δ, we know that for $\Delta = \frac{1}{2}$, it holds

$$\log(1/\rho_\Delta) \ge 2I_\mu(T),$$

as long as c_0 is a sufficiently small constant. Therefore, Proposition 8 gives that

$$\mathbb{E}^{\mu,\mathsf{ALG}}\Big[\|\widehat{\pi} - \mathsf{normalize}(\theta^\star)\|^2\Big] = \mathbb{E}^{\mu,\mathsf{ALG}}[\tilde{g}^{M_\theta}(\pi)] \ge \frac{1}{4}.$$

This completes the proof of the first inequality.

Finally, using the fact that $\mathbb{P}_{\theta^\star\sim\mu}(\|\theta^\star\| \le c_1 r) \le \frac{1}{100}$ for a small absolute constant c_1, we can conclude that

$$\sup_{M^\star\in\mathcal{M}} \mathbb{E}^{M^\star,\mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \ge \mathbb{E}^{\mu,\mathsf{ALG}}[g^{M_\theta}(\pi)] \ge \frac{c_1 r}{8} = \Omega\left(\min\left\{\frac{d}{\sqrt{T}}, 1\right\}\right).$$

This is the desired result. □

E Additional Results from Section 3.2

In addition to the reward-maximization setting (Example 1), we also introduce a slightly more general setting. In this setting, we assume that for each model $M \in \mathcal{M}$, the risk function is $g^M(\pi) = f^M(\pi_M) - f^M(\pi)$, but f^M is not assumed to be the expected reward function (Example 1). Instead, we only require f^M satisfying the following assumption, where $\mathcal{M}^+$ is a pre-specified model class of reference models that contains $\mathrm{co}(\mathcal{M})$ (following Foster et al. [38]).

Assumption 4. *Let $\mathcal{M}^+ \subseteq (\Pi \to \Delta(\mathcal{O}))$ be a given class of reference models, such that $\mathrm{co}(\mathcal{M}) \subseteq \mathcal{M}^+$. For any $M \in \mathcal{M}$, the risk function takes form $g^M(\pi) = f^M(\pi_M) - f^M(\pi)$ for some functional $f^M : \Pi \to \mathbb{R}$, so that f^M can be extended to $\mathcal{M}^+$, such that for any model $M \in \mathcal{M}$ and reference model $\overline{M} \in \mathcal{M}^+$ we have*

$$|f^M(\pi) - f^{\overline{M}}(\pi)| \le L_{\mathrm{r}} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big), \qquad \forall \pi \in \Pi. \tag{29}$$

In some cases, consider a larger reference model class can be convenient for proving lower bounds, see e.g., Appendix B.1 and Appendix I.5.

E.1 Recovering DEC-based regret lower bounds

In this section, we demonstrate how our general lower bound approach recovers the regret lower bounds of Foster et al. [38], Glasgow and Rakhlin [40]. We first state our lower bound in terms of constrained DEC in the following theorem.

Theorem E.1. *Under the reward maximization setting (Example 1), for any T-round algorithm ALG, there exists $M^\star \in \mathcal{M}$ such that*

$$\mathbf{Reg}_{\mathsf{DM}} \ge \frac{T}{2} \cdot \Big(\mathsf{r\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) - 6\underline{\varepsilon}(T)\Big) - 1$$

with probability at least 0.01 under $\mathbb{P}^{M^\star,\mathsf{ALG}}$, where $\underline{\varepsilon}(T) = \frac{1}{100\sqrt{T}}$.

Theorem E.1 immediately yields an in-expectation regret lower bound in terms of constrained DEC. It also shaves off the unnecessary logarithmic factors in the lower bound of Foster et al. [38, Theorem 2.2].

For the remainder of this section, we sketch how we prove Theorem E.1 in a slightly more general setting (Assumption 4), following Appendix F.3. Before providing our regret lower bounds, we first present several important definitions.

Definition of quantile regret-DEC We note that it is possible to directly modify the definition of quantile PAC-DEC (11), and then apply Theorem 6 to obtain an analogous regret lower bound immediately. However, as Foster et al. [38] noted, the "correct" notion of regret-DEC (cf. Eq. (7)) turns out to be more sophisticated. Therefore, we define the quantile version of regret-DEC similarly, as follows.

Throughout the remainder of this section, we fix the integer T. Define

$$\Pi_T = \left\{ \widehat{\pi} : \widehat{\pi} = \frac{1}{T} \sum_{t=1}^{T} \delta_{\pi_t}, \text{ where } \pi_1, \cdots, \pi_T \in \Pi \right\} \subseteq \Delta(\Pi),$$

i.e., Π_T is the class of all T-round mixture decision. We introduce the mixture decision space Π_T here to handle the average of T-round profile $(\pi_1, \cdots, \pi_T)$ of the algorithm. In particular, when Π is convex, we may regard $\Pi_T = \Pi$.

Next, we define the quantile regret-DEC as

$$\textsf{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M}) := \inf_{p \in \Delta(\Pi_T)} \sup_{M \in \mathcal{M}} \left\{ \widehat{g}^{M}_{\delta}(p) \vee \mathbb{E}_{\pi \sim p}[g^{\overline{M}}(\pi)] \,\middle|\, \mathbb{E}_{\pi \sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\}, \tag{30}$$

and define $\textsf{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}) := \sup_{\overline{M} \in \mathcal{M}^+} \textsf{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$.

The following proposition relates our quantile regret-DEC to the constrained regret-DEC (proof in Appendix F.5).

Proposition E.2. *Suppose that Assumption 4 holds for $\mathcal{M}$. Then, for any $\overline{M} \in \mathcal{M}^+$, it holds that*

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le 2 \cdot \textsf{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M}) + c_\delta L_{\mathrm{r}} \varepsilon,$$

where we denote $c_\delta = \max\left\{\frac{\delta}{1-\delta}, 1\right\}$. In particular, it holds that

$$\textsf{r-dec}^{\mathrm{q}}_{\varepsilon,1/2}(\mathcal{M}) \ge \frac{1}{2}\left(\max_{\overline{M} \in \mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) - L_{\mathrm{r}} \varepsilon \right).$$

Lower bound with quantile regret-DEC Now, we prove the following lower bound for the regret of any T-round algorithm, via our general interactive Fano method (Lemma 5). The proof is presented in Appendix F.4.

Theorem E.3. *Suppose that Assumption 4 holds for $\mathcal{M}$. Then, for any T-round algorithm* ALG*, parameters $\varepsilon, \delta, C > 0$, there exists $M \in \mathcal{M}$ such that*

$$\mathbb{P}^{M, \textsf{ALG}}\left(\mathbf{Reg}_{\mathsf{DM}}(T) \ge T \cdot (\textsf{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}) - C L_{\mathrm{r}} \varepsilon) - 1 \right) \ge \delta - \frac{1}{C^2} - \sqrt{14 T \varepsilon^2}.$$

As a corollary, there exists $M^\star \in \mathcal{M}$ such that

$$\begin{aligned} \mathbf{Reg}_{\mathsf{DM}}(T) &\ge \frac{T}{2} \cdot \left(\max_{\overline{M} \in \mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) - 4 L_{\mathrm{r}} \underline{\varepsilon}(T) \right) - 1 \\ &\ge \frac{T}{2} \cdot \left(\textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) - 4 L_{\mathrm{r}} \underline{\varepsilon}(T) \right) - 1 \end{aligned}$$

with probability at least 0.01 under $\mathbb{P}^{M^\star, \textsf{ALG}}$, where $\underline{\varepsilon}(T) = \frac{1}{100\sqrt{T}}$.

Theorem E.1 is now an immediate corollary, because for reward-maximization setting, we always have $L_{\mathrm{r}} = \sqrt{2}$ in Assumption 4.

E.2 Results for interactive estimation

More generally, we show that for a fairly different task of interactive estimation (Example 3), we also have an equivalence between quantile PAC-DEC with constrained PAC-DEC.

Recall that in this setting, each model $M \in \mathcal{M}$ is assigned with a parameter $\theta_M \in \Theta$, which is the parameter that the agent want to estimate. The decision space $\Pi = \Pi_0 \times \Theta$, where each decision $\pi \in \Pi$ consists of $\pi = (\pi_0, \theta)$, where π_0 is the *explorative* decision to interact with the model, and θ

is the estimator of the model parameter. The risk function is then defined as $g^M(\pi) = \mathrm{Dist}(\theta_M, \theta)$, for certain distance $\mathrm{Dist}(\cdot, \cdot)$.

In interactive estimation, we can show that the quantile DEC is in fact lower bounded the constrained DEC, as follows (proof in Appendix F.6).

Proposition E.4. *Consider the setting of Example 3. Then as long as $\delta < \frac{1}{2}$, it holds that*

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \le 2 \cdot \textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}).$$

In particular, for such a setting (which encompasses the model estimation task considered in Chen et al. [19]), Theorem 6 provides a lower bound of estimation error in terms of constrained PAC-DEC. This is significant because the constrained PAC-DEC upper bound in Theorem 4 is actually not restricted to Example 1, and we have hence shown that

$$\textsf{p-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \lesssim \inf_{\mathsf{ALG}} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, \mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \lesssim \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}),$$

where $\underline{\varepsilon}(T) \asymp \sqrt{1/T}$ and $\bar{\varepsilon}(T) \asymp \sqrt{\log|\mathcal{M}|/T}$. Therefore, for interactive estimation, constrained PAC-DEC is also a *nearly tight* complexity measure.

Remark E.5. The $\log|\mathcal{M}|$-gap between the lower and upper bound can further be closed for convex model class, utilizing the upper bounds in Appendix G. More specifically, we consider a convex model class $\mathcal{M}$, where $M \mapsto \theta_M$ is a convex function on $\mathcal{M}$. Then, a suitable instantiation of ExO^+ (Algorithm 1) achieves

$$\mathbf{Risk}_{\mathsf{DM}}(T) \lesssim \Delta + \inf_{\gamma > 0} \left(\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) + \frac{\log N(\Theta, \Delta) + \log(1/\delta)}{T} \right),$$

where $N(\Theta, \Delta)$ is the Δ-covering number of Θ, because we have $\log \mathsf{Ddim}_{\Delta}(\mathcal{M}) \le \log N(\Theta, \Delta)$ by considering the prior $q = \mathrm{Unif}(\Theta_0)$ for a minimal Δ-covering of Θ. Similar to Theorem I.4, we can upper bound $\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M})$ by $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$. Taking these pieces together, we can show that under the assumption that $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$ is of moderate decay, ExO^+ achieves

$$\mathbf{Risk}_{\mathsf{DM}}(T) \lesssim \textsf{p-dec}^{\mathrm{c}}_{\varepsilon(T)}(\mathcal{M}),$$

where $\varepsilon(T) \asymp \sqrt{\log N(\Theta, 1/T)/T}$.

In particular, for the (non-interactive) *functional estimation* problem (see e.g. Polyanskiy and Wu [60]), the parameter space $\Theta \subset \mathbb{R}$, and hence by considering covering number, we have $\log|\Theta| = \widetilde{O}(1)$. Therefore, for convex $\mathcal{M}$, under mild assumption that the DEC is of moderate decaying (Assumption 3), the minimax risk is then characterized by (up to logarithmic factors)

$$\inf_{\mathsf{ALG}} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, \mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \asymp \textsf{p-dec}^{\mathrm{c}}_{\sqrt{1/T}}(\mathcal{M}).$$

This result can be regarded as a generalization of Polyanskiy and Wu [60] to the interactive estimation setting.

F Proofs from Section 3.2 and Appendix E

Additional notations For notational simplicity, for any distribution $q \in \Delta(\Pi)$ and reference model $\overline{M}$, we denote the localized model class around $\overline{M}$ as

$$\mathcal{M}_{q,\varepsilon}(\overline{M}) := \left\{ M \in \mathcal{M} : \mathbb{E}_{\pi \sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\}.$$

F.1 Proof of Lemma 5

We show that

$$\Delta^\star_{\mathsf{ALG},\delta} = \sup_{\Delta > 0} \left\{ \Delta \,|\, \sup_{M \in \mathcal{M}} \mathbb{P}^{M, \mathsf{ALG}}(g^M(\widehat{\pi}) \ge \Delta) > \delta \right\}, \tag{31}$$

i.e. $\Delta^\star_{\mathsf{ALG},\delta}$ is the maximum risk of ALG over the model class $\mathcal{M}$, measured in terms of the δ-quantile.

Note that by the definition of $\Delta^\star_{\mathsf{ALG},\delta}$, we have

$$\Delta^\star_{\mathsf{ALG},\delta} \ge \sup_{\overline{M} = M \in \mathcal{M}} \sup_{\Delta > 0} \{\Delta : p_{M,\mathsf{ALG}}(\pi : g^M(\pi) \ge \Delta) > \delta\}$$

$$= \sup_{\Delta>0} \left\{ \Delta : \sup_{M\in\mathcal{M}} \mathbb{P}^{M,\mathtt{ALG}}(g^M(\pi) \geq \Delta) > \delta \right\}.$$

On the other hand, using the chain rule of Hellinger distance (Lemma C.1), we have

$$D_{\mathrm{TV}}\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{P}^{\overline{M},\mathtt{ALG}}\right) \geq \sqrt{14T\mathbb{E}_{\pi\sim q_{\overline{M},\mathtt{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)}.$$

Therefore, we know

$$\begin{aligned}
\Delta^\star_{\mathtt{ALG},\delta} &:= \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \sup_{M\in\mathcal{M}} \sup_{\Delta\geq 0} \left\{ \Delta : p_{\overline{M},\mathtt{ALG}}(\pi : g^M(\pi) \geq \Delta) > \delta + \sqrt{14T\mathbb{E}_{\pi\sim q_{\overline{M},\mathtt{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)} \right\} \\
&\leq \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \sup_{M\in\mathcal{M}} \sup_{\Delta\geq 0} \left\{ \Delta : p_{\overline{M},\mathtt{ALG}}(\pi : g^M(\pi) \geq \Delta) > \delta + D_{\mathrm{TV}}\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{P}^{\overline{M},\mathtt{ALG}}\right) \right\}.
\end{aligned}$$

Notice that for any $\overline{M} \in \mathrm{co}(\mathcal{M})$, $M \in \mathcal{M}$, we can take $L(M, X) = g^M(\pi)$ (recall that $X = (\mathcal{H}^T, \widehat{\pi})$ is the whole trajectory), and $\mu_M \in \Delta(\mathcal{M})$ supported on M and $\mathbb{Q} = \mathbb{P}^{\overline{M},\mathtt{ALG}}$ in Theorem 1, and hence

$$\begin{aligned}
\Delta^\star_{\mathtt{ALG},\delta} &\leq \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \sup_{M\in\mathcal{M}} \sup_{\Delta\geq 0} \left\{ \Delta : D_{\mathrm{TV}}\left(1-\delta, \rho_{\Delta,\mathbb{P}^{\overline{M},\mathtt{ALG}}}\right) > D_{\mathrm{TV}}\left(\mathbb{P}^{M,\mathtt{ALG}}, \mathbb{P}^{\overline{M},\mathtt{ALG}}\right) \right\} \\
&\leq \sup_{M\in\mathcal{M}} \sup_{\Delta>0} \left\{ \Delta : \mathbb{P}^{M,\mathtt{ALG}}(\widehat{\pi} : g^M(\widehat{\pi}) \geq \Delta) > \delta \right\},
\end{aligned}$$

where the last line follows from the claim in Appendix D.1. The proof of (31) is hence completed. □

F.2 Proof of Theorem 6

Fix any algorithm ALG and abbreviate $\varepsilon = \underline{\varepsilon}(T)$. Take an arbitrary parameter $\Delta_0 < \textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M})$. Then there exists $\overline{M}$ such that $\Delta_0 < \textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$. Hence, by the definition (11), we know that

$$\Delta_0 < \sup_{M\in\mathcal{M}} \left\{ \widehat{g}^M_\delta(p_{\overline{M},\mathtt{ALG}}) \,\middle|\, \mathbb{E}_{\pi\sim q_{\overline{M},\mathtt{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \leq \varepsilon^2 \right\}.$$

Therefore, there exists $M \in \mathcal{M}$ such that

$$\mathbb{E}_{\pi\sim q_{\overline{M},\mathtt{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \leq \varepsilon^2, \qquad \mathbb{P}_{\pi\sim p_{\overline{M},\mathtt{ALG}}}(g^M(\pi) \geq \Delta_0) \geq \delta.$$

This immediately implies

$$p_{\overline{M},\mathtt{ALG}}(\pi : g^M(\pi) \geq \Delta) > \delta_1 + \sqrt{14T\mathbb{E}_{\pi\sim q_{\overline{M},\mathtt{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)},$$

where $\delta_1 = \delta - \sqrt{14T\underline{\varepsilon}^2}$. Notice that $\delta_1 > \frac{\delta}{2}$, and hence applying Lemma 5 shows that there exists $M \in \mathcal{M}$ such that $\mathbb{P}^{M,\mathtt{ALG}}\big(g^M(\widehat{\pi}) \geq \Delta_0\big) \geq \frac{\delta}{2}$. Letting $\Delta_0 \to \textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M})$ completes the proof. □

F.3 Proof of Proposition 7

In this section, we prove Proposition 7 under the slightly more general setting of Assumption 4.

Proposition F.1. *Under Assumption 4, for any reference model $\overline{M} \in \mathcal{M}^+$ and $\varepsilon > 0, \delta \in [0, 1)$, it holds that*

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M}) \leq \textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M}) + \frac{2\varepsilon L_{\mathrm{r}}}{1-\delta}.$$

For Example 1, we always have $L_{\mathrm{r}} \leq \sqrt{2}$, and hence Proposition 7 follows immediately from Proposition F.1.

Proof of Proposition F.1. Fix a reference model $\overline{M}$ and a $\Delta_0 > \textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$. Then, we pick a pair $(\bar{p}, \bar{q})$ such that

$$\Delta_0 > \sup_{M\in\mathcal{M}} \left\{ \widehat{g}^M_\delta(\bar{p}) \,\middle|\, \mathbb{E}_{\pi\sim\bar{q}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \leq \varepsilon^2 \right\},$$

whose existence is guaranteed by the definition of $\textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$ in (11). In other words, we have

$$\mathbb{P}_{\pi\sim\bar{p}}(g^M(\pi) \le \Delta_0) \ge 1-\delta, \qquad \forall M \in \mathcal{M}_{\bar{q},\varepsilon}(\overline{M})$$

Consider $q = \frac{\bar{p}+\bar{q}}{2}$ and $\varepsilon' = \frac{\varepsilon}{\sqrt{2}}$. Also let

$$\tilde{M} := \operatorname*{arg\,max}_{M\in\mathcal{M}_{q,\varepsilon'}(\overline{M})} f^M(\pi_M).$$

Now, consider $p \in \Delta(\Pi)$ given by

$$p(\cdot) = \bar{p}\big(\cdot | g^{\tilde{M}}(\pi) \le \Delta_0\big).$$

By definition, for $\pi \sim p$ we have $f^{\tilde{M}}(\pi) \ge f^{\tilde{M}}(\pi_{\tilde{M}}) - \Delta_0$, and hence

$$\begin{aligned}
\mathbb{E}_{\pi\sim p}[g^M(\pi)] &= f^M(\pi_M) - \mathbb{E}_{\pi\sim p}[f^M(\pi)] \\
&\le f^M(\pi_M) - \mathbb{E}_{\pi\sim p}\Big[f^{\tilde{M}}(\pi)\Big] + L_{\mathrm{r}} \cdot \mathbb{E}_{\pi\sim p} D_{\mathrm{H}}\Big(M(\pi), \tilde{M}(\pi)\Big) \\
&\le f^M(\pi_M) - f^{\tilde{M}}(\pi_{\tilde{M}}) + \Delta_0 + L_{\mathrm{r}} \cdot \mathbb{E}_{\pi\sim p} D_{\mathrm{H}}\Big(M(\pi), \tilde{M}(\pi)\Big).
\end{aligned}$$

Notice that for any $M \in \mathcal{M}_{q,\varepsilon'}(\overline{M})$, we have $f^M(\pi_M) \le f^{\tilde{M}}(\pi_{\tilde{M}})$ and also

$$\begin{aligned}
\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}\Big(M(\pi), \tilde{M}(\pi)\Big) &\le \frac{1}{\bar{p}(g^{\tilde{M}}(\pi) \le \Delta_0)} \mathbb{E}_{\pi\sim\bar{p}} D_{\mathrm{H}}\Big(M(\pi), \tilde{M}(\pi)\Big) \\
&\le \frac{1}{1-\delta}\Big(\mathbb{E}_{\pi\sim\bar{p}} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big) + \mathbb{E}_{\pi\sim\bar{p}} D_{\mathrm{H}}\Big(\tilde{M}(\pi), \overline{M}(\pi)\Big)\Big) \\
&\le \frac{2\varepsilon}{1-\delta}.
\end{aligned}$$

Combining these inequalities gives

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon'}(\mathcal{M}, \overline{M}) \le \sup_{M\in\mathcal{M}} \left\{ \mathbb{E}_{\pi\sim p}[g^M(\pi)] \mid \mathbb{E}_{\pi\sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \frac{\varepsilon^2}{2} \right\} \le \Delta_0 + \frac{2\varepsilon L_{\mathrm{r}}}{1-\delta}.$$

Letting $\Delta_0 \to \textsf{p-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$ completes the proof. □

F.4 Proof of Theorem E.3

Our proof adopts the analysis strategy originally proposed by Glasgow and Rakhlin [40].

Fix a $0 < \Delta < \textsf{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M})$ and a parameter $c \in (0,1)$. Then there exists $\overline{M} \in \mathcal{M}^+$ such that $\textsf{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M}) > \Delta$.

Fix a T-round algorithm $\texttt{ALG}$ with rules $p_1, \cdots, p_T$, we consider a modified algorithm $\texttt{ALG}'$: for $t = 1, \cdots, T$, and history $\mathcal{H}^{(t-1)}$, we set $p'_t(\cdot|\mathcal{H}^{(t-1)}) = p_t(\cdot|\mathcal{H}^{(t-1)})$ if $\sum_{s=1}^{t-1} g^{\bar{M}}(\pi^s) < T\Delta - 1$, and set $p'_t(\cdot|\mathcal{H}^{(t-1)}) = 1_{\pi_{\bar{M}}}$ if otherwise. By our construction, it holds that under $\texttt{ALG}'$, we have $\sum_{t=1}^{T} g^{\bar{M}}(\pi^t) < T\Delta$ almost surely. Furthermore, we can define the stopping time

$$\tau = \inf\left\{ t : \sum_{s=1}^{t} g^{\bar{M}}(\pi^s) \ge T\Delta - 1 \text{ or } t = T+1 \right\}.$$

If $\tau \le T$, then it holds that $\sum_{t=1}^{\tau} g^{\bar{M}}(\pi^t) \ge T\Delta - 1$.

Now, we consider $p = \mathbb{P}^{\bar{M},\texttt{ALG}'}(\frac{1}{T}\sum_{t=1}^{T} \pi^t = \cdot) \in \Delta(\Pi_T)$. Using our definition of $\textsf{r-dec}^{\mathrm{q}}$, we know that $\mathbb{E}_{\pi\sim p} g^{\bar{M}}(\pi) < \Delta$ by our construction, and hence there exists $M \in \mathcal{M}$ such that

$$\mathbb{P}_{\widehat{\pi}\sim p}(g^M(\widehat{\pi}) \ge \Delta) > \delta, \qquad \mathbb{E}_{\widehat{\pi}\sim p} D^2_{\mathrm{H}}\left(M(\widehat{\pi}), \overline{M}(\widehat{\pi})\right) \le \varepsilon^2.$$

By definition of p and Lemma C.1, we have

$$\mathbb{P}^{\bar{M},\texttt{ALG}'}\left(\sum_{t=1}^{T} g^M(\pi^t) \ge T\Delta\right) > \delta, \qquad D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\texttt{ALG}'}, \mathbb{P}^{\bar{M},\texttt{ALG}'}\right) \le 7T\varepsilon^2. \tag{32}$$

We also know

$$\mathbb{E}^{\overline{M},\mathtt{ALG}'}\left[\frac{1}{T}\sum_{t=1}^{T}|f^{M}(\pi^t)-f^{\bar{M}}(\pi^t)|^2\right] \le \mathbb{E}^{\overline{M},\mathtt{ALG}'}\left[\frac{1}{T}\sum_{t=1}^{T}L_{\mathrm{r}}^2 D_{\mathrm{H}}^2\left(M(\pi^t),\overline{M}(\pi^t)\right)\right]$$
$$= L_{\mathrm{r}}^2\mathbb{E}_{\widehat{\pi}\sim p}D_{\mathrm{H}}^2\left(M(\widehat{\pi}),\overline{M}(\widehat{\pi})\right) \le L_{\mathrm{r}}^2\varepsilon^2,$$

and hence by Markov inequality,

$$\mathbb{P}^{\bar{M},\mathtt{ALG}'}\left(\frac{1}{T}\sum_{t=1}^{T}|f^{M}(\pi^t)-f^{\bar{M}}(\pi^t)| \ge CL_{\mathrm{r}}\varepsilon\right) \le \frac{1}{C^2}.$$

In the following, we consider events

$$\mathcal{E}_1 := \left\{\sum_{t=1}^{T} g^{M}(\pi^t) \ge T\Delta\right\},$$

and the random variable $X := \sum_{t=1}^{T}|f^{M}(\pi^t)-f^{\bar{M}}(\pi^t)|$. By definition, $\mathbb{P}^{\bar{M},\mathtt{ALG}'}(\mathcal{E}_1) > \delta$, $\mathbb{P}^{\bar{M},\mathtt{ALG}'}(X \ge CTL_{\mathrm{r}}\varepsilon) \le \frac{1}{C^2}$. We have the following claim.

Claim: Under the event $\mathcal{E}_1 \cap \{\tau \le T\}$, we have

$$\sum_{t=1}^{\tau} g^{M}(\pi^t) \ge T\Delta - X - 1.$$

To prove the claim, we bound

$$\begin{aligned}
\sum_{t=1}^{\tau} g^{M}(\pi^t) &= \sum_{t=1}^{T} g^{M}(\pi^t) - \sum_{t=\tau+1}^{T} g^{M}(\pi^t) \\
&\ge T\Delta - \sum_{t=\tau+1}^{T}\left[f^{M}(\pi_{M}) - f^{M}(\pi^t)\right] \\
&\ge T\Delta - (T-\tau)f^{M}(\pi_{M}) + \sum_{t=\tau+1}^{T} f^{\bar{M}}(\pi^t) - X \\
&= T\Delta - (T-\tau)\cdot\left(f^{M}(\pi_{M}) - f^{\bar{M}}(\pi_{\bar{M}})\right) - X,
\end{aligned}$$

where the first inequality follows from $\mathcal{E}_1$, and the second inequality follows from $\sum_{t=\tau+1}^{T}|f^{M}(\pi^t)-f^{\bar{M}}(\pi^t)| \le X$. On the other hand, we can also bound

$$\begin{aligned}
\sum_{t=1}^{\tau} g^{M}(\pi^t) &= \sum_{t=1}^{\tau}\left[f^{M}(\pi_{M}) - f^{M}(\pi^t)\right] \\
&\ge \tau f^{M}(\pi_{M}) - \sum_{t=1}^{\tau} f^{\bar{M}}(\pi^t) - X \\
&= \tau\cdot\left(f^{M}(\pi_{M}) - f^{\bar{M}}(\pi_{\bar{M}})\right) + \sum_{t=1}^{\tau} g^{\bar{M}}(\pi^t) - X \\
&\ge \tau\cdot\left(f^{M}(\pi_{M}) - f^{\bar{M}}(\pi_{\bar{M}})\right) + T\Delta - 1 - X,
\end{aligned}$$

where the first inequality follows from $\sum_{t=1}^{\tau}|f^{M}(\pi^t)-f^{\bar{M}}(\pi^t)| \le X$, and the second inequality is because $\sum_{t=1}^{\tau} g^{\bar{M}}(\pi^t) \ge T\Delta - 1$ given $\tau \le T$, which follows from the definition of the stopping time τ. Therefore, taking maximum over the above two inequalities proves our claim.

Now, using the claim, we know

$$\mathbb{P}^{\bar{M},\mathtt{ALG}'}\left(\sum_{t=1}^{\tau\wedge T} g^{M}(\pi^t) \ge T(\Delta - C\varepsilon) - 1\right) \ge \mathbb{P}^{\bar{M},\mathtt{ALG}'}(\mathcal{E}_1 \cap \{X \le CT\varepsilon\}) \ge \delta - \frac{1}{C^2}.$$

Notice that $D_{\mathrm{H}}^2\left(\mathbb{P}^{M,\mathtt{ALG}'}, \mathbb{P}^{\overline{M},\mathtt{ALG}'}\right) \le 7T\varepsilon^2$, and hence for any event $\mathcal{E}$, it holds $\mathbb{P}^{M,\mathtt{ALG}'}(\mathcal{E}) \ge \mathbb{P}^{\overline{M},\mathtt{ALG}'}(\mathcal{E}) - \sqrt{14T\varepsilon^2}$. In particular, we have

$$\mathbb{P}^{M,\mathtt{ALG}'}\left(\sum_{t=1}^{\tau\wedge T} g^M(\pi^t) \ge T(\Delta - CL_{\mathrm{r}}\varepsilon) - 1\right) \ge \delta - \frac{1}{C^2} - \sqrt{14T\varepsilon^2}.$$

Finally, we note that $\mathtt{ALG}$ and $\mathtt{ALG}'$ agree on the first $\tau \wedge T$ rounds (formally, $\mathtt{ALG}$ and $\mathtt{ALG}'$ induce the same distribution of $(\pi^1, \cdots, \pi^{\tau\wedge T})$), and hence

$$\mathbb{P}^{M,\mathtt{ALG}}\left(\sum_{t=1}^{\tau\wedge T} g^M(\pi^t) \ge T(\Delta - CL_{\mathrm{r}}\varepsilon) - 1\right) \ge \delta - \frac{1}{C^2} - \sqrt{14T\varepsilon^2}.$$

The proof is hence complete by noticing that $\sum_{t=1}^{\tau\wedge T} g^M(\pi^t) \le \sum_{t=1}^{T} g^M(\pi^t) = \mathbf{Reg}_{\mathsf{DM}}(T)$ and taking $\Delta \to \mathsf{r\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M})$.

F.5 Proof of Proposition E.2

Fix a $\overline{M} \in \mathcal{M}^+$, and $\Delta > \mathsf{r\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$. Choose $p \in \Delta(\Pi_T)$ such that

$$\widehat{g}^M_\delta(p) \vee \mathbb{E}_{\pi\sim p}[g^{\overline{M}}(\pi)] \le \Delta, \qquad \forall M \in \mathcal{M}_{p,\varepsilon}(\overline{M}).$$

The existence of p is guaranteed by the definition (30). In other words, we have $\mathbb{E}_{\pi\sim p}[g^{\overline{M}}(\pi)] \le \Delta$ and

$$\mathbb{P}_{\pi\sim p}(g^M(\pi) \ge \Delta) \le \delta, \qquad \forall M \in \mathcal{M}_{p,\varepsilon}(\overline{M}).$$

We then has the following claim.

Claim. Suppose that $M \in \mathcal{M}_{p,\varepsilon}(\overline{M})$. Then it holds that

$$\mathbb{E}_{\pi\sim p}[g^M(\pi)] \le \mathbb{E}_{\pi\sim p}[g^{\overline{M}}(\pi)] + \Delta + c_\delta L_{\mathrm{r}} \mathbb{E}_{\pi\sim p} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big). \tag{33}$$

Fix any $M \in \mathcal{M}_{p,\varepsilon}(\overline{M})$, we prove (33) as follows. Consider the event $\mathcal{E} = \{\pi : g^M(\pi) \le \Delta\}$. Then,

$$\begin{aligned} p(\mathcal{E})\big(f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}})\big) &= \mathbb{E}_{\pi\sim p}\mathbf{1}\left\{\mathcal{E}\right\}\big(g^M(\pi) - g^{\overline{M}}(\pi) + f^{\overline{M}}(\pi) - f^M(\pi)\big) \\ &\le p(\mathcal{E})\Delta + L_{\mathrm{r}}\mathbb{E}_{\pi\sim p}\mathbf{1}\left\{\mathcal{E}\right\} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big), \end{aligned}$$

where the inequality uses $g^M(\pi) \le \Delta$ for $\pi \in \mathcal{E}$ and Assumption 4. Therefore,

$$\begin{aligned} &\mathbb{E}_{\pi\sim p} g^M(\pi) = \mathbb{E}_{\pi\sim p}\mathbf{1}\left\{\mathcal{E}\right\} g^M(\pi) + \mathbb{E}_{\pi\sim p}\mathbf{1}\left\{\mathcal{E}^c\right\} g^M(\pi) \\ &\le p(\mathcal{E})\Delta + \mathbb{E}_{\pi\sim p}\mathbf{1}\left\{\mathcal{E}^c\right\}\big(f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) + f^{\overline{M}}(\pi) - f^M(\pi) + g^{\overline{M}}(\pi)\big) \\ &\le 2\Delta + \frac{p(\mathcal{E}^c)L_{\mathrm{r}}}{p(\mathcal{E})}\mathbb{E}_{\pi\sim p}\mathbf{1}\left\{\mathcal{E}\right\} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big) + L_{\mathrm{r}}\mathbb{E}_{\pi\sim p}\mathbf{1}\left\{\mathcal{E}^c\right\} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big) \\ &\le 2\Delta + \max\left\{\frac{p(\mathcal{E}^c)}{p(\mathcal{E})}, 1\right\} L_{\mathrm{r}}\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big). \end{aligned}$$

This completes the proof of our claim.

Therefore, using (33) with $\mathbb{E}_{\pi\sim p}[g^{\overline{M}}(\pi)] \le \Delta$ yields

$$\mathbb{E}_{\pi\sim p}[g^M(\pi)] \le 2\Delta + c_\delta L_{\mathrm{r}}\varepsilon, \qquad \forall M \in \mathcal{M}_{p,\varepsilon}(\overline{M}).$$

This immediately implies

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le 2\Delta + c_\delta L_{\mathrm{r}}\varepsilon.$$

Finally, taking $\Delta \to \mathsf{r\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$ completes the proof. □

F.6 Proof of Proposition E.4

Fix a reference model $\overline{M}$ and let $\Delta_0 > \mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$. Then there exists $p, q \in \Delta(\Pi)$ such that

$$\sup_{M \in \mathcal{M}} \left\{ \widehat{g}^M_\delta(p) \,|\, \mathbb{E}_{\pi\sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\} < \Delta_0.$$

Therefore, it holds that

$$\mathbb{P}_{\pi\sim p}(g^M(\pi) \le \Delta_0) \ge 1-\delta, \qquad \forall M \in \mathcal{M}_{q,\varepsilon}(\overline{M}).$$

If the constrained set $\mathcal{M}_{q,\varepsilon}(\overline{M})$ is empty, then we immediately have $\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) = 0$, and the proof is completed. Therefore, in the following we may assume $\mathcal{M}_{q,\varepsilon}(\overline{M})$ is non-empty, and $\widehat{M} \in \mathcal{M}_{q,\varepsilon}(\overline{M})$.

Claim. Let $\widehat{\theta} = \theta_{\widehat{M}}$ and $\widehat{\pi} = (\pi_0, \widehat{\theta})$ for an arbitrary π_0, it holds that

$$g^M(\widehat{\pi}) \le \Delta_0, \qquad \forall M \in \mathcal{M}_{q,\varepsilon}(\overline{M}).$$

This is because for any $M \in \mathcal{M}_{q,\varepsilon}(\overline{M})$, it holds that

$$\mathbb{P}_{\pi\sim p}(g^M(\pi) \le \Delta_0, g^{\widehat{M}}(\pi) \le \Delta_0) \ge 1 - 2\delta > 0.$$

Hence, there exists $\theta \in \Theta$ such that $\mathrm{Dist}(\theta_M, \theta) \le \Delta_0$ and $\mathrm{Dist}(\theta_{\widehat{M}}, \theta) \le \Delta_0$ holds. Therefore, it must hold that $\mathrm{Dist}(\theta_M, \widehat{\theta}) \le 2\Delta_0$ for any $M \in \mathcal{M}_{q,\varepsilon}(\overline{M})$.

The above claim immediately implies that

$$\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le \sup_{M\in\mathcal{M}} \left\{ g^M(\widehat{\pi}) \,|\, \mathbb{E}_{\pi\sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\} \le 2\Delta_0.$$

Letting $\Delta_0 \to \mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$ yields $\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le 2\mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$, which is the desired result. □

G Exploration-by-Optimization Algorithm

In this section, we present a slightly modified version of the Exploration-by-Optimization Algorithm (ExO^+) developed by Foster et al. [37], built upon Lattimore and Szepesvári [52], Lattimore and Gyorgy [49]. The original ExO^+ algorithm has an *adversarial* regret guarantee for any model class $\mathcal{M}$, scaling with $\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathcal{M}))$, the offset DEC of the mode class $\mathrm{co}(\mathcal{M})$, and $\log|\Pi|$, the log-cardinality of the decision space. For our purpose, we adapt the original ExO^+ algorithm by using a prior $q \in \Delta(\Pi)$ not necessarily the uniform prior, and with a suitably chosen prior q, ExO^+ then achieves a regret guarantee scaling with $\log \mathsf{Ddim}_{\Delta}(\mathcal{M})$, instead of $\log|\Pi|$ (cf. Foster et al. [37]), which is always an upper bound of $\log \mathsf{Ddim}_{\Delta}(\mathcal{M})$.

Offset DEC for regret. We first recall the following (original) definition of DEC [36]:

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) := \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[g^M(\pi)] - \gamma \mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right), \tag{34}$$

and $\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) := \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M})$. Through the Estimation-to-Decision (E2D) algorithm [36], offset regret-DEC provides an upper bound of $\mathbf{Reg}_{\mathsf{DM}}$ for any learning problem, and it is also closely related to the complexity of adversarial decision making.

As discussed in Foster et al. [38], in the reward maximization setting (Example 1), the constrained regret-DEC $\mathsf{r\text{-}dec}^{\mathrm{c}}$ can always be upper bounded in terms of the offset DEC $\mathsf{r\text{-}dec}^{\mathrm{o}}$. Conversely, in the same setting, we also show that the offset DEC can also be upper bounded in terms of the constrained DEC (Theorem I.4), and hence the two concepts can be regarded as equivalent under mild assumptions (e.g. moderate decaying, Assumption 3).

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

Input: Problem $(\mathcal{M}, \Pi)$, prior $q \in \Delta(\Pi)$, parameter $T \geq 1$, $\gamma > 0$.

1: Set $q^1 = q$.
2: **for** $t = 1, \cdots, T$ **do**
3: Solve the *exploration-by-optimization* objective

$$(p^t, \ell^t) \leftarrow \operatorname*{arg\,min}_{p \in \Delta(\Pi), \ell \in \mathcal{L}} \Gamma_{q^t, \gamma}(p, \ell)$$

4: Sample $\pi^t \sim p^t$, execute π^t and observe o^t
5: Update

$$q^{t+1}(\pi) \propto_\pi q^t(\pi) \exp(\ell^t(\pi; \pi^t, o^t))$$

Exploration-by-Optimization algorithm. The algorithm, ExO^+, is restated in Algorithm 1. At each round t, the algorithm maintains a reference distribution $q^t \in \Delta(\Pi)$, and use it to obtain a decision distribution $p^t \in \Delta(\Pi)$ and an estimation function $\ell^t \in \mathcal{L} := (\Pi \times \Pi \times \mathcal{O} \to \mathbb{R})$, by solving a joint minimax optimization problem based on the *exploration-by-optimization* objective: Defining

$$\begin{aligned}\Gamma_{q,\gamma}(p, \ell; M, \pi^\star) = \mathbb{E}_{\pi \sim p}[f^M(\pi^\star) - f^M(\pi)] \\ - \gamma \mathbb{E}_{\pi \sim p} \mathbb{E}_{o \sim M(\pi)} \mathbb{E}_{\pi' \sim q}[1 - \exp\left(\ell(\pi'; \pi, o) - \ell(\pi^\star; \pi, o)\right)],\end{aligned} \tag{35}$$

and

$$\Gamma_{q,\gamma}(p, \ell) = \sup_{M \in \mathcal{M}, \pi^\star \in \Pi} \Gamma_{q,\gamma}(p, \ell; M, \pi^\star), \tag{36}$$

the algorithm solve $(p^t, \ell^t) \leftarrow \arg\min_{p \in \Delta(\Pi), \ell \in \mathcal{L}} \Gamma_{q^t,\gamma}(p, \ell)$. The algorithm then samples $\pi^t \sim p^t$, executes π^t and observes o^t from the environment. Finally, the algorithm updates the reference distribution by performing the exponential weight update with weight function $\ell^t(\cdot; \pi^t, o^t)$.

Guarantee of ExO^+. Following Foster et al. [37], we define

$$\mathsf{exo}_{1/\gamma}(\mathcal{M}, q) := \inf_{p \in \Delta(\Pi), \ell \in \mathcal{L}} \Gamma_{q,\gamma}(p, \ell), \tag{37}$$

and $\mathsf{exo}_{1/\gamma}(\mathcal{M}) = \sup_{q \in \Delta(\Pi)} \mathsf{exo}_{1/\gamma}(\mathcal{M}, q)$. The following theorem is deduced from Foster et al. [37, Theorem 3.1 and 3.2].

Theorem G.1. *Under the reward maximization setting*[3] *(Assumption 4), it holds that*

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/4}(\mathrm{co}(\mathcal{M}))) \leq \mathsf{exo}_{1/\gamma}(\mathcal{M}) \leq \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/8}(\mathrm{co}(\mathcal{M}))), \qquad \forall \gamma > 0.$$

Now, we present the main guarantee of Algorithm 1, which has the desired dependence on the prior $q \in \Delta(\Pi)$.

Theorem G.2. *It holds that with probability at least* $1 - \delta$*,*

$$\mathbf{Reg}_{\mathsf{DM}} \leq T\Big(\Delta + \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/8}(\mathrm{co}(\mathcal{M}))\Big) + \gamma \log\left(\frac{1}{\delta \cdot q(\pi : f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi) \leq \Delta)}\right)$$

Proof. Consider the set $\Pi^\star := \{\pi : f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi) \leq \Delta\}$ and the distribution $q^\star = q(\cdot|\Pi^\star)$. Following Proposition G.3, we consider

$$X_t(\pi^t, o^t) := \mathbb{E}_{\pi \sim q^\star}\left[\ell^t(\pi; \pi^t, o^t)\right] - \log \mathbb{E}_{\pi \sim q^t}\left[\exp\left(\ell^t(\pi; \pi^t, o^t)\right)\right],$$

and Proposition G.3 implies that

$$\sum_{t=1}^{T} X_t(\pi^t, o^t) \leq \log(1/q(\Pi^\star)).$$

[3] We remark that their proof actually applies to a broader setting, e.g. the setting of interactive estimation (Example 3, and see also Remark E.5).

Applying Lemma C.3, we have with probability at least $1-\delta$,

$$\sum_{t=1}^{T} -\log \mathbb{E}_{t-1}\left[\exp\left(-X_t(\pi^t, o^t)\right)\right] \le \sum_{t=1}^{T} X_t(\pi^t, o^t) + \log(1/\delta).$$

Notice that

$$\mathbb{E}_{t-1}\left[\exp\left(-X_t(\pi^t, o^t)\right)\right] = \mathbb{E}_{\pi\sim p^t}\mathbb{E}_{o\sim M^\star(\pi)}\mathbb{E}_{\pi'\sim q^t}\left[\exp\left(\ell^t(\pi';\pi,o) - \mathbb{E}_{\pi^\star\sim q^\star}\ell^t(\pi^\star;\pi,o)\right)\right].$$

Using the fact that $1-x \le -\log x$ and Jensen's inequality, we have

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^\star\sim q^\star}\mathrm{Err}(p^t, \ell^t; q^t, M^\star, \pi^\star) \le \log(1/q(\Pi^\star)) + \log(1/\delta),$$

where we denote

$$\mathrm{Err}(p, \ell; q, M^\star, \pi^\star) := \mathbb{E}_{\pi\sim p}\mathbb{E}_{o\sim M^\star(\pi)}\mathbb{E}_{\pi'\sim q}[1 - \exp\left(\ell(\pi';\pi,o) - \ell(\pi^\star;\pi,o)\right)].$$

Therefore, it holds that

$$\begin{aligned}
\mathbf{Reg}_{\mathsf{DM}} &= \sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \\
&\le \sum_{t=1}^{T} \Delta + \mathbb{E}_{\pi^\star\sim q^\star}\mathbb{E}_{\pi^t\sim p^t}\left[f^{M^\star}(\pi^\star) - f^{M^\star}(\pi^t)\right] \\
&= T\Delta + \gamma\sum_{t=1}^{T} \mathbb{E}_{\pi^\star\sim q^\star}\mathrm{Err}(p^t, \ell^t; q^t, M^\star, \pi^\star) \\
&\quad + \sum_{t=1}^{T} \mathbb{E}_{\pi^\star\sim q^\star} \underbrace{\left[\mathbb{E}_{\pi^t\sim p^t}\left[f^{M^\star}(\pi^\star) - f^{M^\star}(\pi^t)\right] - \gamma\mathrm{Err}(p^t, \ell^t; q^t, M^\star, \pi^\star)\right]}_{=\Gamma_{q^t,\gamma}(p^t,\ell^t;M^\star,\pi^\star)} \\
&\le T\Delta + \gamma(\log(1/q(\Pi^\star)) + \log(1/\delta)) + \sum_{t=1}^{T} \Gamma_{q^t,\gamma}(p^t, \ell^t) \\
&\le T\big(\Delta + \mathsf{exo}_{1/\gamma}(\mathcal{M})\big) + \gamma(\log(1/q(\Pi^\star)) + \log(1/\delta)).
\end{aligned}$$

Applying Theorem G.1 completes the proof. □

Proposition G.3. *For any $q' \in \Delta(\Pi)$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_{\pi\sim q'}[\ell^t(\pi;\pi^t,o^t)] - \log \mathbb{E}_{\pi\sim q^t}\left[\exp\left(\ell^t(\pi;\pi^t,o^t)\right)\right] \le D_{\mathrm{KL}}(q' \,\|\, q).$$

Proof. This is essentially the standard guarantee of exponential weight updates. For simplicity, we assume Π is discrete. Then, by definition,

$$q^t(\pi) = \frac{q(\pi)\exp\left(\sum_{s=1}^{t}\ell^s(\pi;\pi^s,o^s)\right)}{\sum_{\pi'\in\Pi} q(\pi')\exp\left(\sum_{s=1}^{t-1}\ell^s(\pi';\pi^s,o^s)\right)},$$

and hence

$$\begin{aligned}
\log \mathbb{E}_{\pi\sim q^t}\left[\exp\left(\ell^t(\pi;\pi^t,o^t)\right)\right] = \log \mathbb{E}_{\pi\sim q}\exp\left(\sum_{s=1}^{t}\ell^s(\pi;\pi^s,o^s)\right) \\
- \log \mathbb{E}_{\pi\sim q}\exp\left(\sum_{s=1}^{t-1}\ell^s(\pi;\pi^s,o^s)\right).
\end{aligned}$$

Therefore, taking summation over $t = 1, \cdots, T$, we have

$$-\sum_{t=1}^{T} \log \mathbb{E}_{\pi \sim q^t}\left[\exp\left(\ell^t(\pi; \pi^t, o^t)\right)\right] = -\log \mathbb{E}_{\pi \sim q}\left[\exp\left(\sum_{t=1}^{T} \ell^t(\pi; \pi^t, o^t)\right)\right].$$

The proof is then completed by the following basic fact of KL divergence: for any function $h : \Pi \to \mathbb{R}$,

$$\mathbb{E}_{\pi \sim q'}[h(\pi)] \le \log \mathbb{E}_{\pi \sim q} \exp(h(\pi)) + D_{\mathrm{KL}}(q' \,\|\, q).$$

□

H Additional discussion and results from Section 4

H.1 Additional discussion from Section 4.2

Connection to the maximin volume. Hanneke and Yang [41] propose *maximin volume*, a complexity measure that tightly characterizes the complexity of learning *noiseless binary-valued* structured bandit problems. For such problem classes, the decision dimension is exactly the inverse of the maximin volume. While the decision dimension can be viewed as a generalization of the maximin volume in this sense, we emphasize that the decision dimension directly arises from our general lower bound framework, and is applicable to general decision making problems in the DMSO framework.

Noise distribution. We note that the upper bound in (19) applies to any reward distribution with sub-Gaussian noise (cf. Appendix I.2). Meanwhile, since the lower bound in Corollary 12 is specialized to Gaussian noise, it acts as a lower bound for the broader class of sub-Gaussian noise distributions as well. We expect the lower bound to extend to other "reasonable" noise distributions.

H.2 Application: Structured bandits

We now instantiate our general results to give tighter guarantees for structured bandits, improving the upper bounds in Section 4.2.

DEC for structured bandits. We consider the same structured bandit protocol as in Section 4.2; recall that $\mathcal{H}$ denotes the reward function class and $\mathcal{M}_{\mathcal{H}}$ denotes the induced model class. In what follows, we simplify the results in Theorem 15 to be stated purely in terms of $\mathcal{H}$. For a reference value function $\bar{h} : \mathcal{C} \times \mathcal{A} \to [0, 1]$, we define

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H}, \bar{h}) := \inf_{p \in \Delta(\Pi)} \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_{\pi \sim p}\big[h(\pi_h) - h(\pi)\big] \;\middle|\; \mathbb{E}_{\pi \sim p}(h(\pi) - \bar{h}(\pi))^2 \le \varepsilon^2 \right\},$$

where we recall that $\pi_h := \max_{\pi \in \Pi} h(\pi)$. We then define the DEC for $\mathcal{H}$ as

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H}) = \sup_{\bar{h} \in \mathrm{co}(\mathcal{H})} \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H} \cup \{\bar{h}\}, \bar{h}).$$

As a corollary of Theorem G.2, the $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H})$ and $\log \mathsf{Ddim}_{\Delta}(\mathcal{H})$ together provide an upper bound for structured bandits with $\mathcal{H}$.

Theorem H.1. *Let $\mathcal{H}$ be given. Suppose that Π is finite, and that $\varepsilon \mapsto \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H}))$ satisfies moderate decay as a function of ε (Assumption 3) with constant c_{reg}. Let $\bar{\varepsilon}(T) \asymp \sqrt{\log \mathsf{Ddim}_{\Delta}(\mathcal{H})/T}$. The Algorithm 1 ensures that high probability,*

$$\mathbf{Reg}_{\mathsf{DM}} \le T \cdot \Delta + O(c_{\mathrm{reg}} T \sqrt{\log T}) \cdot \mathsf{r\text{-}dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathrm{co}(\mathcal{H})).$$

As a corollary, the minimax sample complexity of structured bandit learning with $\mathcal{H}$ is bounded as

$$\max\left\{T^{\mathsf{DEC}}(\mathcal{H}, \Delta), \log \mathsf{Ddim}_{2\Delta}(\mathcal{H})\right\} \lesssim T^{\star}(\mathcal{M}_{\mathcal{H}}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{H}), \Delta) \cdot \log \mathsf{Ddim}_{\Delta/2}(\mathcal{H}), \quad (38)$$

where we denote $T^{\mathsf{DEC}}(\mathcal{H}, \Delta) = \inf_{\varepsilon \in (0,1)}\{\varepsilon^{-2} : \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H}) \le \Delta\}$ (following (14)) and omit logarithmic factors and dependence on the constant c_{reg}.

There are many standard structured bandit problems where the value function class $\mathcal{H}$ is convex, including multi-armed bandits, linear bandits, and non-parametric bandits (with smoothness [65], or concavity [48], or sub-modularity [58], or etc.). For these problem classes, the complexity of no-regret learning is completely characterized by the DEC of $\mathcal{H}$ and the decision dimension $\mathsf{Ddim}_{\Delta}(\mathcal{H})$ (up to a quadratic factor).

We also note that the lower bound of (38) is proven for Gaussian noise, while our upper bound applies to a much more general class of reward distributions (with bounded variance).

H.3 Application: Contextual bandits with general function approximation

Next, we instantiate our general results for stochastic contextual bandits with general function approximation, generalizing the structured bandit problem. We consider the stochastic contextual bandit problem with context space $\mathcal{C}$, action space $\mathcal{A}$, and a reward function class $\mathcal{H} \subseteq (\mathcal{C} \times \mathcal{A} \to [0,1])$. This problem is a special case of the DMSO setting with decision space $\Pi = (\mathcal{C} \to \mathcal{A})$, and the environment is specified by a tuple $(h_\star \in \mathcal{H}, \nu_\star \in \Delta(\mathcal{C}))$. The protocol is as follows: For each round t, the environment draws $c^t \sim \nu$, and the learner takes action $a^t = \pi^t(c^t)$ based on the decision $\pi^t : \mathcal{C} \to \mathcal{A}$, and receives a reward $r^t \sim \mathsf{N}(h_\star(c^t, a^t), 1)$.

We can formulate the model class as follows. For a reward function $h \in \mathcal{H}$ and context distribution $\nu \in \Delta(\mathcal{C})$, the corresponding model $M_{h,\nu}$ is specified as

$$(c, a, r) \sim M_{h,\nu}(\pi): \quad c \sim \nu, a = \pi(a), r \sim \mathsf{N}(h(c,a), 1).$$

Let $\mathcal{M}_{\mathcal{H}} = \{M_{h,\nu} : h \in \mathcal{H}, \nu \in \Delta(\mathcal{C})\}$ be the induced model class of contextual bandits. Following Appendix H.2, we instantiate Theorem 15 to provide characterization of learning $\mathcal{M}_{\mathcal{H}}$.

DEC for contextual bandits. For any context $c \in \mathcal{C}$, the value function class $\mathcal{H}$ induces a restricted value function class $\mathcal{H}|_c = \{h(c,\cdot) : h \in \mathcal{H}\}$, which corresponds to a (non-contextual) bandit function class. We define the following variant of the DEC

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H}) := \sup_{c \in \mathcal{C}} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H}|_c),$$

which corresponds to the maximum of the *per-context* DEC over all contexts. We also define $T^{\mathsf{DEC}}(\mathcal{H}, \Delta) = \inf_{\varepsilon \in (0,1)} \{\varepsilon^{-2} : \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H}) \le \Delta\}$, following (14).

Decision dimension for contextual bandits. Specializing the decision dimension to contextual bandits, we define

$$\mathsf{Ddim}_{\Delta}(\mathcal{H}) := \inf_{p \in \Delta(\Pi)} \sup_{h \in \mathcal{H}, \nu \in \Delta(\mathcal{C})} \frac{1}{p(\pi : \mathbb{E}_{c \sim \nu}[h(c, \pi_h(c)) - h(c, \pi(c))] \le \Delta)}, \tag{39}$$

where $\pi_h \in \Pi$ is defined via $\pi_h(c) := \arg\max_{a \in \mathcal{A}} h(c,a)$ for $c \in \mathcal{C}$.

Intuitively, the value of the decision dimension $\log \mathsf{Ddim}_{\Delta}(\mathcal{H})$ for contextual bandits captures the difficulty of estimating optimal actions, but also the difficulty of generalizing across contexts. For example, when we consider the *unstructured* contextual bandit problems (i.e., $\mathcal{H} = (\mathcal{C} \times \mathcal{A} \to [0,1])$), it holds that $\log \mathsf{Ddim}_{\Delta}(\mathcal{H}) = |\mathcal{C}| \log |\mathcal{A}|$, but in general we can have $\log \mathsf{Ddim}_{\Delta}(\mathcal{H}) \ll \log |\Pi| = |\mathcal{C}| \log |\mathcal{A}|$.

As a corollary of Theorem 15, we derive the following upper and lower bounds on the complexity of contextual bandit learning with $\mathcal{H}$.

Theorem H.2. *Let $\mathcal{H}$ be given. Suppose that both the context space $\mathcal{C}$ and the action space $\mathcal{A}$ are finite, and that $\varepsilon \mapsto \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H}))$ satisfies moderate decay as a function of ε (Assumption 3) with constant c_{reg}. Let $\bar{\varepsilon}(T) \asymp \sqrt{\log \mathsf{Ddim}_{\Delta}(\mathcal{H})/T}$. Then Algorithm 1 ensures that with high probability, ;*

$$\mathbf{Reg}_{\mathsf{DM}} \le T \cdot \Delta + O(c_{\mathrm{reg}} T \sqrt{\log T}) \cdot \textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathrm{co}(\mathcal{H})).$$

As a corollary, the complexity of learning $\mathcal{M}_{\mathcal{H}}$ is bounded by

$$\max\left\{T^{\mathsf{DEC}}(\mathcal{H}, \Delta), \frac{\log \mathsf{Ddim}_{2\Delta}(\mathcal{H})}{\log |\mathcal{C}|}\right\} \lesssim T^{\star}(\mathcal{M}_{\mathcal{H}}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{H}), \Delta) \cdot \log \mathsf{Ddim}_{\Delta/2}(\mathcal{H}), \tag{40}$$

omitting dependence on c_{reg} and logarithmic factors.

By definition, we have $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H})) = \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H})$ if the *per-context* value function class $\mathcal{H}|_c$ is convex for every context $c \in \mathcal{C}$. Natural settings in which the per-context value function class $\mathcal{H}|_c$ is convex include contextual linear bandits [24], contextual non-parametric bandits [18], and contextual concave bandits [48]. For these problem classes, the complexity of no-regret learning is completely characterized by the DEC of $\mathcal{H}$ and the newly proposed $\mathsf{Ddim}_{\Delta}(\mathcal{H})$ (up to a quadratic factor and a factor of $\log |\mathcal{C}|$).

As a concrete example. we can derive upper bounds based on the decision dimension for finite-action contextual bandits as follows.

Corollary H.3. *For any value function class $\mathcal{H}$, Algorithm 1 ensures the following regret bound with high probability.*

$$\mathbf{Reg}_{\mathsf{DM}}(T) \le T \cdot \Delta + O\Big(\sqrt{T|\mathcal{A}| \cdot \log \mathsf{Ddim}_{\Delta}(\mathcal{H})}\Big).$$

Compared to the well-known regret bound of $O(\sqrt{T|\mathcal{A}| \cdot \log |\mathcal{H}|})$ for learning any with any finite contextual bandit class $\mathcal{H}$ [34, 69], this result above always provides a tighter upper bound, as $\log \mathsf{Ddim}_{\Delta}(\mathcal{H}) \le \log |\mathcal{H}|$. For certain (very simple) function classes $\mathcal{H}$, the quantity $\log \mathsf{Ddim}_{\Delta}(\mathcal{H})$ can be much smaller than $\log |\mathcal{H}|$ (for details, see Example 7). More importantly, $\log \mathsf{Ddim}_{\Delta}(\mathcal{H})$ leads to lower bounds for *any* contextual bandit function class (Theorem H.2). By contrast, lower bounds for structured contextual bandits in prior work have been proven in a case-by-case fashion (for specific value function classes $\mathcal{H}$).

I Proofs from Section 4 and Appendix H

In this section, we mainly focus on no-regret learning, and we present the regret upper and lower bounds in terms of DEC and $\log \mathsf{Ddim}_{\Delta}(\mathcal{M})$. The results can be generalized immediately to PAC learning.

I.1 Proof of Theorem 10

Fix an arbitrary reference model $\overline{M} \in (\Pi \to \Delta(\mathcal{O}))$ such that Assumption 2 holds. We remark that $\overline{M}$ is not necessarily in $\mathcal{M}$ or $\mathrm{co}(\mathcal{M})$.

We only need to prove the following fact.

Fact. If $T < \frac{\log \mathsf{Ddim}_{\Delta}(\mathcal{M})-2}{2C_{\mathrm{KL}}}$, then for any T-round algorithm ALG, there exists a model $M \in \mathcal{M}$ such that $\mathbf{Risk}_{\mathsf{DM}}(T) \ge \Delta$ with probability at least $\frac{1}{2}$ under $\mathbb{P}^{M,\mathsf{ALG}}$.

Proof. By the definition (16) of $\mathsf{Ddim}_{\Delta}(\mathcal{M})$, we know

$$\frac{1}{\mathsf{Ddim}_{\Delta}(\mathcal{M})} := \sup_{p \in \Delta(\Pi)} \inf_{M \in \mathcal{M}} p(\pi : g^M(\pi) \le \Delta).$$

Therefore, we have

$$\inf_{M \in \mathcal{M}} p_{\overline{M},\mathsf{ALG}}(\pi : g^M(\pi) \le \Delta) \le \frac{1}{\mathsf{Ddim}_{\Delta}(\mathcal{M})},$$

and hence there exists $M \in \mathcal{M}$ such that

$$T < \frac{\log\big(1/p_{\overline{M},\mathsf{ALG}}(\pi : g^M(\pi) \le \Delta)\big) - 2}{2C_{\mathrm{KL}}}.$$

Notice that by the chain rule of KL divergence, we have

$$D_{\mathrm{KL}}(\mathbb{P}^{M,\mathsf{ALG}} \,\|\, \mathbb{P}^{\overline{M},\mathsf{ALG}}) = \mathbb{E}^{M,\mathsf{ALG}}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(M(\pi^t) \,\|\, \overline{M}(\pi^t))\right] \le TC_{\mathrm{KL}}.$$

Hence, using data-processing inequality,

$$\begin{aligned} D_{\mathrm{KL}}(p_{M,\mathsf{ALG}} \,\|\, p_{\overline{M},\mathsf{ALG}}) &< \frac{\log\big(1/p_{\overline{M},\mathsf{ALG}}(\pi : g^M(\pi) \le \Delta)\big) - 2}{2} \\ &\le D_{\mathrm{KL}}(1/2 \,\|\, p_{\overline{M},\mathsf{ALG}}(\pi : g^M(\pi) \le \Delta)). \end{aligned}$$

This immediately implies $p_{M,\mathsf{ALG}}(\pi : g^M(\pi) \le \Delta) < \frac{1}{2}$ (by Theorem 1, or more directly, by Lemma D.1). □

Remark I.1. For simplicity, we present the above proof that does not go through our algorithmic Fano's inequality (Proposition 8). However, it is not difficult to see how Theorem 10 can be derived from Proposition 8, as we have

$$\inf_{\mu \in \Delta(\mathcal{M})} \sup_{\pi \in \Pi} \mu(M : g^M(\pi) \le \Delta) = \sup_{p \in \Delta(\Pi)} \inf_{M \in \mathcal{M}} p(\pi : g^M(\pi) \le \Delta),$$

as long as the Minimax theorem can be applied (e.g. when Π is finite or $\mathcal{M}$ is finite).

I.2 Proof of Theorem 11

In this section, we present an algorithm based on reduction (Algorithm 2) that achieves the desired upper bound. For the application to bandits with Gaussian rewards, we relax the assumption $R : \mathcal{O} \to [0, 1]$ as follows.

Assumption 5. *For any $M \in \mathcal{M}$ and $\pi \in \Pi$, the random variable $R(o)$ is 1-sub-Gaussian under $o \sim M(\pi)$.*

Suppose that $\Delta > 0$ is given, and fix a distribution $p^\star_\Delta$ that attains the infimum of (16). Based on $p^\star_\Delta$, we consider a reduced decision space $\Pi_{\mathsf{sub}} \subset \Pi$, generated as

$$\Pi_{\mathsf{sub}} = \{\pi_1, \cdots, \pi_N\}, \qquad \pi_1, \cdots, \pi_N \sim p^\star_\Delta \text{ independently},$$

where we set $N = \mathsf{Ddim}_\Delta(\mathcal{M}) \log(1/\delta)$. Then the space Π_{sub} is guaranteed to contain a near-optimal decision, as follows.

Lemma I.2. *With probability at least $1 - \delta$, there exists $\pi \in \Pi_{\mathsf{sub}}$ such that $g^{M^\star}(\pi) \le \Delta$.*

Therefore, we can then regard $M^\star$ as a N-arm bandit instance with action space $\mathcal{A} = \Pi_{\mathsf{sub}}$, and for each pull of an arm $\pi \in \mathcal{A}$, the stochastic reward r is generated as $r = R(o), o \sim M^\star(\pi)$. Then, we pick a standard bandit algorithm $\mathsf{BanditALG}$, e.g. the UCB algorithm (see e.g. Lattimore and Szepesvári [51]), and apply it to the multi-arm bandit instance $M^\star_{\mathsf{Bandit}}$, and the guarantee of $\mathsf{BanditALG}$ yields

$$\sum_{t=1}^{T} \max_{\pi' \in \Pi_{\mathsf{sub}}} f^{M^\star}(\pi') - f^{M^\star}(\pi^t) \le O\Big(\sqrt{TN \log(T/\delta)}\Big).$$

with probability at least $1 - \delta$. Therefore, we have

$$\begin{aligned}\mathbf{Reg}_{\mathsf{DM}}(T) &\le T \cdot (f^{M^\star}(\pi_{M^\star}) - \max_{\pi' \in \Pi_{\mathsf{sub}}} f^{M^\star}(\pi')) + O\Big(\sqrt{TN \log(T/\delta)}\Big) \\ &\le T \cdot \Delta + O\Big(\sqrt{TN \log(T/\delta)}\Big),\end{aligned}$$

with probability at least $1 - 2\delta$. This gives the desired upper bound, and we summarize the full algorithm in Algorithm 2. □

Proof of Lemma I.2. By definition,

$$\begin{aligned}\mathbb{P}\big(\forall i \in [N], g^{M^\star}(\pi_i) > \Delta\big) &\le p^\star_\Delta\big(\pi : g^{M^\star}(\pi) > \Delta\big)^N \\ &\le \left(1 - \frac{1}{\mathsf{Ddim}_\Delta(\mathcal{M})}\right)^N \\ &\le \exp\left(-\frac{N}{\mathsf{Ddim}_\Delta(\mathcal{M})}\right) \le \delta.\end{aligned}$$

□

I.3 Proof of Theorem 14

We first state the following more general result, and Theorem 14 is then a direct corollary (under Assumption 3).

Theorem I.3. *With suitably chosen parameter $\gamma > 0$ and prior $q \in \Delta(\Pi)$, ExO^+ (Algorithm 1) achieves with probability at least $1 - \delta$:*

$$\begin{aligned}\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} &\le \Delta + C \inf_{\gamma > 0}\left(\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma/8}(\mathrm{co}(\mathcal{M})) + \gamma \frac{\log \mathsf{Ddim}_\Delta(\mathcal{M}) + \log(1/\delta)}{T}\right) \\ &\le \Delta + C\sqrt{\log(T)} \cdot \overline{\mathsf{r\text{-}dec}}^{\mathrm{c}}_{\bar\varepsilon(T)}(\mathrm{co}(\mathcal{M})),\end{aligned}$$

where C is an absolute constant, $\bar\varepsilon(T) = \sqrt{\frac{\log \mathsf{Ddim}_\Delta(\mathcal{M}) + \log(1/\delta)}{T}}$, and the modified version of constrained DEC is defined as

$$\overline{\mathsf{r\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M})) := \varepsilon \cdot \sup_{\varepsilon' \in [\varepsilon, 1]} \frac{\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon'}(\mathrm{co}(\mathcal{M}))}{\varepsilon'}. \tag{42}$$

Algorithm 2 A reduction algorithm based on the decision dimension

Input: Problem $(\mathcal{M}, \Pi)$, parameter $\Delta, \delta > 0, T \geq 1$, Algorithm BanditALG for multi-arm bandits.

1: Set

$$p^\star_\Delta = \arg \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \frac{1}{p(\pi : g^M(\pi) \leq \Delta)}. \tag{41}$$

2: Set $N = \mathsf{Ddim}_\Delta(\mathcal{M}) \log(1/\delta)$ and sample the decision subspace $\Pi_{\mathsf{sub}} = \{\pi_1, \cdots, \pi_N\} \subset \Pi$ as

$$\pi_1, \cdots, \pi_N \sim p^\star_\Delta \text{ independently.}$$

3: Run the bandit algorithm BanditALG on the instance $M^\star_{\mathsf{Bandit}}$ for T rounds.

Proof of Theorem I.3. By the definition (16) of $\mathsf{Ddim}_\Delta(\mathcal{M})$, we know

$$\frac{1}{\mathsf{Ddim}_\Delta(\mathcal{M})} := \sup_{p \in \Delta(\Pi)} \inf_{M \in \mathcal{M}} p(\pi : g^M(\pi) \leq \Delta).$$

Therefore, there exists $q \in \Delta(\Pi)$ such that

$$\inf_{M \in \mathcal{M}} q(\pi : g^M(\pi) \leq \Delta) \geq \frac{1}{\mathsf{Ddim}_\Delta(\mathcal{M})},$$

We then instantiate Algorithm 1 with such a prior q. Theorem I.3 follows immediately by combining Theorem G.2 with the following structural result that relates offset DEC to constrained DEC. □

Theorem I.4. *Suppose that Assumption 4 holds for the model class $\mathcal{M}$. Then for any $\varepsilon \in (0, 1]$, it holds that*

$$\inf_{\gamma > 0} \left(\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}) + \gamma \varepsilon^2 \right) \leq \left(3\sqrt{\lfloor \log_2(2/\varepsilon) \rfloor} + 2 \right) \cdot \left(\overline{\mathsf{r\text{-}dec}}^{\mathrm{c}}_\varepsilon(\mathcal{M}) + L_{\mathrm{r}} \varepsilon \right).$$

Proof. Fix a $\varepsilon \in (0, 1]$ and $\overline{M} \in \mathrm{co}(\mathcal{M})$. We only need to prove the following result:

Claim. Suppose that $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon'}(\mathcal{M}, \overline{M}) \leq D\varepsilon'$ for all $\varepsilon' \in [\varepsilon, 1]$. Then there exists $\gamma = \gamma(D, \varepsilon)$ such that

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_\gamma(\mathcal{M}) + \gamma \varepsilon^2 \leq \left(3\sqrt{\lfloor \log_2(2/\varepsilon) \rfloor} + 2 \right) \cdot (D + L_{\mathrm{r}}) \varepsilon.$$

Set $K = \lfloor \log_2(1/\varepsilon) \rfloor + 1$ and fix a parameter $c = c(\varepsilon) \in (0, \frac{1}{2}]$ to be specified later in proof. Define $\varepsilon_i := 2^{-i}$ for $i = 0, \cdots, K-1$ and $\varepsilon_K = \varepsilon$. We also define $\lambda_i := c\varepsilon \cdot 2^i$ for $i = 0, \cdots, K-1$, and $\lambda_K = 1 - \sum_{i=0}^{K-1} \lambda_i \geq c$.

Define $\Delta_i = \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})$, and let p_i attains the $\inf_p$. In the following, we choose $\gamma = \gamma(D, \varepsilon) = \frac{9(D + L_{\mathrm{r}})}{8c\varepsilon}$.

By definition of p_i, it holds that

$$\mathbb{E}_{\pi \sim p_i}[g^M(\pi)] \leq \Delta_i, \qquad \forall M \in \mathcal{M} \cup \{\overline{M}\} : \mathbb{E}_{\pi \sim p_i} D^2_{\mathrm{H}} \left(M(\pi), \overline{M}(\pi) \right) \leq \varepsilon_i^2.$$

In particular, we may abbreviate $\mathcal{M}_i := \{M \in \mathcal{M} : \mathbb{E}_{\pi \sim p_i} D^2_{\mathrm{H}} \left(M(\pi), \overline{M}(\pi) \right) \leq \varepsilon_i^2\}$, and it holds

$$f^M(\pi_M) \leq f^{\overline{M}}(\pi_{\overline{M}}) + \Delta_i + L_{\mathrm{r}} \varepsilon_i, \qquad \forall M \in \mathcal{M}_i.$$

Next, we choose $p = \sum_{i=0}^K \lambda_i p_i \in \Delta(\Pi)$, and we know

$$\mathbb{E}_{\pi \sim p}[g^{\overline{M}}(\pi)] \leq \sum_{i=0}^K \lambda_i \mathbb{E}_{\pi \sim p}[g^{\overline{M}}(\pi)] \leq \sum_{i=0}^K \lambda_i \Delta_i =: \Delta.$$

Fix a $M \in \mathcal{M}$. Let $j \in \{0, \cdots, K\}$ be the maximum index such that $M \in \mathcal{M}_j$. Such a j must exists because $\mathcal{M} = \mathcal{M}_0$. Now,

$$\begin{aligned}\mathbb{E}_{\pi\sim p}[g^M(\pi)] &= f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) + \mathbb{E}_{\pi\sim p}[g^{\overline{M}}(\pi)] + \mathbb{E}_{\pi\sim p}[f^{\overline{M}}(\pi) - f^M(\pi)] \\ &\le \Delta_j + L_{\mathrm{r}}\varepsilon_j + \Delta + L_{\mathrm{r}}\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big).\end{aligned}$$

Case 1: $j = K$. Then, using AM-GM inequality, we have

$$\mathbb{E}_{\pi\sim p}[g^M(\pi)] - \gamma\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \le \Delta_K + \varepsilon_K + \Delta + \frac{L_{\mathrm{r}}^2}{4\gamma}.$$

Case 2: $j < K$. Then for each $i > j$, it holds that $\mathbb{E}_{\pi\sim p_j} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) > \varepsilon_j^2$, and hence

$$\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \ge \sum_{i=j+1}^{K} \lambda_j \mathbb{E}_{\pi\sim p_j} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \ge \sum_{i=j+1}^{K} \lambda_j \varepsilon_j^2 \ge \frac{c\varepsilon\cdot\varepsilon_j}{2}.$$

Therefore, using AM-GM inequality,

$$\begin{aligned}&\mathbb{E}_{\pi\sim p}[g^M(\pi)] - \gamma\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \\ &\le \Delta_j + L_{\mathrm{r}}\varepsilon_j + \Delta + \frac{9L_{\mathrm{r}}^2}{4\gamma} - \frac{8}{9}\gamma\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \\ &\le \Delta_j + L_{\mathrm{r}}\varepsilon_j + \Delta + \frac{9L_{\mathrm{r}}^2}{4\gamma} - \frac{8c\gamma\varepsilon}{9}\varepsilon_j.\end{aligned}$$

By our choice of γ, we have $\gamma\varepsilon \ge \frac{9}{8c}\left(\frac{\Delta_j}{\varepsilon_j} + L_{\mathrm{r}}\right)$, and hence in both cases, we have

$$\mathbb{E}_{\pi\sim p}[g^M(\pi)] - \gamma\mathbb{E}_{\pi\sim p} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \le \Delta + (D + L_{\mathrm{r}})\varepsilon + \frac{9L_{\mathrm{r}}^2}{4\gamma}.$$

Note that by definition, we have $\Delta \le (cK+1)D\varepsilon$ and $\gamma(\varepsilon)\cdot\varepsilon = \frac{9}{8c}(D + L_{\mathrm{r}})$, and hence

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma(\varepsilon)}(\mathcal{M}, \overline{M}) \le (2D + L_{\mathrm{r}} + cKD + 2cL_{\mathrm{r}})\varepsilon.$$

Thus,

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma(\varepsilon)}(\mathcal{M}, \overline{M}) + \gamma(\varepsilon)\varepsilon^2 \le \left(2D + L_{\mathrm{r}} + cK(D + L_{\mathrm{r}}) + \frac{9(D + L_{\mathrm{r}})}{8c}\right)\varepsilon_K.$$

Balancing c and re-arranging yields the desired result. □

I.4 Proof of Theorem 15

Note that the minimax-optimal sample complexity $T^\star(\mathcal{M}, \Delta)$ is just a way to better illustrate our minimax regret upper and lower bounds. By the definition of $T^\star(\mathcal{M}, \Delta)$, we have

$$\frac{1}{T}\mathbf{Reg}^\star_T = \sup\{\Delta : T^\star(\mathcal{M}, \Delta) \le T\}.$$

Under Assumption 3, the regret upper bound in Theorem 14 implies (up to c_{reg}, C_{KL} and logarithmic factors)

$$\frac{1}{T}\mathbf{Reg}^\star_T \lesssim \textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}).$$

And the regret lower bound Theorem E.1 implies (up to c_{reg} and logarithmic factors)

$$\textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \lesssim \frac{1}{T}\mathbf{Reg}^\star_T.$$

By the definition of $T^\star(\mathcal{M}, \Delta)$ and $T^{\mathsf{DEC}}(\mathcal{M}, \Delta)$, we then have

$$T^{\mathsf{DEC}}(\mathcal{M}, \Delta) \lesssim T^\star(\mathcal{M}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{M}), \Delta)\cdot\log \textsf{Ddim}_{\Delta/2}(\mathcal{M}).$$

Together with Theorem 10, we prove that

$$\max\left\{T^{\mathsf{DEC}}(\mathcal{M}, \Delta), \frac{\log \textsf{Ddim}_{\Delta}(\mathcal{M})}{C_{\mathrm{KL}}}\right\} \lesssim T^\star(\mathcal{M}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{M}), \Delta)\cdot\log \textsf{Ddim}_{\Delta/2}(\mathcal{M}).$$

□

I.5 Proof of Theorem H.1

For the upper bound, we work with more general noise structure (beyond Gaussian noises). We define $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ to be the class of all bandits models with mean reward function in $\mathcal{H}$ and variance bounded by 1. Specifically, for any $M \in \mathcal{M}_{\mathcal{H},\mathbb{V}}$, it is associated with a value function $h^M \in \mathcal{H}$, such that for any decision $\pi \in \Pi$, the distribution $M(\pi)$ of the random reward r has mean $h^M(\pi)$ and variance at most 1.

We also recall that the subclass $\mathcal{M}_{\mathcal{H}} \subseteq \mathcal{M}_{\mathcal{H},\mathbb{V}}$ is the bandit problem class with the standard Gaussian noise.

Proof of Theorem H.1: lower bound of (38). The lower bound with $\log \mathsf{Ddim}_\Delta(\mathcal{H})$ is exactly Corollary 12.

To prove the lower bound with $T^{\mathsf{DEC}}(\mathcal{H}, \Delta)$, we need to lower bound the DEC of $\mathcal{M}_{\mathcal{H}}$ in terms of the DEC of $\mathcal{H}$, as follows.

Lemma I.5. *Consider $\mathcal{M}^+ = \mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$ as the class of all reference models (Appendix E). Then,*

$$\max_{\overline{M} \in \mathcal{M}^+} \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\mathcal{H}} \cup \{\overline{M}\}, \overline{M}) \geq \mathsf{r\text{-}dec}^{\mathrm{c}}_{2\sqrt{2}\varepsilon}(\mathcal{H}). \tag{43}$$

Notice that for $\mathcal{M}^+$, Assumption 4 holds with $L_{\mathrm{r}} = \sqrt{10}$ (by Lemma C.5). Therefore, as a corollary of Theorem E.3: for any T-round algorithm $\mathtt{ALG}$, there exists $M^\star \in \mathcal{M}_{\mathcal{H}}$ such that

$$\mathbf{Reg}_{\mathsf{DM}}(T) \geq \frac{T}{2} \cdot (\mathsf{r\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{H}) - 5\underline{\varepsilon}(T)) - 1 \tag{44}$$

with probability at least 0.01 under $\mathbb{P}^{M^\star, \mathtt{ALG}}$, where $\underline{\varepsilon}(T) = \frac{1}{50\sqrt{T}}$. Therefore, the lower bound in terms of $T^{\mathsf{DEC}}(\mathcal{H}, \Delta)$ follows immediately (using regularity condition Assumption 3).

Combining both lower bounds completes the proof. □

Proof of Theorem H.1: upper bound. We apply Theorem I.3 similar to the proof of Theorem 15 (in Appendix I.2).

Using Theorem I.3, we know that ExO^+ can be suitably instantiated on the model class $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ so that with probability at least $1-\delta$,

$$\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} \leq \Delta + C\sqrt{\log(T)} \cdot \overline{\mathsf{r\text{-}dec}}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}})),$$

where C is an absolute constant, $\bar{\varepsilon}(T) = \sqrt{\frac{\log \mathsf{Ddim}_\Delta(\mathcal{H}) + \log(1/\delta)}{T}}$. We only need to upper bound the $\overline{\mathsf{r\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}}))$ (defined in (42)) in terms of the DEC of $\mathrm{co}(\mathcal{H})$.

Lemma I.6. *For any $\varepsilon \geq 0$, it holds that*

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\mathcal{H},\mathbb{V}}) \leq \mathsf{r\text{-}dec}^{\mathrm{c}}_{\sqrt{10}\varepsilon}(\mathcal{H})$$

We also note that $\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}}) \subseteq \mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$ because the model class $\mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$ is convex and it contains $\mathcal{M}_{\mathcal{H},\mathbb{V}}$. Therefore, we know

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}})) \leq \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}) \leq \mathsf{r\text{-}dec}^{\mathrm{c}}_{\sqrt{10}\varepsilon}(\mathrm{co}(\mathcal{H})).$$

Using the regularity of $\varepsilon \mapsto \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H}))$, we know

$$\overline{\mathsf{r\text{-}dec}}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}})) \leq c_{\mathrm{reg}} \cdot \mathsf{r\text{-}dec}^{\mathrm{c}}_{\sqrt{10}\varepsilon}(\mathrm{co}(\mathcal{H})).$$

This gives the desired upper bound. □

I.5.1 Proof of Lemma I.5

Fix a $\varepsilon \in [0,1]$, we denote $\varepsilon_1 = 2\sqrt{2}\varepsilon$ and take any $\Delta < \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon_1}(\mathcal{H})$. We pick $\bar{h} \in \mathrm{co}(\mathcal{H})$ such that $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon_1}(\mathcal{H}, \bar{h}) > \Delta$. Then, it holds that

$$\inf_{p \in \Delta(\Pi)} \sup_{h \in \mathcal{H} \cup \{\bar{h}\}} \left\{ \mathbb{E}_{\pi \sim p}[h(\pi_h) - h(a)] \,|\, \mathbb{E}_{\pi \sim p}(h(a) - \bar{h}(a))^2 \leq \varepsilon_1^2 \right\} \geq \Delta.$$

Suppose that $\bar{h} \in \mathrm{co}(\mathcal{H})$ is given by $\bar{h} = \mathbb{E}_{h\sim\mu}[h]$ with $\mu \in \Delta(\mathcal{H})$. Then, consider the reference model $\overline{M} \in \mathcal{M}^+$ with mean reward function $\bar{h}$ and Gaussian noise, i.e. $\overline{M}(\pi) = \mathsf{N}\big(\bar{h}(\pi), 1\big)$. Then, we know that for $\mathcal{M} = \mathcal{M}_{\mathcal{H}}$,

$$
\begin{aligned}
&\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \\
&= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}\cup\{\overline{M}\}} \left\{ \mathbb{E}_{\pi\sim p}[g^M(\pi)] \,|\, \mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\} \\
&= \inf_{p\in\Delta(\Pi)} \sup_{h\in\mathcal{H}\cup\{\bar{h}\}} \left\{ \mathbb{E}_{\pi\sim p}[h(\pi_h) - h(\pi)] \,|\, \mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(\mathsf{N}(h(\pi),1), \mathsf{N}\big(\bar{h}(\pi),1\big)\right) \le \varepsilon^2 \right\} \\
&\ge \inf_{p\in\Delta(\Pi)} \sup_{h\in\mathcal{H}\cup\{\bar{h}\}} \left\{ \mathbb{E}_{\pi\sim p}[h(\pi_h) - h(\pi)] \,|\, \mathbb{E}_{\pi\sim p} (h(\pi) - \bar{h}(\pi))^2 \le 8\varepsilon^2 \right\} \ge \Delta,
\end{aligned}
$$

where the last line follows from Lemma C.5. Taking $\Delta \to \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_1}(\mathcal{H})$ completes the proof of (43). □

I.5.2 Proof of Lemma I.6

Fix a reference model $\overline{M} \in \mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}})$. By definition, we know the mean reward function $h^{\overline{M}}$ of $\overline{M}$ belongs to $\mathrm{co}(\mathcal{H})$, i.e. $\overline{M} \in \mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$. Therefore, for any model $M \in \mathcal{M}_{\mathcal{H},\mathbb{V}}$ and decision $\pi \in \Pi$, by Lemma C.5,

$$
D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \ge \frac{1}{10} |h^M(\pi) - h^{\overline{M}}(\pi)|^2.
$$

Therefore, for $\mathcal{M} = \mathcal{M}_{\mathcal{H},\mathbb{V}}$,

$$
\begin{aligned}
&\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \\
&= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}\cup\{\overline{M}\}} \left\{ \mathbb{E}_{\pi\sim p}[g^M(\pi)] \,|\, \mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\} \\
&\ge \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}\cup\{\overline{M}\}} \left\{ \mathbb{E}_{\pi\sim p}[g^M(\pi)] \,|\, \mathbb{E}_{\pi\sim p} |h^M(\pi) - h^{\overline{M}}(\pi)|^2 \le 10\varepsilon^2 \right\} \\
&= \inf_{p\in\Delta(\Pi)} \sup_{h\in\mathcal{H}\cup\{\bar{h}\}} \left\{ \mathbb{E}_{\pi\sim p}[h(\pi_h) - h(\pi)] \,|\, \mathbb{E}_{\pi\sim p} (h(\pi) - \bar{h}(\pi))^2 \le 8\varepsilon^2 \right\} \\
&= \textsf{r-dec}^{\mathrm{c}}_{\sqrt{10}\varepsilon}(\mathcal{H} \cup \{\bar{h}\}, \bar{h}),
\end{aligned}
$$

where the second equality follows from the fact that when $= h$, we have $g^M(\pi) = h(\pi_h) - h(\pi)$. Taking supremum over $\overline{M}$ completes the proof. □

I.6 Proof of Theorem H.2

Similar to Appendix I.5, we consider a larger model class $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ of models with general noise structure. A model $M \in \mathcal{M}_{\mathcal{H},\mathbb{V}}$ is specified by a context distribution $\nu_M \in \Delta(\mathcal{C})$, a reward function $h^M \in \mathcal{H}$, and a reward distribution $\mathsf{R}^M(\cdot|\cdot,\cdot)$, such that for any $c \in \mathcal{C}, a \in \mathcal{A}$, $r \sim \mathsf{R}^M(\cdot|c,a)$ has mean $h^M(c,a)$ and variance at most 1. The model M is then given by

$$
(c,a,r) \sim M(\pi): \quad c \sim \nu_M, a = \pi(c), r \sim \mathsf{R}^M(\cdot|c,a).
$$

The model class $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ is defined to be the set of all possible models described above.

Proof of Theorem H.2: lower bound. The lower bound with $\log \mathsf{Ddim}_{\Delta}(\mathcal{H})$ follows immediately by applying Theorem 10 to the class $\mathcal{M}_{\mathcal{H}}$, which admits $C_{\mathrm{KL}} = O(\log|\mathcal{C}|)$ in Assumption 2 (as shown in Example 6).

On the other hand, the lower bound with $T^{\mathsf{DEC}}(\mathcal{H},\Delta)$ follows from the reduction to the *per-context* bandits problem. Specifically, for a fixed context $c \in \mathcal{C}$, $\mathcal{H}|_c$ corresponds to a structure bandits class $\mathcal{M}_{\mathcal{H}|_c}$. Notice that we can naturally regard $\mathcal{M}_{\mathcal{H}|_c} \subset \mathcal{M}_{\mathcal{H}}$ by viewing $\mathcal{M}_{\mathcal{H}|_c}$ as a contextual bandits class with the fixed context c. Therefore, by Theorem H.1 (specifically (44)):

$$
\frac{1}{T}\mathbf{Reg}^\star_T \ge \frac{1}{T}\mathbf{Reg}^\star_T \gtrsim \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{H}|_c) - 6\underline{\varepsilon}(T), \qquad \underline{\varepsilon}(T) = \frac{1}{50\sqrt{T}}.
$$

Taking maximum over $c \in \mathcal{C}$ yields

$$
\frac{1}{T}\mathbf{Reg}^\star_T \gtrsim \textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{H}) - 6\underline{\varepsilon}(T).
$$

This gives the desired lower bound with $T^{\mathsf{DEC}}(\mathcal{H},\Delta)$.

Combining both lower bounds completes the proof. □

Proof of Theorem H.2: upper bound. We follow the proof strategy of Appendix I.5. By Theorem I.3, ExO$^+$ can be suitably instantiated on the problem class $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ so that with probability at least $1-\delta$:

$$\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} \le \Delta + C \inf_{\gamma>0}\left(\textsf{r-dec}^{\mathrm{o}}_{\gamma/8}(\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}})) + \gamma\frac{\log \mathsf{Ddim}_{\Delta}(\mathcal{M}) + \log(1/\delta)}{T}\right).$$

We also note that $\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}}) \subseteq \mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$. Therefore, it remains to upper bound the offset DEC of $\mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$.

Lemma I.7. *For $\gamma > 0$, it holds that*

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\mathcal{H},\mathbb{V}}) \le \sup_{c\in\mathcal{C}} \textsf{r-dec}^{\mathrm{o}}_{\gamma/2}(\mathcal{M}_{\mathcal{H}|_c,\mathbb{V}}).$$

Then, we can apply the result of Theorem I.4. From the proof of Theorem I.4, it is not hard to see that: for any $\varepsilon > 0$, there exists $\gamma = \gamma(\varepsilon)$ such that for any $c \in \mathcal{C}$,

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma/2}(\mathcal{M}_{\mathcal{H}|_c,\mathbb{V}}) + \gamma\varepsilon^2 \lesssim \sqrt{\log(2/\varepsilon)}\cdot(c_{\mathrm{reg}}\cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H})) + \varepsilon),$$

where we also use the regularity condition of $\varepsilon \mapsto \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H}))$. This immediately gives

$$\mathbf{Reg}_{\mathsf{DM}} \le T\Delta + \mathcal{O}(c_{\mathrm{reg}}T\sqrt{\log T})\cdot \textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathrm{co}(\mathcal{H})),$$

where $\bar{\varepsilon}(T) = \sqrt{\frac{\log \mathsf{Ddim}_{\Delta}(\mathcal{H})+\log(1/\delta)}{T}}$. This is the desired upper bound. □

I.6.1 Proof of Lemma I.7

Fix a reference model $\overline{M} \in \mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}})$, and then $\overline{M} \in \mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$ by definition. In particular, $\overline{M}$ has mean value function $h^{\overline{M}} \in \mathcal{H}$ and context distribution $\bar{\nu} \in \Delta(\mathcal{C})$. We also know that for each $c \in \mathcal{C}$, $h^{\overline{M}}(x,\cdot) \in \mathrm{co}(\mathcal{H}|_c)$.

Then, by Lemma C.4, we also have

$$2D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \ge \mathbb{E}_{c\sim\nu_M, a=\pi(c)} D^2_{\mathrm{H}}\left(\mathsf{R}^M(r=\cdot|c,a), \mathsf{R}^{\overline{M}}(r=\cdot|c,a)\right).$$

Thus, we adopt the following notations: For each $c \in \mathcal{C}$ and model $M \in \mathcal{M}_{\mathcal{H},\mathbb{V}}$, we define $M_c \in \mathcal{M}_{\mathcal{H}|_c,\mathbb{V}}$ to be a bandit model such that for every action $a \in \mathcal{A}$, $M_c(a) = \mathsf{R}^M(r=\cdot|c,a)$. Then by definition, it holds that

$$2D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \ge \mathbb{E}_{c\sim\nu_M, a=\pi(c)} D^2_{\mathrm{H}}\left(M_c(a), \overline{M}_c(a)\right).$$

Now, combining the inequalities above, we have

$$\begin{aligned}
&\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\mathcal{H},\mathbb{V}}, \overline{M}) \\
&= \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}_{\mathcal{H},\mathbb{V}}} \mathbb{E}_{\pi\sim p}[g^M(\pi)] - \gamma\mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \\
&\le \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}_{\mathcal{H},\mathbb{V}}} \mathbb{E}_{\pi\sim p}\mathbb{E}_{c\sim\nu_M, a=\pi(c)}\left[h^M(c,\pi_M(c)) - h^M(c,a) - \frac{\gamma}{2}D^2_{\mathrm{H}}\left(M_c(a), \overline{M}_c(a)\right)\right] \\
&\overset{(1)}{=} \inf_{p=(p_c),p_c\in\Delta(\mathcal{A})}\sup_{M\in\mathcal{M}_{\mathcal{H},\mathbb{V}}} \mathbb{E}_{c\sim\nu_M, a\sim p_c}\left[h^M(c,\pi_M(c)) - h^M(c,a) - \frac{\gamma}{2}D^2_{\mathrm{H}}\left(M_c(a), \overline{M}_c(a)\right)\right] \\
&\overset{(2)}{\le} \inf_{p=(p_c),p_c\in\Delta(\mathcal{A})}\sup_{M\in\mathcal{M}_{\mathcal{H},\mathbb{V}}}\sup_{c\in\mathcal{C}} \mathbb{E}_{a\sim p_c}\left[h^M(c,\pi_M(c)) - h^M(c,a) - \frac{\gamma}{2}D^2_{\mathrm{H}}\left(M_c(a), \overline{M}_c(a)\right)\right] \\
&\overset{(3)}{=} \inf_{p=(p_c),p_c\in\Delta(\mathcal{A})}\sup_{c\in\mathcal{C}}\sup_{M_c\in\mathcal{M}_{\mathcal{H}|_c,\mathbb{V}}} \mathbb{E}_{a\sim p_c}\left[h^{M_c}(\pi_{M_c}) - h^{M_c}(a) - \frac{\gamma}{2}D^2_{\mathrm{H}}\left(M_c(a), \overline{M}_c(a)\right)\right] \\
&\overset{(4)}{=} \sup_{c\in\mathcal{C}}\inf_{p_c\in\Delta(\mathcal{A})}\sup_{M_c\in\mathcal{M}_{\mathcal{H}|_c,\mathbb{V}}} \mathbb{E}_{a\sim p_c}\left[h^{M_c}(\pi_{M_c}) - h^{M_c}(a) - \frac{\gamma}{2}D^2_{\mathrm{H}}\left(M_c(a), \overline{M}_c(a)\right)\right] \\
&= \sup_{c\in\mathcal{C}} \textsf{r-dec}^{\mathrm{o}}_{\gamma/2}(\mathcal{M}_{\mathcal{H}|_c,\mathbb{V}}, \overline{M}_c) \le \sup_{c\in\mathcal{C}} \textsf{r-dec}^{\mathrm{o}}_{\gamma/2}(\mathcal{M}_{\mathcal{H}|_c,\mathbb{V}}),
\end{aligned}$$

where the equality (1) is because for a sequence $(p_c \in \Delta(\mathcal{A}))_{c\in\mathcal{C}}$, there is a corresponding $p \in \Delta(\Pi)$ such that for $\pi \sim p$, we have $\pi(c) \sim p_c$ independently; in inequality (2) we bound the expectation

over $c \sim \nu_M$ by the supremum $\sup_{c \in \mathcal{C}}$; the equality (3) follows from the fact that $M_c \in \mathcal{M}_{\mathcal{H}|_c, \mathbb{V}}$ is a bandit model with mean reward function $h^{M_c}(\cdot) = h^M(c, \cdot)$; and the equality (4) is because we can choose p_c separately for every $c \in \mathcal{C}$. By the arbitrariness of $\overline{M} \in \mathrm{co}(\mathcal{M})$, we now have

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\mathcal{H}, \mathbb{V}}) \le \sup_{c \in \mathcal{C}} \mathrm{dec}^{\mathrm{o}}_{\gamma/2}(\mathcal{M}_{\mathcal{H}|_c, \mathbb{V}}).$$

□

I.7 Proof of Corollary H.3

We follow the notations of Appendix I.6. By Lemma I.7, we have

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\mathcal{H}, \mathbb{V}}) \le \frac{1}{\gamma} + \sup_{c \in \mathcal{C}} \textsf{r-dec}^{\mathrm{o}}_{\gamma/4}(\mathcal{M}_{\mathcal{H}|_c, \mathbb{V}}).$$

Notice that for each $c \in \mathcal{C}$, $\mathcal{M}_{\mathcal{H}|_c, \mathbb{V}}$ is a class of $|\mathcal{A}|$-arm bandits, and hence by Foster et al. [36, Proposition 5.1] and Lemma C.5, we have

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\mathcal{H}|_c, \mathbb{V}}) \le \frac{8|\mathcal{A}|}{\gamma}.$$

Therefore, Theorem I.3 implies that $\textsf{ExO}^+$ achieves with probability at least $1 - \delta$:

$$\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} \le \Delta + O\left(\frac{|\mathcal{A}|}{\gamma} + \gamma \frac{\log \textsf{Ddim}_{\Delta}(\mathcal{H}) + \log(1/\delta)}{T}\right).$$

Balancing $\gamma > 0$ gives the desired upper bound. □

As a remark, we provide an example of function class $\mathcal{H}$ with $\log \textsf{Ddim}_{\Delta}(\mathcal{H}) \ll \log |\mathcal{H}|$.

Example 7. Suppose that $\mathcal{A} = \{0, 1\}$, and the function class $\mathcal{H} = \{h_x\}_{x \in \mathcal{C}}$, where

$$h_x(c, 0) = \frac{1}{2}, \qquad h_x(c, 1) = \begin{cases} 1, & c = x, \\ 0, & c \neq x. \end{cases}.$$

Clearly, we have $\log |\mathcal{H}| = \log |\mathcal{C}|$.

On the other hand, we consider a distribution p over policies, such that $\pi \sim p$ is generated as $\pi(c) \sim \mathrm{Bern}(\varepsilon)$, independently over all $c \sim \mathcal{C}$. Then, for any $h = h_x \in \mathcal{H}$ and $\nu \in \Delta(\mathcal{C})$, we have

$$\mathbb{E}_{c \sim \nu}[h(c, \pi_h(c)) - h(c, \pi(c))] = \nu(x) \cdot \frac{1}{2}\mathbf{1}\{\pi(x) = 1\} + \frac{1}{2}\mathbb{E}_{c \sim \nu}[\mathbf{1}\{c \neq x, \pi(c) = 1\}].$$

Notice that $\pi(x) = 1$ with probability Δ, and conditional on the event $\{\pi(x) = 1\}$,

$$\mathbb{E}_{\pi \sim p}[\mathbb{E}_{c \sim \nu}[\mathbf{1}\{c \neq x, \pi(c) = 1\}] | \pi(x) = 1] \le \Delta.$$

Hence,

$$p(\pi : \mathbb{E}_{c \sim \nu}[h(c, \pi_h(c)) - h(c, \pi(c))] \le \Delta) \ge \frac{\Delta}{2},$$

which implies $\log \textsf{Ddim}_{\Delta}(\mathcal{H}) \le \log(2/\Delta)$.

Therefore, for unbounded context space $\mathcal{C}$, we have $\log \textsf{Ddim}_{\Delta}(\mathcal{H}) \ll \log |\mathcal{H}|$ for the function class $\mathcal{H}$ defined above.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. The claims are validated by detailed proofs.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes] .

Justification: The paper discusses the limitations of the work performed.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] .

   Justification: The paper provides detailed assumptions and proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA] .

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not

including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes] .

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theoretical work. There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.