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

Fan Chen* fanchen@mit.edu
Dylan J. Foster† dylanfoster@microsoft.com
Yanjun Han‡ yanjunhan@nyu.edu
Jian Qian* jianqian@mit.edu
Alexander Rakhlin* rakhlin@mit.edu
Yunbei Xu§ 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 method, 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. [2021, 2023b] 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 *Fractional Covering Number*, which facilitates the new lower bounds for interactive decision making that extend the DEC methodology by incorporating the complexity of estimation. Using the fractional covering number, we (i) provide a unified characterization of learnability for *any* stochastic bandit problem, (ii) close the remaining gap between the upper and lower bounds in Foster et al. [2021, 2023b] (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. For an algorithm ALG that collects data (either passively or interactively) from the model M, the minimax criterion (stated somewhat informally here) is

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

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

*Massachusetts Institute of Technology.
†Microsoft Research.
‡New York University.
§National University of Singapore.

statistics, the minimax approach was pioneered by A. Wald [Wald, 1945], who made the connection to von Neumann's theory of games [Von Neumann and Morgenstern, 1944] and unified statistical estimation and hypothesis testing under the umbrella of *statistical decision theory*. Minimax optimality and minimax rates of convergence of estimators have since become a central object in the modern non-asymptotic statistics [van de Geer, 2000, Wainwright, 2019]; 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 method, and Assouad's method. These techniques entail constructing "difficult" subsets of the class $\mathcal{M}$. Le Cam's method focuses on two hypotheses, while Assouad's method and Fano's method involve multiple hypotheses indexed by the vertices of a hypercube and a simplex, respectively. The relationships between these methods are explored in Yu [1997].

Classical statistical estimation is a purely passive task. A parallel line of research [Lattimore and Szepesvári, 2020a] 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 [Agarwal et al., 2012, Raginsky and Rakhlin, 2011a,b], 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. [2021, 2023b] shows that a complexity measure known as the *Decision-Estimation Coefficient* (DEC) leads to both lower and upper bounds on the minimax rates for a general class of interactive decision making problems. Interestingly, the lower bound techniques in Foster et al. [2021] 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. Yet, while the setting of interactive decision making encompasses statistical estimation, the DEC lower bound does *not* recover Fano's method or Assouad's lemma. Intuitively, DEC captures the complexity of *exploration* for the interactive problems, which is complementary to the complexity of *estimation*, typically captured by Fano's method and Assouad's lemma. As a result, the DEC only characterizes the minimax rates for interactive decision making up to a gap related to the complexity of a certain induced estimation problem; see Section 2.3 for detailed discussion.

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. [2021, 2023b], 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 (estimation-based) gaps between the upper and lower bounds on the minimax rates for interactive decision making left open by Foster et al. [2023b].

1.1 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 Decision-Estimation Coefficient-based lower bounds for interactive decision making [Foster et al., 2021, 2023b].

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's method to a novel algorithm-dependent condition by introducing the concept of "ghost data" generated from a reference distribution. This technique recovers Le Cam's two-point method (and the convex hull method), Assouad's method, and Fano's method as special cases. Further, by virtue of being algorithm-dependent in nature, the interactive Fano method also seamlessly recovers DEC-based lower bounds for interactive decision making as a special case, and leads to refined quantile-based variants.

Fractional covering number 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 *fractional covering number*, 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 fractional covering number provides the first complete characterization for finite-time learnability of any structured bandit problem, albeit up to an exponential gap in quantitative rates. As a secondary result, we use the fractional covering number to close the remaining gap between the upper and lower bounds in Foster et al. [2021, 2023b], up to polynomial factors, for any interactive decision making problem in which the underlying model class is convex.

1.2 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

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

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. We first introduce the framework and show how it subsumes statistical estimation (Section 2.1) and interactive decision making (Section 2.2), then give brief background on existing lower bound techniques and gaps in understanding (Section 2.3).

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 $\mathsf{ALG} \in \mathcal{D}$ 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 ALG: $X \sim \mathbb{P}^{M,\mathsf{ALG}}$. The performance of the algorithm 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 ALG. As described in the Introduction, the best possible expected risk the learner may achieve is the following *minimax risk*:

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

While our main results concern the general problem formulation in Eq. (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

In statistical decision theory [Wald, 1945, Bickel and Doksum, 2001, Berger, 1985], 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 incurs the loss $L(\theta^\star, A)$. This general framework subsumes most statistical estimation problems.

Example 1 (Mean estimation)**.** For the mean estimation task, the parameter space $\Theta \subseteq \mathbb{R}^d$, and for each $\theta \in \Theta$, $P_\theta = \mathcal{N}(\theta, I_d)$. The goal is to estimate the ground truth parameter $\theta^\star$, i.e., the decision space is $\mathcal{A} = \mathbb{R}^d$, and the loss is given by $L(\theta, A) = \|\theta - A\|$, where $\|\cdot\|$ is a norm over $\mathbb{R}^d$.

Example 2 (Functional estimation)**.** In the functional estimation task, a function $T : \Theta \to \mathbb{R}$ is given, and the goal is to estimate the value of $T(\theta^\star)$, i.e., the decision space is $\mathcal{A} = \mathbb{R}$, and the loss is $L(\theta, A) = |T(\theta) - A|$.

Example 3 (Density estimation)**.** In the density estimation task, the goal is to estimate $P_{\theta^\star}$, i.e., the decision space $\mathcal{A} \subseteq \Delta(\mathcal{Y})$, and the loss is given by $L(\theta, A) = D(P_\theta, A)$, where D is a certain divergence (e.g., TV distance or KL divergence).

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

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

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 [Foster et al., 2021], 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 $o^t \in \mathcal{O}$ sampled 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 A.

To each model $M \in \mathcal{M}$, we associate a *risk* function $L(M, \cdot) : \Pi \to \mathbb{R}_{\geq 0}$, which measures the performance of a decision in Π under M. 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

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

  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

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

With an appropriate choice for L, the setting captures reward maximization (regret minimization) [Foster et al., 2021, 2023b], model estimation and preference-based learning [Chen et al., 2022a], multi-agent decision making and partial monitoring [Foster et al., 2023a], and various other tasks. In the main text of this paper, we focus on reward maximization and defer the results for more general choices L to the appendices (cf. Appendix A).

Example 4 (Reward maximization)**.** In the reward-maximization task, $R : \mathcal{O} \to [0, 1]$ is 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. For any $M \in \mathcal{M}$, we let $\pi_M \in \arg\max_{\pi \in \Pi} f^M(\pi)$ be an optimal decision under M, and the risk function is defined by $L(M, \pi) = f^M(\pi_M) - f^M(\pi)$, measuring the sub-optimality of the decision π under model M.

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

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 X of the form $X = (\mathcal{H}^T, \widehat{\pi})$, where $\widehat{\pi}$ is a final decision. An algorithm $\mathsf{ALG} = \{q^t\}_{t\in[T]} \cup \{p\}$ 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 $L(M^\star, X) = L(M^\star, \widehat{\pi})$ for PAC learning (4), and $L(M^\star, X) = \sum_{t=1}^T L(M^\star, \pi^t)$ for no-regret learning (3). For any algorithm ALG and model M, $\mathbb{P}^{M,\mathsf{ALG}}(\cdot)$ is the distribution of $X = (\mathcal{H}^T, \widehat{\pi})$ generated by the algorithm ALG under the model M, and we let $\mathbb{E}^{M,\mathsf{ALG}}[\cdot]$ to be the corresponding expectation.

2.3 Background on Lower Bound Techniques

To motivate our results, we briefly survey the most relevant lower bound techniques for estimation and decision making.

Minimax bounds for statistical estimation. There is a vast body of literature on minimax risk bounds for statistical estimation, including Hasminskii and Ibragimov [1979], Bretagnolle and Huber [1979], Birgé [1986], Donoho and Liu [1991a], Cover and Thomas [1999], Ibragimov and Has'Minskii [1981], Tsybakov [2008] as well as references therein. For minimax lower bounds, the most widely applied techniques are Le Cam's two-point method and convex-hull method [LeCam, 1973], Assouad's lemma [Assouad, 1983], and Fano's method [Cover and Thomas, 1999]. Variants and applications of these three methods abound [Acharya et al., 2021, Chen et al., 2016, Polyanskiy and Wu, 2019, Duchi and Wainwright, 2013]; Fano's method 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. [2016], which is recovered by our interactive Fano method (cf. Proposition 3). Another celebrated thread, starting from the seminal work of Donoho and Liu [1987], 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 [Donoho and Liu, 1991a,b, Le Cam and Yang, 2000, Polyanskiy and Wu, 2019].

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 [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Simchowitz et al., 2017, Lattimore and Szepesvári, 2020a, Kleinberg et al., 2019, etc.], contextual bandits [Rigollet and Zeevi, 2010, Foster et al., 2020, etc.], Markov Decision Processes (MDPs) [Osband and Van Roy, 2016, Domingues et al., 2021, Zhou et al., 2021, Weisz et al., 2021, Wang et al., 2021, etc.], partially observable RL [Krishnamurthy et al., 2016, Liu et al., 2022, Chen et al., 2023, 2024, etc.], dynamical systems and control [Simchowitz et al., 2018, Jedra and Proutiere, 2019, Simchowitz and Foster, 2020, Wagenmaker and Jamieson, 2020, Ziemann and Sandberg, 2024, etc.], and offline RL [Rashidinejad et al., 2021, Xie et al., 2021, Wagenmaker et al., 2021, Jin et al., 2021, Chen et al., 2022b, Li et al., 2024, Wagenmaker et al., 2024, etc.]. Most of these lower bounds are proven in a case-by-case basis, as the constructions of hard instances are specialized to the specific settings.

Decision-Estimation Coefficient. Toward a unifying understanding of the minimax complexity for interactive decision making problems, Foster et al. [2021, 2023b] 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 [Donoho and Liu, 1987], 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 general decision making settings [Foster et al., 2021, 2022, Chen et al., 2022a, Foster et al., 2023b,a, Glasgow and Rakhlin, 2023], including adversarial decision making [Foster et al., 2022], PAC decision making [Chen et al., 2022a, Foster et al., 2023b], reward-free learning and preference-based learning [Chen et al., 2022a], and multi-agent decision making and partial monitoring [Foster et al., 2023a].

However, there is a remaining gap between the DEC lower and upper bounds [Foster et al., 2021, 2023b], which closely relates to the complexity of *estimation*. Specifically, through the DEC framework, the sample complexity (number of rounds required to achieve ε-risk) is characterized as

$$T^{\mathsf{DEC}}(\mathcal{M},\varepsilon) \lesssim \# \text{ sample complexity} \lesssim T^{\mathsf{DEC}}(\mathcal{M},\varepsilon) \times \mathsf{Est}(\mathcal{M}),$$

where $T^{\mathsf{DEC}}(\mathcal{M},\varepsilon)$ is a quantity measuring the complexity of *exploration*,[2] and $\mathsf{Est}(\mathcal{M})$ is a measure of the complexity of *online estimation* over $\mathcal{M}$. The dependency on $\mathsf{Est}(\mathcal{M})$ can be necessary: For example, in linear bandits, the optimal sample complexity scales as d^2/ε^2, while $T^{\mathsf{DEC}}(\mathcal{M},\varepsilon) \asymp d/\varepsilon^2$, and $\mathsf{Est}(\mathcal{M}) \asymp d$. However, the complexity of estimation $\mathsf{Est}(\mathcal{M})$ is missing from the DEC lower bound. This gap remains one of the main open questions in the DEC approach.

One potential reason is that the DEC lower bound does *not* recover Fano's method or Assouad's lemma, as it essentially generalizes Le Cam's two-point method. More specifically, while the statistical estimation task is subsumed by the DMSO framework, the DEC lower bound specialized to this setting at best recovers Le Cam's two-point method. On the other hand, the $\Omega(d^2/\varepsilon^2)$ lower bound for linear bandits is typically proven through Assouad's lemma [Bubeck et al., 2015] or Fano's method [Rajaraman et al., 2023], similar to its statistical estimation analog. Therefore, to close the remaining gap in the DEC approach, it is necessary to have a deeper understanding of the latter two methods in the interactive setting.

Additional related work. A large portion of the aforementioned lower bounds for interactive learning are proven using (variants of) the two-point method and can be recovered by the DEC lower bound approach [Foster et al., 2021, 2023b]. Beyond the two-point method, comparatively fewer lower bounds for interactive learning have been established using Assouad's lemma or Fano's method [Castro and Nowak, 2008, Rigollet and Zeevi, 2010, Agarwal et al., 2009, Raginsky and Rakhlin, 2011b, Foster et al., 2020, Simchowitz and Foster, 2020, Rajaraman et al., 2023, etc.]. The approaches in these papers are specialized to the specific settings under consideration, and there is not a general principle through which Fano's method or Assouad's lemma can be lifted to handle interactivity. Indeed, the challenge of applying Fano's method in interactive contexts has been highlighted in various prior works, e.g., Arias-Castro et al. [2012, Section 1.3] and Rajaraman et al. [2023, Section 1.5.4].

The DEC is also closely related to a Bayesian complexity measure known as the information ratio [Russo and Van Roy, 2016, 2018, Lattimore and Szepesvári, 2019, Lattimore and Gyorgy, 2021, 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) [Xu and Zeevi, 2023], developed for frequentist algorithms. Additionally, the DEC is connected to asymptotic instance-dependent complexity, as explored by Wagenmaker and Foster [2023].

[2]Formally, the quantity $T^{\mathsf{DEC}}(\mathcal{M},\varepsilon)$ here is the sample complexity implied by DEC (cf. Section 4).

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.

Background: Fano's method. To motivate our approach, we which can be viewed as a generalization of the classical Fano method [Cover and Thomas, 1999], let us briefly recall the classical approach and highlight some shortcomings. The classical Fano method applies to the statistical estimation setting Section 2.1 (a special case of ISDM), and takes the following form.

Proposition 1 (Classical Fano method)**.** *Consider the statistical estimation setting (Section 2.1) with parameter space Θ. Suppose that there exist $\theta_1, \ldots, \theta_m \in \Theta$ such that the following separation condition holds:*

$$L(\theta_i, a) + L(\theta_j, a) \geq 2\Delta, \quad \forall i \neq j \in [m], \forall a \in \mathcal{A}. \tag{5}$$

Let μ be the uniform distribution over $\{\theta_1, \cdots, \theta_m\}$, and let $I_\mu(\theta; Y)$ denote the mutual information of $(\theta, Y) \sim \mathbb{P}_\mu$ generated by $\theta \sim \mu$ and $Y = (Y_1, \ldots, Y_n) \overset{\text{i.i.d}}{\sim} P_\theta$. Then for any algorithm ALG, we have

$$\mathbb{E}_{\theta\sim\mu, Y\sim P_\theta}\left[L(\theta, \mathsf{ALG}(Y))\right] \geq \Delta \cdot \sup_{\delta>0}\{\delta : I_\mu(\theta; Y) < \mathrm{kl}(1-\delta \,\|\, 1/m)\}, \tag{6}$$

where the binary KL divergence is defined as $\mathrm{kl}(p \,\|\, q) = D_{\mathrm{KL}}(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(q))$. This implies the minimax lower bound

$$\inf_{\mathsf{ALG}} \sup_{\theta\in\Theta} \mathbb{E}_{Y\sim P_\theta}\left[L(\theta, \mathsf{ALG}(Y))\right] \geq \Delta\left(1 - \frac{I_\mu(\theta; Y) + \log 2}{\log m}\right). \tag{7}$$

The $\log m$ factor in Eq. (7) reflects the complexity of estimation in the parameter space, which is a key concept we aim to incorporate into interactive decision making. Looking deeper, the "estimation complexity" term $\log m$ in Eq. (7) arises from the "quantile" parameter $1/m$ appearing in the Eq. (6). This parameter reflects the fact that under the separation condition (5), the following *quantile probability* is at most $1/m$ for any distribution $Y \sim \mathbb{Q}$:

$$\mathbb{P}_{\theta\sim\mu, Y\sim\mathbb{Q}}(L(\theta, \mathsf{ALG}(Y)) < \Delta) \leq \sup_a \mathbb{P}_{\theta\sim\mu}(L(\theta, a) < \Delta) \leq \frac{1}{m}. \tag{8}$$

Note that in this expression, θ is drawn from the uniform prior μ and Y is drawn *independently* of θ. To deduce Eq. (6), it suffices to choose $\mathbb{Q} = \mathbb{E}_{\theta\sim\mu} P_\theta$ and apply data-processing inequality:

$$\begin{aligned} I_\mu(\theta; Y) &\geq \mathrm{kl}(\mathbb{P}_\mu(L(\theta, \mathsf{ALG}(Y)) < \Delta) \,\|\, \mathbb{P}_{\theta\sim\mu, Y\sim\mathbb{Q}}(L(\theta, \mathsf{ALG}(Y)) < \Delta)) \\ &\geq \mathrm{kl}(\mathbb{P}_\mu(L(\theta, \mathsf{ALG}(Y)) < \Delta) \,\|\, 1/m). \end{aligned}$$

In particular, for any $\delta \in (0, 1)$ such that $I_\mu(\theta; Y) \leq \mathrm{kl}(1-\delta \,\|\, 1/m)$, we have $\mathbb{P}_\mu(L(\theta, \mathsf{ALG}(Y)) < \Delta) \leq 1-\delta$, using the monotonicity of the KL divergence. This argument gives Eq. (6) immediately, and by choosing $\delta_\star = 1 - \frac{I_\mu(\theta;Y)+\log 2}{\log m}$ in Eq. (6), we arrive in the canonical statement in Eq. (7).

To summarize, the structure of the classical Fano lower bound involves (i) a **prior** μ, (ii) a **reference distribution** $\mathbb{Q} = \mathbb{E}_{\theta\sim\mu} P_\theta$, and (iii) a **quantile parameter** δ determined by the (iv) **separation condition** (5) in the argument above. Crucially, we understand that the complexity of estimation

$\log m$ arises from the *quantile probability* $\mathbb{P}_{\theta\sim\mu,Y\sim\mathbb{Q}}(L(\theta,\mathsf{ALG}(Y)) < \Delta)$, and the only use of the traditional separation condition (5) is to further bound this quantile by $1/m$ as in Eq. (8).

Having gained these insights into classical Fano method, we would like to point out several limitations:

- First, in the form above, it is only applicable to statistical estimation rather than general interactive settings.
- Second, it relies on mutual information due to the choice of the reference distribution, which depends on the evolution of the algorithm over all T rounds in interactive settings, making it difficult to analyze in many interactive problems.
- Third, and perhaps most importantly, the separation condition (5) must hold for an arbitrary decision a. This "hard" separation condition is unlikely to hold for general model classes, as noted throughout the DEC approach [Foster et al., 2023b, Remark 2.3] line of work.

To address these shortcomings, we make use of core concepts (prior, reference distribution, quantile parameter, separation condition) above, but adopt a new perspective that emphasizes and generalizes the role of the quantile probability $\mathbb{P}_{\theta\sim\mu,Y\sim\mathbb{Q}}(L(\theta,\mathsf{ALG}(Y)) < \Delta)$.

The interactive Fano method. We now present our new lower bound approach, the interactive Fano method. The core idea here is to relax the separation condition by introducing a general reference distribution $\mathbb{Q}$. We also consider a a general prior μ and a general f-divergence D_f.

Theorem 2 (Interactive Fano method)**.** *Fix a f-divergence D_f. Let ALG be a given algorithm, $\mu \in \Delta(\mathcal{M})$ be a given prior distribution over models, and $\Delta > 0$ be a given risk level. For any reference distribution $\mathbb{Q} \in \Delta(\mathcal{X})$, we define*

$$\rho_{\Delta,\mathbb{Q}} = \mathbb{P}_{M\sim\mu,X\sim\mathbb{Q}}(L(M,X) < \Delta). \tag{9}$$

Then, the following quantile lower bound holds:

$$\mathbb{P}_{M\sim\mu,X\sim\mathbb{P}^{M,\mathsf{ALG}}}(L(M,X) \geq \Delta) \geq \sup_{\mathbb{Q}\in\Delta(\mathcal{X}),\delta\in[0,1]} \left\{\delta : \mathbb{E}_{M\sim\mu}[D_f(\mathbb{P}^{M,\mathsf{ALG}},\mathbb{Q})] < \mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}})\right\}, \tag{10}$$

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.

In particular, Eq. (10) implies the following in-expectation lower bound:

$$\begin{aligned}\sup_{M\in\mathcal{M}} \mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}[L(M,X)] &\geq \mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}[L(M,X)] \\ &\geq \Delta\cdot \sup_{\mathbb{Q}\in\Delta(\mathcal{X}),\delta\in[0,1]} \left\{\delta : \mathbb{E}_{M\sim\mu}[D_f(\mathbb{P}^{M,\mathsf{ALG}},\mathbb{Q})] < \mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}})\right\}.\end{aligned}$$

This result generalizes the classical Fano method in the prequel (as well as more sophisticated variants [Zhang, 2006, Duchi and Wainwright, 2013, Chen et al., 2016]) 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.
- It makes use of an arbitrary user-specified f-divergence D_f and a general prior μ rather than the discrete uniform prior, both of which are generalizations of the Fano method that have appeared in previous works [Zhang, 2006, Duchi and Wainwright, 2013, Chen et al., 2016].

- The most important and novel change is that Theorem 2 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 Eq. (9). 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 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. [2021, 2023b].

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 the DEC approach [Foster et al., 2021, 2023b] to derive new lower bounds (see Section 3.2).

3.1 Recovering non-interactive lower bounds

We begin by applying Theorem 2 to recover classical non-interactive lower bounds for statistical estimation. Since a goal of our paper is to integrate the Fano and Assouad methods (which provide dimensional insights but are typically challenging to apply in interactive settings) 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.

3.1.1 Generalized Fano method

We begin by recovering a generalized version of the classical Fano method which subsumes Proposition 1, as well as other prior generalizations [Zhang, 2006, Duchi and Wainwright, 2013, Chen et al., 2016] developed in statistical estimation.

Proposition 3 (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* $\Delta \geq 0$*:*

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}\left[L(M,X)\right] \geq \Delta\left(1 + \frac{I_{\mu,\mathsf{ALG}}(M;X) + \log 2}{\log \sup_x \mu(M : L(M,x) < \Delta)}\right). \tag{11}$$

This result recovers the classical Fano method for statistical estimation (Proposition 1), which corresponds to the special case where $\Theta = \{1, 2, \ldots, m\}$, $\mu = \mathrm{Unif}(\Theta)$ is the uniform prior, and $\sup_a \mu(\theta : L(\theta, a) < \Delta) \leq 1/m$ under the separation condition (5). In its narrowest form, Fano's inequality—which is often used as a lemma in classical Fano's method—is a further specialization obtained by considering $\mathcal{A} = \Theta$ and the indicator function $\mathbb{1}(\theta \neq a)$. We note that Proposition 3 is stated for *interactive* setting, while the generalized Fano's inequalities in previous works [Zhang, 2006, Duchi and Wainwright, 2013, Chen et al., 2016] is only stated for statistical estimation.

Proof of Proposition 3. Fix the parameter $\Delta > 0$ and let $\mu \in \Delta(\mathcal{M})$ be given. To apply Theorem 2, we consider KL divergence (corresponding to $f(x) = x \log x$) and choose the reference distribution

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

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

$$\mathbb{E}_{M\sim\mu}D_{\mathrm{KL}}(\mathbb{P}^{M,\mathsf{ALG}} \,\|\, \mathbb{Q}) = I_{\mu,\mathsf{ALG}}(M;X),$$

and by definition, we have

$$\rho_{\Delta,\mathbb{Q}} = \mathbb{P}_{M\sim\mu,X'\sim\mathbb{Q}}(L(M,X') < \Delta) \le \sup_x \mu(M : L(M,x) < \Delta), \tag{12}$$

By Theorem 2, for any $\delta \in (0,1)$ such that $I_{\mu,\mathsf{ALG}}(M;X) < \mathrm{kl}(1-\delta \,\|\, \rho_{\Delta,\mathbb{Q}})$, we have

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

In particular, we may choose

$$\delta_\star := 1 + \frac{I_{\mu,\mathsf{ALG}}(M;X) + \log 2}{\log \sup_x \mu(M : L(M,x) < \Delta)}.$$

As long as $\delta_\star > 0$, we have $I_{\mu,\mathsf{ALG}}(M;X) < \mathrm{kl}(1-\delta_\star \,\|\, \rho_{\Delta,\mathbb{Q}})$, and hence $\mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}[L(M,X)] \ge \delta_\star\Delta$. This gives the desired lower bound (note that if $\delta_\star \le 0$, there is nothing to prove). □

Note that in Proposition 3, the term $\log\sup_x \mu(M \in \mathcal{M} : L(M,x) < \Delta)$ in the denominator of Eq. (11) 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 to derive tight guarantees for interactive decision making in general. The Decision-Estimation Coefficient, which we define in Section 3.2, more precisely accounts for the role of decisions selected by the algorithm on the loss $L(M,X)$ (as well as the information acquired).

3.1.2 Le Cam's two-point method and Assouad's method

To recover Le Cam's two-point method and Assouad's method from Theorem 2, we appeal to the following result, which recovers a lower bound known as the Le Cam convex hull method [LeCam, 1973, Yu, 1997] which generalizes both approaches.

Proposition 4 (Recovering Le Cam's convex hull method)**.** *For a parameter space Θ and observation space $\mathcal{Y}$, consider a class of distributions $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\} \subseteq \Delta(\mathcal{Y})$ indexed by Θ. Let $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$ be a loss function. Let $\Theta_0 \subseteq \Theta$ and $\Theta_1 \subseteq \Theta$ be subsets that satisfy the* separation condition

$$L(\theta_0, a) + L(\theta_1, a) \ge 2\Delta, \quad \forall a \in \mathcal{A}, \theta_0 \in \Theta_0, \theta_1 \in \Theta_1 \tag{13}$$

for a parameter $\Delta > 0$. Then it holds that

$$\inf_{\mathsf{ALG}} \sup_{\theta\in\Theta} \mathbb{E}_{Y\sim P_\theta} L(\theta, \mathsf{ALG}(Y)) \ge \frac{\Delta}{2} \max_{\nu_0\in\Delta(\Theta_0),\nu_1\in\Delta(\Theta_1)} \left(1 - D_{\mathrm{TV}}\left(P_{\nu_0}^{\otimes n}, P_{\nu_1}^{\otimes n}\right)\right),$$

where the infimum is taken over all algorithms $\mathsf{ALG} : \mathcal{Y}^{\otimes n} \to \mathcal{A}$, and $P_{\nu_i}^{\otimes n}$ is the distribution on $\mathcal{Y}^{\otimes n}$ induced by $\theta \sim \nu_i, Y = (Y_1, \ldots, Y_n) \overset{\text{i.i.d}}{\sim} P_\theta$ for $i \in \{0,1\}$.

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 [Yu, 1997]. It is important to note that the classical Fano's method, e.g. in the form of Proposition 3, cannot recover Proposition 4. This is because of fundamental differences between the divergences (KL versus TV) used in the traditional Fano method and the convex hull method.

Proof of Proposition 4. We recover this result by applying Theorem 2 with TV distance. We first frame the problem in the ISDM framework. Consider the "enlarged" model class $\mathcal{M} = \{M_\nu : \nu \in \Delta(\Theta)\}$, where for any algorithm $\mathsf{ALG} : \mathcal{Y}^{\otimes n} \to \mathcal{A}$, the distribution $\mathbb{P}^{M_\nu,\mathsf{ALG}}$ is given by

$$X = (Y, \mathsf{ALG}(Y)) \sim \mathbb{P}^{M_\nu,\mathsf{ALG}} : \ \theta \sim \nu, Y = (Y_1, \cdots, Y_n) \overset{\text{i.i.d}}{\sim} P_\theta.$$

In other words, $\mathbb{P}^{M_\nu,\mathsf{ALG}}$ is the distribution induced by the prior ν and the algorithm ALG. We then extend the loss function to $L : \mathcal{M} \times \mathcal{X} \to \mathbb{R}_+$ as

$$L(M_\nu, X) := \inf_{\theta \in \mathrm{supp}(\nu)} L(\theta, \mathsf{ALG}(Y)), \qquad \forall X = (Y, \mathsf{ALG}(Y)),\ \nu \in \Delta(\Theta).$$

By the separation condition (13), we have $L(M_{\nu_0}, X) + L(M_{\nu_1}, X) \geq 2\Delta$ for any $\nu_0 \in \Delta(\Theta_0)$, $\nu_1 \in \Delta(\Theta_1)$. Therefore, choosing the prior $\mu = \mathrm{Unif}(\{M_{\nu_0}, M_{\nu_1}\})$ and the reference distribution $\mathbb{Q} = \mathbb{E}_{M\sim\mu}\mathbb{P}^{M,\mathsf{ALG}}$ gives

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

and by the data-processing inequality,

$$\begin{aligned}\mathbb{E}_{M\sim\mu}[D_{\mathrm{TV}}(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{Q})] &= \frac{1}{2}\big(D_{\mathrm{TV}}\left(\mathbb{P}^{M_{\nu_0},\mathsf{ALG}}, \mathbb{Q}\right) + D_{\mathrm{TV}}\left(\mathbb{P}^{M_{\nu_1},\mathsf{ALG}}, \mathbb{Q}\right)\big) \\ &\leq \frac{1}{2} D_{\mathrm{TV}}\left(\mathbb{P}^{M_{\nu_0},\mathsf{ALG}}, \mathbb{P}^{M_{\nu_1},\mathsf{ALG}}\right) \leq \frac{1}{2} D_{\mathrm{TV}}\left(P_{\nu_0}^{\otimes n}, P_{\nu_1}^{\otimes n}\right).\end{aligned}$$

Therefore, for any $0 \leq \delta < \frac{1}{2} - \frac{1}{2} D_{\mathrm{TV}}\left(P_{\nu_0}^{\otimes n}, P_{\nu_1}^{\otimes n}\right)$, we have

$$\mathbb{E}_{M\sim\mu}[D_{\mathrm{TV}}(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{Q})] \leq \mathsf{d}_{\mathrm{TV},\delta}(\rho_{\Delta,\mathbb{Q}}),$$

and hence applying Theorem 2 gives

$$\mathbb{E}_{\theta\sim\frac{\nu_0+\nu_1}{2}} \mathbb{E}_{Y\sim P_\theta}[L(\theta, \mathsf{ALG}(Y))] \geq \mathbb{E}_{M\sim\mu}\mathbb{E}_{X\sim\mathbb{P}^{M,\mathsf{ALG}}}[L(M, X)] \geq \frac{\Delta}{2}\big(1 - D_{\mathrm{TV}}\left(P_{\nu_0}^{\otimes n}, P_{\nu_1}^{\otimes n}\right)\big).$$

Taking the supremum over $\nu_0 \in \Delta(\Theta_0)$ and $\nu_1 \in \Delta(\Theta_1)$ gives the desired result. □

We can use the Le Cam convex hull method to recover the classic two-point method and Assouad's method, as follows.

Example 5 (Le Cam's two-point method)**.** In Proposition 4, we can take the distribution ν_0 (ν_1) to be supported on a single point in Θ_0 (Θ_1), to recover the classical two-point method. Concretely, under the setting and assumption of Proposition 4, we have the following two-point lower bound:

$$\inf_{\mathsf{ALG}} \sup_{\theta\in\Theta} \mathbb{E}_{Y\sim P_\theta} L(\theta, \mathsf{ALG}(Y)) \geq \frac{\Delta}{2} \max_{\theta_0\in\Theta_0, \theta_1\in\Theta_1} \left(1 - D_{\mathrm{TV}}\left(P_{\theta_0}^{\otimes n}, P_{\theta_1}^{\otimes n}\right)\right),$$

where $P_\theta^{\otimes n}$ is the distribution of $Y = (Y_1, \cdots, Y_n) \overset{\text{i.i.d}}{\sim} P_\theta$. ◁

Example 6 (Assouad's method)**.** Suppose that $\Theta = \{-1, 1\}^d$ for some $d \geq 1$, and that the loss function has the following coordinate-wise structure:

$$L(\theta, a) = \sum_{j=1}^{d} L_j(\theta, a), \qquad \forall \theta \in \Theta, a \in \mathcal{A}.$$

We write $\theta \sim_j \theta'$ if θ and θ' differ only in the j-th coordinate. Assume that the following separation condition holds for all $j \in [d]$:

$$L_j(\theta, a) + L_j(\theta', a) \geq 2\Delta, \quad \forall a \in \mathcal{A}, \theta \sim_j \theta'.$$

To apply Proposition 4, we consider $\Theta_i^j = \{\theta : \theta_j = i\}$ for $i \in \{0,1\}$ and $j \in [d]$. Then, for each $j \in [d]$, the separation condition (13) holds for the loss L_j and the subsets Θ_0^j, Θ_1^j. Therefore, applying Proposition 4 for each $j \in [d]$ with $\nu_0^j = \mathrm{Unif}(\Theta_0^j)$ and $\nu_1^j = \mathrm{Unif}(\Theta_1^j)$ gives the following Assouad-type lower bound:

$$\inf_{\mathsf{ALG}} \sup_{\theta \in \Theta} \mathbb{E}_{Y \sim P_\theta} L(\theta, \mathsf{ALG}(Y)) \geq \frac{d\Delta}{2} \min_{\exists j : \theta \sim_j \theta'} \left(1 - D_{\mathrm{TV}}\left(P_\theta^{\otimes n}, P_{\theta'}^{\otimes n}\right)\right).$$

◁

3.2 Recovering DEC-based lower bounds for interactive decision making

Within the DMSO framework (Section 2.2), Foster et al. [2021, 2023b] introduced the *Decision-Estimation Coefficient* (DEC) as a complexity measure governing the statistical complexity of interactive decision making, 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. [2021, 2023b] through Theorem 2. We focus on the lower bounds from Foster et al. [2023b], which are based on a variant of the DEC called the *constrained DEC*, and provide the tightest guarantees from prior work.

3.2.1 Background on the Decision-Estimation Coefficient

Consider the reward maximization variant of the DMSO setting (Example 4). 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

$$\mathsf{r\text{-}dec}_\varepsilon^{\mathrm{c}}(\mathcal{M}, \overline{M}) := \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[L(M, \pi)] \mid \mathbb{E}_{\pi \sim p} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \leq \varepsilon^2 \right\}, \tag{14}$$

and define the constrained PAC-DEC via

$$\mathsf{p\text{-}dec}_\varepsilon^{\mathrm{c}}(\mathcal{M}, \overline{M}) := \inf_{p, q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[L(M, \pi)] \mid \mathbb{E}_{\pi \sim q} D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right) \leq \varepsilon^2 \right\}. \tag{15}$$

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

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

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

Based on these complexity measures, Foster et al. [2023b] (see also Glasgow and Rakhlin [2023]) provide the following lower and upper bounds on optimal risk and regret, under mild growth conditions on the DECs.

Theorem 5 (Informal; Foster et al. [2023b]). *Consider the reward maximization variant of the DMSO setting (Example 4). 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. Informally, the $\log|\mathcal{M}|$ factor in the upper bound arises from the complexity of estimating the underlying model $M^\star$, which is not captured by the DEC itself. There exist classes $\mathcal{M}$ for which the upper and lower bounds are tight individually, but both can be loose in general.

3.2.2 A new complexity measure: The quantile Decision-Estimation Coefficient

We recover the DEC-based lower bounds from Foster et al. [2023b] through lower bounds based on a new variant we refer to as the *quantile DEC*. To do so, we briefly recount the proof technique used by Foster et al. [2023b] to prove the lower bounds in Theorem 5.

Let us focus on PAC guarantees for the DMSO setting. Given an algorithm 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 ALG itself) such that $D_{\mathrm{TV}}(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{P}^{\overline{M},\mathsf{ALG}})$ is small, yet 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 2 will be critical. We now apply the reasoning above within Theorem 2. For any model M, we consider the following distributions over decisions:

$$q_{M,\mathsf{ALG}} = \mathbb{E}^{M,\mathsf{ALG}}\left[\frac{1}{T}\sum_{t=1}^{T} q^t(\cdot \mid \mathcal{H}^{t-1})\right] \in \Delta(\Pi), \quad p_{M,\mathsf{ALG}} = \mathbb{E}^{M,\mathsf{ALG}}\left[p(\mathcal{H}^T)\right] \in \Delta(\Pi). \tag{16}$$

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 2 with the Hellinger distance. We will use the sub-additivity of Hellinger distance (Lemma B.1), which allows us to bound

$$D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{P}^{\overline{M},\mathsf{ALG}}\right) \le 7T \cdot \mathbb{E}_{\pi\sim p_{\overline{M},\mathsf{ALG}}}\left[D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)\right]. \tag{17}$$

Theorem 2 then yields the following intermediate result.

Lemma 6 (Recovering interactive two-point method)**.** *Let* $\delta \in [0,1]$ *be given, and consider an algorithm* ALG*. Define*

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

Then there exists $M \in \mathcal{M}$ *such that* $\mathbb{P}^{M,\mathsf{ALG}}\Big(L(M,\widehat{\pi}) \ge \Delta^\star_{\mathsf{ALG},\delta}\Big) \ge \delta$*.*

Proof of Lemma 6. To apply Theorem 2, we consider the squared Hellinger distance (which we recall is a f-divergence corresponding to $f(x) = \frac{1}{2}(\sqrt{x}-1)^2$). Consider a fixed tuple $(\overline{M}, M, \Delta)$ with $M \in \mathcal{M}$, $\overline{M} \in \mathrm{co}(\mathcal{M})$, and $\Delta \geq 0$ that satisfies

$$\sqrt{p_{\overline{M},\mathsf{ALG}}(\pi : L(M,\pi) \geq \Delta)} > \sqrt{\delta} + \sqrt{14T\,\mathbb{E}_{\pi\sim q_{\overline{M},\mathsf{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)}. \tag{18}$$

We choose the reference distribution to be $\mathbb{Q} = \mathbb{P}^{\overline{M},\mathsf{ALG}}$, and take μ to be the singleton distribution supported on M. Recall that for the DMSO framework, the observation is $X = (\mathcal{H}^T, \widehat{\pi})$, and the loss function is $L(M, X) = L(M, \widehat{\pi})$ (Section 2.2). Then, by definition, we have

$$\rho_{\Delta,\mathbb{Q}} = \mathbb{P}_{X\sim\mathbb{Q}}(L(M,X) < \Delta) = p_{\overline{M},\mathsf{ALG}}(\pi : L(M,\pi) < \Delta).$$

Further, using the sub-additivity of Hellinger distance (17), we have

$$D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{P}^{\overline{M},\mathsf{ALG}}\right) \leq 7T \cdot \mathbb{E}_{\pi\sim q_{\overline{M},\mathsf{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right).$$

Therefore, using the condition (18), we have

$$\frac{1}{2}\left(\sqrt{\delta} - \sqrt{1-\rho_{\Delta,\mathbb{Q}}}\right)^2 > D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{P}^{\overline{M},\mathsf{ALG}}\right).$$

Hence, it holds that $D^2_{\mathrm{H}}(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{Q}) < D^2_{\mathrm{H}}(1-\delta, \rho_{\Delta,\mathbb{Q}})$, and applying Theorem 2 gives

$$\mathbb{P}^{M,\mathsf{ALG}}(L(M,\widehat{\pi}) \geq \Delta) \geq \delta.$$

Taking supremum over all pairs $(\overline{M}, M, \Delta)$ satisfying Eq. (18) gives the desired lower bound. ☐

The quantile Decision-Estimation Coefficient. Using Lemma 6, 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. [2023b].

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

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

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:

$$\mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M}) := \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{ \widehat{L}_\delta(M,p) \,\middle|\, \mathbb{E}_{\pi\sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \leq \varepsilon^2 \right\}, \tag{19}$$

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 6, it is immediate to see that the quantile PAC-DEC provides a lower bound on the PAC risk.

Theorem 7 (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*

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

Proof of Theorem 7. Fix any algorithm ALG and abbreviate $\underline{\varepsilon} = \underline{\varepsilon}_\delta(T)$. Take an arbitrary parameter $\Delta_0 < \textsf{p-dec}^{\mathrm{q}}_{\underline{\varepsilon},\delta}(\mathcal{M})$. Then there exists $\overline{M}$ such that $\Delta_0 < \textsf{p-dec}^{\mathrm{q}}_{\underline{\varepsilon},\delta}(\mathcal{M}, \overline{M})$. Hence, by the definition (19), we know that

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

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

$$\mathbb{E}_{\pi\sim q_{\overline{M},\mathsf{ALG}}} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \underline{\varepsilon}^2, \qquad \mathbb{P}_{\pi\sim p_{\overline{M},\mathsf{ALG}}}(L(M,\pi) \ge \Delta_0) \ge \delta.$$

This immediately implies

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

where $\sqrt{\delta_1} = \sqrt{\delta} - \sqrt{14T\underline{\varepsilon}^2}$. Notice that $\delta_1 > \frac{\delta}{2}$ by the choice of $\underline{\varepsilon} = \frac{1}{14}\sqrt{\frac{\delta}{T}}$, and hence applying Lemma 6 shows that there exists $M \in \mathcal{M}$ such that $\mathbb{P}^{M,\mathsf{ALG}}(L(M,\widehat{\pi}) \ge \Delta_0) \ge \frac{\delta}{2}$. Letting $\Delta_0 \to \textsf{p-dec}^{\mathrm{q}}_{\underline{\varepsilon},\delta}(\mathcal{M})$ completes the proof. □

Unlike the original constrained DEC lower bounds (Theorem 5), which are restricted to the reward maximization variant of the DMSO setting (Example 4), the quantile DEC lower bound in Theorem 7 holds for *any risk function* L. As a result, the lower bound applies to a broader range of generalized PAC learning tasks, including model estimation [Chen et al., 2022a] and multi-agent decision making [Foster et al., 2023a], 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 10) in Appendix D.2. Nonetheless, Theorem 7 is powerful enough to recover the lower bounds for reward maximization in Theorem 5, as we will now show.

3.2.3 Recovering DEC-based lower bounds using the quantile DEC

At first glance, Theorem 7 might appear to be weaker than the constrained PAC-DEC lower bound in Theorem 5, since a conversion from quantile risk to expected risk will yield a loose lower bound. However, by specializing to reward maximization (Example 10) and leveraging the structure of the risk function L, we show that quantile PAC-DEC is equivalent to its constrained counterpart for this setting, leading to a tight lower bound.

Proposition 8 (Recovering the PAC DEC lower bound)**.** *Under the reward maximization setting (Example 4), for any $\varepsilon > 0$ and $\delta \in [0, 1)$ it holds that*

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \le \textsf{p-dec}^{\mathrm{q}}_{\sqrt{2}\varepsilon,\delta}(\mathcal{M}) + \frac{4\varepsilon}{1-\delta}.$$

As a corollary, we may choose $\delta = \frac{1}{2}$ and $\underline{\varepsilon}(T) = \frac{1}{20\sqrt{T}}$ in Theorem 7, so that

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

Thus, the quantile PAC-DEC lower bound indeed recovers the constrained PAC-DEC lower bound in Theorem 5.

Recovering DEC lower bounds for regret. Our quantile DEC lower bound extends to regret with minor modifications, allowing us to recover the regret lower bounds of Foster et al. [2023b] (Theorem 5). We defer the details to the Appendix D.1 (Theorem D.1).

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

Beyond the DEC methodology, a number of prior works have proven lower bounds for interactive decision making using extensions of the classical Fano method, typically in a somewhat case-by-case fashion [Castro and Nowak, 2008, Agarwal et al., 2009, Raginsky and Rakhlin, 2011b, etc.]. Notably, recent work of Rajaraman et al. [2023] provides tight lower bounds for linear and ridge bandit problem through a variant of Fano method which involves directly analyzing the cumulative mutual information acquired by the decision making algorithm of the course of T rounds of interaction. This approach achieves tight dependence on the problem dimension, which is not recovered by the standard DEC lower bound in Foster et al. [2021, 2023b].

In the following, we apply Theorem 2 to provide a mutual information-based approach for interactive decision making, recovering Rajaraman et al. [2023].

Proposition 9 (Mutual information-based lower bound for interactive decision making)**.** *Consider the DMSO setting. For any $T \ge 1$ 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 we recall that $I_{\mu,\mathsf{ALG}}(M; \mathcal{H}^T)$ is the mutual information between M and $\mathcal{H}^T$ under $M \sim \mu$ and $\mathcal{H}^T \sim \mathbb{P}^{M,\mathsf{ALG}}$, and the supremum is taken over all possible DMSO algorithms ALG. Then for any algorithm ALG,

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

Proof of Proposition 9. Recall that we frame the DMSO setting as an instance of ISDM in Section 2.2, where the observation is given by $X = (\mathcal{H}^T, \widehat{\pi})$, and the loss is $L(M, X) = L(M, \widehat{\pi})$. In particular, for any prior $\mu \in \Delta(\mathcal{M})$, we have

$$\sup_X \mu(M : L(M,X) < \Delta) = \sup_{\pi\in\Pi} \mu(M : L(M,\pi) < \Delta),$$

and by definition, $I_{\mu,\mathsf{ALG}}(M; X) \le I_\mu(T)$. In particular, for any pair (Δ, μ) such that

$$\sup_\pi \mu(M : L(M,\pi) \le \Delta) \le \frac{1}{4}\exp(-2I_\mu(T)), \tag{20}$$

we have $\frac{I_{\mu,\mathsf{ALG}}(M;X)+\log 2}{\log \sup_x \mu(M:L(M,x)<\Delta)} \ge -\frac{1}{2}$, and hence applying Proposition 3 gives $\sup_{M\in\mathcal{M}} \mathbb{E}^{M,\mathsf{ALG}}[L(M,\widehat{\pi})] \ge \frac{\Delta}{2}$. Taking supremum over all pairs (Δ, μ) satisfying (20) gives the desired lower bound. □

Proposition 9 is an important departure from the classical (non-interactive) Fano method (Proposition 1), replacing the non-interactive mutual information by the maximum (interactive) mutual information that can be achieved by a T-round algorithm. It allows us to analyze the T-round evolution of the mutual information under interactions, which can lead to tighter bounds than any analysis that proceeds on a *per-round*.

Using Proposition 9, along with mutual information bounds from Rajaraman et al. [2023], we recover a $\Omega(d/\sqrt{T})$ PAC lower bound for d-dimensional linear bandits, which in turn recovers the $\Omega(d\sqrt{T})$ regret lower bound [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Lattimore and Szepesvári, 2020a, etc.].

Corollary 10. *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 Gaussian rewards. The model class is $\mathcal{M} = \{M_\theta\}_{\theta \in \Theta}$, where for each $\theta \in \Theta$, the model M_θ is given by $M_\theta(\pi) = \mathcal{N}(\langle \pi, \theta \rangle, 1)$. Then Proposition 9 implies a minimax risk lower bound:*

$$\inf_{\mathsf{ALG}} \sup_{M \in \mathcal{M}} \mathbb{E}^{M,\mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] \geq \Omega\left(\min\{d/\sqrt{T}, 1\}\right). \tag{21}$$

Notably, the DEC-based lower bounds in Foster et al. [2021, 2023b] only scale as $\sqrt{d/T}$ for this problem setting; informally, this is because one $\sqrt{d}$ factor in Eq. (21) arises from hardness of *exploration* (which is captured by the DEC), but the other $\sqrt{d}$ factor arises from hardness of *estimation* (which is not captured by the DEC).

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 2) to derive new lower and upper bounds for interactive decision making that go beyond the previous results based on the Decision-Estimation Coefficient [Foster et al., 2021, 2023b] by incorporating hardness of estimation. First, in Section 4.1, we introduce a new complexity measure, the *fractional covering number*, which serves as a lower and upper bound for interactive decision making with any model class $\mathcal{M}$ satisfying a certain regularity condition; informally, the fractional covering number reflects the complexity of estimating a near-optimal decision. In Section 4.2, we specialize this result to the problem of structured bandits and complement our lower bounds based on the fractional covering number with an upper bound, thereby establishing that the fractional covering number gives the first complete characterization for structured bandit *learnability* (albeit, with an exponential gap between the upper and lower bounds). Finally, in Section 4.3, we show how to combine the fractional covering number with the DEC to derive tighter lower bounds for general decision making. In the process, we close the remaining gap between the upper and lower bounds in [Foster et al., 2021, 2023b] up to a quadratic factor for any convex model class $\mathcal{M}$.

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 5 can be closed. To highlight this gap in a more interpretable fashion, we re-state Theorem 5 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:

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

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

$$T^\star(\mathcal{M}, \Delta) := \inf_{T \geq 1} \{T : \mathbf{Reg}^\star(\mathcal{M}, T) \leq T\Delta\} \tag{22}$$

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 $\mathbf{Reg}^\star(\mathcal{M}, T)$.

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

$$T^{\mathsf{DEC}}(\mathcal{M}, \Delta) = \inf_{\varepsilon \in (0,1)} \{\varepsilon^{-2} : \mathsf{r\text{-}dec}^{\mathsf{c}}_{\varepsilon}(\mathcal{M}) \leq \Delta\}. \tag{23}$$

With this definition, Theorem 5 is equivalent to the following characterization of the minimax sample complexity $T^{\star}(\mathcal{M}, \Delta)$:

$$T^{\mathsf{DEC}}(\mathcal{M}, \Delta) \lesssim T^{\star}(\mathcal{M}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathcal{M}, \Delta) \cdot \log|\mathcal{M}|. \tag{24}$$

That is, Theorem 5 characterizes the minimax sample complexity up to a multiplicative $\log|\mathcal{M}|$ factor. Our main result in this section will be to use the fractional covering number and interactive Fano method (Theorem 2), to (i) tighten the above characterization (24) 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 fractional covering number

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

$$\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta) := \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \frac{1}{p(\pi : L(M, \pi) \le \Delta)}. \tag{25}$$

Informally, the fractional covering number $\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)$ 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, fractional covering number naturally arises as a lower bound on optimal risk/regret 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 $D_{\mathrm{KL}}(M(\pi) \,\|\, \overline{M}(\pi)) \le C_{\mathrm{KL}}$ 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 A.1.

Our main lower bound based on the fractional covering number follows by specializing Theorem 2 to KL divergence.

Theorem 11 (Fractional covering number lower bound)**.** *Suppose that $\mathcal{M}$ satisfies Assumption 2 with parameter $C_{\mathrm{KL}} > 0$. Then for any algorithm ALG and $\Delta > 0$, unless*

$$T \ge \frac{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M}, 2\Delta) - 2}{2C_{\mathrm{KL}}},$$

there exists $M^{\star} \in \mathcal{M}$ such that

$$\mathbb{P}^{M^{\star}, \mathsf{ALG}}[L(M^{\star}, \widehat{\pi}) \ge \Delta] \ge \frac{1}{2}.$$

In particular, for (generalized) no-regret learning, fractional covering number also implies a regret lower bound through Theorem 11.[3]

Corollary 12. *Suppose that Assumption 2 holds. Then for any $\Delta > 0$, it holds that*

$$T^{\star}(\mathcal{M}, \Delta) \ge \frac{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M}, 2\Delta) - 2}{2C_{\mathrm{KL}}}.$$

[3]For any T-round no-regret algorithm ALG, we can take the output decision $\widehat{\pi} \sim \mathrm{Unif}(\pi^1, \cdots, \pi^T)$ (*online-to-batch* conversion), and then $\mathbb{E}^{M, \mathsf{ALG}}[\mathbf{Risk}_{\mathsf{DM}}(T)] = \frac{1}{T}\mathbb{E}^{M, \mathsf{ALG}}[\mathbf{Reg}_{\mathsf{DM}}(T)]$ for any model M.

That is, for any algorithm to achieve ΔT-regret, it is necessary to have $T = \Omega(\log \mathsf{N}_{\text{frac}}(\mathcal{M}, 2\Delta))$. Combining this with Theorem 5, we conclude that boundedness of both the DEC and the fractional covering number is necessary for learning with any model class $\mathcal{M}$.

Upper bounds based on the fractional covering number. We now complement Theorem 11 by showing that for any reward maximization instance of the DMSO setting (Example 4), boundedness of the fractional covering number 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 $\mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta)$, the upper bound will be polynomial.

Theorem 13 (Fractional covering number upper bound)**.** *Consider the reward maximization task (Example 4). There exists an algorithm that for any class $\mathcal{M}$ and $\Delta > 0$, ensures that with probability at least $1 - \delta$,*

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

Combining Theorem 11 and Theorem 13 yields the following bounds on $T^{\star}(\mathcal{M}, \Delta)$ (omitting poly-logarithmic factors):

$$\frac{\log \mathsf{N}_{\text{frac}}(\mathcal{M}, 2\Delta)}{C_{\text{KL}}} \lesssim T^{\star}(\mathcal{M}, \Delta) \lesssim \frac{\mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta/2)}{\Delta^2}. \tag{26}$$

The gap between the lower and upper bounds of Eq. (26) is exponential; However, for model classes with $C_{\text{KL}} = O(1)$, Eq. (26) is enough to provide a characterization of *finite-time learnability.* As a special case, we now show that fractional covering number characterizes the learnability of any structured bandit problem.

4.1.1 Properties of the fractional covering number

Before proceeding to applications, let us briefly discuss some connections between the fractional covering number and classical notions of covering number considered in the context of statistical estimation.

To start, we recall that for many standard statistical estimation tasks such as regression and non-parametric estimation, the risk function L is given by a (pseudo) metric (e.g., ℓ_2 distance between parameters or mean-squared error in predictions). The following lemma shows that in this case, the fractional covering number coincides with the classical covering number induced by the metric, e.g., in location estimation (Example 1), density estimation (Example 3), etc.

Lemma 14 (Connection to classical covering numbers)**.** *Suppose the decision space Π is equipped with a pseudo-metric $\rho : \Pi \times \Pi \to \mathbb{R}_+$ and there is a map $M \mapsto \pi_M \in \Pi$ such that the risk function is given by*

$$L(M, \pi) = \rho(\pi_M, \pi), \qquad \forall M \in \mathcal{M}, \pi \in \Pi. \tag{27}$$

Let $\Pi_{\mathcal{M}} := \{\pi_M : M \in \mathcal{M}\} \subseteq \Pi$, and define $\mathsf{N}(\Pi_{\mathcal{M}}, \Delta)$ to be the Δ-covering number of $\Pi_{\mathcal{M}}$ under ρ. Then

$$\mathsf{N}(\Pi_{\mathcal{M}}, 2\Delta) \leq \mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta) \leq \mathsf{N}(\Pi_{\mathcal{M}}, \Delta).$$

Duality between fractional covering and fractional packing For classical covering numbers with respect to a pseudo-metric (as in Lemma 14), it is known that there is a duality between covering and packing. In spite of a lack of metric structure for the general setting we study, we can show that fractional covering number naturally admits a dual representation in terms of a *fractional packing number*. Specifically, it holds that

$$\inf_{\mu\in\Delta(\mathcal{M})}\sup_{\pi\in\Pi}\mu(M:L(M,\pi)\leq\Delta)=\sup_{p\in\Delta(\Pi)}\inf_{M\in\mathcal{M}}p(\pi:L(M,\pi)\leq\Delta),$$

as long as the minimax theorem can be applied (e.g. when Π or $\mathcal{M}$ are finite or satisfy appropriate compactness conditions). Therefore, in this case, we have

$$\mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)=\sup_{\mu\in\Delta(\mathcal{M})}\inf_{\pi\in\Pi}\frac{1}{\mu(M:L(M,\pi)\leq\Delta)} \tag{28}$$

which can be interpreted as a fractional packing number. We mention in passing that using this interpretation, it is possible to derive Theorem 11 directly from Proposition 9.

Recovering the Yang-Barron method. As a simple example of the fractional covering number, we recover and further generalize the well-known Yang-Barron method [Yang and Barron, 1999] for statistical estimation problems (see also [Wainwright, 2019, Section 15.3.5]).

Example 7 (Yang-Barron method)**.** For a statistical estimation problem with model class $\mathcal{M}$, we define the KL covering number of $\mathcal{M}$ as

$$\mathsf{N}_{\mathrm{KL}}(\mathcal{M},\varepsilon):=\min\left\{k:\exists M^1,\cdots,M^k\in\mathcal{M},\text{such that }\forall M\in\mathcal{M},\min_{i\in[k]}D_{\mathrm{KL}}(M\parallel M^i)\leq\varepsilon^2\right\}.$$

For a fixed parameter $\varepsilon>0$, we can pick $k=\mathsf{N}_{\mathrm{KL}}(\mathcal{M},\varepsilon)$ and $M^1,\cdots,M^k\in\mathcal{M}$ such that $\min_{i\in[k]}D_{\mathrm{KL}}(M\parallel M^i)\leq\varepsilon^2$ for all $M\in\mathcal{M}$. Then, let us consider the localized sub-class $\mathcal{M}_i:=\{M\in\mathcal{M}:D_{\mathrm{KL}}(M\parallel M^i)\leq\varepsilon^2\}$ for each $i\in[k]$. It is clear that Assumption 2 holds for each $\mathcal{M}_i$ with $C_{\mathrm{KL}}\leq\varepsilon^2$. Further, using $\mathcal{M}=\bigcup_{i=1}^n\mathcal{M}_i$, we have

$$\log\mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)\leq\max_{i\in[k]}\log\mathsf{N}_{\mathrm{frac}}(\mathcal{M}_i,\Delta)+\log k. \tag{29}$$

For details, see Appendix G.4. Therefore, applying Theorem 11 to $\mathcal{M}_i$ and take supremum over $i\in[k]$ and $\varepsilon>0$ gives the following result: *For any algorithm* ALG *to achieve* $\sup_{\theta\in\Theta}\mathbb{E}^{\theta,\mathsf{ALG}}L(M,\pi)\leq\Delta$*, it is necessary that*

$$\log\mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)\leq\inf_{\varepsilon>0}\left(2T\varepsilon^2+\log\mathsf{N}_{\mathrm{KL}}(\mathcal{M},\varepsilon)\right)+2.$$

When the risk function L is given by a metric (cf. Eq. (27)), this inequality coincides with the Yang-Barron method formulated in Wainwright [2019, Section 15.3.5], as the fractional covering number $\mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)$ can be lower bounded by the covering number (Lemma 14).

Recovering the local packing lower bound for statistical estimation. As a simple example of the fractional covering number, we recover the well-known local packing-based lower bound [Birgé, 1986] for the classical problem of location estimation [Wainwright, 2019].

Example 8 (Local packing lower bound for location estimation)**.** In the location estimation task (Example 1), recall that the model class is given by $\mathcal{M} = \{M_\theta : \theta \in \Theta\}$, where $M_\theta = \mathcal{N}(\theta, I_d)$. Consider the local packing number of Θ around $\theta^\star \in \Theta$, which is given by

$$\mathsf{N}_{\mathrm{loc}}(\Theta, \Delta; \theta^\star) := \max\left\{ k : \exists \theta^1, \cdots, \theta^k \in \Theta, \left\|\theta^i - \theta^\star\right\| \le \Delta, \left\|\theta^i - \theta^j\right\| > \frac{\Delta}{2}, \forall i \ne j \right\}.$$

Then, for the localized sub-class $\mathcal{M}_{\varepsilon,\theta^\star} := \{M_\theta : \|\theta - \theta^\star\| \le \varepsilon\} \subseteq \mathcal{M}$, Assumption 2 holds with $C_{\mathrm{KL}} \le \frac{1}{2}\varepsilon^2$, and we also have $\mathsf{N}_{\mathrm{frac}}(\mathcal{M}_{\varepsilon,\theta^\star}, \varepsilon/4) \ge \mathsf{N}_{\mathrm{loc}}(\Theta, \varepsilon; \theta^\star)$. Therefore, we can apply Theorem 11 to $\mathcal{M}_{\varepsilon,\theta^\star}$ and take supremum over all $\theta^\star \in \Theta$ and $\varepsilon \ge 8\Delta$ to show the following result: *For any algorithm* ALG *to achieve* $\sup_{\theta \in \Theta} \mathbb{E}^{\theta, \mathsf{ALG}} \|\widehat{\pi} - \theta\| \le \Delta$*, it is necessary that*

$$T \ge \sup_{\varepsilon \ge 8\Delta} \frac{\log \mathsf{N}_{\mathrm{loc}}(\Theta, \varepsilon) - 2}{\varepsilon^2},$$

where $\mathsf{N}_{\mathrm{loc}}(\Theta, \varepsilon) = \sup_{\theta^\star \in \Theta} \mathsf{N}_{\mathrm{loc}}(\Theta, \varepsilon; \theta^\star)$ is the local packing number of Θ. This lower bound is known to be tight in general [Birgé, 1983, 1986, Le Cam, 1986, etc.]. ◁

4.2 Application: Bandit learnability

As our main application of the fractional covering number, we provide a new characterization of *learnability* (i.e., asymptotic achievability of non-trivial sample complexity) for structured bandits with general function approximation. In the literature on statistical learning, there is a long line of work which characterizes learnability of hypothesis classes in terms of abstract complexity measures. Examples include the Vapnik-Chervonenkis dimension for binary classification [Vapnik and Chervonenkis, 1971, Blumer et al., 1989], the Littlestone dimension [Littlestone, 1988] for online classification [Ben-David et al., 2009] and differentially private classification [Bun et al., 2020, Alon et al., 2022], and their real-valued counterparts (e.g. scale-sensitive dimensions) for regression [Bartlett et al., 1994, Alon et al., 1997].

Beyond the settings above—particularly for interactive settings—learnability is less well understood. The question of what complexity measure characterizes bandit learnability has been explored in Russo and Van Roy [2013], Abernethy et al. [2013], Simchowitz et al. [2017], Hashimoto et al. [2018, etc.], but a complete resolution has yet to be reached. Remarkably, Ben-David et al. [2019] demonstrate that there exists a simple and natural learning task for which it is impossible to characterize learnability through any *combinatorial* dimension. More recently, Hanneke and Yang [2023] provide similar impossibility results for characterizing the learnability of noiseless structured noiseless bandits with real-valued rewards.

Our characterization bypasses the impossibility results of Hanneke and Yang [2023]. Specifically, Hanneke and Yang [2023] 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* [Ben-David et al., 2019] for the problem class. Our characterization is compatible with this result because the argument of Hanneke and Yang [2023] relies on the noiseless nature of the bandit problem, and hence does not preclude a characterization for the noisy setting.

Structured bandit setting and learnability characterization. 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 $r^t \sim \mathcal{N}\big(h_\star(\pi^t), 1\big)$

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 $\mathcal{M}_\mathcal{H} = \{\pi \mapsto \mathcal{N}(h(\pi), 1) \mid h \in \mathcal{H}\}$.

$$\mathcal{M}_\mathcal{H} = \{\pi \mapsto \mathcal{N}(h(\pi), 1) \mid h \in \mathcal{H}\}.$$

We define the fractional covering number for $\mathcal{H}$ via

$$\mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta) := \mathsf{N}_{\mathrm{frac}}(\mathcal{M}_\mathcal{H}, \Delta) = \inf_{p \in \Delta(\Pi)} \sup_{h \in \mathcal{H}} \frac{1}{p(\pi : h(\pi_h) - h(\pi) \leq \Delta)}, \tag{30}$$

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 [2023], 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 [2023] in more detail at the end of this section.

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

Corollary 15 (Lower bound for stochastic bandits)**.** *For the bandit model class $\mathcal{M}_\mathcal{H}$ defined as above, it holds that $T^\star(\mathcal{M}_\mathcal{H}, \Delta) \geq 2 \log \mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta) - 2$.*

Combining this result with the upper bound in Theorem 13, we obtain the following bounds on the minimax-optimal sample complexity for the structure bandit problem with class $\mathcal{H}$:

$$\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H}, 2\Delta) \lesssim T^\star(\mathcal{M}_\mathcal{H}, \Delta) \lesssim \frac{\mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta/2)}{\Delta^2}. \tag{31}$$

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

Theorem 16 (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 $\mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta) < +\infty$ for all $\Delta > 0$.*

We remark that the lower and upper bound in Eq. (31) cannot be improved in terms of the fractional covering number alone:

- For K-armed bandits, we have $\mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta) \leq K$, meaning the upper bound is tight.
- For d-dimensional linear bandits, we have $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta) = \Omega(d)$, meaning the lower bound is nearly tight.

Nevertheless, the exponential gap in Eq. (31) can be partly mitigated by combining the fractional covering number with the DEC, as we will show in Section 4.3.

Remark 17 (Noise distribution)**.** We note that the upper bound in Eq. (31) applies to any reward distribution with sub-Gaussian noise (cf. Appendix G.2). Meanwhile, since the lower bound in Corollary 15 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.

Connection to the maximin volume. Complementary to the hardness results, Hanneke and Yang [2023] propose a complexity measure called the *maximin volume* which tightly characterizes the complexity of learning *noiseless binary-valued* structured bandit problems. For such problem classes, the fractional covering number is exactly the inverse of the maximin volume. While the fractional covering number can be viewed as a generalization of the maximin volume in this sense, we emphasize that the fractional covering number directly arises from our general lower bound framework, and is applicable to general decision making problems in the DMSO framework.

4.3 Improved upper bounds for general decision making

To close this section, we derive tighter upper bounds that scale with $\log \mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta)$ by combining the fractional covering number with the Decision-Estimation Coefficient.We focus on regret minimization, but upper bounds for PAC can be derived similarly.

To state our upper bound in the simplest form possible, we focus on the the reward maximization setting variant of the DMSO setting, and make the following regularity assumption on the DEC for $\mathcal{M}$.[4]

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".For a broad range of classes $\mathcal{M}$ (see, e.g., Foster et al. [2023b]), we have $\mathsf{r\text{-}dec}^{\mathsf{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M})) \asymp L\varepsilon^{\rho}$ for some problem-dependent parameter $L > 0$ and $\rho \in (0,1]$, so that Assumption 3 is automatically satisfied.

We now state our upper bound, which tightens Theorem 5 by replacing the $\log|\mathcal{M}|$ dependence in the upper bound with $\log \mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta)$ (with the caveat that the upper bound scales with the DEC for the *convexified* model class $\mathrm{co}(\mathcal{M})$).

Theorem 18 (Upper bound with DEC and fractional covering number)**.** *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*[5]*. Let* $\bar{\varepsilon}(T) \asymp \sqrt{\log \mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta)/T}$*. Then for any* $\Delta > 0$*, Algorithm 1 (see Appendix F.1) 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 19. *For any class* $\mathcal{M}$ *that satisfies Assumption 2 and 3, we have*

$$\max\left\{ T^{\mathsf{DEC}}(\mathcal{M}, \Delta), \frac{\log \mathsf{N}_{\text{frac}}(\mathcal{M}, 2\Delta)}{C_{\mathrm{KL}}} \right\} \lesssim T^{\star}(\mathcal{M}, \Delta) \lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{M}), \Delta) \cdot \log \mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta/2), \quad (32)$$

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 19 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 Eq. (32) is always tighter than Eq. (24) (and also tighter than the result in Foster et al. [2022]), as by definition we have

$$\log \mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta) \leq \log \mathsf{N}_{\text{frac}}(\mathcal{M}, 0) \leq \min\left\{ \log|\mathcal{M}|, \log|\Pi| \right\}, \qquad \forall \Delta > 0.$$

Furthermore, $\log \mathsf{N}_{\text{frac}}(\mathcal{M}, \Delta)$ can be significantly smaller than $\min\left\{ \log|\mathcal{M}|, \log|\Pi| \right\}$:

[4] Both restrictions can be removed; the fully general upper bound is detailed in Appendix F.1 and Appendix G.5. Note that a similar growth assumption is also required in the regret upper bound in Theorem 5.

[5] The finiteness assumption on the decision space Π can be relaxed, e.g. to require that Π admits finite covering number.

- For d-dimensional concave bandits (see e.g. Foster et al. [2021, Section 6.1.2]), the log covering number of $\mathcal{M}$ is $e^{\Omega(d)}$, while we have $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)\leq \tilde{O}(d)$.
- For any class of structured contextual bandits (Appendix F.2), we have $\log|\Pi| = |\mathcal{C}|\log|\mathcal{A}|$ and the log covering number of $\mathcal{M}$ is $\Omega(|\mathcal{C}|)$, while $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)$ can be bounded even when the context space $\mathcal{C}$ is infinite.

This reflects the fact that the fractional covering number adapts to the intrinsic complexity of estimation in interactive decision making.

Remark 20 (Necessity of convex hull)**.** We note that in general, we cannot replace the $T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{M}),\varepsilon)$ by $T^{\mathsf{DEC}}(\mathcal{M},\varepsilon)$ in the upper bound of Eq. (32). Specifically, Chen et al. [2023] construct a class $\mathcal{M}$ of partially observable MDPs, such that $T^{\star}(\mathcal{M},\varepsilon)$ can be arbitrarily is not polynomial in $T^{\mathsf{DEC}}(\mathcal{M},\varepsilon)$ and $\log|\Pi|$ (see also the discussion in Chen et al. [2023, Appendix I]). Therefore, the upper bound of Theorem 18 cannot be improved to scale with the DEC for $\mathcal{M}$.

Remark 21 (Regularity condition)**.** In Theorem 18 and Theorem 19, we assume the regularity of DEC (Assumption 3) for a clean presentation of the upper bound. Without the regularity condition, the upper bound of Eq. (32) becomes slightly worse (with an extra Δ^{-1} factor). A detailed discussion is deferred to Appendix G.5 (Remark G.4).

4.3.1 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 19 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

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

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 F.2, the $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H})$ and $\log\mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta)$ together provide an upper bound for structured bandits with $\mathcal{H}$.

Theorem 22. *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 $\bar{\varepsilon}(T)\asymp\sqrt{\log\mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta)/T}$. The Algorithm 1 ensures that high probability,*

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

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

$$\max\left\{T^{\mathsf{DEC}}(\mathcal{H},\Delta),\log\mathsf{N}_{\mathrm{frac}}(\mathcal{H},2\Delta)\right\}\lesssim T^{\star}(\mathcal{M}_{\mathcal{H}},\Delta)\lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{H}),\Delta)\cdot\log\mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta/2), \qquad (33)$$

where we denote $T^{\mathsf{DEC}}(\mathcal{H},\Delta)=\inf_{\varepsilon\in(0,1)}\{\varepsilon^{-2}:\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{H})\leq\Delta\}$ (following Eq. (23)) 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 [Rigollet and Zeevi, 2010], or concavity [Lattimore, 2020], or sub-modularity [Nie et al., 2022], or etc.). For these problem classes, the complexity of no-regret learning is completely characterized by the DEC of $\mathcal{H}$ and the fractional covering number $\mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta)$ (up to a quadratic factor).

We also note that the lower bound of Eq. (33) is proven for Gaussian noise, while our upper bound applies to a much more general class of reward distributions (with bounded variance).

Additional result: Structured contextual bandits We also apply the results of Theorem 18 and Theorem 19 to structured contextual bandits, providing new characterization in terms of the DEC and the fractional covering number. The detailed discussion is presented in Appendix F.2.

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

Jacob Abernethy, Kareem Amin, Michael Kearns, and Moez Draief. Large-scale bandit problems and kwik learning. In *International Conference on Machine Learning*, pages 588–596, 2013.

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

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.

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.

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.

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.

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.

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

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.

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

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

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

P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Number v. 1 in Mathematical Statistics: Basic Ideas and Selected Topics. Prentice Hall, 2001. ISBN 9780138503635.

Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237, 1983.

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

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.

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

Sébastien Bubeck, Ofer Dekel, Tomer Koren, and Yuval Peres. Bandit convex optimization:$\sqrt{T}$ regret in one dimension. In *Conference on Learning Theory*, pages 266–278. PMLR, 2015.

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.

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.

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

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.

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*, 2022a.

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, 2022b.

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

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.

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

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.

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

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

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.

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

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

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

John C Duchi. Lecture notes on statistics and information theory. 2024.

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

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, 2023a.

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.

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.

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

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

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, 2023b.

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.

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

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

Tatsunori Hashimoto, Steve Yadlowsky, and John Duchi. Derivative free optimization via repeated classification. In *International Conference on Artificial Intelligence and Statistics*, pages 2027–2036. PMLR, 2018.

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.

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

TS Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 562–573. Springer, 2009.

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.

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.

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

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

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

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

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020a.

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

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer New York, NY, 1986.

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

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

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.

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

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

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.

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

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.

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

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

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.

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.

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

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

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26, pages 2256–2264, 2013.

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

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

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.

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

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.

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.

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

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

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.

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

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

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.

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.

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.

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

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

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

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

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.

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

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

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

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

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.

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

Contents of Appendix

A Additional Background on DMSO Framework

The DMSO framework (Section 1.2) encompasses a wide range of learning goals beyond the reward maximization setting [Foster et al., 2021, 2023b], including reward-free learning, model estimation, and preference-based learning [Chen et al., 2022a], and also multi-agent decision making and partial monitoring [Foster et al., 2023a]. We provide two examples below for illustration.

Example 9 (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 $L(M, \pi) = \max_{\pi^\star} \mathbb{C}^M(\pi^\star, \pi)$. Chen et al. [2022a] provide lower and upper bounds for this setting in terms of Preference-based DEC (PBDEC).

Example 10 (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[6], and θ is the estimator of the model parameter. In this setting, we define $L(M, \pi) = \mathrm{Dist}(\theta_M, \theta)$ 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. [2022a]. Natural examples include estimating some coordinates of the parameter θ in linear bandits. We provide nearly tight guarantee for this setting in Appendix D.2.

Applicability of our results Our general interactive Fano method Lemma 6 applies to any generalized no-regret / PAC learning goal (Section 1.2). Therefore, our risk lower bound in terms of quantile PAC-DEC Theorem 7 and fractional covering number lower bound Theorem 11 both apply to any generalized learning goal. For a concrete example, see Appendix D.2 for the application to interactive estimation.

A.1 Examples for 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 11 (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 $h^M \in \mathcal{H}$:

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

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

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

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

Example 12 (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

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

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

Example 12 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 13 (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 $r \sim \mathcal{N}(h^M(c, a), 1)$.

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

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 $M \in \mathcal{M}_{\mathcal{H},\mathbb{V}}$,

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

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

The factor of $\log |\mathcal{C}|$ in Example 13 is due to the definition (45) of $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta)$, 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).

B Technical Tools

The following lemma is the "chain rule" of Hellinger distance [Jayram, 2009] (see also Duchi [2024, Lemma 11.5.3] and Foster et al. [2024, Lemma D.2]).

Lemma B.1 (Sub-additivity for squared Hellinger distance)**.** *Let $(\mathcal{X}_1, \mathcal{F}_1), \dots, (\mathcal{X}_T, \mathcal{F}_T)$ 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 $X_1, \dots, X_T$ 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

$$D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \le 7\ \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^T D^2_{\mathrm{H}}\big(\mathbb{P}^t(\cdot \mid X_{1:t-1}), \mathbb{Q}^t(\cdot \mid X_{1:t-1})\big)\right]. \tag{34}$$

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, $\mathbb{P}^{M,\mathsf{ALG}}(X_t = \cdot \mid X_{1:t-1})$ 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 B.1 to $D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{P}^{\overline{M},\mathsf{ALG}}\right)$ gives the following corollary.

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

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

Lemma B.3 (Foster et al. [2021, 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 $t \le T$,*

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

Lemma B.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 B.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* $D^2_{\mathrm{H}}(\mathbb{P},\mathbb{Q}) \geq \frac{1}{10}|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2$.

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

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

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

□

C Proofs from Section 3

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

C.1 Proof of Theorem 2

In the following, we fix a prior $\mu \in \Delta(\mathcal{M})$, parameter $\Delta > 0$, f-divergence D_f, and an algorithm ALG. For simplicity, we denote $D_f(x,y) = D_f(\mathrm{Bern}(x), \mathrm{Bern}(y))$ for $x, y \in [0,1]$.

We only need to prove the following claim.

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

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

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

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

Therefore, using $\mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}}) > \mathbb{E}_{M\sim\mu} D_f(\mathbb{P}^{M,\mathsf{ALG}}, \mathbb{Q})$, we know that $\mathsf{d}_{f,\delta}(\rho_{\Delta,\mathbb{Q}}) > D_f(\bar{\rho}_{\Delta}, \rho_{\Delta,\mathbb{Q}})$. In particular, this implies $\rho_{\Delta,\mathbb{Q}} < 1-\delta$, and

$$D_f(\bar{\rho}_{\Delta}, \rho_{\Delta,\mathbb{Q}}) < D_f(1-\delta, \rho_{\Delta,\mathbb{Q}})$$

Hence, we consider two cases: (1) $\bar{\rho}_{\Delta} \leq \rho_{\Delta,\mathbb{Q}}$, and (2) $\bar{\rho}_{\Delta} > \rho_{\Delta,\mathbb{Q}}$. For case (1), we have $\bar{\rho}_{\Delta} \leq \rho_{\Delta,\mathbb{Q}} < 1-\delta$. For case (2), we can use the monotone property of D_f (Lemma C.1), which also implies $\bar{\rho}_{\Delta} < 1-\delta$.

Therefore, it holds that $\bar{\rho}_\Delta < 1 - \delta$, and

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

Hence, the proof of Eq. (10) is completed. The in-expectation lower bounds then follows from the fact that

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

□

Lemma C.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 C.1 Fix any $1 \geq x > z \geq y \geq 0$, we define

$$p = y \cdot \frac{x-z}{x-y} \in [0,1], \qquad q = 1 - (1-y) \cdot \frac{x-z}{x-y} \in [0,1].$$

Then, by definition,

$$p(1-x) + qx = z, \qquad p(1-y) + qy = y,$$

and hence for the channel P from $\{0,1\}$ to itself given by $P(\cdot|0) = \mathrm{Bern}(p), P(\cdot|1) = \mathrm{Bern}(q)$, it holds that

$$P \circ \mathrm{Bern}(x) = \mathrm{Bern}(z), \qquad P \circ \mathrm{Bern}(y) = \mathrm{Bern}(y).$$

Therefore, by data-processing inequality, we have

$$D_f(\mathrm{Bern}(z), \mathrm{Bern}(y)) \leq D_f(\mathrm{Bern}(x), \mathrm{Bern}(y)).$$

This is the desired result. □

C.2 Proof of Corollary 10

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 $\pi^t \in \{\pi \in \mathbb{R}^d : \|\pi\|_2 \leq 1\}$ and receives a Gaussian reward $r^t \sim \mathcal{N}(\langle \theta^\star, \pi^t \rangle, 1)$. 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 C.2. *For any $r > 0$, we define the prior μ_r over $\mathbb{B}^d(r)$ by*

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

Then for any algorithm ALG*, we have*

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

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

$$I_{\mu,\mathsf{ALG}}(\theta^\star; \mathcal{H}^T) \leq \frac{d}{2} \log\left(1 + \frac{\lambda T}{d^2}\right). \tag{35}$$

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

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

which is exactly Eq. (35).

Next we deduce the claimed result from Eq. (35). 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 Eq. (35), 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_d^2 random variable, we arrive at the desired statement. □

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

Theorem C.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 9 implies that*

$$\mathbb{E}^{\mu,\mathsf{ALG}}\left[\left\| \widehat{\pi} - \frac{\theta^\star}{\|\theta^\star\|} \right\|^2 \right] \ge \frac{1}{4}.$$

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 9 to the following risk function

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

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

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

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

$$\rho_\Delta := \sup_\pi \mu\Big(\theta : \tilde{L}(M_\theta, \pi) \leq \Delta\Big) = O\Big(\sqrt{d}\Delta^{(d-1)/2}\Big). \tag{36}$$

To see so, by symmetry of Gaussian distribution, we know for fixed any $\pi \in \mathbb{R}^d$,

$$\mu\Big(\theta : \tilde{L}(M_\theta, \pi) \leq \Delta\Big) = \mathbb{P}_{\theta \sim \mathrm{Unif}(\mathbb{S}^{d-1})}\Big(\theta : \|\theta - \pi\|^2 \leq \Delta\Big),$$

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

$$\Big\{\theta \in \mathbb{S}^{d-1} : \|\theta - \pi\|_2^2 \leq \Delta\Big\} = \left\{\theta \in \mathbb{S}^{d-1} : \theta_1 \geq \frac{x^2 + 1 - \Delta}{2x}\right\} \subseteq \Big\{\theta \in \mathbb{S}^{d-1} : \theta_1 \geq \sqrt{1 - \Delta}\Big\}.$$

By Bubeck et al. [2016, Section 2], for $\theta \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, 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,

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

With the upper bound (36) of ρ_Δ, we know that for $\Delta = \frac{1}{2}$, it holds

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

as long as c_0 is a sufficiently small constant. Therefore, Proposition 9 gives that

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

This completes the proof of the first inequality.

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

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

This is the desired result. $\square$

D Additional Results from Section 3.2

In addition to the reward-maximization setting (Example 4), 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 $L(M,\pi) = f^M(\pi_M) - f^M(\pi)$, but f^M is not assumed to be the expected reward function (Example 4). Instead, we only require f^M satisfying the following assumption, where $\mathcal{M}^+ \subseteq (\Pi \to \Delta(\mathcal{O}))$ is a pre-specified model class of reference models that contains $\mathrm{co}(\mathcal{M})$ (following Foster et al. [2023b]). The lower bound we prove can be stronger by allowing $\mathcal{M}^+$ to be a larger model class.

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 $L(M,\pi) = f^M(\pi_M) - f^M(\pi)$ 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*

$$|f^M(\pi) - f^{\overline{M}}(\pi)| \le C_{\mathrm{r}} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big), \qquad \forall \pi \in \Pi. \tag{37}$$

In some cases, considering a larger reference model class $\mathcal{M}^+$ can be convenient for proving lower bounds, see e.g., Appendix A.1 and Appendix G.7.

D.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. [2023b], Glasgow and Rakhlin [2023]. We first state our lower bound in terms of constrained DEC in the following theorem.

Theorem D.1. *Under the reward maximization setting (Example 4), for any T-round algorithm* ALG*, there exists $M^\star \in \mathcal{M}$ such that*

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

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

Theorem D.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. [2023b, Theorem 2.2].

For the remainder of this section, we sketch how we prove Theorem D.1 in a slightly more general setting (Assumption 4), following Appendix E.1. 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 (19), and then apply Theorem 7 to obtain an analogous regret lower bound immediately. However, as Foster et al. [2023b] noted, the "correct" notion of regret-DEC (cf. Eq. (14)) 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

$$\text{r-dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M}) := \inf_{p\in\Delta(\Pi_T)} \sup_{M\in\mathcal{M}} \left\{ \widehat{L}_\delta(M,p) \vee \mathbb{E}_{\pi\sim p}[L(\overline{M},\pi)] \,\middle|\, \mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \le \varepsilon^2 \right\}, \tag{38}$$

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

The following proposition relates our quantile regret-DEC to the constrained regret-DEC (proof in Appendix E.3).

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

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

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

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

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 6). The proof is presented in Appendix E.2.

Theorem D.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*

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

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

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

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

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

D.2 Results for interactive estimation

More generally, we show that for a fairly different task of interactive estimation (Example 10), 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 $L(M, \pi) = \rho(\theta_M, \theta)$, for certain distance $\rho(\cdot, \cdot)$.

In interactive estimation, we can show that the quantile DEC is in fact lower bounded the constrained DEC, as follows (proof in Appendix E.4).

Proposition D.4. *Consider the setting of Example 10. Then as long as $\delta < \frac{1}{2}$, it holds that*

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

In particular, for such a setting (which encompasses the model estimation task considered in Chen et al. [2022a]), Theorem 7 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 5 is actually not restricted to Example 4, and we have hence shown that

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

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 D.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 F.1. 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 $\textsf{ExO}^+$ (Algorithm 1) achieves

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

where $N(\Theta, \Delta)$ is the Δ-covering number of Θ, because we have $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta) \leq \log N(\Theta, \Delta)$ by considering the prior $q = \mathrm{Unif}(\Theta_0)$ for a minimal Δ-covering of Θ. Similar to Theorem G.5, 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, $\textsf{ExO}^+$ achieves

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

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 [2019]), 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)

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

This result can be regarded as a generalization of Polyanskiy and Wu [2019] to the interactive estimation setting.

E Proofs from Section 3.2 and Appendix D

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

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

E.1 Proof of Proposition 8

In this section, we prove Proposition 8 under the slightly more general setting of Assumption 4.

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

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

For Example 4, we always have $C_{\mathrm{r}} \leq \sqrt{2}$, and hence Proposition 8 follows immediately from Proposition E.1.

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

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

whose existence is guaranteed by the definition of $\mathsf{p\text{-}dec}^{\mathrm{q}}_{\varepsilon,\delta}(\mathcal{M}, \overline{M})$ in (19). In other words, we have

$$\mathbb{P}_{\pi\sim\bar{p}}(L(M, \pi) \leq \Delta_0) \geq 1 - \delta, \qquad \forall M \in \mathcal{M}_{\bar{q},\varepsilon}(\overline{M})$$

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

$$p(\cdot) = \bar{p}\Big(\cdot | L(\tilde{M}, \pi) \leq \Delta_0\Big).$$

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

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

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

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

Combining these inequalities gives

$$\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon'}(\mathcal{M}, \overline{M}) \leq \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi\sim p}[L(M, \pi)] \,|\, \mathbb{E}_{\pi\sim q} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right) \leq \frac{\varepsilon^2}{2} \right\} \leq \Delta_0 + \frac{2\varepsilon C_{\mathrm{r}}}{1-\delta}.$$

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

E.2 Proof of Theorem D.3

Our proof adopts the analysis strategy originally proposed by Glasgow and Rakhlin [2023].

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

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

$$\tau = \inf \left\{ t : \sum_{s=1}^{t} L(\overline{M}, \pi)s) \geq T\Delta - 1 \text{ or } t = T + 1 \right\}.$$

If $\tau \leq T$, then it holds that $\sum_{t=1}^{\tau} L(\overline{M}, \pi^t) \geq T\Delta - 1$.

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

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

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

$$\mathbb{P}^{\overline{M},\mathsf{ALG}'}\left(\sum_{t=1}^{T} L(M, \pi^t) \geq T\Delta\right) > \delta, \qquad D^2_{\mathrm{H}}\left(\mathbb{P}^{M,\mathsf{ALG}'}, \mathbb{P}^{\overline{M},\mathsf{ALG}'}\right) \leq 7T\varepsilon^2. \tag{39}$$

We also know

$$\begin{aligned}\mathbb{E}^{\overline{M},\mathsf{ALG}'}\left[\frac{1}{T}\sum_{t=1}^{T} |f^M(\pi^t) - f^{\overline{M}}(\pi^t)|^2\right] &\leq \mathbb{E}^{\overline{M},\mathsf{ALG}'}\left[\frac{1}{T}\sum_{t=1}^{T} C_{\mathrm{r}}^2 D^2_{\mathrm{H}}\left(M(\pi^t), \overline{M}(\pi^t)\right)\right] \\ &= C_{\mathrm{r}}^2 \mathbb{E}_{\widehat{\pi}\sim p} D^2_{\mathrm{H}}\left(M(\widehat{\pi}), \overline{M}(\widehat{\pi})\right) \leq C_{\mathrm{r}}^2 \varepsilon^2,\end{aligned}$$

and hence by Markov inequality,

$$\mathbb{P}^{\overline{M},\mathsf{ALG}'}\left(\frac{1}{T}\sum_{t=1}^{T} |f^M(\pi^t) - f^{\overline{M}}(\pi^t)| \geq CC_{\mathrm{r}}\varepsilon\right) \leq \frac{1}{C^2}.$$

In the following, we consider events

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

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

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

$$\sum_{t=1}^{\tau} L(M, \pi^t) \geq T\Delta - X - 1.$$

To prove the claim, we bound

$$\begin{aligned}\sum_{t=1}^{\tau} L(M,\pi^t) &= \sum_{t=1}^{T} L(M,\pi^t) - \sum_{t=\tau+1}^{T} L(M,\pi^t) \\ &\geq T\Delta - \sum_{t=\tau+1}^{T} [f^M(\pi_M) - f^M(\pi^t)] \\ &\geq T\Delta - (T-\tau) f^M(\pi_M) + \sum_{t=\tau+1}^{T} f^{\overline{M}}(\pi^t) - X \\ &= T\Delta - (T-\tau)\cdot\big(f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}})\big) - X,\end{aligned}$$

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

$$\begin{aligned}\sum_{t=1}^{\tau} L(M,\pi^t) &= \sum_{t=1}^{\tau}[f^M(\pi_M) - f^M(\pi^t)] \\ &\geq \tau f^M(\pi_M) - \sum_{t=1}^{\tau} f^{\overline{M}}(\pi^t) - X \\ &= \tau\cdot\big(f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}})\big) + \sum_{t=1}^{\tau} L(\overline{M},\pi^t) - X \\ &\geq \tau\cdot\big(f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}})\big) + T\Delta - 1 - X,\end{aligned}$$

where the first inequality follows from $\sum_{t=1}^{\tau}|f^M(\pi^t) - f^{\overline{M}}(\pi^t)| \leq X$, and the second inequality is because $\sum_{t=1}^{\tau} L(\overline{M},\pi^t) \geq T\Delta - 1$ given $\tau \leq T$, 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

$$\mathbb{P}^{\overline{M},\mathsf{ALG}'}\left(\sum_{t=1}^{\tau\wedge T} L(M,\pi^t) \geq T(\Delta - C\varepsilon) - 1\right) \geq \mathbb{P}^{\overline{M},\mathsf{ALG}'}(\mathcal{E}_1 \cap \{X \leq CT\varepsilon\}) \geq \delta - \frac{1}{C^2}.$$

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

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

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

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

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

E.3 Proof of Proposition D.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

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

The existence of p is guaranteed by the definition (38). In other words, we have $\mathbb{E}_{\pi\sim p}[L(\overline{M},\pi)] \le \Delta$ and

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

We then has the following claim.

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

$$\mathbb{E}_{\pi\sim p}[L(M,\pi)] \le \mathbb{E}_{\pi\sim p}[L(\overline{M},\pi)] + \Delta + c_\delta C_{\mathrm{r}} \mathbb{E}_{\pi\sim p} D_{\mathrm{H}}\big(M(\pi), \overline{M}(\pi)\big). \tag{40}$$

Fix any $M \in \mathcal{M}_{p,\varepsilon}(\overline{M})$, we prove (40) as follows. Consider the event $\mathcal{E} = \{\pi : L(M,\pi) \le \Delta\}$. Then,

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

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

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

This completes the proof of our claim.

Therefore, using (40) with $\mathbb{E}_{\pi\sim p}[L(\overline{M},\pi)] \le \Delta$ yields

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

This immediately implies

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

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

E.4 Proof of Proposition D.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

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

Therefore, it holds that

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

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

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

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

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

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

The above claim immediately implies that

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

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. □

F Additional discussion and results from Section 4

F.1 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. [2022], built upon Lattimore and Szepesvári [2020b], Lattimore and Gyorgy [2021]. 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 $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)$, instead of $\log|\Pi|$ (cf. Foster et al. [2022]), which is always an upper bound of $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)$.

Offset DEC for regret. We first recall the following (original) definition of DEC [Foster et al., 2021]:

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) := \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[L(M, \pi)] - \gamma \mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right), \tag{41}$$

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 [Foster et al., 2021], 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. [2023b], in the reward maximization setting (Example 4), 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 G.5), 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

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

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

$$\begin{aligned} \Gamma_{q,\gamma}(p, \ell; M, \pi^\star) = \mathbb{E}_{\pi \sim p}[f^M(\pi^\star) - f^M(\pi)] \\ - \gamma \mathbb{E}_{\pi \sim p} \mathbb{E}_{o \sim M(\pi)} \mathbb{E}_{\pi' \sim q}\big[1 - \exp\big(\ell(\pi'; \pi, o) - \ell(\pi^\star; \pi, o)\big)\big], \end{aligned} \tag{42}$$

and

$$\Gamma_{q,\gamma}(p, \ell) = \sup_{M \in \mathcal{M}, \pi^\star \in \Pi} \Gamma_{q,\gamma}(p, \ell; M, \pi^\star), \tag{43}$$

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. [2022], we define

$$\mathsf{exo}_{1/\gamma}(\mathcal{M}, q) := \inf_{p \in \Delta(\Pi), \ell \in \mathcal{L}} \Gamma_{q,\gamma}(p, \ell), \tag{44}$$

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. [2022, Theorem 3.1 and 3.2].

Theorem F.1. *Under the reward maximization setting*[7] *(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 F.2. *It holds that with probability at least* $1 - \delta$,

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

[7] We remark that their proof applies as long as f^M can be linearly extended to $\mathrm{co}(\mathcal{M})$.

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

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

and Proposition F.3 implies that

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

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

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

Notice that

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

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

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

Therefore, it holds that

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

Applying Theorem F.1 completes the proof. □

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

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

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 $h : \Pi \to \mathbb{R}$,

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

□

F.2 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 $r^t \sim \mathcal{N}\big(h_\star(c^t,a^t),1\big)$.

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

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

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 Section 4.3.1, we instantiate Theorem 19 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 $T^{\mathsf{DEC}}(\mathcal{H},\Delta) = \inf_{\varepsilon\in(0,1)}\{\varepsilon^{-2} : \textsf{r-dec}^{\mathsf{c}}_{\varepsilon}(\mathcal{H}) \le \Delta\}$, following Eq. (23).

Fractional covering number for contextual bandits. Specializing the fractional covering number to contextual bandits, we define

$$\mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta) := \inf_{p\in\Delta(\Pi)} \sup_{h\in\mathcal{H},\nu\in\Delta(\mathcal{C})} \frac{1}{p(\pi : \mathbb{E}_{c\sim\nu}[h(c,\pi_h(c)) - h(c,\pi(c))] \le \Delta)}, \tag{45}$$

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 fractional covering number $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta)$ 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 $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta) = |\mathcal{C}|\log|\mathcal{A}|$, but in general we can have $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta) \ll \log|\Pi| = |\mathcal{C}|\log|\mathcal{A}|$.

As a corollary of Theorem 19, we derive the following upper and lower bounds on the complexity of contextual bandit learning with $\mathcal{H}$.

Theorem F.4. *Let $\mathcal{H}$ be given. Suppose that both the context space $\mathcal{C}$ and the action space $\mathcal{A}$ are finite, and that $\varepsilon \mapsto \textsf{r-dec}^{\mathsf{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H}))$ satisfies moderate decay as a function of ε (Assumption 3) with constant c_{reg}. Let $\bar{\varepsilon}(T) \asymp \sqrt{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta)/T}$. Then Algorithm 1 ensures that with high probability, ;*

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

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

$$\max\left\{T^{\mathsf{DEC}}(\mathcal{H},\Delta), \frac{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},2\Delta)}{\log|\mathcal{C}|}\right\} \lesssim T^{\star}(\mathcal{M}_{\mathcal{H}},\Delta) \lesssim T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{H}),\Delta)\cdot\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta/2), \tag{46}$$

omitting dependence on c_{reg} and logarithmic factors.

By definition, we have $\textsf{r-dec}^{\mathsf{c}}_{\varepsilon}(\mathrm{co}(\mathcal{H})) = \textsf{r-dec}^{\mathsf{c}}_{\varepsilon}(\mathcal{H})$ if the *per-context* value function class $\mathcal{H}|_c$ is convex for every context $c \in \mathcal{C}$. Natural settings in which $\mathcal{H}|_c$ is convex include contextual linear bandits [Chu et al., 2011], contextual non-parametric bandits [Cesa-Bianchi et al., 2017], contextual concave bandits [Lattimore, 2020], etc. For these problem classes, the complexity of no-regret learning is completely characterized by the DEC of $\mathcal{H}$ and the newly proposed $\mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta)$ (up to a quadratic factor and a factor of $\log|\mathcal{C}|$).

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

Corollary F.5. *For any value function class $\mathcal{H}$, Algorithm 1 ensures the following regret bound with high probability.*

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

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 $\mathcal{H}$ [Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2020], this result above always provides a tighter upper bound, as $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta) \le \log|\mathcal{H}|$. For certain (very simple) function classes $\mathcal{H}$, the quantity $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta)$ can be much smaller than $\log |\mathcal{H}|$ (for details, see Example 14). More importantly, $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta)$ leads to lower bounds for *any* contextual bandit function class (Theorem F.4). 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 $\mathcal{H}$).

G Proofs from Section 4 and Appendix F

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

G.1 Proof of Theorem 11

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

Proof. By the definition (25) of $\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)$, we know

$$\frac{1}{\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)} := \sup_{p \in \Delta(\Pi)} \inf_{M \in \mathcal{M}} p(\pi : L(M, \pi) \leq \Delta).$$

Therefore, we have

$$\inf_{M \in \mathcal{M}} p_{\overline{M},\mathsf{ALG}}(\pi : L(M, \pi) \leq \Delta) \leq \frac{1}{\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)},$$

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

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

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

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

Hence, using data-processing inequality,

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

This immediately implies $p_{M,\mathsf{ALG}}(\pi : L(M, \pi) \leq \Delta) < \frac{1}{2}$ by the monotonicity of KL divergence. $\square$

G.2 Proof of Theorem 13

In this section, we present an algorithm based on reduction to multi-arm bandits (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 (25). 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 $N = \mathsf{N}_{\rm frac}(\mathcal{M}, \Delta) \log(1/\delta)$. Then the space Π_{sub} is guaranteed to contain a near-optimal decision, as follows.

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

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 [2020a]), 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 G.1. By definition,

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

□

G.3 Proof of Lemma 14

Proof of the upper bound. Take a minimal Δ-covering of $\Pi_{\mathcal{M}}$, i.e., a set $\{\pi^1, \cdots, \pi^n\} \subseteq \Pi$ such that for all $M \in \mathcal{M}$, there exists $i \in [n]$ such that $\rho(\pi_M, \pi^i) \le \Delta$. Therefore, we may consider

Algorithm 2 A reduction algorithm based on the fractional covering number

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

1: Set

$$p^\star_\Delta = \arg \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \frac{1}{p(\pi : L(M, \pi) \leq \Delta)}. \tag{47}$$

2: Set $N = \mathsf{N}_{\rm frac}(\mathcal{M}, \Delta) \log(1/\delta)$ 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 $\mathsf{BanditALG}$ on the instance $M^\star_{\mathsf{Bandit}}$ for T rounds.

---

the distribution $p = \mathrm{Unif}(\{\pi^1, \cdots, \pi^n\})$, which guarantee

$$\mathsf{N}_{\rm frac}(\mathcal{M}, \Delta) \leq \sup_{M \in \mathcal{M}} \frac{1}{p(\pi : \rho(\pi_M, \pi) \leq \Delta)} \leq n = \mathsf{N}(\Pi_{\mathcal{M}}, \Delta).$$

Proof of the lower bound. Consider the maximal 2Δ-packing of $\Pi_{\mathcal{M}}$, i.e., let $\{\pi^1, \cdots, \pi^m\} \subseteq \Pi_{\mathcal{M}}$ be a maximal set such that $\rho(\pi^i, \pi^j) > 2\Delta$ for any $i \neq j$. Then, by the duality between packing and covering, the set $\{\pi^1, \cdots, \pi^m\}$ form a 2Δ-covering of $\Pi_{\mathcal{M}}$, and hence we have $m \geq \mathsf{N}(\Pi_{\mathcal{M}}, 2\Delta)$. On the other hand, the sets $\Pi^i := \{\pi : \rho(\pi, \pi^j) \leq \Delta\}$ are pairwise disjoint, and hence for any $p \in \Delta(\Pi)$, we have

$$m \cdot \inf_{M \in \mathcal{M}} p(\pi : \rho(\pi_M, \pi) \leq \Delta) \leq \sum_{i=1}^{m} p(\pi : \rho(\pi^i, \pi) \leq \Delta) \leq 1.$$

Therefore, it holds that $\mathsf{N}_{\rm frac}(\mathcal{M}, \Delta) \geq m \geq \mathsf{N}(\Pi_{\mathcal{M}}, 2\Delta)$. □

G.4 Proof of Example 7

It remains to prove Eq. (29). More generally, we prove the following lemma.

Lemma G.2. *For model class $\mathcal{M} = \bigcup_{i=1}^{n} \mathcal{M}_i$, it holds that*

$$\mathsf{N}_{\rm frac}(\mathcal{M}, \Delta) \leq \sum_{i=1}^{n} \mathsf{N}_{\rm frac}(\mathcal{M}_i, \Delta).$$

Proof of Lemma G.2 For each $i \in [n]$, we define $\lambda_i = \frac{\mathsf{N}_{\rm frac}(\mathcal{M}_i, \Delta)}{\sum_{j=1}^{n} \mathsf{N}_{\rm frac}(\mathcal{M}_j, \Delta)}$.

Fix $p_1, \cdots, p_n \in \Delta(\Pi)$. Then, let us consider the distribution $p = \sum_{i=1}^{n} \lambda_i p_i \in \Delta(\Pi)$. For any model $M \in \mathcal{M}$, there exists $i \in \mathcal{M}_i$, and hence $p(\pi : L(M, \pi) \leq \Delta) \geq \lambda_i \min_{M_i \in \mathcal{M}_i} p_i(\pi : L(M_i, \pi) \leq \Delta)$. Therefore, it holds that

$$\inf_{M \in \mathcal{M}} p(\pi : L(M, \pi) \leq \Delta) \geq \min_{i \in [n]} \lambda_i \inf_{M \in \mathcal{M}_i} p_i(\pi : L(M, \pi) \leq \Delta).$$

In other words,

$$\sup_{M\in\mathcal{M}} \frac{1}{p(\pi : L(M,\pi)\le\Delta)} \ge \max_{i\in[n]} \frac{1}{\lambda_i} \sup_{M\in\mathcal{M}_i} \frac{1}{p_i(\pi : L(M,\pi)\le\Delta)}.$$

Taking infimum over $p_1,\cdots,p_n\in\Delta(\Pi)$ gives

$$\begin{aligned}
\mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta) = & \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \frac{1}{p(\pi : L(M,\pi)\le\Delta)} \ge \max_{i\in[n]} \frac{1}{\lambda_i} \sup_{M\in\mathcal{M}_i} \frac{1}{p_i(\pi : L(M,\pi)\le\Delta)} \\
\le & \inf_{p_1,\cdots,p_n\in\Delta(\Pi)} \max_{i\in[n]} \frac{1}{\lambda_i} \sup_{M\in\mathcal{M}_i} \frac{1}{p_i(\pi : L(M,\pi)\le\Delta)} \\
= & \max_{i\in[n]} \frac{1}{\lambda_i} \inf_{p_i\in\Delta(\Pi)} \sup_{M\in\mathcal{M}_i} \frac{1}{p_i(\pi : L(M,\pi)\le\Delta)} \\
= & \max_{i\in[n]} \frac{1}{\lambda_i} \cdot \mathsf{N}_{\mathrm{frac}}(\mathcal{M}_i,\Delta) = \sum_{i=1}^n \mathsf{N}_{\mathrm{frac}}(\mathcal{M}_i,\Delta),
\end{aligned}$$

where the last line follows from the definition of $\lambda_1,\cdots,\lambda_n$. This is the desired result. □

G.5 Proof of Theorem 18

We first state the following more general result, and Theorem 18 is then a direct corollary (under Assumption 3). Analoguous guarantees also hold for PAC learning.

Theorem G.3. *Let $T\ge 1, \delta\in(0,1)$. With suitably chosen prior $q\in\Delta(\Pi)$, ExO^+ (Algorithm 1) achieves with probability at least $1-\delta$:*

$$\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} \le \Delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma/8}(\mathrm{co}(\mathcal{M})) + \gamma\frac{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)+\log(1/\delta)}{T}. \tag{48}$$

In particular, when $\mathcal{M}$ is a reward-maximization problem class (Example 4), ExO^+ achieves (with a suitable parameter γ) that with probability at least $1-\delta$:

$$\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} \le \Delta + C\sqrt{\log(T)}\cdot \overline{\textsf{r-dec}}^{\mathrm{c}}_{\bar\varepsilon(T)}(\mathrm{co}(\mathcal{M})), \tag{49}$$

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

$$\overline{\textsf{r-dec}}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M})) := \varepsilon\cdot\sup_{\varepsilon'\in[\varepsilon,1]} \frac{\textsf{r-dec}^{\mathrm{c}}_{\varepsilon'}(\mathrm{co}(\mathcal{M}))}{\varepsilon'}. \tag{50}$$

Remark G.4 (Upper bound without regularity condition)**.** In Theorem 18 (and Eq. (49)), we assume that (1) $\mathcal{M}$ is a reward-maximization problem, and (2) the constrained regret-DEC of $\mathrm{co}(\mathcal{M})$ satisfies certain regularity condition (Assumption 3). As we have noted in Remark 21, we can relax these two assumptions and obtain a weaker upper bound.

Specifically, we may only assume that $f^M:\Pi\to[0,1]$ is affine with respect to $M\in\mathrm{co}(\mathcal{M})$ (cf. Theorem F.1). In this case, we can still bound the regret of ExO^+ as

$$\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} \le \Delta + \textsf{r-dec}^{\mathrm{c}}_{\sqrt{\gamma/8}}(\mathrm{co}(\mathcal{M})) + \gamma\frac{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{M},\Delta)+\log(1/\delta)}{T}, \tag{51}$$

which follows from Eq. (48) and the fact that (by definition)

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathcal{M})) \le \textsf{r-dec}^{\mathrm{c}}_{\sqrt{\gamma}}(\mathrm{co}(\mathcal{M})), \qquad \forall \gamma > 0.$$

In particular, using Eq. (51) above, we can show that (omitting poly-logarithmic factors)

$$T^{\star}(\mathcal{M}, \Delta) \lesssim \frac{1}{\Delta} \cdot T^{\mathsf{DEC}}(\mathrm{co}(\mathcal{M}), \Delta/3) \cdot \log \mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta/2).$$

This is worse than the upper bound of Theorem 19 by (roughly) a factor of Δ^{-1}.

Proof of Theorem G.3. By the definition (25) of $\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)$, we know

$$\frac{1}{\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)} := \sup_{p \in \Delta(\Pi)} \inf_{M \in \mathcal{M}} p(\pi : L(M, \pi) \le \Delta).$$

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

$$\inf_{M \in \mathcal{M}} q(\pi : L(M, \pi) \le \Delta) \ge \frac{1}{\mathsf{N}_{\mathrm{frac}}(\mathcal{M}, \Delta)},$$

We then instantiate Algorithm 1 with such a prior q, and Eq. (48) follows immediately from Theorem F.2. To prove Eq. (49), we invoke the following structural result that relates offset DEC to constrained DEC.

Theorem G.5. *Suppose that Assumption 4 holds for the model class $\mathcal{M}$. Then for any $\varepsilon \in (0, 1]$, it holds that*

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

Under the assumption that $\varepsilon \mapsto \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M}))$ is of moderate decay with a constant c_{reg}, we have

$$\overline{\textsf{r-dec}}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M})) \le c_{\mathrm{reg}} \textsf{r-dec}^{\mathrm{c}} \varepsilon(\mathcal{M}), \qquad \forall \varepsilon \in (0, 1].$$

Hence, Eq. (49) follows from (48) as long as the parameter γ is chosen according to Eq. (G.5). □

G.5.1 Proof of Theorem G.5

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

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

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 $\lambda_K = 1 - \sum_{i=0}^{K-1} \lambda_i \ge c$.

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

By definition of p_i, it holds that

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

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

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

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

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

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,

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

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

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

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

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

Therefore, using AM-GM inequality,

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

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

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

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

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma(\varepsilon)}(\mathcal{M}, \overline{M}) \le (2D + C_{\mathrm{r}} + cKD + 2cC_{\mathrm{r}})\varepsilon.$$

Thus,

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

Balancing c and re-arranging yields the desired result. $\square$

G.6 Proof of Theorem 19

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

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

Under Assumption 3, the regret upper bound in Theorem 18 implies (up to $c_{\mathrm{reg}}, C_{\mathrm{KL}}$ and logarithmic factors)

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

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

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

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

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

Together with Theorem 10, we prove that

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

□

G.7 Proof of Theorem 22

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 22: lower bound of (33). The lower bound with $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta)$ is exactly Corollary 15.

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 G.6. *Consider $\mathcal{M}^+=\mathcal{M}_{\mathrm{co}(\mathcal{H}),\mathbb{V}}$ as the class of all reference models (Appendix D). Then,*

$$\max_{\overline{M}\in\mathcal{M}^+}\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\mathcal{H}}\cup\{\overline{M}\},\overline{M})\ge \mathsf{r\text{-}dec}^{\mathrm{c}}_{2\sqrt{2}\varepsilon}(\mathcal{H}). \tag{52}$$

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

$$\mathbf{Reg}_{\mathsf{DM}}(T) \geq \frac{T}{2} \cdot (\textsf{r-dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{H}) - 5\underline{\varepsilon}(T)) - 1 \tag{53}$$

with probability at least 0.01 under $\mathbb{P}^{M^\star,\mathsf{ALG}}$, 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 22: upper bound. We apply Theorem G.3 similar to the proof of Theorem 19 (in Appendix G.2).

Using Theorem G.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 $1-\delta$,

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

where C is an absolute constant, $\bar{\varepsilon}(T) = \sqrt{\frac{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta) + \log(1/\delta)}{T}}$. We only need to upper bound the $\overline{\textsf{r-dec}}^{\mathrm{c}}_{\varepsilon}(\mathrm{co}(\mathcal{M}_{\mathcal{H},\mathbb{V}}))$ (defined in (50)) in terms of the DEC of $\mathrm{co}(\mathcal{H})$.

Lemma G.7. *For any $\varepsilon \geq 0$, it holds that*

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

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

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

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

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

This gives the desired upper bound. □

G.7.1 Proof of Lemma G.6

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

$$\begin{aligned} &\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \\ =& \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}\cup\{\overline{M}\}} \left\{ \mathbb{E}_{\pi\sim p}[L(M,\pi)] \,|\, \mathbb{E}_{\pi\sim p} D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right) \leq \varepsilon^2 \right\} \end{aligned}$$

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

where the last line follows from Lemma B.5. Taking $\Delta \to \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_1}(\mathcal{H})$ completes the proof of (52). □

G.7.2 Proof of Lemma G.7

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 B.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 $\mathcal{M} = \mathcal{M}_{\mathcal{H},\mathbb{V}}$,

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

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

G.8 Proof of Theorem F.4

Similar to Appendix G.7, 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 F.4: lower bound. The lower bound with $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta)$ follows immediately by applying Theorem 11 to the class $\mathcal{M}_{\mathcal{H}}$, which admits $C_{\mathrm{KL}} = O(\log|\mathcal{C}|)$ in Assumption 2 (as shown in Example 13).

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 22 (specifically (53)):

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

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

$$\frac{1}{T}\mathbf{Reg}^\star(\mathcal{M}_{\mathcal{H}}, T) \gtrsim \mathsf{r\text{-}dec}^{\mathrm{c}}_{\underline{\varepsilon}(T)}(\mathcal{H}) - 6\underline{\varepsilon}(T).$$

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

Combining both lower bounds completes the proof. □

Proof of Theorem F.4: upper bound. We follow the proof strategy of Appendix G.7. By Theorem G.3, ExO^+ can be suitably instantiated on the problem class $\mathcal{M}_{\mathcal{H},\mathbb{V}}$ so that with probability at least $1-\delta$:

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

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 G.8. *For $\gamma > 0$, it holds that*

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

Then, we can apply the result of Theorem G.5. From the proof of Theorem G.5, it is not hard to see that: for any $\varepsilon > 0$, there exists $\gamma = \gamma(\varepsilon)$ such that for any $c \in \mathcal{C}$,

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

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

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

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

G.8.1 Proof of Lemma G.8

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 B.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

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\mathcal{H},\mathbb{V}}, \overline{M})$$

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

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

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

□

G.9 Proof of Corollary F.5

We follow the notations of Appendix G.8. By Lemma G.8, we have

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

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. [2021, Proposition 5.1] and Lemma B.5, we have

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

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

$$
\frac{1}{T}\mathbf{Reg}_{\mathsf{DM}} \le \Delta + \frac{16|\mathcal{A}|}{\gamma} + \gamma \frac{\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H},\Delta) + \log(1/\delta)}{T}.
$$

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

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

Example 14. Suppose that $\mathcal{A} = \{0, 1\}$, and the function class $\mathcal{H} = \{h_x\}_{x\in\mathcal{C}}$, where

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

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 $\{\pi(x) = 1\}$,

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

Hence,

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

which implies $\log \mathsf{N}_{\mathrm{frac}}(\mathcal{H}, \Delta) \leq \log(2/\Delta)$.

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