On the Hardness of Bandit Learning

Nataly Brukhim* NBRUKHIM@PRINCETON.EDU
Institute for Advanced Study

Aldo Pacchiano* PACCHIAN@BU.EDU
Boston University & Broad Institute of MIT and Harvard

Miroslav Dudik MDUDIK@MICROSOFT.COM
Microsoft Research

Robert Schapire SCHAPIRE@MICROSOFT.COM
Microsoft Research

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the task of bandit learning, also known as best-arm identification, under the assumption that the true reward function f belongs to a known, but arbitrary, function class $\mathcal{F}$. While many instances of this problem are well understood, we seek a general theory of bandit learnability, akin to the PAC framework for classification. Our investigation is guided by the following two fundamental questions: (1) *which* classes $\mathcal{F}$ are learnable, and (2) *how* they are learnable. For example, in the case of binary PAC classification, learnability is fully determined by a combinatorial dimension, namely, the VC dimension, and can be attained via a simple algorithmic principle, namely, empirical risk minimization (ERM).

In contrast to classical learning-theoretic results, our findings reveal fundamental limitations of learning in structured bandits, offering new insights into the boundaries of bandit learnability. First, for the question of "*which*", we show that the paradigm of identifying the learnable classes via a dimension-like quantity fails for bandit learning. We give a simple proof demonstrating that no combinatorial dimension can characterize bandit learnability, even in finite classes, following a standard definition of dimension introduced by Ben-David et al. (2019).

For the question of "*how*", we prove a computational hardness result: we construct a reward function class for which at most two queries are needed to find the optimal action, yet no algorithm can do so in polynomial time, unless $\mathrm{RP} = \mathrm{NP}$. Perhaps surprisingly, we also prove that this class admits efficient algorithms for standard (albeit possibly computationally hard) algorithmic operations often considered in learning theory, such as an ERM. Therefore, this implies that computational hardness is in this case inherent to the task of bandit learning.

Beyond these results, we investigate additional themes such as learning under noise, trade-offs between noise models, and the relationship between query complexity and regret minimization.

1. Introduction

In statistical learning theory, the *probably approximately correct* (PAC) framework (Valiant, 1984) is central to understanding binary classification learnability. A key result shows that PAC learnability is fully determined by the VC dimension (Vapnik and Chervonenkis, 1974; Blumer et al., 1989), elegantly linking learnability and sample complexity. Similar characterizations exist for widely diverse variants of statistical and online learning (e.g., Bartlett et al., 1994; Littlestone, 1988). The

* Equal contribution.

appeal of combinatorial characterizations is often in their simplicity, reducing learnability to a single parameter, and offering useful insights into algorithm design and problem structure.

In contrast, a similarly tight characterization of *bandit learning*, and in particular of the problem known as *best-arm identification*, is still lacking. In this setting, there is a set of *actions* (or arms) $\mathcal{A}$, and an unknown *reward function* $f^* : \mathcal{A} \to [0, 1]$. A learner repeatedly queries actions $a \in \mathcal{A}$ and observes their corresponding reward, a random variable with mean $f^*(a)$. The aim of a learning algorithm is to identify a near-optimal action using as few queries as possible. Analogous to classic learning settings, one may assume the rewards are *realizable* by a known, but arbitrary, class of reward functions $\mathcal{F} \subseteq [0, 1]^{\mathcal{A}}$. A key focus of study in this context is the optimal *query complexity* associated with a given class.

The pursuit of VC-dimension-like parameters for bandit learning has drawn considerable attention (Amin et al., 2011; Russo and Van Roy, 2013; Foster et al., 2021; Brukhim et al., 2023; Foster et al., 2023; Hanneke and Wang, 2024). However, existing parameters are often non-combinatorial in nature, rather complex, and in the general case exhibit substantial gaps between upper and lower bounds (see Section 1.1 for further discussion). We start our investigation by asking whether there exists a combinatorial characterization of bandit learning. Somewhat disappointingly, we prove that no such characterization of bandit learnability exists. Specifically, we use the definition of a combinatorial dimension introduced by Ben-David et al. (2019) that encompasses all standard notions of dimension in both statistical and online learning. Using a rather simple argument, we demonstrate that no such dimension can universally characterize bandit learnability, even for finite classes.

We then shift our focus to exploring algorithmic approaches to the problem. Specifically, we examine reward function classes with *small* optimal query complexity and seek a general algorithmic principle that achieves it, guided by the question:

When is a class $\mathcal{F}$ of a *bounded* query complexity, *efficiently* bandit-learnable?

There are several algorithmic oracle assumptions commonly considered in the context of computational efficiency. For example, in statistical learning theory, the gold standard is the simple *empirical risk minimization* (ERM) principle which determines that it suffices to find any function in the class that is consistent with past observations. In interactive settings, an estimation algorithm is often used both to find a consistent function and to produce future predictions (see, e.g., Foster et al., 2023; Brukhim et al., 2023). One might assume that a class which admits efficient algorithms as above might also be efficiently bandit-learnable. Interestingly, we prove a hardness result showing that is not the case. Specifically, we construct a reward function class for which at most *two queries* are needed to find the optimal action, yet no algorithm can do so in polynomial time, unless $RP = NP$. Moreover, we prove that this class admits efficient algorithms to the aforementioned tasks, demonstrating that the hardness is inherent to the bandit setting.

An important aspect of bandit learnability is the noise model being considered. In the absence of noise, learning is less constrained and is therefore simpler. However, it is also more brittle, as it relies heavily on the precise function values that define the structure of the class $\mathcal{F}$. In contrast, under sufficiently noisy conditions, bandit learnability exhibits a form of robustness, allowing it to be characterized by simple parameters and algorithms, as shown in recent work by Hanneke and Wang (2024). However, while some works (Hanneke and Yang, 2023; Amin et al., 2011) focused on the noise-free regime, Hanneke and Wang (2024) considered a highly complex family of distributions of arbitrary noise, leaving intermediate noise regimes largely unaddressed (see further discussion in Section 1.1).

In this work, we partially address this gap and examine the effect of noise on the query complexity of bandit learning. We focus on a Gaussian noise model and study the relationship between the noise variance and the query complexity. For instance, we show that certain function classes have a query complexity of 1 when $\sigma = 0$ but become unlearnable (i.e., infinite query complexity) when $\sigma = 1$. Moreover, we identify an upper bound $\bar{\sigma}$ on σ such that, for any function class, the query complexity for any $\sigma \leqslant \bar{\sigma}$ is upper bounded by the query complexity for $\sigma = 0$. This observation implies that the query complexity in the low-noise regime can be captured by the noise-free setting.

Additionally, we prove that for a specific family of function classes, there exist class-dependent thresholds for σ, that separate distinct learning regimes. Above a certain noise level, the query complexity is governed by a simple parameter γ known as the *generalized maximin volume*, introduced by Hanneke and Wang (2024). Below a different threshold the query complexity is 1, exhibiting a large gap from γ. Understanding the broader interplay between noise variance and query complexity across arbitrary function classes remains an open and interesting direction for future research.

Finally, we examine an alternative notion of learnability in bandits via the lens of regret minimization and study its relationship with query complexity of best-arm identification. Specifically, we prove that any algorithm which achieves the optimal query complexity d, must also incur regret that is linear in d, and is *not* regret-optimal for time horizon $T = O(d)$. This result establishes that no single algorithm can simultaneously achieve both optimal query complexity and optimal regret.

1.1. Related work

The PAC framework and related combinatorial characterizations have played a crucial role in providing quantitative insights into learnability across statistical learning theory. However, bandit learning, particularly best-arm identification (BAI), lacks a unifying framework and remains largely a collection of case-specific analyses (see, e.g., Bubeck et al., 2012, and references within). Moreover, most prior BAI work (e.g., Garivier and Kaufmann, 2016; Kaufmann et al., 2016) assume that the mean rewards lie in some fixed bounded product space, e.g., $\mathcal{F} = [0,1]^K$, and so pulling one of the K arms provides no information about others. In contrast, the focus of this work is the setting in which observations can possibly reveal additional information, based on the structure of the class $\mathcal{F} \subsetneq [0,1]^K$.

Indeed, the approach of studying the structure of the class itself has gained attention in recent years (Foster et al., 2021, 2023; Hanneke and Yang, 2023; Hanneke and Wang, 2024). A notable proposed parameter for capturing interactive decision making is the decision-estimation coefficient (DEC) (Foster et al., 2021, 2023). However, it suffers from arbitrarily large gaps between upper and lower bounds (Foster et al., 2023) and fails to characterize learnability in stochastic bandits (see Hanneke and Wang, 2024).

More recently, Hanneke and Wang (2024) introduced a characterization for stochastic bandits with *arbitrary* noise, but it exhibits an exponential gap between upper and lower bounds and does not seamlessly extend to standard noise models, e.g., Gaussian noise. In Section 5, we further analyze their generalized maximin volume parameter, showing that under moderate-variance Gaussian noise, it can diverge arbitrarily from the optimal query complexity.

Finally, we establish that no combinatorial dimension fully characterizes bandit learnability. While Hanneke and Yang (2023) demonstrated a related result using complex set-theoretic arguments, their proof relies on the cardinality of the continuum and does not directly address combi-

natorial dimensions. In contrast, we provide a rather simple, direct argument showing that no such dimension exists, within the standard model of set theory, without any additional assumptions.

2. Query complexity of bandit learning

In this work, we study query complexity of bandit learning. Specifically, we focus on the following problem. Let $\mathcal{A}$ be an action set, $\mathcal{F}$ a set of reward functions $f : \mathcal{A} \to [0, 1]$, and $f^* \in \mathcal{F}$ the target reward function. In each round $t = 1, \ldots, T$, the learner queries an action $a_t \in \mathcal{A}$ and receives reward $r_t \in [0, 1]$ with $\mathbb{E}[r_t|a_t] = f^*(a_t)$. The goal is *best-arm identification*: for a given $\epsilon \in [0, 1]$, using as few queries as possible, identify an ϵ-optimal action. We consider both the *noise-free* setting, where $r_t = f^*(a_t)$, and the *noisy* setting, where in each round $t = 1, ..., T$, the learner observes $r_t = f^*(a_t) + \xi$ for some zero-mean random variable ξ. Throughout the paper, unless stated otherwise, we will assume a Gaussian noise model, i.e., $\xi \sim \mathcal{N}(0, \sigma^2)$.

We say that a class of reward functions $\mathcal{F} \subseteq [0, 1]^{\mathcal{A}}$ is *bandit-learnable* if there is a (possibly-randomized) algorithm Alg and a function $m : (0, 1)^2 \to \mathbb{N}$ such that for any $f \in \mathcal{F}$, when given any $\epsilon, \delta > 0$ and after having made at most $m(\epsilon, \delta)$ queries a_t to f and observed r_t (under the appropriate noise model), algorithm Alg outputs $\hat{a}$ such that with probability at least $1 - \delta$,

$$f(\hat{a}) \geqslant \sup_{a \in \mathcal{A}} f(a) - \epsilon.$$

The function $m(\cdot, \cdot)$ is the query complexity of Alg. We often denote $m^{\sigma}_{\text{Alg}}(\cdot, \cdot)$ when considering noisy feedback, for the appropriate choice of σ. We then define the query complexity of a given class $\mathcal{F}$, for any fixed choice of parameters, as follows.

Definition 1 *Given $\epsilon, \delta \in [0, 1]$, the (ϵ, δ)-query complexity for class $\mathcal{F} \subseteq [0, 1]^{\mathcal{A}}$ under a Gaussian noise model with $\xi \sim \mathcal{N}(0, \sigma^2)$, denoted $\text{QC}^{\sigma}_{\epsilon,\delta}(\mathcal{F})$, is the minimum over all $m^{\sigma}_{\text{Alg}}(\epsilon, \delta)$, where m_{Alg} is the query complexity of a bandit learning algorithm* Alg *for the class $\mathcal{F}$.*

3. No combinatorial dimension can characterize bandit learnability

A fundamental result of statistical learning theory is the characterization of PAC learnability in terms of the VC dimension of a class. Similar combinatorial characterizations exist for diverse variants of statistical learning (Vapnik, 1989; Natarajan and Tadepalli, 1988; Bartlett et al., 1994; Ben-David et al., 1992; Brukhim et al., 2022) as well as online learning (Littlestone, 1988; Ben-David et al., 2009; Rakhlin et al., 2015; Daniely et al., 2015).

All standard notions of dimension in the aforementioned learning settings can be abstracted as a function $\mathfrak{D}$ that maps a class of functions $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ to $\mathbb{N} \cup \{\infty\}$, while satisfying the following requirements: (1) *learnability characterization*: a class $\mathcal{F}$ is learnable if and only if $\mathfrak{D}(\mathcal{F}) < \infty$, and (2) *finite character*: for every integer d and $\mathcal{F}$, the statement "$\mathfrak{D}(\mathcal{F}) \geqslant d$" can be demonstrated by a finite set of domain points and a finite collection of members of $\mathcal{F}$. We will next give a more formal definition of the finite character property. First, we define the notion of a shattered set. In the definition and throughout this section, we write $\mathcal{F}|_X$ to denote the set of all functions in $\mathcal{F}$ restricted to points in X.

Definition 2 (Shattered sets) *For every $d \in \mathbb{N}$ let $V_d : \mathcal{X}^d \times 2^{\mathcal{Y}^d} \mapsto \{YES, NO\}$ be a shattering function. A set $X \in \mathcal{X}^d$ is shattered by hypothesis class $\mathcal{F}$, with respect to V_d if and only if $\mathcal{F}|_X$ is of finite cardinality and $V_d(X, \mathcal{F}|_X) =$ YES.*

Definition 3 (Finite character property) *We say that a dimension $\mathfrak{D}$ satisfies the finite character property if for every $d \in \mathbb{N}$ there exists a shattering function V_d such that $\mathfrak{D}(\mathcal{F}) \geqslant d$ if and only if there exists a shattered set of size at least d.*

This property was first defined by Ben-David et al. (2019), who gave a formal definition of the notion of "combinatorial dimension" or "complexity measure", satisfied by all previously proposed dimensions in statistical learning theory. The intuition is that a finite character property can be checked by probing finitely many elements of $\mathcal{X}$ and $\mathcal{F}$. For example, the classic VC dimension (Vapnik and Chervonenkis, 1974; Vapnik, 1989) satisfies the finite character property since the statement "$\mathrm{VC}(\mathcal{F}) \geqslant d$" can be verified with a finite set of points $X = \{x_1 \ldots, x_d\} \subseteq \mathcal{X}$ and a finite set of classifiers $h_1, \ldots, h_{2^d} \in \mathcal{F}|_X$ that shatter X.

In a similar manner to the statistical setting, a dimension capturing *bandit* learnability can be abstracted as a function $\mathfrak{D}$ that maps a class of reward functions $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ to $\mathbb{N} \cup \{\infty\}$. We say the dimension $\mathfrak{D}$ satisfies the *finite character* property if Definition 3 holds. We say the dimension *characterizes bandit learnability* if for every integer d and $\epsilon, \delta > 0$, there exists integers m, M so that for every $\mathcal{F}$ the following holds: (1) if $\mathfrak{D}(\mathcal{F}) \geqslant d$, then $\mathrm{QC}^0_{\epsilon,\delta}$ is at least m, and (2) if $\mathfrak{D}(\mathcal{F}) < d$, then $\mathrm{QC}^0_{\epsilon,\delta}$ is at most M. The integers m, M tend to ∞ as d tends to ∞.

Perhaps the most well-known example of a combinatorial dimension in the context of bandit learning is the eluder dimension (Russo and Van Roy, 2013). Over more than a decade, it has been a central technique in the context of bandits as well as reinforcement learning (RL) (Li et al., 2022; Wang et al., 2020; Jin et al., 2021). It can be easily verified that the eluder dimension does satisfy the finite character property. However, it is also known there are arbitrarily large gaps between bounds obtained via the eluder dimension and related combinatorial measures (Brukhim et al., 2023).

The following theorem shows that there is no non-trivial dimension that satisfies the finite-character property and also characterizes bandit learnability. Our result holds regardless of the assumed cardinality of the continuum, and within standard ZFC set theory. Our findings complement the celebrated result from (Hanneke and Yang, 2023), that demonstrates a particular reward function class for which bandit learnability (or EMX learnability; Ben-David et al., 2019) depends on the cardinality of the continuum and is therefore independent of the standard set theory ZFC axioms. Our result implies that even when we restrict our attention to classes for which bandit learnability is provable within ZFC, there cannot exist a dimension with the finite-character property characterizing bandit learnability.

Theorem 4 (No finite-character dimension for bandits) *Let $\mathcal{X}, \mathcal{Y}$ be arbitrary (possibly infinite) sets, of size $|\mathcal{Y}| \geqslant |\mathcal{X}| \geqslant d+1$ for some integer $d > 2$. Let $\mathfrak{D}$ be a dimension for bandit classes in $\mathcal{Y}^{\mathcal{X}}$ that satisfies the finite character property, and such that $\exists \mathcal{F}$ with $\mathfrak{D}(\mathcal{F}) \geqslant d$. Then, for any $\epsilon, \delta \geqslant 0$, there exist $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ for which $\mathfrak{D}(\mathcal{F}) \geqslant d$, but the query complexity of bandit-learning $\mathcal{F}$ is bounded by $\mathrm{QC}^0_{\epsilon,\delta}(\mathcal{F}) \leqslant 2$. In particular, $\mathfrak{D}$ does not characterize bandit learnability.*

Proof Consider a class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $\mathfrak{D}(\mathcal{F}) \geqslant d > 2$. By the finite-character assumption, and since $\mathfrak{D}(\mathcal{F}) \geqslant d$, there exists a shattering function V_d, a set $X = \{x_1, ..., x_d\}$ and a set of vectors $F = \{v_1, ..., v_n\} \in \mathcal{F}|_X$ for some integer n, that is, $F \in (\mathcal{Y}^d)^n$, such that $V_d(X, F) = \mathrm{YES}$.

Since $|\mathcal{X}| > d$, there must exist a point $x_0 \in \mathcal{X}$ such that $x_0 \notin X$. We define a new class $\mathcal{F}'$ over the domain $\mathcal{X}$ as follows. For any $f \in \mathcal{F}$, we define $f' \in \mathcal{F}'$ such that:

$$f'(x) = \begin{cases} f(x) & \text{if } x \neq x_0, \\ \arg\max_{x \in \mathcal{X}} f(x) & \text{if } x = x_0. \end{cases}$$

Then, the new class $\mathcal{F}' \subseteq \mathcal{Y}^{\mathcal{X}}$ consists of all functions f' of the above form. We have that $|\mathcal{F}'| \leqslant |\mathcal{F}|$. We now want to show the following two properties hold: (1) the query complexity of $\mathcal{F}'$ is at most 2, and (2) $\mathfrak{D}(\mathcal{F}') \geqslant d$. It suffices to show (1) and (2) to complete the proof.

First, to show (1), notice that any algorithm can first query x_0 and obtain the value $x_1 := \arg\max_{x \in \mathcal{X}} f(x)$. Then, querying x_1 either immediately attains the optimal value of f', or it may hold that $f(x_1) \leqslant x_1$ in which case x_0 is the optimal value, since $f'(x_0) = x_1$. Thus, at most 2 queries are needed to determine the optimal value of any $f' \in \mathcal{F}'$ up to any $\epsilon \geqslant 0$.

Next, to show (2), simply observe that the set F above is contained in $\mathcal{F}'|_X$. Then, since $V_d(X, F) = V_d(X, \mathcal{F}'|_X) =$ YES we get that X is also shattered by hypothesis class $\mathcal{F}'$ with respect to V_d and so $\mathfrak{D}(\mathcal{F}') > d$. ■

Remark 5 ("Reverse" finite character property) *The property in Theorem 3 requires that a* lower *bound on the dimension be demonstrated by finitely many domain points X and members of $\mathcal{F}|_X$. Indeed, as observed by Ben-David et al. (2019), all standard notions of dimensions in statistical and online learning satisfy this property. One may also consider an alternative property which requires that an* upper *bound on the dimension be demonstrated by finitely many domain points X and members of $\mathcal{F}|_X$. However, one can easily show that there cannot exist a dimension satisfying both this property and characterizing bandit learnability, for any infinite class.*

4. Hardness of bandit learning

In this section, we study the computational efficiency of bandit learning in comparison to standard (albeit possibly computationally hard) algorithmic operations often considered in learning theory. A fundamental example is empirical risk minimization (ERM), which can be used to find a hypothesis consistent with the observed data (as is sufficient, for instance, for PAC learnability). In interactive learning settings, estimation algorithms are often used both to select consistent hypotheses and to make predictions (see, e.g., Foster et al., 2023; Brukhim et al., 2023). Given these, one might naturally expect that if a function class supports efficient algorithms for such tasks, it should also be efficiently learnable in the bandit setting.

Quite surprisingly, we prove that this intuition fails. We construct a class of reward functions where the optimal action can be identified with just two queries, yet no polynomial-time algorithm can achieve this, unless $\mathrm{RP} = \mathrm{NP}$. Furthermore, we show that this class does admit efficient algorithms for standard learning tasks, highlighting that in this case the computational hardness arises solely from the nature of the bandit-learning task.

Commonly used algorithmic procedures Below we give 3 definitions of the relevant algorithmic procedures we will consider in the main theorem presented in this section, Theorem 9. Specifically, we formally define a consistency (ERM) algorithm, an online estimation algorithm, and a maximization algorithm, as follows.

Definition 6 (Consistency (ERM) algorithm) *An algorithm* Alg *is a* consistency (ERM) algorithm *for a class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ if for every $f \in \mathcal{F}$ and for every set $S = \{(a_1, f(a_1)), \ldots, (a_m, f(a_m))\}$, where each $a_i \in \mathcal{A}$, when given S as input,* Alg *returns $\hat{f} \in \mathcal{F}$ such that for all $i = 1, \ldots, m$ it holds that $f(x_i) = \hat{f}(x_i)$.*

Definition 7 (Online estimation algorithm) *An* online estimation algorithm *for $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ is an algorithm that at each round $t = 1, \ldots, T$, when given a sequence of past observations $(a_1, f(a_1)), \ldots, (a_{t-1}, f(a_{t-1}))$, for some $f \in \mathcal{F}$, it returns an estimator $\hat{f}_t \in \mathcal{F}$. The algorithm has* decaying estimation error *if there exists $\boldsymbol{EST}(T) \geqslant 0$ growing sublinearly in T, that is, $\boldsymbol{EST}(T) = o(T)$, such that for any sequence $a_1, \ldots, a_T \in \mathcal{A}$, we have*

$$\sum_{t=1}^{T} \left(\hat{f}_t(a_t) - f(a_t)\right)^2 \leqslant \boldsymbol{EST}(T). \tag{1}$$

Definition 8 (Maximizing algorithm) *An algorithm* Alg *for a class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ is a* maximizing algorithm *if for every $f \in \mathcal{F}$ and every $\epsilon > 0$, it returns $\hat{a} \in \mathcal{A}$ such that*

$$f(\hat{a}) \geqslant \sup_{a \in \mathcal{A}} f(a) - \epsilon.$$

A maximizing algorithm for a class $\mathcal{F}$ over a finite action set $\mathcal{A}$ is said to be efficient if each function $f \in \mathcal{F}$ has a concise representation using $O(\text{poly-log}(|A|))$ bits, and the algorithm has running time that is polynomial in the size of the input, i.e., $\text{poly}(\log(|A|), 1/\epsilon)$.

Hardness of bandit learning Recall that the complexity class RP (randomized polynomial time; Gill III, 1974; Valiant and Vazirani, 1985) is the class of decision problems solvable in polynomial time by a probabilistic Turing machine such that: if the answer is "yes", at least $1/2$ of computation paths accept; if the answer is "no", all computation paths reject.

The following theorem demonstrates a reduction from the NP-complete problem of Boolean satisfiability to bandit learning, using a construction of a function class which at the same time allows efficient algorithms for standard learning algorithms. This establishes hardness of bandit learning, under the assumption that $\text{RP} \neq \text{NP}$.

Theorem 9 (Hardness of bandit learning) *For every $n \in \mathbb{N}$, there exists a finite function class $\mathcal{F}_n \subseteq [0,1]^{\mathcal{A}_n}$ over action set $\mathcal{A}_n$ of size $2^{n+1} + 1$, such that for every $\epsilon, \delta \geqslant 0$,*

$$\text{QC}^0_{\epsilon,\delta}(\mathcal{F}_n) \leqslant 2,$$

and such that the following holds. If there exists a bandit learning algorithm for every $\mathcal{F}_n$ with running time that is polynomial in n, then $\text{RP} = \text{NP}$.

Moreover, each class $\mathcal{F}_n$ admits efficient deterministic algorithms as follows:

- *The class $\mathcal{F}_n$ admits a consistency (ERM) algorithm, of runtime $O(n^2)$.*
- *The class $\mathcal{F}_n$ admits an online estimation algorithm, of runtime $O(n^2)$ and $\boldsymbol{EST}(T) = O(1)$.*
- *The class $\mathcal{F}_n$ admits a maximizing algorithm, of runtime $\tilde{O}(n^2)$, for every $\epsilon \geqslant 0$.*

Remark 10 *We remark that although Theorem 9 is stated in the noise-free setting, a similar result can also be proved in the noisy setting. First, it can be shown that Gaussian noise model with sufficiently low variance $\sigma \approx 1/2^n$ is not qualitatively different from the noise-free case.*[1] *In particular,*

1. The proof follows similarly to that of Theorem 16. See Section 5 for further discussion.

for such small values of σ, we obtain $\mathrm{QC}^{\sigma}_{\epsilon,\delta}(\mathcal{F}_n) \leqslant 2$ as well as all other statements from Theorem 9, where the guarantees for efficient algorithms now hold with high probability. More generally, our construction exhibits a trade-off between the optimal query complexity and the variance of the noise model, such that a large-variance noise model can be incorporated while increasing the optimal query complexity. In particular, Theorem 9 could be extended to the noisy setting under Gaussian noise with large, constant *variance (e.g., $\sigma = 1$), but query complexity of order $\mathrm{QC}^{\sigma}_{\epsilon,\delta} = \tilde{O}(n^2)$, for every ϵ, δ. Thus, although the optimal QC is polynomial in n, a similar construction as shown below demonstrates that there is no bandit learning algorithm that runs in polynomial time, unless $\mathrm{RP} = \mathrm{NP}$. See Section 5 for related results and discussion.*

Proof [Proof of Theorem 9] Throughout the proof, we fix $n \in \mathbb{N}$ and simply denote $\mathcal{A}_n, \mathcal{F}_n$ by $\mathcal{A}, \mathcal{F}$, for brevity. We start by defining $\mathcal{A}$ as follows: $\mathcal{A} = \{\star\} \cup \mathcal{A}^{(2)} \cup \mathcal{A}^{(3)}$, such that $\mathcal{A}^{(2)} = \{0,1\}^n$, and $\mathcal{A}^{(3)} = [2^n]$. We will construct a class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ which is best thought of as represented by a query tree of the following structure: $\star$ corresponds to the root node, and actions in $\mathcal{A}^{(2)}$ and $\mathcal{A}^{(3)}$ correspond to nodes of the second and third layer of the tree, respectively. Before defining the class $\mathcal{F}$, we consider the following set:

$$\Phi = \{\text{all 3CNF formulas } \phi \text{ on } n \text{ variables and at most } n^2 \text{ clauses}\}.$$

For every $\phi \in \Phi$ that is satisfiable, we denote by a^*_ϕ the satisfying assignment for ϕ that is minimal according to the natural ordering on $\mathcal{A}^{(2)}$. Define $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ as follows: $\mathcal{F} = \mathcal{F}^{sat} \cup \mathcal{F}^{all}$ where,

$$\mathcal{F}^{all} = \{f_\phi : \forall \phi \in \Phi\}, \qquad \text{and} \qquad \mathcal{F}^{sat} = \{f_{\phi,c} : \forall\ \phi \in \Phi \text{ s.t. } \phi \text{ is satisfiable},\ c \in \mathcal{A}^{(3)}\},$$

and where the functions of the form $f_{\phi,c}$ and f_ϕ are defined as follows:

$$f_{\phi,c}(a) = \begin{cases} \text{encode}(\phi) & \text{if } a = \star \\ \frac{1}{2^{n+1}} \cdot c & \text{if } a = a^*_\phi \in \mathcal{A}^{(2)} \\ 1 & \text{if } a = c \in \mathcal{A}^{(3)} \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad f_\phi(a) = \begin{cases} \text{encode}(\phi) & \text{if } a = \star \\ 0 & \text{otherwise,} \end{cases}$$

where $\text{encode}(\phi)$ encodes the formula by some value in $[\frac{1}{4}, \frac{1}{2}]$. For example, $\text{encode}(\cdot)$ can be implemented as follows. Each literal can first be encoded using $\log(n) + 1$ bits (the variable index plus 1 bit for negation). The full formula requires $O(n^2 \log(n))$ bits, and this binary string could be embedded in $[0, 1/5)$ by writing it after the decimal point. Lastly, this value could then be shifted by $1/4$ so that it lies in the desired range $[\frac{1}{4}, \frac{1}{2}]$. This encoding can be easily decoded by any learner if there is no noise added to the encoded value $f_{\phi,c}(\star)$.

Query complexity 2: Let us argue that the query complexity of this class $\mathcal{F}$ is indeed at most 2. Specifically, we will describe a deterministic algorithm Alg for $\mathcal{F}$ such that for any $f \in \mathcal{F}$ it requires at most 2 queries to recover the optimal action. First, Alg queries $a = \star$ and observes the encoding $\text{encode}(\phi)$, which allows it to recover the formula ϕ. Then, by brute force search over all assignments $a \in \mathcal{A}^{(2)}$, it can obtain a^*_ϕ, if there is one, and if there is none, the optimal action is simply $\star$. If ϕ is satisfiable, Alg will query $a = a^*_\phi$ and if the value is 0, then the optimal action is again $\star$. Otherwise, Alg observes $\frac{1}{2^{n+2}} \cdot c$. Thus, it has recovered the optimal action c for the function f, with only 2 queries. The third query of $a = c$ will then yield the optimal value.

Hardness: We prove hardness for any bandit learning algorithm B for $\mathcal{F}$. Fix $\epsilon = 1/10$, and note that by the construction of the class, any algorithm that finds an ϵ-optimal action, has actually found an optimal one. Let a_B denote the final query submitted by any algorithm B. Assume towards contradiction that there exists an algorithm B such that for every $f \in \mathcal{F}$, by using only $\text{poly}(n)$ runtime, it outputs an action a_B such that:

$$\mathbb{P}_B \left[f(a_B) = \max_{a \in \mathcal{A}} f(a) \right] \geqslant 3/4. \tag{2}$$

We will prove this solves the SAT decision problem in $\text{poly}(n)$ time and in a probabilistic manner, demonstrating that this NP-complete problem is in RP, in contradiction to the assumption that $\text{RP} \neq \text{NP}$. Specifically, we will describe how, given access to B, one can construct an algorithm so that for every $\phi \in \Phi$, if it is satisfiable the algorithm accepts (declares "yes") with probability at least $1/2$, and if not - the algorithm always rejects (declares "no").

Given any formula ϕ, we simulate running the algorithm B by responding exactly as if $f_\phi \in \mathcal{F}^{all}$ would respond. Specifically, for each query made by B we respond as follows: if the query is $\star$, we respond with encode($\star$), and for any other we respond with 0, until either B queries for $a \in \mathcal{A}^{(2)}$ which is a satisfying assignment for ϕ (which we can easily verify efficiently), in which case we halt the simulation, or until B terminates and returns its final query a_B.

We will now show that our simulation can solve the SAT decision problem with a one-sided error with constant probability, as detailed next. First, assume ϕ is not satisfiable. Then, B can never query for $a \in \mathcal{A}^{(2)}$ which is a satisfying assignment for ϕ, and so we would run the simulation until B terminates, in $\text{poly}(n)$ time, after which we declare that ϕ is not satisfiable. This always occurs, thus whenever ϕ is not satisfiable then with probability 1 we reject.

Now, assume ϕ is satisfiable. We have the following lemma, whose proof is deferred to the appendix.

Lemma 11 *Let $\mathcal{F}$ as constructed above, and let B be any bandit learner for $\mathcal{F}$ as above (i.e., for every $f \in \mathcal{F}$ its output satisfies Equation (2)). Fix any $\phi \in \Phi$ that is satisfiable. Then, there exists $c \in \mathcal{A}^{(3)}$ such that if B is being run with $f_{\phi,c}$ and $a_1, \ldots, a_m$ denotes its query sequence during that run, it holds that:*

$$\mathbb{P}_B \left[\exists i \in [m], \quad a_i \textit{ is a satisfying assignment for } \phi \ \wedge \ \forall j < i,\ a_j \neq c \right] \geqslant \frac{3}{4} - \frac{2m}{2^n}.$$

Then, by Lemma 11 we have that there exists some $c \in \mathcal{A}^{(3)}$ such that when B interacting with $f_{\phi,c}$, and $a_1, \ldots, a_m$ denotes its query sequence during that run, then :

$$\mathbb{P}_B \left[\exists i \in [m], \quad a_i \text{ is a satisfying assignment for } \phi \ \wedge \ \forall j \leqslant i,\ a_j \neq c \right] \geqslant \frac{1}{2},$$

since m is some polynomial in n, then for all sufficiently large n we have $\frac{2m}{2^n} \leqslant 1/10$. Importantly, we do not need to know what this c is during simulation. The reason is that, by the above, we have that with probability at least $1/2$ we will be able to simulate a response sequence by $f_{\phi,c}$ since it will be identical to the response sequence by f_ϕ, until we observe a satisfying assignment, in which case we halt. Thus, with probability at least $1/2$ we will observe a satisfying assignment, and declare "yes". Notice that with probability $< 1/2$ our simulation will not be consistent with $f_{\phi,c}$ but that is

of no concern to us, as we may reject in this case. It is, however, crucial that B runs in poly(n) time even when interacting with f_ϕ rather than with $f_{\phi,c}$, which indeed holds as $f_\phi \in \mathcal{F}$.

The proof of the theorem is then concluded by proving the existence of efficient algorithms for the class $\mathcal{F}$, which holds by Lemma 18, given in the appendix. ■

5. Noise-free vs. noisy setting query complexity

The first question we address is whether there exists any provable relationship between the noise-free query complexity $\mathrm{QC}^0_{\epsilon,\delta}(\mathcal{F})$ and $\mathrm{QC}^\sigma_{\epsilon,\delta}(\mathcal{F})$. First we show there exist function classes for which their noise-free query complexity is constant but such that (ϵ,δ)-complexity is unbounded.

Proposition 12 *Given $\epsilon \in [0,1/2)$, there exists a function class $\mathcal{F}$ such that $\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F}) = 1$ for all $\delta' \in [0,1)$ but $\mathrm{QC}^\sigma_{\epsilon,\delta}(\mathcal{F}) = \infty$ for all $\delta \in [0,1/2)$ and all $\sigma > 0$.*

The function class $\mathcal{F}$ used to prove Proposition 12 is based on a "informative action" construction where the action space equals $\{0\} \cup \mathbb{N}$. The optimal action of any function in $\mathcal{F}$ is indexed by $n \in \mathbb{N}$. Action 0 is "informative" because its mean reward reveals the identity of the optimal action so that $\mathrm{QC}_{\epsilon,\delta'}(\mathcal{F}) = 1$. Nonetheless in the noisy setting when n goes to infinity, estimating the reward of action 0 or finding the optimal action via enumeration of $\mathbb{N}$ requires a number of queries growing with n. The formal proof is given in Appendix B.1.

Upper and lower bound bounds for $\mathrm{QC}^1_{\epsilon,\delta}(\mathcal{F})$ were derived by Hanneke and Wang (2024) for the high-noise regime (i.e., when σ is of order 1 for functions with values in $[0,1]$) based on the generalized maximin volume $\gamma_{\mathcal{F},\epsilon}$ of $\mathcal{F}$ (see definition below).

Definition 13 (Generalized maximin volume; Hanneke and Wang, 2024) Generalized maximin volume *of a function class $\mathcal{F}$ is defined as*

$$\gamma_{\mathcal{F},\epsilon} = \sup_{p \in \Delta(\mathcal{A})} \inf_{f \in \mathcal{F}} \mathbb{P}_{a \sim p}\left(\sup_{a^*} f(a^*) - f(a) \leqslant \epsilon\right), \tag{3}$$

where $\Delta(\mathcal{A})$ is the set of all distributions on $\mathcal{A}$.

Theorem 1 of Hanneke and Wang (2024) presents an elegant and insightful result, establishing that $\mathrm{QC}^1_{\epsilon,\delta}(\mathcal{F})$ can be lower bounded by $\Omega\left(\log(1/\gamma_{\mathcal{F},\epsilon})\right)$ and upper bounded (up to constant and logarithmic factors) by $1/(\gamma_{\mathcal{F},\epsilon/2} \cdot \epsilon^2)$. In this work we explore the low-noise regime where these results break down. In Theorem 14 we show among other things that for any $K \in \mathbb{N}$ and $\epsilon \in [0,1/2)$ there is a function class $\mathcal{F}$ such that $\gamma_{\mathcal{F},\epsilon} = K$ but there exist values of $\sigma > 0$ where $\mathrm{QC}^\sigma_{\epsilon,1/4}(\mathcal{F}) = 1 < \log(1/\gamma_{\mathcal{F},\epsilon})$. This result shows $\mathrm{QC}^\sigma_{\mathcal{F},\epsilon}(\mathcal{F})$ behaves fundamentally differently in the low-noise and high-noise regimes, highlighting the need for better theories to understand this phase transition.

Theorem 14 *There exist universal constants $c, \bar{c} > 0$ such that for every integer $K \geqslant 2$ there exists a function class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ with action space $|\mathcal{A}| = K+1$ such that for every $\epsilon \in [0,1/2)$ it holds that $\gamma_{\mathcal{F},\epsilon} = 1/K$ and if $\sigma^2 \geqslant \frac{1}{\bar{c}K^{2/3}}$ then*

$$\bar{c}K^{2/3}\sigma^2 \leqslant \mathrm{QC}^\sigma_{\epsilon,1/4}(\mathcal{F}) \leqslant c\log^{2/3}(K)K^{2/3}\sigma^2.$$

In particular,

$$\bar{c}\log(1/\gamma_{\mathcal{F},\epsilon})\sigma^2 \leqslant \mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}),$$

and if $\sigma^2 \leqslant \frac{1}{c\log^{2/3}(K)K^{2/3}}$ *then*

$$\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) = \mathrm{QC}^{0}_{\epsilon,0}(\mathcal{F}) = 1.$$

Similar to Proposition 12, the function class $\mathcal{F}$ in Theorem 14 is constructed around an "informative action" structure, where the action space is given by $\{0\} \cup [K]$. The optimal action belongs to $[K]$, while the mean reward of action 0 (the informative action) reveals its identity. This construction differs from Proposition 12 in its encoding representation. Specifically, our design ensures that strategies leveraging the information encoded in the mean reward of action 0 achieve greater efficiency compared to the $\mathcal{O}(\sigma^2 K)$ queries needed by a strategy that individually estimates the mean rewards of all actions in $[K]$. Interestingly, in the large-variance regime the optimal data collection strategy that achieves the lower bound rate works by querying action 0 sufficiently to narrow down the optimal action choices to $\mathcal{O}(K^{2/3})$ actions. When the noise variance is sufficiently small, the mean reward encoded by action 0 can be inferred from a noisy sample with probability of error at most $1/4$, leading to a query complexity of just 1. These results highlight the intricate balance between exploiting the information structure of the function class—encoded here by action 0—and relying on brute-force exploration by following the policy dictated by the generalized maximin volume in Equation 3. The formal proof is given in Appendix B.3.

Building on these results we introduce the (ϵ,δ)-gap of a function class $\mathcal{F}$, denoted $\mathrm{Gap}_{\epsilon,\delta}(\mathcal{F})$, which we then use to derive sufficient conditions on σ to guarantee $\mathrm{QC}^{\sigma}_{\epsilon,\delta}(\mathcal{F}) \lesssim \mathrm{QC}^{0}_{\epsilon,\delta}(\mathcal{F})$.

Definition 15 (Informal: Gap of $\mathcal{F}$) *Let $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ be a finite function class and action space. Given $\epsilon,\delta \in [0,1]$, we define $\mathrm{Gap}_{\epsilon,\delta}(\mathcal{F})$ as the smallest difference between achievable function values for an action that an (ϵ,δ)-optimal algorithm might play, given any positive probability history.*

Definition 21 in Appendix B.1 formalizes the description above. Our next result establishes that when σ is small, the (ϵ,δ)-query complexity of $\mathcal{F}$ is not too different from that of the noise-free (ϵ,δ')-query complexity of $\mathcal{F}$ provided that $\delta' < \delta$.

Theorem 16 *Let $\delta,\delta' \in (0,1)$ such that $\delta > \delta' \geqslant 0$. For any finite class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ over a finite action space the noisy query complexity with zero-mean Gaussian noise with variance σ^2 such that $\sigma^2 < \frac{\mathrm{Gap}^2_{\epsilon,\delta'}(\mathcal{F})}{4\log(2\mathrm{QC}^{0}_{\epsilon,\delta'}(\mathcal{F})/(\delta-\delta'))}$ satisfies:*

$$\mathrm{QC}^{\sigma}_{\epsilon,\delta}(\mathcal{F}) \leqslant \mathrm{QC}^{0}_{\epsilon,\delta'}(\mathcal{F})$$

To prove theorem 16 we show that we can construct a noisy feedback algorithm Alg based on an (ϵ,δ')-optimal noise-free algorithm Alg' that is guaranteed to have an error probability of at most δ. Alg uses nearest neighbors to transform noisy rewards into mean reward values. When the noise is small Alg recovers the correct mean rewards with an error probability at most $\delta - \delta'$. These rewards are fed into a copy of Alg' and the suggested action exploration policies are executed. The resulting algorithm achieves the same query complexity as Alg' with a slightly degraded error upper bound of δ. The inequality $\delta' < \delta$ is required because for any level of non-zero gaussian noise ($\sigma > 0$) translating noisy rewards into their mean reward values will necessarily produce an irreducible error probability. The proof of Theorem 16 is given in Appendix B.1.

6. Separation between regret and query complexity

We study the separation between regret and query complexity, both in the noise-free and noisy settings. The regret of an algorithm Alg with action space $\mathcal{A}$, interacting for T rounds by producing an action $a_t \in \mathcal{A}$ for $t = 1, \ldots, T$ and observing rewards generated by f^* is defined as,

$$\mathrm{Regret}_{\mathrm{Alg}}(T) = \sum_{t=1}^{T} \max_{a \in \mathcal{A}} f^*(a) - f^*(a_t).$$

A typical objective in the bandit online learning and reinforcement learning literature is to design algorithms that satisfy a sublinear regret bound such that $\lim_{T\to\infty} \frac{\mathrm{Regret}(T)}{T} = 0$. In this section, we explore whether achieving low query complexity and low regret are compatible objectives. We show negative results in this regard in the noise-free (Appendix C.1) and noisy settings (Section 6.1). In each of these scenarios, we show that it is impossible to construct algorithms that achieve optimal query complexity while also incurring sublinear regret. This holds because, in certain problems, any optimal algorithm for ϵ-arm identification must allocate a significant number of queries to actions that, while highly informative, result in substantial regret.

6.1. Regret vs. QC: noisy case

In this section we explore the compatibility of the optimal query complexity and regret minimization in noisy feedback problems. Similar to our results in the noise-free setting in Theorem 17 we show there are problems where the goal of finding an optimal action cannot be achieved without paying a regret scaling linearly with the query complexity; however, for the same function classes, there is an algorithm achieving regret scaling as the square root of the number of time-steps.

Theorem 17 *Let $d, T \in \mathbb{N}$. There exists a function class $\mathcal{F}$ over action space $\mathcal{A}$ with unit-variance Gaussian noise such that $d \leqslant \mathrm{QC}^1_{0,1/4}(\mathcal{F}) \leqslant 80d$ and any algorithm Alg such that $m^1_{\mathrm{Alg}}(0, 1/4) \leqslant T$ satisfies*

$$\max_{f \in \mathcal{F}} \mathbb{E}_{\mathrm{Alg}}[\mathrm{Regret}(T, f)] \geqslant \frac{d}{128}.$$

Moreover, there is an algorithm Alg' that satisfies $\max_{f\in\mathcal{F}} \mathbb{E}_{\mathrm{Alg}'}[\mathrm{Regret}(T, f)] \leqslant 8\sqrt{2T\log(T)}$ for all $T \in \mathbb{N}$.

Theorem 17 suggests there exists a function class with query complexity $\mathcal{O}(d)$ such that when $T = \mathcal{O}(d^\alpha)$ for $\alpha < 2$ then no algorithm Alg that is able to find an optimal action in T queries can also satisfy an $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound. Nonetheless, for the same function class, there are algorithms that achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret bounds. Theorem 17 is closely related to Theorem 1 of Bubeck et al. (2011). While Theorem 1 of Bubeck et al. (2011) rules out the existence of algorithms that achieve optimal regret and simple regret simultaneously, Theorem 17 establishes that no algorithm can achieve both optimal query complexity and optimal regret. Although related, the notions of simple regret and query complexity are different. The *simple regret* for a fixed horizon T is the expected gap between the algorithm's output arm a_T and the optimal arm a^*. In contrast, *query complexity* can be thought of as the minimum horizon T such that the optimal simple regret is at most ϵ.

The function class $\mathcal{F}$ used to prove Theorem 17 has an "information lock" structure. The action space is divided into two sets $\mathcal{A}_1$ and $\mathcal{A}_2$. The values of the mean rewards of actions in $\mathcal{A}_1$ can be used to infer the identity of the mean optimal action. Actions in $\mathcal{A}_1$ have large regret and their mean rewards are equal to $1/2 + \epsilon_1$ or $1/2 - \epsilon_1$ while the mean rewards of actions in $\mathcal{A}_2$ are equal to 1 or $1 - \epsilon_2$ for parameters $\epsilon_1, \epsilon_2 \in [0, 1]$ such that $\epsilon_1 \geqslant \epsilon_2$.

To prove Theorem 17 we first establish that $\mathrm{QC}^1_{0,1/4}(\mathcal{F}) = \Theta(1/\epsilon_1^2)$. Second, we show that when $\epsilon_2 \approx \epsilon_1^2$, then any algorithm Alg such that $m^1_{\mathrm{Alg}}(0, 1/4) \leqslant T$ must also incur regret satisfying $\max_{f\in\mathcal{F}} \mathbb{E}[\mathrm{Regret}(T, f)] \geqslant \Omega(1/\epsilon_1^2)$. The proof of Theorem 17 follows by setting $\epsilon_1 \approx 1/\sqrt{d}$. Finally, since the problem in this class is an instance of multi-armed bandits, the UCB algorithm is guaranteed to collect sublinear regret. The formal proof of Theorem 17 can be found in Appendix C.2. In Appendix C.1, we establish analogous results for the noise-free setting. Notably, these findings do *not* follow directly from Theorem 17. While Theorem 17 is stated for $\sigma = 1$, the query complexity in this construction approaches 1 as σ tends to zero, preventing a straightforward extension to the noise-free case.

7. Conclusion

In this work, we have presented new insights into the study of the learnability of structured bandit problems, shedding light on the interaction between their statistical and computational properties. Our main results highlight fundamental distinctions between classical learnability as studied in statistical learning theory and learnability in the bandit setting.

We show that there cannot exist a combinatorial finite-character dimension that fully characterizes bandit learnability, a result that sets apart this setting from standard PAC learnability. We also prove there cannot exist query optimal algorithms that are computationally tractable even with access to standard algorithmic primitives such as empirical risk minimization and function maximization oracles.

We also investigated the effects of observation noise on the query complexity of bandit problems. We show that there are function classes where a small amount of observation noise leaves query complexity unaffected, alongside classes where any amount of noise makes bandit learnability impossible. Finally, we demonstrate that there is a sharp distinction between algorithms adapted for query complexity and regret minimization: There are no algorithms that can simultaneously minimize these two objectives.

We hope that by documenting these phenomena, we can help advance the community's understanding of bandit learnability and guide future algorithmic design and theoretical advancements in this area.

Acknowledgments

A.P. thanks Alessio Russo for helpful discussions.

References

Kareem Amin, Michael Kearns, and Umar Syed. Bandits, query learning, and the haystack dimension. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 87–106. JMLR Workshop and Conference Proceedings, 2011.

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, Nicolo Cesa-Bianchi, and Philip M Long. Characterizations of learnability for classes of {O,. . . , n}-valued functions. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 333–340, 1992.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 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.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *Annual Symposium on Foundations of Computer Science, FOCS*, 2022.

Nataly Brukhim, Miro Dudik, Aldo Pacchiano, and Robert E Schapire. A unified model and dimension for interactive estimation. *Advances in Neural Information Processing Systems*, 36: 64589–64617, 2023.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

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

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

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

John T Gill III. Computational complexity of probabilistic turing machines. In *Proceedings of the sixth annual ACM symposium on Theory of computing*, pages 91–95, 1974.

Steve Hanneke and Kun Wang. A complete characterization of learnability for stochastic noisy bandits. *arXiv preprint arXiv:2410.09597*, 2024.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

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

Gene Li, Pritish Kamath, Dylan J Foster, and Nati Srebro. Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022.

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

Balas K. Natarajan and Prasad Tadepalli. Two new frameworks for learning. In *ICML*, pages 402–415, 1988.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015.

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

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Leslie G Valiant and Vijay V Vazirani. Np is as easy as detecting unique solutions. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 458–463, 1985.

Vladimir Vapnik. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). In *COLT*, pages 3–21, 1989.

Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

Appendix A. Missing proof of Section 4

Lemma 18 *Let $\mathcal{F}$ as constructed above. Then, the following holds:*

- *The class $\mathcal{F}_n$ admits a maximizing algorithm, of runtime $\tilde{O}(n^2)$ for every $\epsilon \geqslant 0$.*
- *The class $\mathcal{F}_n$ admits a consistency (ERM) algorithm, of runtime $O(n^2)$.*
- *The class $\mathcal{F}_n$ admits an online estimation algorithm, of runtime $O(n^2)$ and $\boldsymbol{EST}(T) = O(1)$.*

Proof [Proof of Lemma 18] Consider the class $\mathcal{F}$ as constructed in the proof of Theorem 9.

Consistency (ERM) algorithm: The class $\mathcal{F}$ admits an efficient consistency algorithm, as follows. Let $S = \{(a_1, r_1), ..., (a_k, r_k)\}$ denote a sequence of $k \geqslant 0$ queries and their corresponding values, consistent with some $f \in \mathcal{F}$, such that S is given as input to the consistency algorithm. There are 3 types of actions for which the feedback could be non-zero, and so there are $2^3 = 8$ total possibilities to consider. For the trivial case of $r_i = 0$ for all $i \in [k]$, then any $f \in \mathcal{F}^{all}$ is consistent.

Then, if the only action a_i for which $r_i \neq 0$ is $a_i = \star$, we can simply return f_ϕ for ϕ recovered from $r_i = \text{encode}(\star)$. If S contains both $a_i = \star$, and $a_j \in \mathcal{A}^{(2)}$ such that $r_j \neq 0$, then the true function can be fully recovered from S (since $r_j = c/2^{n+1}$), and so the algorithm can return $f_{\phi,c}$. Similarly, if S contains both $a_i = \star$, and $a_j \in \mathcal{A}^{(3)}$ such that $r_j \neq 0$, then the true function can again be fully recovered from S (since $a_j = c$).

Next, if $\star$ was not queried, but S indicates that some a_i is the minimal satisfying assignment, then we can construct a formula ϕ so that $\phi(x)$ is true if and only if $x = a_i$, so that a_i is ϕ's only satisfying assignment. Constructing such a boolean formula is straightforward: for each clause $j = 1, ...n$ we take a disjunction over the literal b_j such that the assignment $a_i(j)$ is true, repeated 3 times. Then, we let ϕ be the conjunction of those n clauses.

If the only non-zero observed value is $f(a) = 1$ then we can take any $\phi \in \Phi$ and $c = a$, and return $f_{\phi,c}$. If both $f(a) = 1$ and $f(a') \neq 0$ have been observed, then we can repeat the construction of ϕ as above, for which a' the only satisfying assignment, and set $c = a$. Thus, we have described an efficient consistency (ERM) algorithm for the class $\mathcal{F}$.

Online estimation algorithm: The class $\mathcal{F}$ admits an online estimation algorithm, as follows. First, by running a consistency procedure as described above, then for any sequence of past observations we get $\hat{f} \in \mathcal{F}$ consistent with it. Thus, in every round $t = 1, ..., T$, we obtain a function $\hat{f}_t \in \mathcal{F}$ consistent with all observations up to time $t - 1$. Then, since for any ground truth function $f \in \mathcal{F}$ used to generate the data sequence there are at most 3 types of actions for which the feedback could be non-zero, and since all values are bounded in $[0, 1]$, the predicting error of the algorithm is at most 3. Thus, for any integer T, $\textbf{EST}(T) \leqslant 3$.

Maximization algorithm: Lastly, notice that $f_{\phi,c}$ has a concise representation using at most $O(n^2 \cdot \log(n))$ bits, used to represent both the formula ϕ and the value c. Thus, the class $\mathcal{F}$ admits an efficient maximizing algorithm - when given its concise representation once can easily recover the optimal action c by reading it from the function description.

■

A.1. Proof of Theorem 11

Since ϕ is satisfiable, we know that the optimal query for any $f_{\phi,c}$ is c. We also denote by a_B^c the final output of B after having interacted with $f_{\phi,c}$, for some fixed c. Denote the event that "$\exists i \in [m],$ s.t. a_i is a satisfying assignment for ϕ" by B-good and its converse by B-bad. Then, for any fixed c, if B is being run with some $f_{\phi,c}$ it holds that,

$$\begin{aligned}\mathbb{P}_B\left[f(a_B) = \max_{a\in\mathcal{A}} f(a)\right] &= \mathbb{P}_B\left[a_B^c = c\right] \\ &= \mathbb{P}_B\left[a_B^c = c \ \wedge\ B\text{-good}\right] + \mathbb{P}_B\left[a_B^c = c \ \wedge\ B\text{-bad}\right] \\ &\leqslant \mathbb{P}_B\left[B\text{-good}\right] + \mathbb{P}_B\left[\exists i \in [m],\ a_i = c \ \wedge\ B\text{-bad}\right].\end{aligned} \tag{4}$$

Next, we consider the uniform distribution U over all $c \in \mathcal{A}^{(3)}$. We use $R \sim \mathcal{R}$ to denote the internal randomization used by the algorithm B. We may then view B's computation of its final output a_B as a fixed and deterministic function of R, and of the observed values from $f_{\phi,c}$. Then,

$$\begin{aligned}&\mathbb{E}_{c\sim U}\Big[\ \mathbb{P}_B\left[\exists i \in [m],\ a_i = c \ \wedge\ B\text{-bad}\right]\Big] \\ &\quad= \mathbb{E}_{R\sim\mathcal{R}}\left[\mathbb{P}_{c\sim U}\left[\exists i \in [m],\ a_i = c \ \wedge\ B\text{-bad} \mid R\right]\right] \\ &\quad= \mathbb{E}_{R\sim\mathcal{R}}\left[\mathbb{P}_{c\sim U}\left[\exists i \in [m],\ a_i = c \mid R, B\text{-bad}\right]\cdot \mathbb{P}_{c\sim U}[B\text{-bad}|R]\right] \\ &\quad\leqslant \frac{m}{2^n},\end{aligned}$$

where the equalities follow from linearity of expectation and law of total expectation, and the inequality follows from the following fact; When R is fixed (so B is viewed as a deterministic function) and conditioned on the event that B is B-bad for c, then notice that during the interaction of B with $f_{\phi,c}$ it only observed 0 for all of its queries (except $\star$). Thus, a_B is only a deterministic function of ϕ and a sequence of queries $(a_i, 0)$, unless any query happened to "hit" c. Since c is chosen at random independently of ϕ, the probability of any deterministic choice out of all elements in $\mathcal{A}^{(3)}$ will happen to be c is $1/2^n$, and for m such choices to probability is at most $m/2^n$. Then, this implies that there exists a *particular* $c \in \mathcal{A}^{(3)}$ for which,

$$\mathbb{P}_B\left[\exists i \in [m],\ a_i = c \ \wedge\ B\text{-bad}\right] \leqslant \frac{m}{2^n}. \tag{5}$$

Next, combining this inequality with Equation (2) and Equation (4) we get,

$$\mathbb{P}_B\left[B\text{-good}\right] \geqslant \frac{3}{4} - \frac{m}{2^n}. \tag{6}$$

Observe that the same reasoning used for the proof of Equation (5) also holds for any $m' \leqslant m$ (where B-bad is modified to only consider queries up to m' as well). Therefore, for all $m' \leqslant m$,

$$\mathbb{P}_B\left[\exists i \in [m'],\ a_i = c \ \wedge\ \forall i \in [m'],\ a_i \text{ is not a satisfying assignment for } \phi\right] \leqslant \frac{m'}{2^n}. \tag{7}$$

Then, combining Equation (6) and Equation (7) we get:

$$\begin{aligned}\frac{3}{4} - \frac{m}{2^n} &\leqslant \mathbb{P}[\text{B-good}] = \mathbb{P}[\exists i \in [m], a_i \text{ satisfying } \phi] \\ &= \mathbb{P}[\exists i \in [m], a_i \text{ satisfying } \phi \wedge \ \forall a_i \text{ satisfying } \phi, \exists\ j < i, a_j = c] \qquad (c \text{ queried before } a_i) \\ &\quad + \mathbb{P}\left[\exists i \in [m], \quad a_i \text{ satisfying } \phi \ \ \wedge \ \ \forall j < i,\ a_j \neq c\right] \qquad (c \text{ not queried before } a_i) \\ &\leqslant \frac{m}{2^n} + \mathbb{P}\left[\exists i \in [m], \quad a_i \text{ satisfying } \phi \ \ \wedge \ \ \forall j < i,\ a_j \neq c\right].\end{aligned}$$

Re-arranging yields the desired claim.

Appendix B. Missing proofs of Section 5

B.1. Proof of Proposition 12

Proposition 12 *Given $\epsilon \in [0, 1/2)$, there exists a function class $\mathcal{F}$ such that $\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F}) = 1$ for all $\delta' \in [0, 1)$ but $\mathrm{QC}^\sigma_{\epsilon,\delta}(\mathcal{F}) = \infty$ for all $\delta \in [0, 1/2)$ and all $\sigma > 0$.*

Let $\mathcal{A} = \{0\} \cup \mathbb{N}$. Consider a function class indexed by $i \in \mathbb{N}$ such that,

$$f_i(a) = \begin{cases} \frac{1}{2i} & \text{if } a = 0 \\ 1 & \text{if } a = i \\ 0 & \text{o.w.} \end{cases}$$

It is clear that $\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F}) = 1$ since the query action 0 produces an observation $\frac{1}{2i}$ containing enough information to identify the optimal action.

We'll consider a companion empty problem,

$$f_0(a) = 0 \ \forall a \in \mathcal{A}$$

with zero mean unit Gaussian noise. Let's consider a (possibly randomized) algorithm Alg and its interaction with f_0 and f_j over n rounds. We will show that Alg cannot succeed at identifying an optimal arm with probability at least $1 - \delta$ after n steps for all $f \in \mathcal{F}$. We prove this by way of contradiction. Assume $m^\sigma_{\mathrm{Alg}}(\epsilon, \delta) \ \leqslant \ n$ for a finite natural number $n \in \mathbb{N}$.

Let $k(n)$ be a random variable specifying Alg's guess for the optimal action at time n. We allow for Alg to guess a distribution over candidate optimal actions at time n. In this case $k(n)$ captures the realization of a sample from this distribution. For any $j \in \mathbb{N}$ define the event $\mathcal{E}_j$ as,

$$\mathcal{E}_j = \{k(n) = j\}$$

When $m^\sigma_{\mathrm{Alg}}(\epsilon, \delta) \ \leqslant \ n$ it follows that for all $i \in \mathbb{N}$,

$$\mathbb{P}_{\mathrm{Alg}, f_i}(\mathcal{E}_i) \ \geqslant \ 1 - \delta. \tag{8}$$

The divergence decomposition of bandit problems (Lemma 30) implies,

$$\mathrm{KL}(\mathbb{P}_{\mathrm{Alg}, f_0} \parallel \mathbb{P}_{\mathrm{Alg}, f_j}) = \mathbb{E}_{\mathrm{Alg}, f_0}[T_0(n)] \left(\frac{1}{2j}\right)^2 + \mathbb{E}_{\mathrm{Alg}, f_0}[T_j(n)]$$

where $T_m(n)$ equals the random variable equal to the number of times algorithm Alg tried action m up to time n. Let $i \in \mathbb{N}$ be such that $i \geqslant 16$ and $i \geqslant 8n$. Order the indices $[i, 2i-1]$ as $\mathrm{I}_1, \cdots, \mathrm{I}_i$ such that $\mathbb{E}_{\mathrm{Alg},f_0}[T_{\mathrm{I}_1}(n)] \leqslant \cdots \leqslant \mathbb{E}_{\mathrm{Alg},f_0}[T_{\mathrm{I}_i}(n)]$. It follows that for all $j \in [1, \cdots, i/2]$,

$$\mathbb{E}_{\mathrm{Alg},f_0}[T_{\mathrm{I}_j}(n)] \leqslant 2n/i$$

This is because if this was not the case, then we would have $\sum_{j=i/2+1}^{i} \mathbb{E}_{\mathrm{Alg},f_0}[T_{\mathrm{I}_j}(n)] > i/2 * 2n/i = n$, a contradiction.

Now observe that $\mathbb{E}_{\mathrm{Alg},f_0}[T_0(n)] \leqslant n$, and by asssumption $n \leqslant i/8$. Therefore for all $j \in [1, \cdots, i/2]$:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\mathrm{Alg},f_0} \parallel \mathbb{P}_{\mathrm{Alg},f_{\mathrm{I}_j}}) &= \mathbb{E}_{\mathrm{Alg},f_0}[T_0(n)] \left(\frac{1}{2I_j}\right)^2 + \mathbb{E}_{\mathrm{Alg},f_0}[T_j(n)] \\
&\overset{(i)}{\leqslant} \frac{n}{4i^2} + \frac{1}{4} \\
&\leqslant 1/4 + 1/32 \\
&= 9/32.
\end{aligned}$$

Where inequality (i) follows because $I_j \in [i, \cdots, 2i-1]$. Recall that $k(n)$ is a random variable specifying Alg's guess for the optimal action at time n and $\mathcal{E}_i$ is the event that Alg guessed action $k(n) = i$ as optimal at time n. Let $\hat{i}$ denote the action such that,

$$\hat{j} = \underset{j \in [1,\cdots,i/2]}{\arg\min}\ \mathbb{P}_{\mathrm{Alg},f_0}\left(\mathcal{E}_{\mathrm{I}_j}\right)$$

Since $\sum_{j=1}^{\infty} \mathbb{P}_{\mathrm{Alg},f_0}(\mathcal{E}_j) = 1$ it follows that $\sum_{j=1}^{i/2} \mathbb{P}_{\mathrm{Alg},f_0}\left(\mathcal{E}_{\mathrm{I}_j}\right) \leqslant 1$ and therefore,

$$\mathbb{P}_{\mathrm{Alg},f_0}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right) \leqslant 2/i$$

Pinsker's inequality implies,

$$2\left(\mathbb{P}_{\mathrm{Alg},f_0}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right) - \mathbb{P}_{\mathrm{Alg},f_{\mathrm{I}_{\hat{j}}}}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right)\right)^2 \leqslant \mathrm{KL}(\mathbb{P}_{\mathrm{Alg},f_0} \parallel \mathbb{P}_{\mathrm{Alg},f_{\mathrm{I}_{\hat{j}}}}) \leqslant 9/32.$$

this in turn implies,

$$\left|\mathbb{P}_{\mathrm{Alg},f_0}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right) - \mathbb{P}_{\mathrm{Alg},f_{\mathrm{I}_{\hat{j}}}}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right)\right| \leqslant 3/8.$$

Finally, since $\mathbb{P}_{\mathrm{Alg},f_0}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right) \leqslant 2/i$ we conclude that,

$$\mathbb{P}_{\mathrm{Alg},f_{\mathrm{I}_{\hat{j}}}}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right) \leqslant \frac{2}{i} + \left|\mathbb{P}_{\mathrm{Alg},f_0}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right) - \mathbb{P}_{\mathrm{Alg},f_{\mathrm{I}_{\hat{j}}}}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right)\right| \leqslant 3/8 + \frac{2}{i}.$$

Since $i > 16$ it follows that,

$$\mathbb{P}_{\mathrm{Alg},f_{\mathrm{I}_{\hat{j}}}}\left(\mathcal{E}_{\mathrm{I}_{\hat{j}}}\right) < 1/2$$

thus contradicting Equation 8 when $\delta < 1/2$.

B.2. Proof of Theorem 14

In the proof of Theorem 14 we define a family of function classes indexed by $K \in \mathbb{N}$. We split the proof of our main result by first showing in Lemma 19 an upper bound on the query complexity of these function classes, and second proving a matching lower bound in Lemma 20.

Lemma 19 *There exists a universal constant $c > 0$ such that for any $K \in \mathbb{N}$ satisfying $K \geqslant 2$ the function class $\mathcal{F}$ with action space $\mathcal{A} = \{a_i\}_{i=0}^K$ defined as $\mathcal{F} = \{f_i\}_{i=1}^K$ with,*

$$f_i(a) = \begin{cases} \frac{i}{4K} & \text{if } a = a_0 \\ 1/2 & \text{if } a \neq i \\ 1 & \text{o.w.} \end{cases}$$

satisfies that for any $\epsilon \in [0, 1/2)$, if $\sigma^2 > \frac{1}{c \log^{2/3}(K) K^{2/3}}$,

$$\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) \leqslant c \log^{2/3}(K) K^{2/3} \sigma^2,$$

and if $\sigma^2 \leqslant \frac{1}{c \log^{2/3}(K) K^{2/3}}$ then,

$$\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) = \mathrm{QC}^{0}_{\epsilon,0}(\mathcal{F}) = 1.$$

Proof

Upper bound To prove the desired bound we consider the following algorithm,

1. Estimate the mean reward of action a_0 up to $0 < \alpha \leqslant 1$ accuracy. We will be choosing an appropriate value of α at the end.
2. Use the estimator from 1) to identify a candidate set of $2\alpha K$ functions that agree with the observed data.

Lemma 27 implies step 1. can be achieved with an error probability of at most δ_0 by constructing an empirical mean estimator $\hat{\mu}_0$ of the mean reward of action a_0 using at most $\frac{2\sigma^2 \log(2/\delta_0)}{\alpha^2}$ samples.

In step 2 we define a candidate set of models $\mathcal{I} = \{i \text{ s.t. } |\hat{\mu}_0 - f_i(a_0)| \leqslant \alpha\}$ that agree with the estimator $\hat{\mu}_0$. The size of $\tilde{I}$ satisfies $|\tilde{I}| \leqslant 2\alpha K$.

We will analyze two cases:

1. $\alpha < 1/K$ so that $|\mathcal{I}| \leqslant 2\alpha K < 2$.
2. $\alpha \geqslant 1/K$.

Case 1: when $\alpha < 1/K$ and $2\alpha K < 2$ the set $\mathcal{I}$ satisfies $|\mathcal{I}| = 1$ and therefore with probability at least $1 - \delta_0$ the algorithm can read the identify of the optimal action. In this case it is sufficient to set $\delta_0 = 1/4$ and $\alpha = \frac{1}{2K}$ and the number of queries is upper bounded by $\max(8 \log(8) \sigma^2 K^2, 1)$.

Case 2: when $\alpha \geqslant 1/K$ the algorithm then estimates up to $1/4$-accuracy each action in $\mathcal{I}$ ensuring an error probability per action of at most δ. Lemma 27 implies this can be done by using $32\sigma^2 \log(2/\delta)$ samples for each of the actions in $\mathcal{I}$.

The union bound implies that in this case the action with the greatest empirical mean reward is guaranteed to be the best action with an error probability of error at most $\delta_0 + 2\alpha K \delta$.

Setting $\delta_0 = 1/8$ and $\delta = \frac{1}{16\alpha K}$ we conclude the algorithm can achieve an error probability of at most $1/4$ with a total number of queries upper bounded by,

$$\begin{aligned}\frac{2\sigma^2\log(2/\delta_0)}{\alpha^2} + 2\alpha K \cdot 32\sigma^2\log(2/\delta) &= \frac{2\sigma^2\log(16)}{\alpha^2} + 2\alpha K \cdot 32\sigma^2\log(32\alpha K) \\ &\leqslant \frac{2\sigma^2\log(16)}{\alpha^2} + 2\alpha K \cdot 32\sigma^2\log(32K).\end{aligned}$$

If the algorithm uses $\alpha = \left(\frac{\log(16)}{32K\log(32K)}\right)^{1/3}$ its query complexity can be upper bounded by,

$$\max\left(2 \cdot \frac{\log^{1/3}(16)}{(32)^{1/3}}\log^{2/3}(32K)K^{2/3}\sigma^2, 1\right)$$

Up to $\mathcal{O}(\cdot)$ notation Case 2's query complexity is smaller than the query upper bound from Case 1. We conclude there exists a universal constant $c > 0$ such that,

$$\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) \leqslant \begin{cases} 1 & \text{if } \sigma^2 \leqslant \frac{c}{\log^{2/3}(K)K^{2/3}} \\ c\log^{2/3}(K)K^{2/3}\sigma^2 & \text{o.w.} \end{cases}$$

We conclude the proof by noting $\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F})$ must be at least 1 so that when $\sigma^2 \leqslant \frac{c}{\log^{2/3}(K)K^{2/3}}$ we conclude $\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) = 1$.

■

Lemma 20 *For any $K \in \mathbb{N}$ satisfying $K \geqslant 2$ the function class $\mathcal{F}$ with action space $\mathcal{A} = \{a_i\}_{i=0}^K$ defined as $\mathcal{F} = \{f_i\}_{i=1}^K$ with,*

$$f_i(a) = \begin{cases} \frac{i}{4K} & \textit{if } a = a_0 \\ 1/2 & \textit{if } a \neq i \\ 1 & \textit{o.w.} \end{cases}$$

satisfies that for any $\epsilon \in [0, 1/2)$,

$$\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) \geqslant \max\left(\frac{1}{3}K^{2/3}\sigma^2, 1\right).$$

Proof

Lower bound In our proof we will use the helper function $\bar{f}_K$ defined as,

$$\bar{f}_K(a) = \begin{cases} \frac{1}{4} & \text{if } a = a_0 \\ 1/2 & \text{o.w.} \end{cases}$$

Consider any algorithm Alg interacting with problems in $\mathcal{F}$ and with $\bar{f}_K$ over n rounds. We will show in case $m^{\sigma}_{\mathrm{Alg}}(\epsilon, 1/4) \leqslant n$, then n must be lower bounded by $\max\left(\frac{1}{3}K^{2/3}\sigma^2, 1\right)$.

It is clear that any algorithm with an error probability of at most $1/4$ must make at least one single query, thus if $m^{\sigma}_{\mathrm{Alg}}(\epsilon, 1/4) \leqslant n$ then $n \geqslant 1$. To complete our result we want to show that $n \geqslant \frac{1}{3}K^{2/3}\sigma^2$ as well. We dedicate the rest of the argument to prove this lower bound.

Our argument is based on analyzing the hypothetical interaction between Alg and $\bar{f}_K$ and using it to derive properties of the interactions between Alg and functions in $\mathcal{F}$. Throughout this exploration we assume that Alg satisfies $m^{\sigma}_{\text{Alg}}(\epsilon, 1/4) \leqslant n$ so that,

$$\mathbb{P}_{\text{Alg},f_i}(k(n) = i) \geqslant 3/4. \tag{9}$$

Where $k(n)$ is the random variable equal to Alg's guess for the optimal action at time n. Let's start by considering the expected number of queries of action a_0 by algorithm Alg when interacting with problem $\bar{f}_K$:

$$L = \mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_0(n)\right],$$

where $T_0(n)$ denotes the (random) number of queries of action a_0 by Alg up to time n and the expectation is taken w.r.t the randomness of Alg and the measurement noise during its interaction with $\bar{f}_K$, Define $\mathcal{I} = \{i \in [K] : |\bar{f}_K(a_0) - f_i(a_0)| \leqslant \frac{\sigma}{4\sqrt{L}}\}$ be the set of indices of functions in $\mathcal{F}$ having a_0 function values close to $\bar{f}_K(a_0) = 1/4$. The size of $\mathcal{I}$ satisfies,

$$|\mathcal{I}| \geqslant \frac{\sigma}{4\sqrt{L}} \cdot K.$$

Consider $\{(\mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_i(n)\right], \mathbb{P}_{\text{Alg},\bar{f}_K}\left(k(n) = i\right))\}_{i\in\mathcal{I}}$. Lemma 28 implies there is an index $\tilde{i}$ such that,

$$\mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_{\tilde{i}}(n)\right] \leqslant \frac{3n}{|\mathcal{I}|} \leqslant \frac{12n\sqrt{L}}{K\sigma} \tag{10}$$

$$\mathbb{P}_{\text{Alg},\bar{f}_K}\left(k(n) = \tilde{i}\right) \leqslant \frac{3}{|\mathcal{I}|} \tag{11}$$

Let's now compute the KL between $\mathbb{P}_{\text{Alg},\bar{f}_K}$ and $\mathbb{P}_{\text{Alg},f_{\tilde{i}}}$. This quantity satisfies the following inequalities,

$$\begin{aligned}
\text{KL}\left(\mathbb{P}_{\text{Alg},\bar{f}_K} \,\|\, \mathbb{P}_{\text{Alg},f_{\tilde{i}}}\right) &= \mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_0(n)\right]\left(f_{\tilde{i}}(a_0) - \bar{f}_K(a_0)\right)^2 \cdot \frac{1}{\sigma^2} + \mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_{\tilde{i}}(n)\right] \cdot \frac{1}{4\sigma^2} \\
&\overset{(i)}{\leqslant} \mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_0(n)\right]\left(\frac{\sigma}{4\sqrt{L}}\right)^2 \cdot \frac{1}{\sigma^2} + \mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_{\tilde{i}}(n)\right] \cdot \frac{1}{4\sigma^2} \\
&\overset{(ii)}{\leqslant} \frac{1}{16} + \frac{6n\sqrt{L}}{K\sigma^3}
\end{aligned} \tag{12}$$

where inequality (i) holds because $\tilde{i} \in \mathcal{I}$ and therefore $\left|f_{\tilde{i}}(a_0) - \bar{f}_K(a_0)\right| \leqslant \frac{\sigma}{4\sqrt{L}}$ and (ii) because $\mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_0(n)\right] = L$ and $\mathbb{E}_{\text{Alg},\bar{f}_K}\left[T_{\tilde{i}}(n)\right] \leqslant \frac{12n\sqrt{L}}{K\sigma}$ as implied by equation 10. In order to obtain a lower bound for n we consider two cases.

1. Case 1. $\frac{\sigma}{4\sqrt{L}} \leqslant \frac{1}{6K}$.
2. Case 2. $\frac{\sigma}{4\sqrt{L}} > \frac{1}{6K}$.

In Case 1 we have $\sqrt{L} \geqslant \frac{3}{2}\sigma K$ and therefore $n \geqslant L \geqslant \frac{9}{4}\sigma^2K^2$. In Case 2 we have $|\mathcal{I}| \geqslant 6$ and therefore equation 11 implies

$$\mathbb{P}_{\text{Alg},\bar{f}_K}\left(k(n) = \tilde{i}\right) \leqslant \frac{3}{|\mathcal{I}|} \leqslant \frac{1}{2}.$$

Equation 9 implies that $\mathbb{P}_{\mathrm{Alg},f_{\tilde{i}}}\left(k(n)=\tilde{i}\right) \geqslant \frac{3}{4}$. Combining the last two inequalities,

$$\frac{1}{4} \leqslant \left|\mathbb{P}_{\mathrm{Alg},f_{\tilde{i}}}\left(k(n)=\tilde{i}\right)-\mathbb{P}_{\mathrm{Alg},\bar{f}_K}\left(k(n)=\tilde{i}\right)\right|.$$

The probability gap lower bound above can be combined with Pinsker inequality and inequality 12 to obtain,

$$\begin{aligned}\frac{1}{8} &\leqslant 2\left(\mathbb{P}_{\mathrm{Alg},\bar{f}_K}\left(k(n)=\tilde{i}\right)-\mathbb{P}_{\mathrm{Alg},f_{\tilde{i}}}\left(k(n)=\tilde{i}\right)\right)^2\\ &\leqslant \mathrm{KL}\left(\mathbb{P}_{\mathrm{Alg},\bar{f}_K}\,\|\,\mathbb{P}_{\mathrm{Alg},f_{\tilde{i}}}\right)\\ &\leqslant \frac{1}{16}+\frac{6n\sqrt{L}}{K\sigma^3}\\ &\leqslant \frac{1}{16}+\frac{6n^{3/2}}{K\sigma^3}\end{aligned}$$

And therefore,

$$n \geqslant \frac{1}{3}K^{2/3}\sigma^2.$$

This finalizes the lower bound showing that,

$$\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) \geqslant \frac{1}{3}K^{2/3}\sigma^2.$$

■

Theorem 14 *There exist universal constants $c,\bar{c}>0$ such that for every integer $K\geqslant 2$ there exists a function class $\mathcal{F}\subseteq[0,1]^{\mathcal{A}}$ with action space $|\mathcal{A}|=K+1$ such that for every $\epsilon\in[0,1/2)$ it holds that $\gamma_{\mathcal{F},\epsilon}=1/K$ and if $\sigma^2 \geqslant \frac{1}{\bar{c}K^{2/3}}$ then*

$$\bar{c}K^{2/3}\sigma^2 \leqslant \mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}) \leqslant c\log^{2/3}(K)K^{2/3}\sigma^2.$$

In particular,

$$\bar{c}\log(1/\gamma_{\mathcal{F},\epsilon})\sigma^2 \leqslant \mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F}),$$

and if $\sigma^2 \leqslant \frac{1}{c\log^{2/3}(K)K^{2/3}}$ then

$$\mathrm{QC}^{\sigma}_{\epsilon,1/4}(\mathcal{F})=\mathrm{QC}^{0}_{\epsilon,0}(\mathcal{F})=1.$$

Proof In order to prove Theorem 17 we use function classes $\mathcal{F}$ and action spaces $\mathcal{A}$ defined by $K\in\mathbb{N}$ and $\epsilon'\in[0,1/2)$ such that $K\geqslant 2$, $\epsilon'>\epsilon$ and where

$$\mathcal{A}=\{a_i\}_{i=0}^{K}$$

All elements in $\mathcal{F}$ are indexed by $i\in[K]$ such that,

$$f_i(a)=\begin{cases}\frac{i}{4K} & \text{if } a=a_0\\ 1/2 & \text{if } a\neq i\\ 1 & \text{o.w.}\end{cases}$$

We assume the noise model satisfies $\xi \sim \mathcal{N}(0, \sigma^2)$. We start by showing that $\gamma_{F,\epsilon} = \frac{1}{K}$ for all $\epsilon < 1/2$. By definition,

$$\gamma_{\mathcal{F}_K,\epsilon} = \sup_{p \in \Delta(\mathcal{A})} \inf_{f_i \in \mathcal{F}_K} \mathbb{P}_{a\sim p} \left(f_i(a_i) - f_i(a) \leqslant \epsilon \right), \tag{13}$$

where $\Delta(\mathcal{A})$ is the set of all distributions on $\mathcal{A}$. This is because when $0 \leqslant \epsilon < 1/2$ action a_i is the unique ϵ-optimal action for problem f_i. Any distribution p over $\mathcal{A}$ satisfies

$$\min_{i \in [K]} p(a_i) \leqslant \frac{1}{K}.$$

Thus it follows that,

$$\mathbb{P}_{a\sim p} \left(f_i(a_i) - f_i(a) \leqslant \epsilon \right) = \min_{i \in [K]} p(a_i) \leqslant \frac{1}{K}.$$

for any $p \in \Delta(\mathcal{A})$. And therefore,

$$\gamma_{\mathcal{F},\epsilon} \leqslant \frac{1}{K}.$$

Finally, the uniform distribution $p = \text{Uniform}(a_1, \cdots, a_K)$ satisfies

$$\inf_{f_i \in \mathcal{F}} \mathbb{P}_{a\sim p} \left(f_i(a_i) - f_i(a) \leqslant \epsilon \right) = \frac{1}{K},$$

showing that $\gamma_{\mathcal{F},\epsilon} = 1/K$. The remainder of this Theorem follows from the results of Lemmas 19 and 20.

■

B.3. Proof of Theorem 16

Definition 21 (Gap of $\mathcal{F}$) *Let $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$. Let $\epsilon, \delta \in [0,1]$ be a finite function class over a finite action space, and* Alg *be any (possibly randomized) algorithm. Let $\Gamma_n(\text{Alg})$ denote all length n partial action-rewards trajectories $((a_1, r_1), ..., (a_n, r_n))$ in the support of the trajectory distribution of* Alg*. If there are no such trajectories (for example when n is larger than the largest trajectory in the support of* Alg*) we define $\Gamma_n(\text{Alg}) = \varnothing$. For any n we define the n-degree Gap of* Alg *and $\mathcal{F}$ to be:*

$$\text{Gap}_n(\text{Alg}, \mathcal{F}) = \inf_{\tau_n \in \Gamma_n(\text{Alg}),\ r \in \{f(a_n) | f \in \mathcal{F}(\tau_{n-1})\} \text{ s.t. } r \neq r_n} |r - r_n|.$$

Where $\tau_i = ((a_1, r_1), ..., (a_i, r_i))$ for all $i \in [n]$ and $\tau_0 = \varnothing$. When $\Gamma_n(\text{Alg}) = \varnothing$ we define $\text{Gap}_n(\text{Alg}, \mathcal{F}) = \infty$. We define the Gap of Alg *and $\mathcal{F}$ to be:*

$$\text{Gap}(\text{Alg}, \mathcal{F}) = \min_{n \in \mathbb{N}} \text{Gap}_n(\text{Alg}, \mathcal{F}).$$

Finally, we define the Gap of $\mathcal{F}$ to be

$$\text{Gap}_{\epsilon,\delta}(\mathcal{F}) = \max_{\text{Alg} \in \mathbb{A}_{\epsilon,\delta}} \text{Gap}(\text{Alg}, \mathcal{F})$$

where $\mathbb{A}_{\epsilon,\delta}$ denotes the set of randomized algorithms with query complexity $\text{QC}^0_{\epsilon,\delta}(\mathcal{F})$.

Theorem 16 *Let $\delta, \delta' \in (0,1)$ such that $\delta > \delta' \geqslant 0$. For any finite class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ over a finite action space the noisy query complexity with zero-mean Gaussian noise with variance σ^2 such that $\sigma^2 < \frac{\mathrm{Gap}^2_{\epsilon,\delta'}(\mathcal{F})}{4\log(2\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})/(\delta-\delta'))}$ satisfies:*

$$\mathrm{QC}^{\sigma}_{\epsilon,\delta}(\mathcal{F}) \leqslant \mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})$$

Proof We prove this result by constructing an algorithm Alg that satisfies $m^{\sigma}_{\mathrm{Alg}}(\epsilon,\delta) \leqslant \mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})$.

Lemma 27 implies that when noise variance satisfies $\sigma^2 < \frac{\mathrm{Gap}^2_{\epsilon,\delta'}(\mathcal{F})}{4\log(2\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})/(\delta-\delta'))}$ then a single noisy query of action $a \in \mathcal{A}$ for function $f \in \mathcal{F}$ produces a value $\hat{r}$ satisfying $|\hat{r} - f(a)| > \frac{\mathrm{Gap}_{\epsilon,\delta'}(\mathcal{F})}{2}$ with probability at most $\frac{\delta-\delta'}{\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})}$.

Pick an algorithm $\widetilde{\mathrm{Alg}}$ realizing $\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})$ and $\mathrm{Gap}_{\epsilon,\delta'}(\mathcal{F})$. We construct algorithm Alg based on $\widetilde{\mathrm{Alg}}$. The difference will be that Alg will map all received noisy rewards to a candidate noiseless reward. The resulting candidate noiseless reward is used to build a reconstructed noiseless action-reward history that is fed into $\widetilde{\mathrm{Alg}}$.

Algorithm Alg plays actions $a_1, \cdots, a_i$, observes rewards $\hat{r}_1, \cdots, \hat{r}_i$ and adjusts them to fit noiseless values $r_1, \cdots, r_i$ to build a partial history $\tau_i = ((a_1, r_1), \cdots, (a_i, r_i))$, where we define $\tau_0 = \varnothing$. We'll explain how Alg will construct $r_1, \cdots, r_i$ from $((a_1, \hat{r}_1), \cdots, (a_i, \hat{r}_i))$ below. This partial trajectory is fed into $\widetilde{\mathrm{Alg}}$. If $\widetilde{\mathrm{Alg}}$ is ready to suggest an ϵ-optimal policy, Alg outputs it as its own guess. In case $\widetilde{\mathrm{Alg}}$ decides to proceed exploring it will propose a distribution over actions $\widetilde{\mathrm{Alg}}(\tau_i)$. Algorithm Alg will play $a_{i+1} \sim \widetilde{\mathrm{Alg}}(\tau_i)$ and observe a noisy reward $\hat{r}_{i+1}$.

It remains to define how Alg after constructing a noiseless trajectory $((a_1, r_1), \cdots, (a_i, r_i))$ will map a noisy observation $\hat{r}_{i+1}$ to a candidate noiseless r_{i+1}. This is done by solving the nearest neighbors problem,

$$r_{i+1} = \operatorname*{arg\,min}_{r \in \{f(a) | f \in \mathcal{F}(\tau_i)\}} |\hat{r}_{i+1} - r|.$$

When $\sigma^2 < \frac{\mathrm{Gap}^2_{\epsilon,\delta'}(\mathcal{F})}{4\log(2\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})/(\delta-\delta'))}$ it follows that when Alg interacts with function $f \in \mathcal{F}$,

$$\mathbb{P}_{\mathrm{Alg},f}\left(r_i = f(a_i) \mid r_j = f(a_j)\ \forall j \leqslant i-1 \text{ and Alg plays } a_i = a\right) \geqslant 1 - \frac{\delta-\delta'}{\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})}. \quad (14)$$

This is because when $r_j = f(a_j)$ for all $j \leqslant i-1$, then the partial trajectory $((a_1, r_1), \cdots, (a_{i-1}, r_{i-1}))$ is in the support of $\widetilde{\mathrm{Alg}}$. Integrating equation 14 over the possibly random choice of a_i,

$$\mathbb{P}_{\mathrm{Alg},f}\left(r_i = f(a_i) \mid r_j = f(a_j)\ \forall j \leqslant i-1\right) \geqslant 1 - \frac{\delta-\delta'}{\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})}$$

Let's consider the event $\mathcal{E}_{\mathrm{bound}}$ when Alg outputs a guess for an ϵ-optimal policy after at most $\mathrm{QC}_{\epsilon,\delta'}(\mathcal{F})$ queries. Let also define $\mathcal{E}_{\mathrm{nn}}$ to be the event that all nearest neighbor estimators r_i during Alg's execution match the correct mean reward. Additionally define $\mathcal{E}_{\mathrm{correct}}$ as the event when Alg

outputs a correct ϵ-optimal policy. A change of measure argument implies that, for any f,

$$\begin{aligned}\mathbb{P}_{\mathrm{Alg},f}(\mathcal{E}_{\mathrm{correct}} \cap \mathcal{E}_{\mathrm{bound}} \cap \mathcal{E}_{\mathrm{nn}}) &\geqslant \mathbb{P}_{\widetilde{\mathrm{Alg}},f}(\mathcal{E}_{\mathrm{correct}} \cap \mathcal{E}_{\mathrm{bound}}) \cdot \left(1 - \frac{\delta - \delta'}{\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})}\right)^{\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})} \\ &\overset{(i)}{\geqslant} \mathbb{P}_{\widetilde{\mathrm{Alg}},f}(\mathcal{E}_{\mathrm{correct}} \cap \mathcal{E}_{\mathrm{bound}}) \cdot \big(1 - (\delta - \delta')\big)\end{aligned}$$

inequality (i) holds because for $a \in [0,1]$ and $n \in \mathbb{N}$ the bernoulli inequality implies $(1-\frac{a}{n})^n \geqslant 1-a$. Finally, $\mathbb{P}_{\widetilde{\mathrm{Alg}},f}(\mathcal{E}_{\mathrm{correct}} \cap \mathcal{E}_{\mathrm{bound}}) \geqslant 1-\delta'$ since $\widetilde{\mathrm{Alg}}$ realizes $\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})$. Combining these inequalities we obtain,

$$\begin{aligned}\mathbb{P}_{\mathrm{Alg},f}(\mathcal{E}_{\mathrm{correct}} \cap \mathcal{E}_{\mathrm{bound}} \cap \mathcal{E}_{\mathrm{nn}}) &\geqslant (1-\delta') \cdot \big(1 - (\delta - \delta')\big) \\ &= 1 - \delta + \delta'\delta - (\delta')^2 \\ &> 1 - \delta.\end{aligned}$$

Since $\mathbb{P}_{\mathrm{Alg},f}(\mathcal{E}_{\mathrm{correct}} \cap \mathcal{E}_{\mathrm{bound}}) \geqslant \mathbb{P}_{\mathrm{Alg},f}(\mathcal{E}_{\mathrm{correct}} \cap \mathcal{E}_{\mathrm{bound}} \cap \mathcal{E}_{\mathrm{nn}})$ we conclude Alg's probability of error when interacting with any function f for at most $\mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F})$ rounds is upper bounded by δ. This in turn implies that,

$$\mathrm{QC}^{\sigma}_{\epsilon,\delta}(\mathcal{F}) \leqslant \mathrm{QC}^0_{\epsilon,\delta'}(\mathcal{F}).$$

■

Appendix C. Missing discussion and proofs of Section 6

C.1. Regret vs. QC: noise-free case

In this section we prove there is a a separation between regret and ϵ-optimality in the noise-free setting. In noise-free problems with finite action spaces, it is possible to achieve finite regret. Any algorithm that is guaranteed to find an optimal action in finitely many queries can find the optimal action and then incur zero regret in any subsequent time-step by playing it. In Theorem 22 we show there are problems where the objective of finding an optimal action cannot be achieved without paying a regret scaling linearly with the query complexity.

Theorem 22 *Let $d \in \mathbb{N}$, $\epsilon \in [0,1)$ and $\gamma \in (0,1)$ such that $\epsilon + \gamma < 1$. There is a function class $\mathcal{F} \subseteq [0,1]^{\mathcal{A}}$ such that $\mathrm{QC}^0_{\epsilon,0}(\mathcal{F}) = d$, and such that for any $T \geqslant d$ and any algorithm Alg such that $m^0_{\mathrm{Alg}}(\epsilon, 0) \leqslant T$ satisfies*

$$\max_{f \in \mathcal{F}} \mathrm{Regret}_{\mathrm{Alg}}(T, f) \geqslant d.$$

Moreover for any T there is an algorithm Alg' such that $\max_{f\in\mathcal{F}} \mathrm{Regret}_{\mathrm{Alg}'}(T,f) \leqslant (\epsilon+\gamma)T$.

The function class $\mathcal{F}$ used to prove Theorem 22 is based on a "breadcrums trail" construction where the action space can be identified with a binary tree of depth $d+1$. The function class is indexed by the 2^d leaf actions. The function identified by a leaf action a_{leaf} achieves an optimal value of 1 at a_{leaf}. The actions along the path from the root to a_{leaf} have rewards of 0 except the root action that achieves a value of $1-\epsilon-\gamma$. All leaf actions different from a_{leaf} have a value of zero and every other action has a reward of $1-\epsilon-\gamma$. It is clear that the algorithm that always

plays the root action achieves a regret of order $(\epsilon+\gamma)T$. We prove the lower bound by proving, first that $\mathrm{QC}^0_{\epsilon,0}(\mathcal{F}) = d$ and by showing that any algorithm achieving $\mathrm{QC}^0_{\epsilon,0}(\mathcal{F})$ must have necessarily selected at least d actions with zero rewards when interacting with one of the functions of the class thus incurring regret at least d despite the existence of a low regret algorithm.

In order to prove Theorem 22 we introduce a family of function classes parameterized by $d \in \mathbb{N}$ and $\Delta \in [0,1]$ such that $\Delta = \epsilon+\gamma$. We define the function class $\mathcal{F}_\Delta$ over action space $\mathcal{A}_{\text{tree}}$ indexed by the nodes of a height $d+1$ binary tree defined here as having $d+1$ levels and $2^{d+1}-1$ nodes. We call $a_{l,i}$ the i-th action of the l-th level of the tree. See Figure 1 to see an example of a tree with depth 3 (in this case $d=2$).

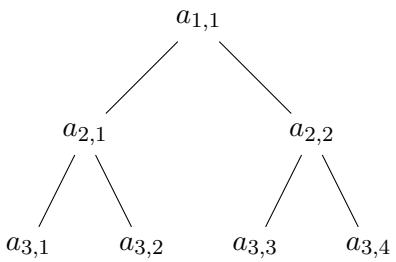

Figure 1: $\mathcal{A}_{\text{tree}}$ action set when $d=2$.

We define the function class $\mathcal{F}_\Delta$ as indexed by paths from the root node in $\mathcal{A}_{\text{tree}}$ to a leaf (alternatively indexed by leaves). For any such path $\mathbf{p} = \{a_{1,1}, a_{2,i_2}, \cdots, a_{d+1,i_{d+1}}\}$ the function $f^{(\mathbf{p})}$ equals,

$$f^{(\mathbf{p})}(a) = \begin{cases} 1 & \text{if } a = a_{d+1,i_{d+1}} \\ 0 & \text{if } a = a_{d+1,j} \text{ and } j \neq i_{d+1} \\ 0 & \text{if } a \in \{a_{2,1}, a_{2,i_2}, \cdots, a_{d,i_d}\} \\ 1-\Delta & \text{o.w.} \end{cases} \tag{15}$$

The function parameterized by path $\mathbf{p}$ achieves an optimal value of one at the leaf of path $\mathbf{p}$, a value of zero at any other leaf different from $a_{d+1,i_{d+1}}$, the endpoint of $\mathbf{p}$ and also a value of zero at any other action in $\mathbf{p}$ different from the endpoint and the root $a_{1,1}$. The function also achieves a "large" reward value of $1-\Delta$ at the root $a_{1,1}$ and every other action not listed above. When $\epsilon < \Delta$, the only ϵ-optimal action in $f^{(\mathbf{p})}$ equals $a_{d+1,i_{d+1}}$. Our first partial result towards proving Theorem 22 is to show the query complexity of $\mathcal{F}_\Delta$ equals d:

Proposition 23 *Let $\epsilon, \Delta \in [0,1]$ such that $\epsilon < \Delta$ and define $\mathcal{F}_\Delta$ as the tree function class defined in equation 15. The query complexity of $\mathcal{F}_\Delta$ satisfies,*

$$\mathrm{QC}^0_{\epsilon,0}(\mathcal{F}_\Delta) = d$$

Proof For any algorithm Alg its query complexity and depth satisfy

$$\mathrm{QC}^0_{\epsilon,0}(\mathcal{F}_\Delta) \leqslant m^0_{\text{Alg}}(\epsilon, 0).$$

Consider an algorithm that queries the actions in $\mathcal{A}_{\text{tree}}$ starting at $a_{2,1}$. If it observes a 0 it continues querying its left-side leaf. If it is querying an interior node and observes a $1-\Delta$, it continues querying the left-leaf of its sibling action[2]. Algorithm Alg is guaranteed to have identified an action with a reward of 1 after at most d queries. The algorithm may not have queried an action of reward 1 after exactly d queries but it will be certain of its identity.

Moreover, we can show $\mathrm{QC}^0_{\epsilon,0}(\mathcal{F}_\Delta) \geqslant d$.

2. Here we define the sibling of an action to be that which is the opposite child of its parent action.

Let Alg be an arbitrary deterministic algorithm. To prove that $\mathrm{QC}^0_{\epsilon,0}(\mathcal{F}_\Delta) \geqslant d$ we start by noting that any action in $\mathcal{A}_{\text{tree}}$ has exactly two possible values $(1-\Delta, 0)$ for inner actions and $(1, 0)$ for leaf actions.

In order to arrive at a contradiction let's assume Alg is able to identify an ϵ-optimal action after only $d-1$ interactions for all $f \in \mathcal{F}_\Delta$. We'll prove that when limiting the interactions to only $d-1$ or less, there will be two functions in $\mathcal{F}_\Delta$ that will produce the same interaction trace.

This is easy to deduce. First observe that histories of size at most $d-1$ give rise to at most 2^{d-1} distinct reward traces. This is because any query action can take at most two reward values. Finally, since the total number of functions in $\mathcal{F}_\Delta$ is 2^d we conclude there must exist two functions that produce the same trace if these are limited to sizes at most $d-1$.

This finalizes the proof.

■

Our second result establishes that any algorithm Alg that is able to find an ϵ-optimal action for $\epsilon < \Delta$ must incur in at least d regret.

Lemma 24 *Let $0 \leqslant \epsilon < \Delta$. For any algorithm* Alg *interacting with $(\mathcal{F}_\Delta, \mathcal{A}_{\text{tree}})$ such that $m^0_{\text{Alg}}(\epsilon, 0) \leqslant T$ there is a problem $f^{\mathbf{p}}$ such that when* Alg *interacts with it,*

$$\textit{Regret}_{\text{Alg}}(T, f^{(\mathbf{p})}) \geqslant d.$$

Proof Let $n_{\mathbf{p}}$ be the number of queries when Alg received a reward of 0 when interacting with $f^{(\mathbf{p})}$ over T rounds.

The algorithm's regret is lower bounded by n_{p}. We proceed to show that when $m^0_{\text{Alg}}(\epsilon, 0) \leqslant T$

$$\max_{\mathbf{p}} n_{\mathbf{p}} \geqslant d$$

.

Algorithm Alg can be understood as producing an action given a partial history. Given a partial history, $a_1, r_1, \cdots, a_t, r_t$ algorithm Alg produces an action a_{t+1}. We say a_{t+1} is a dummy action if for all $f \in \mathcal{F}(a_1, r_1, \cdots, a_t, r_t)$ the value $f(a_{t+1})$ is the same. That is, action a_{t+1} is uninformative because no matter what function f algorithm Alg may be interacting with, this action will not have any useful information for Alg. We say that an algorithm is in a "reduced" form if it never proposes a to play a dummy action.

Given any deterministic algorithm Alg we construct a "reduced" version of Alg, which we call $\widetilde{\text{Alg}}$. This algorithm can be constructed by skipping all dummy actions proposed by Alg.

We work now with $\widetilde{\text{Alg}}$, the reduced version of Alg. Notice that irrespective of the partial history (say $a_1, r_1, \cdots, a_t, r_t$), either $|\mathcal{F}(a_1, r_1, \cdots, a_t, r_t)| = 1$ or when action a_{t+1} is proposed by $\widetilde{\text{Alg}}$ in reaction to this partial history, there is a function $f^{(\mathbf{p})} \in \mathcal{F}(a_1, r_1, \cdots, a_t, r_t)$ such that $f^{(\mathbf{p})}(a_{t+1}) = 0$. This is because for any non-dummy action there must exist two functions in the version space that have different values and any action in $\mathcal{A}_{\text{tree}}$ can achieve two values over all functions in in $\mathcal{F}_\Delta$, one of which is always zero.

When an action with value 0 is observed, the version space at most halves. that is if a_{t+1} is a non-dummy action for history $a_1, r_1, \cdots, r_t, a_t$ then,

$$\mathcal{F}(a_1, r_1, \cdots, a_t, r_t) \leqslant 2\mathcal{F}(a_1, r_1, \cdots, a_t, r_t, a_{t+1}, 0). \tag{16}$$

Finally, let's follow the sequence of actions $\widetilde{\text{Alg}}$ proposes when all observed rewards are 0. Let $a_1, 0, a_2, 0, \cdots$ be the resulting history. Equation 16 implies that

$$\mathcal{F}(a_1, 0, \cdots, a_t, 0) \leqslant 2\mathcal{F}(a_1, 0, \cdots, a_{t+1}, 0)$$

and therefore $\mathcal{F}(a_1, 0, \cdots, a_t, 0) \geqslant \mathcal{F}_\Delta/2^t$. Since Alg, can guess a correct an ϵ-optimal action only whn the version space is of size 1, we conclude that the history generated by $\widetilde{\text{Alg}}$ when all observed rewards are 0 must produce at least d consecutive 0 observations. Let $f^{(\tilde{\mathbf{p}})}$ be the function that generates this "zero" history. It follows that $n_{\tilde{\mathbf{p}}} \geqslant d$. This has to be because the algorithm is assumed to produce a valid ϵ-optimal action at the end of the interaction with any function up to T steps.

This finalizes the desired result.

■

We have the necessary ingredients to prove Theorem 22.

Proof Theorem 22 is a corollary of Lemma 24. The function class in Theorem 22 can be identified with $(\mathcal{F}_\Delta, \mathcal{A}_{\text{tree}})$ with $\Delta > \epsilon$. Setting $\gamma = \Delta - \epsilon$ the algorithm Alg' that always selects action $a_{1,1}$ satisfies $\text{Regret}_{\text{Alg}'}(T, f) = (\epsilon + \gamma)T$ for all $T \in \mathbb{N}$ and all $f \in \mathcal{F}_\Delta$. ■

C.2. Proof of Theorem 17

In order to prove Theorem 17 we use function classes $\mathcal{F}_K$ and action spaces $\mathcal{A}_K$ defined by $K \in \mathbb{N}$ and $\epsilon_1, \epsilon_2 \in [0, 1]$ where

$$\mathcal{A}_K = \mathcal{A}_1 \cup \mathcal{A}_2$$

such that $\mathcal{A}_1 = \{a_1^{(1)}, \cdots, a_{\lceil \log(K) \rceil}^{(1)}\}$ and $\mathcal{A}_2 = \{a_1^{(2)}, \cdots, a_K^{(2)}\}$. All elements in $\mathcal{F}_K$ are indexed by $k \in [K]$ such that,

$$f_k(a) = \begin{cases} 1/2 \pm \epsilon_1 \text{ if } a \in \mathcal{A}_1 & \\ 1 - \epsilon_2 & \text{if } a \neq a_k^{(2)} \\ 1 & \text{o.w.} \end{cases}$$

We also assume that for any $k \in [K]$, the values $\mathbf{1}\left(f_k(a_1^{(1)}) > 1/2\right) \cdots, \mathbf{1}\left(f_k(a_{\lceil \log(K) \rceil} > 1/2\right)$ form the binary representation of k with noise model $\xi \sim \mathcal{N}(0, 1)$. This example is designed to ensure that querying actions in $\mathcal{A}_1$ enough will yield enough information to identify an $\epsilon < \epsilon_2$ optimal action (for any $\epsilon < \epsilon_2$) but trying these actions generates linear regret.

Our first result is to "sandwich" the query complexity of this function class,

Lemma 25 *When $\epsilon_2 \leqslant \epsilon_1$ and $K \geqslant 2$ the query complexity of $\mathcal{F}_K, \mathcal{A}_K$ satisfies,*

$$\text{QC}^1_{0,1/4}(\mathcal{F}_K) \in \left[\frac{\log(4/3)}{2\epsilon_1^2}, \frac{16\log(K)\log\left(4\log(K)\right)}{\epsilon_1^2}\right]$$

Proof

As part of our argument we'll consider the "empty" problem f_0 over action sets $\mathcal{A}_1, \mathcal{A}_2$ defined as,

$$f_0(a) = \begin{cases} 1/2 & \text{if } a \in \mathcal{A}_1 \\ 1 - \epsilon_2 & \text{if } a \in \mathcal{A}_2 \end{cases}$$

Let Alg be an algorithm interacting with $\mathcal{F}_K, \mathcal{A}_K$ over n steps and let's assume that $m^1_{\text{Alg}}(0, 1/4) \leqslant n$ so that when interacting with any function in $\mathcal{F}_K$, after n steps Alg can guess the correct optimal action with probability at least $3/4$.

Let's consider the interaction of Alg with function f_0. We use the notation $T_i^{(j)}(n)$ to denote the (random) number of queries Alg has performed on action $a_i^{(j)} \in \mathcal{A}_j$ for $j \in \{1, 2\}$.

Let $\mathcal{E}_i$ for $i \in [K]$ denote the event that at time n algorithm Alg outputs a guess $k(n)$ for the optimal action satisfying $k(n) = i$. So that,

$$\mathcal{E}_i = \{k(n) = i\}.$$

Since $m^1_{\text{Alg}}(0, 1/4) \leqslant n$ it follows that,

$$\mathbb{P}_{\text{Alg},f_i}(\mathcal{E}_i) \geqslant 3/4. \tag{17}$$

Let's consider the problem f_i. The divergence decomposition Lemma 30 implies,

$$\begin{aligned}
\text{KL}(\mathbb{P}_{\text{Alg},f_0} \parallel \mathbb{P}_{\text{Alg},f_i}) &= \sum_{i' \in [\log(K)]} \mathbb{E}_{\text{Alg},f_0}\left[T_{i'}^{(1)}(n)\right]\epsilon_1^2 + \mathbb{E}_{\text{Alg},f_0}\left[T_i^{(2)}(n)\right]\epsilon_2^2 \\
&\leqslant n\epsilon_1^2 + \mathbb{E}_{\text{Alg},f_0}\left[T_i^{(2)}(n)\right]\epsilon_2^2 \\
&\leqslant 2n\epsilon_1^2.
\end{aligned}$$

The Huber-Bretagnolle inequality (Lemma 29) applied to measures $\mathbb{P}_{\text{Alg},f_0}$ and $\mathbb{P}_{\text{Alg},f_i}$ for $i \in [K]$ implies,

$$\mathbb{P}_{\text{Alg},f_0}(\mathcal{E}_i) + \mathbb{P}_{\text{Alg},f_i}(\mathcal{E}_i^c) \geqslant \exp\left(-2n\epsilon_1^2\right)$$

summing over all $i \in [K]$ we get,

$$\begin{aligned}
K\exp\left(-2n\epsilon_1^2\right) &\leqslant \sum_i \mathbb{P}_{\text{Alg},f_0}(\mathcal{E}_i) + \mathbb{P}_{\text{Alg},f_i}(\mathcal{E}_i^c) \\
&\leqslant 1 + \sum_i \mathbb{P}_{\text{Alg},f_i}(\mathcal{E}_i^c)
\end{aligned}$$

Thus, there is at least one index $\hat{i}$ such that,

$$\mathbb{P}_{\text{Alg},f_{\hat{i}}}\left(\mathcal{E}_{\hat{i}}^c\right) \geqslant \exp\left(-2n\epsilon_1^2\right) - 1/K.$$

Equation 17 implies the error probability $\mathbb{P}_{\text{Alg},f_{\hat{i}}}\left(\mathcal{E}_{\hat{i}}^c\right) \leqslant 1/4$ and therefore,

$$\exp\left(-2n\epsilon_1^2\right) - 1/K \leqslant 1/4.$$

Expanding the last inequality we obtain the condition $\exp\left(-2n\epsilon_1^2\right) \leqslant 3/4$ and therefore $n \geqslant \frac{\log(4/3)}{2\epsilon_1^2}$. We conclude that,

$$\text{QC}^1_{0,1/4}(\mathcal{F}_K) \geqslant \frac{\log(4/3)}{2\epsilon_1^2}.$$

The upper bound $\text{QC}^1_{0,1/4}(\mathcal{F}_K) \leqslant \frac{16\log(4\log(K))}{\epsilon_1^2}$ follows a simple argument. We exhibit an algorithm Alg that can find an optimal action after at most $\frac{16\log(K)\log(4\log(K))}{\epsilon_1^2}$ queries. This procedure consists fo two parts,

1. Explore all informative actions and estimate them up to a $\epsilon_1/4$ error.
2. Use these values to decode the location of the optimal action.

Lemma 27 implies Step 1 can be achieved by estimating the reward of each action in $\mathcal{A}_1$ up to $\epsilon_1/4$ error and probability at least $1 - 1/4\log(K)$. This requires at most $16\log(4\log(K))/\epsilon_1^2$ queries for each of the $\log(K)$ actions in $\mathcal{A}_1$. A union bound implies this procedure is guaranteed to find the correct optimal action with a probability of error of at most $1/4$. We have therefore constructed an algorithm satisfying $m^1_{\text{Alg}}(0, 1/4) \leqslant \frac{16\log(K)\log(4\log(K))}{\epsilon_1^2}$. Thus,

$$\text{QC}^1_{0,1/4}(\mathcal{F}_K) \leqslant \frac{16\log(K)\log\left(4\log(K)\right)}{\epsilon_1^2}.$$

■

Our next supporting result shows that any algorithm that can find an optimal action must incur large regret.

Lemma 26 *Let $T \in \mathbb{N}$ and $\epsilon_1 \in [0, 1]$ such that $\epsilon_1 \leqslant 1/4$ and define $\epsilon_2 = 4\epsilon_1^2$. For any algorithm* Alg *such that $m^1_{\text{Alg}}(0, 1/4) \leqslant T$ there is a problem $f \in \mathcal{F}_K$ such that,*

$$\mathbb{E}_{\text{Alg}}[\textit{Regret}(T, f)] \geqslant \frac{1}{64\epsilon_1^2}.$$

for some universal constant $c > 0$.

Proof We'll again consider an empty problem f'_0 over action sets $\mathcal{A}_1$ and $\mathcal{A}_2$ defined as,

$$f'_0(a) = \begin{cases} 1/2 & \text{if } a \in \mathcal{A}_1 \\ 1 & \text{if } a \in \mathcal{A}_2 \end{cases}$$

Just as in the proof of Lemma 25 let's consider the problem f_i for $i \in [K]$ and Alg's interaction with f'_0 and f_i up to a horizon of T. The divergence decomposition Lemma 30 implies,

$$\begin{aligned} \text{KL}(\mathbb{P}_{\text{Alg},f_i} \parallel \mathbb{P}_{\text{Alg},f'_0}) &= \sum_{i' \in [\log(K)]} \mathbb{E}_{\text{Alg},f_i}\left[T^{(1)}_{i'}(T)\right]\epsilon_1^2 + \sum_{i' \neq i} \mathbb{E}_{\text{Alg},f_i}\left[T^{(2)}_{i'}(T)\right]\epsilon_2^2 \qquad (18) \\ &= \mathbb{E}_{\text{Alg},f_i}\left[T_{\mathcal{A}_1}(T)\right]\epsilon_1^2 + \mathbb{E}_{\text{Alg},f_i}\left[T^{(2)}_{-i}(T)\right]\epsilon_2^2 \qquad (19) \end{aligned}$$

where we define $T_{\mathcal{A}_1}(T) = \sum_{i' \in \mathcal{A}_1} T^{(1)}_{i'}(T)$ as the number of queries from actions in $\mathcal{A}_1$ and $T^{(2)}_{-i}(T) = \sum_{i' \in \mathcal{A}_2 \setminus \{i\}} T^{(2)}_{i'}(T)$. Pinsker's inequality implies,

$$2\left(\mathbb{P}_{\text{Alg},f'_0}(\mathcal{E}_i) - \mathbb{P}_{\text{Alg},f_i}(\mathcal{E}_i)\right)^2 \leqslant \text{KL}(\mathbb{P}_{\text{Alg},f_i} \parallel \mathbb{P}_{\text{Alg},f'_0}) \qquad (20)$$

Where $\mathcal{E}_i$ for $i \in [K]$ denote the event that at time T algorithm Alg outputs a guess $k(T)$ for the optimal action satisfying $k(T) = i$. So that,

$$\mathcal{E}_i = \{k(T) = i\}.$$

Since $\sum_{i=1}^{K} \mathbb{P}_{\text{Alg},f_0'}(\mathcal{E}_i) = 1$, there exists an index $\hat{i}$ such that $\mathbb{P}_{\text{Alg},f_0'}(\mathcal{E}_{\hat{i}}) \leqslant \frac{1}{K} \leqslant 1/2$.

Since $m^1_{\text{Alg}}(0, 1/4) \leqslant T$, it follows that $\mathbb{P}_{\text{Alg},f_{\hat{i}}}\left(\mathcal{E}_{\hat{i}}^c\right) \leqslant \frac{1}{4}$ and therefore $\mathbb{P}_{\text{Alg},f_{\hat{i}}}\left(\mathcal{E}_{\hat{i}}\right) \geqslant \frac{3}{4}$. Thus,

$$\left|\mathbb{P}_{\text{Alg},f_0'}\left(\mathcal{E}_{\hat{i}}\right) - \mathbb{P}_{\text{Alg},f_{\hat{i}}}\left(\mathcal{E}_{\hat{i}}\right)\right| \geqslant 1/4. \tag{21}$$

Combining Equations 19, 20 and 21 we get,

$$\frac{1}{8} \leqslant \text{KL}(\mathbb{P}_{\text{Alg},f_{\hat{i}}} \parallel \mathbb{P}_{\text{Alg},f_0'}) = \mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T_{\mathcal{A}_1}(T)\right]\epsilon_1^2 + \mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T^{(2)}_{-\hat{i}}(T)\right]\epsilon_2^2. \tag{22}$$

Thus, $\max\left(\mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T_{\mathcal{A}_1}(T)\right]\epsilon_1^2, \mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T^{(2)}_{-\hat{i}}(T)\right]\epsilon_2^2\right) \geqslant \frac{1}{16}$ so that at least one of the following two inequalities hold,

A) $\mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T_{\mathcal{A}_1}(T)\right] \geqslant \frac{1}{16\epsilon_1^2}$.

B) $\mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T^{(2)}_{-\hat{i}}(T)\right] \geqslant \frac{1}{16\epsilon_2^2}$.

Finally,

$$\mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[\text{Regret}(T, f_{\hat{i}})\right] \geqslant \frac{1}{4}\mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T_{\mathcal{A}_1}(T)\right] + \mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[T^{(2)}_{-\hat{i}}(T)\right]\epsilon_2.$$

And therefore when case A) holds,

$$\mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[\text{Regret}(T, f_{\hat{i}})\right] \geqslant \frac{1}{64\epsilon_1^2}.$$

When case B) holds,

$$\mathbb{E}_{\text{Alg},f_{\hat{i}}}\left[\text{Regret}(T, f_{\hat{i}})\right] \geqslant \frac{1}{16\epsilon_2} \overset{(i)}{\geqslant} \frac{1}{64\epsilon_1^2}.$$

Inequality (i) follows because $\epsilon_2 = 4\epsilon_1^2$.

■

Finally, we are ready to prove Theorem 17.

Theorem 17 *Let $d, T \in \mathbb{N}$. There exists a function class $\mathcal{F}$ over action space $\mathcal{A}$ with unit-variance Gaussian noise such that $d \leqslant \text{QC}^1_{0,1/4}(\mathcal{F}) \leqslant 80d$ and any algorithm Alg such that $m^1_{\text{Alg}}(0, 1/4) \leqslant T$ satisfies*

$$\max_{f\in\mathcal{F}} \mathbb{E}_{\text{Alg}}[\text{Regret}(T, f)] \geqslant \frac{d}{128}.$$

Moreover, there is an algorithm Alg' that satisfies $\max_{f\in\mathcal{F}} \mathbb{E}_{\text{Alg}'}\left[\text{Regret}(T, f)\right] \leqslant 8\sqrt{2T\log(T)}$ for all $T \in \mathbb{N}$.

Proof To prove this Theorem we leverage our supporting results from Lemma 25 and 26. We consider $(\mathcal{F}_K, \mathcal{A}_K)$ for $K = 2$. Under this choice Lemma 25 implies that,

$$\mathrm{QC}^1_{0,1/4}(\mathcal{F}_K) \in \left[\frac{\log(4/3)}{2\epsilon_1^2}, \frac{16\log(K)\log\left(4\log(K)\right)}{\epsilon_1^2}\right] = \left[\frac{\log(4/3)}{2\epsilon_1^2}, \frac{16\log(2)\log\left(4\log(2)\right)}{\epsilon_1^2}\right]$$

Setting $d = \frac{\log(4/3)}{2\epsilon_1^2}$ so that $\epsilon_1 = \sqrt{\frac{\log(4/3)}{2d}} \leqslant \frac{1}{\sqrt{d}}$ and noting that $16\log(2)\log\left(4\log(2)\right) * 2/\log(4/3) < 80$ we get,

$$\mathrm{QC}^1_{0,1/4}(\mathcal{F}_K) \in [d, 80d].$$

Moreover, Lemma 26 implies that for any algorithm Alg such that $m^1_{\mathrm{Alg}}(0, 1/4) \leqslant T$ there is a function $f \in \mathcal{F}_K$ such that,

$$\max_{f\in\mathcal{F}_K} \mathbb{E}_{\mathrm{Alg}}[\mathrm{Regret}(T, f)] \geqslant \frac{1}{128\epsilon_1^2} \geqslant \frac{d}{128}.$$

To finalize the proof we note that UCB satisfies an expected regret bound of order $8\sqrt{2T\log(T)}$ when using only actions in $\mathcal{A}_2$.

This finalizes the proof. ■

Appendix D. Useful Lemmas

Lemma 27 (Supporting Result) *Let $\delta \in (0, 1)$, $\sigma > 0$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. Then, the probability that,*

$$\mathbb{P}\left(|\xi| \geqslant \sigma\sqrt{2\log(2/\delta)}\right) \leqslant \delta.$$

In addition, let $\hat{\xi} = \frac{1}{n}\sum_{i=1}^n \xi_i$ be the empirical average of n independent samples from $\mathcal{N}(0, \sigma^2)$ then,

$$\mathbb{P}\left(\left|\hat{\xi}\right| \geqslant \sigma\sqrt{\frac{2\log(2/\delta)}{n}}\right) \leqslant \delta$$

Proof The random variable ξ is σ^2-subgaussian. Subgaussian random variables satisfy,

$$\mathbb{P}\left(|\xi| \geqslant \sigma\sqrt{2\log(2/\delta)}\right) \leqslant \delta.$$

the second result follows because the subgaussian parameter of $\hat{\xi}$ is $\frac{\sigma^2}{n}$. The result follows. ■

Lemma 28 *Let $a_1, \cdots, a_K$ and $b_1, \cdots, b_K$ be two sequences of nonnegative numbers. There exists an index $\tilde{i}$ such that,*

$$a_{\tilde{i}} \leqslant \frac{3\sum_{i=1}^K a_i}{K}, \qquad b_{\tilde{i}} \leqslant \frac{3\sum_{i=1}^K b_i}{K}$$

Proof For simplicity let $A = \sum_{i=1}^{K} a_i$ and $B = \sum_{i=1}^{K} b_i$. Let $\sigma(1), \cdots, \sigma(K)$ be the permutation permutation of $[K]$ such that $a_{\sigma(1)} \leqslant a_{\sigma(2)} \leqslant \cdots \leqslant a_{\sigma(K)}$. Each of $a_{\sigma(i)}$ for $i = 1, \cdots, \lfloor K/2 \rfloor$ satisfies,

$$a_{\sigma(i)} \leqslant 2A/K$$

This is because if this was not true then, $a_{\sigma(i)} > 2A/K$ for all $i \geqslant \lfloor K/2 \rfloor$ which would imply

$$\sum_{i=1}^{K} a_i \geqslant \sum_{i=\lfloor K/2 \rfloor + 1}^{K} a_{\sigma(i)} > (K - \lfloor K/2 \rfloor) \cdot 2A/K \geqslant A,$$

a contradiction. Consider now values $b_{\sigma(1)}, \cdots, b_{\sigma(\lfloor K/2 \rfloor)}$. The sum of these $\lfloor K/2 \rfloor$ values satisfies

$$\sum_{i=1}^{\lfloor K/2 \rfloor} b_{\sigma(i)} \leqslant \sum_{i=1}^{K} b_{\sigma(i)} = B.$$

Thus, it follows there is an index $\bar{i} \in \lfloor K/2 \rfloor$ such that,

$$b_{\sigma(\bar{i})} \leqslant \frac{B}{\lfloor K/2 \rfloor} \leqslant 3B/K.$$

Since $a_{\sigma(\bar{i})} \leqslant 2A/K < 3B/K$ the result follows by setting $\tilde{i} = \sigma(\bar{i})$. ■

Lemma 29 (Huber Bretagnolle Inequality) *Let P, Q be two measures and $\mathcal{E}$ be a measurable event. Then,*

$$P(\mathcal{E}) + Q(\mathcal{E}^c) \geqslant \exp\left(-\mathrm{KL}(P \parallel Q)\right)$$

As well as the divergence decomposition for Bandit problems,

Lemma 30 *Let $\nu = (P_1, \cdots, P_k)$ be the reward distributions associated with one k-armed bandit, and let $\nu' = (P'_1, \cdots, P'_k)$ be the reward distributions associated with another k-armed bandit. Fix some policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the probability measures on the canonical bandit model induced by the n-round interconnection of π and ν (respectively π and ν'). Then*

$$\mathrm{KL}(\mathbb{P}_\nu \parallel \mathbb{P}_{\nu'}) = \sum_{i=1}^{k} \mathbb{E}_\nu[T_i(n)]\mathrm{KL}(P_i \parallel P'_i)$$

where $T_i(n)$ is the random variable specifying how many times π selects action i up to round n.

See for example Theorem 14.2 and Lemma 15.1 in Lattimore and Szepesvári (2020) for a reference of Lemmas 29 and 30.