Proceedings of Machine Learning Research vol 195:1–38, 2023 36th Annual Conference on Learning Theory

Bandit Learnability can be Undecidable

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Liu Yang LIU.YANG0900@OUTLOOK.COM
Santai Technology Co., Ltd

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We pursue a general investigation into structured bandits. Specifically, for an abstract space $\mathcal{X}$, we suppose a true reward function f resides in a known, but arbitrary, function class $\mathcal{F}$. The algorithm may then pull a number of arms x (i.e., query for the value $f(x)$), and thereby attempts to identify an arm $\hat{x}$ of near-maximum reward: $f(\hat{x}) \geq \sup_x f(x) - \epsilon$. While special cases of this problem are well understood in the literature, our interest is in the possibility of a fully-general theory of bandit learnability, analogous to the PAC model for classification: that is, a theory which precisely characterizes which function classes $\mathcal{F}$ admit a learning algorithm guaranteed to identify a near-optimal arm within a bounded number of pulls.

Our main result in this regard is an illuminating impossibility result. Namely, there exist well-defined function classes $\mathcal{F}$ such that bandit learnability is *undecidable* within ZFC set theory. While such undecidability results have previously been shown for a certain abstractly-defined learning problem known as EMX, this is the first example of a natural or commonly-encountered learning problem (i.e., bandits) for which learnability can be provably undecidable. Our proof is based on establishing a (rather-sophisticated) equivalence between certain subfamilies of EMX learning problems and corresponding constructed bandit problems.

Despite this general undecidability result, we also establish new general results in special cases. Specifically, we characterize the optimal query complexity in the special case of binary-valued reward functions in terms of a combinatorial complexity measure related to the teaching dimension. We also present an extension to general bounded real-valued rewards, though in this case the upper bound is not always optimal. In the process, we also establish a separation between learnability by deterministic vs randomized learners. We instantiate the new complexity measures for several important families of function classes $\mathcal{F}$.

Keywords: Bandits, Undecidability, Learnability, Zeroth-order Optimization, Regret, Query Complexity

1. Introduction

Within the field of learning theory, the PAC framework has been instrumental in providing an abstract unifying perspective for understanding statistical learning of binary classifiers, and yields beautifully concise characterizations of learnability and sample complexity. Prior to the proposal of this framework (Vapnik and Chervonenkis, 1974; Valiant, 1984), the literature was largely a diverse collection of special case analyses (e.g., Cover, 1965). Once the abstract PAC framework entered the literature, such analyses could be unified, and further developments could easily be understood via abstract complexity measures.

In contrast, the present-day literature on *bandit* learning still remains a fractured menagerie of special cases (see Bubeck and Cesa-Bianchi, 2012), completely lacking any kind of abstract

unifying theory (though there have been recent attempts toward trying to fill this gap, notably Foster, Kakade, Qian, and Rakhlin, 2021a). In the bandit problem (in the well-specified and non-adversarial setting), there is a set of *arms* $\mathcal{X}$, and an unknown *reward function* $f^\star : \mathcal{X} \to [0,1]$. The learner may choose any arm $x_1 \in \mathcal{X}$ to "pull" (i.e., query), receive a reward r_1 (a random variable with mean $f^\star(x_1)$), choose another arm x_2, receive a reward r_2, and so on. The objective is either to identify an arm $\hat{x}$ with $\mathbb{E} f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon$ or to achieve low expected *regret* $T \sup_x f^\star(x) - \mathbb{E}\sum_{t=1}^T r_t = o(T)$. As a first step, in the present work, we will focus on the *noise-free* setting, where rewards for arm x are *equal* to the value $f^\star(x)$ of the reward function.

In analogy to the PAC framework, the natural formulation of an abstract theory of bandit learning is to consider a *function class* $\mathcal{F}$: a set of possible reward functions $\mathcal{X} \to [0,1]$. We say the set $\mathcal{F}$ is *learnable* in the bandit setting if there is an algorithm $\mathcal{A}$ and a function $M : (0,1) \to \mathbb{N}$ such that, for any $\epsilon \in (0,1)$, for any $f^\star \in \mathcal{F}$, after pulling at most $M(\epsilon)$ arms, the bandit algorithm returns an arm $\hat{x} \in \mathcal{X}$ with $\mathbb{E} f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon$. This is analogous to the notion of learnability in the PAC framework, based on having finite sample complexity.[1] We note that this is also equivalent to a *zeroth-order optimization* problem. We refer to the quantity $M(\epsilon)$ as the *query complexity* of the algorithm $\mathcal{A}$ for the function class $\mathcal{F}$.

Similarly, we can say a function class $\mathcal{F}$ is *no-regret learnable* in the bandit setting if there is an algorithm $\mathcal{A}$ and a function $\mathcal{R} : \mathbb{N} \to [0,\infty)$ with $\mathcal{R}(T) = o(T)$ such that, for any reward function $f^\star \in \mathcal{F}$ and any $T \in \mathbb{N}$, $T \sup_x f^\star(x) - \mathbb{E}\sum_{t=1}^T r_t \leq \mathcal{R}(T)$. As we show below, a class $\mathcal{F}$ is learnable if and only if it is no-regret learnable, so for our present discussion we will merely refer to learnability (in the sense of the first definition above).

Given this definition of learnability, the main theoretical question in this framework is the following:

Which classes $\mathcal{F}$ are learnable in the bandit setting?

By far the most commonly studied function class is the set $\mathcal{F} = [0,1]^{\mathcal{X}}$ of *all functions*. The first analysis of this setting is usually credited to Robbins (1952), with the optimal regret first identified by Lai and Robbins (1985). This function class gives rise to natural strategies for no-regret learning, such as the popular *Upper Confidence Bound* (UCB) strategy (Agrawal, 1995), which elegantly balances the opposing interests of *exploration* (searching for better arms) and *exploitation* (pulling arms that are known to give high rewards).

At this point, some readers may be thinking, "If we already have a theory of bandit learning for the class of *all* possible reward functions, isn't that already the most general theory we could ask for?" To answer this, note that *learnability* is a property of *classes* of functions. So a theory of learnability that only covers *a single* class $\mathcal{F}$ of functions is actually *not* general at all, even if that class is the largest possible class. Instead, we would want a theory of learnability to cover all possible classes of functions. Analogously, in the PAC framework for classification, we could easily prove theorems about learnability of the concept class of all possible binary functions $\mathcal{X} \to \{0,1\}$. But such theorems would be quite uninteresting. It would merely say that the class is learnable if and only if the space $\mathcal{X}$ is *finite*, and that the sample complexity scales linearly in $|\mathcal{X}|$. In contrast,

1. Note that, as in the abstract "VC theory" of PAC learnability, here we focus only on the rewards and number of arms pulled, setting aside the more nuanced issue of the computational complexity of selecting the arms. Indeed, our use of the term "algorithm" should be interpreted as merely requiring a sequence of mapping functions (possibly randomized), rather than intending some particular model of computation in which such a function could be implemented.

PAC/VC theory gives a *more general* theory of learnability, since it covers *all possible concept classes*, showing that any well-behaved concept class is learnable if and only if its VC dimension is finite (Vapnik and Chervonenkis, 1974). This way, even if the space $\mathcal{X}$ is infinite, there are still many interesting concept classes that are PAC learnable, such as linear separators or neural networks, rectangle classifiers, low-rank decision trees, etc. We know these are learnable because of the general theory relating finiteness of the VC dimension to learnability. Moreover, even when $\mathcal{X}$ is finite, having a sample complexity scaling linearly in $|\mathcal{X}|$ is often still unacceptable (e.g., when $\mathcal{X}$ is the Boolean cube), and thus the more-general theory of PAC learning with general function classes also provides a more satisfying *quantitative* analysis, replacing a linear dependence on $|\mathcal{X}|$ with a linear dependence on the VC dimension of the concept class (the two coincide only for the concept class of all binary functions).

Similarly, in the bandit setting, it is an easy observation that the class $\mathcal{F} = [0,1]^{\mathcal{X}}$ of all functions is only learnable if $\mathcal{X}$ is *finite*, and indeed that the query complexity, $M(\epsilon)$, scales linearly in $|\mathcal{X}|$. This remains as unsatisfying of a theory in the bandit setting as in the classification setting. Thus, a general theory of learnability requires us to broaden our perspective, to allow for *any* function class $\mathcal{F}$.[2]

There have been a few other function classes commonly studied in the literature, beside the class $[0,1]^{\mathcal{X}}$ of all functions. Perhaps the next most studied class $\mathcal{F}$ is the class of *linear* functions on a compact set $\mathcal{X} \subset \mathbb{R}^d$ (see Bubeck and Cesa-Bianchi, 2012, Chapter 5). In this case, the dependence on $|\mathcal{X}|$ is replaced by a dependence on d, the dimension. Another family of function classes studied in a substantial number of works are the classes $\mathcal{F}$ of *smooth* functions on a metric space $\mathcal{X}$ (Kleinberg, 2004; Kleinberg, Slivkins, and Upfal, 2008; Bubeck, Munos, Stoltz, and Szepesvári, 2011; Minsker, 2013) (typically, Lipschitz or Hölder classes). In this case, the optimal query complexity is quantified in terms of the smoothness parameters for the functions in $\mathcal{F}$ together with a notion of dimension for the metric space $\mathcal{X}$. Our general approach below may be viewed as related to certain techniques from some of these works, in that they involve estimation of level sets as a component in the optimization algorithm (see e.g., Minsker, 2013).

A somewhat more-recent attempt at a more-general perspective, allowing a broader family of function classes $\mathcal{F}$, is the work of Hashimoto, Yadlowsky, and Duchi (2018). Similarly to some of the above works on smoothness, the essential strategy of Hashimoto, Yadlowsky, and Duchi (2018) is to reduce the bandit problem to the problem of classifying points in $\mathcal{X}$ by whether they are included in a level set of the reward function $f^\star$. Their strategy requires this classification to be essentially *perfect*, so that they identify a subset of $\mathcal{X}$ that *definitely* contains all points where $f^\star$ is greater than some threshold τ, which they adjust over time in their algorithm. Based on this approach, they derive a query complexity bound based on the VC dimension and disagreement coefficient of the level sets of functions in $\mathcal{F}$. Their general theorem is restricted to the case of finite $\mathcal{X}$, and indeed their query complexity explicitly depends on $|\mathcal{X}|$. Nevertheless, their approach and techniques hint at a more general theory, for which they are able to state some special cases.

Very recently, a series of preprints by Foster, Kakade, Qian, and Rakhlin (2021a); Foster, Golowich, Qian, Rakhlin, and Sekhari (2022); Foster, Golowich, and Han (2023) have explored an impressively general approach to bandit learning they term *Estimation-to-Decisions*. Similar to the present work, their interest is in approaching a *general* theory of bandit learnability and learn-

2. Indeed, this same issue propagates to related settings, such as contextual bandits and reinforcement learning, where even in the works proposing general theories allowing general classes of reward functions, the regret bounds still exhibit a dependence on the number of possible *actions* (e.g., Foster, Rakhlin, Simchi-Levi, and Xu, 2021b).

ing complexity (either query complexity or regret). Their general analysis is expressed in terms of a quantity they term the *decision estimation coefficient*. We discuss this work in detail below, but for now mention that there remain gaps between their upper and lower bounds, both quantitative and qualitative, with potentially infinite gaps in query complexity in some cases. Thus, while impressively general, their work does not provide a *complete* characterization of bandit learnability.

Each of the above particular families of function classes has required a separate analysis. In the present work, we ask whether it is possible to answer the question of learnability in a *fully general* theory that captures *all* function classes $\mathcal{F}$: that is, whether it is possible to formulate a unified abstract theory of bandit learnability.

Main Result and Interpretation: As the main result of this work, we find that, in a sense, such a fully general theory is *impossible*: precisely, we prove that bandit learnability can be *undecidable* within the ZFC axioms. Intuitively, what this means is that the above state of affairs in the literature is, in a sense, unavoidable: that is, it seems likely that there simply *cannot* be a characterization of bandit learnability that is simultaneously *complete*, *explicit*, and *simple*, since any complete characterization of bandit learnability would need to be so complicated or implicitly-specified as to even be sensitive to esoteric considerations in set theoretic axioms. In light of this fact, it seems the direction of the literature may be forced to pivot away from seeking fully-general theories of bandit learnability, rather focusing on identifying important *subfamilies* of bandit learning problems for which complete characterizations of learnability are possible. In this work, we provide one (extremely simple) illustrative example of this: namely, *binary-valued* bandits. In addition to being an important observation for the study of bandit learnability, we note that our undecidability result is also the first example of a natural or commonly-encountered learning problem that can be provably undecidable, and as such also advances our understanding of undecidability of learning problems more broadly.[3]

1.1. Main Results

We formally state the main result of this work as follows.

Theorem 1 *For $\mathcal{X} = \mathbb{R}$, there is a bandit problem $(\mathcal{X}, \mathcal{F})$ such that, whether or not $(\mathcal{X}, \mathcal{F})$ is learnable is independent of the ZFC axioms.*

The interpretation of this is that there are concretely-definable bandit learning problems for which, if we augment the ZFC axioms with certain additional (compatible) axioms, then the problem is learnable, whereas if we augment ZFC with certain other (compatible) axioms, then the problem is *not* learnable. Thus, any theory based purely on ZFC cannot provide an answer to the learnability of such classes, and hence cannot be fully general.

Overview of the Proof: A detailed outline of the proof of Theorem 1 is provided in Section 2, followed by the formal proof in Section 3. Here we provide a brief non-technical overview. The broad approach of the proof of Theorem 1 is to build on earlier work of Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a), who argued undecidability of learnability for an instance of a

3. Previous work of Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a) found that an abstract problem they call *EMX learning* is undecidable. We discuss this at length below. Other recent works of Hanneke, Kontorovich, Sabato, and Weiss (2021) and Caro (2021) show a number of learnability questions are undecidable or unprovable within ZFC. However, the natures of those results are fundamentally different, as we discuss at length in Section 1.7.

learning problem they call *EMX*, for Expectation Maximization. Our proof of Theorem 1 provides a sophisticated construction of a bandit problem corresponding to any given EMX problem from a subfamily studied by Ben-David et al. (2019a), known as *union-bounded* classes. By establishing equivalences of learnability between corresponding instances from these two subfamilies of learning problems, the undecidability result of Ben-David et al. (2019a) carries over to the bandit setting.

The relation between the EMX and bandit problems at first seems non-obvious. EMX is a learning problem aiming to identify a binary function $\hat{h}$ in a class $\mathcal{H}$ guaranteeing $\mathbb{E}_{X\sim P}[\hat{h}(X)] \geq \sup_{h\in\mathcal{H}} \mathbb{E}_{X\sim P}[h(X)] - \epsilon$, based on a number $N(\epsilon)$ of i.i.d. samples $X_1, \ldots, X_{N(\epsilon)}$ from an unknown distribution P. In contrast, bandit learning aims to identify an arm $\hat{x} \in \mathcal{X}$ guaranteeing reward $f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon$, based on observed rewards $f^\star(x_t)$ for adaptively chosen arms $x_1, \ldots, x_n$, as the only source of information about the unknown reward function $f^\star \in \mathcal{F}$. The basic idea in our construction is that, for each possible distribution P in the EMX problem, we define an appropriate reward function in the corresponding bandit problem, so that each $h \in \mathcal{H}$ has a corresponding arm x_h in the bandit problem with $f^\star(x_h) \propto \mathbb{E}_{X\sim P}[h(X)] + O(1)$. This sets up a correspondence between the unknown objects in the two problems.

However, there are several complicating factors, which require a significantly more nuanced construction. To convert a learner for the EMX problem into a learner for the bandit problem, we need the ability to generate i.i.d. samples from the distribution P corresponding to the bandit problem's reward function. The bandit problem (as studied here) has inherently *deterministic* rewards $f^\star(x)$, which presents a challenge. We must therefore rely on the randomness of the learner to generate these samples. Toward this end, we construct an additional set of arms x_w whose $f^\star(x_w)$ reward values are such that, if a learner randomly samples one of these arms (with a carefully-chosen distribution), the reward value can be transformed to the value of a random sample from the distribution P (or a distribution close to P). On the other hand, to convert a bandit learner into an EMX learner, we must be careful to ensure that there is no additional structure in the bandit problem beyond this ability to sample from P: i.e., even for a learner that adaptively chooses its queries among x_w arms, we can still simulate the $f^\star(x_w)$ reward values using only the i.i.d. samples available in the EMX problem. To achieve this, we in fact construct an entire family of reward functions corresponding to each P, in a careful way guaranteeing that if we choose a reward function at random from this family, the rewards observed by the learner have an induced distribution that is again nearly the same as i.i.d. samples from P.

In addition to the above, another subtlety arising in establishing this second direction of the equivalence is that the arms x_h must have precise reward values $f^\star(x_h)$ in order to maintain (for the purpose of the first direction) the exact correspondence between the near-maximality of the reward $f^\star(\hat{x})$ of the returned arm $\hat{x} = x_{\hat{h}}$ in the bandit problem and the near-maximality of the score $\mathbb{E}_{X\sim P}[\hat{h}(X)]$ of the corresponding $\hat{h} \in \mathcal{H}$ for the EMX problem (noting that, since we are concerned with learnability, we need a single ϵ-independent construction $\mathcal{F}$ that establishes the correspondence simultaneously for all $\epsilon \in (0, 1)$). The problem with this is that, since these $f^\star(x_h)$ rewards must have a precise (hence non-random) correspondence to the scores $\mathbb{E}_{X\sim P}[h(X)]$, if the bandit learner attempts to pull an arm x_h in the process of learning (i.e., not merely as its return value), we have no way to simulate the exact reward value $f^\star(x_h)$ using only the i.i.d. samples available in the EMX problem; for instance, since we have no restrictions on the bandit learner in this reduction, we cannot assume it would respond reasonably to replacing $f^\star(x_h)$ by a random approximation based on estimating $\mathbb{E}_{X\sim P}[h(X)]$ from a finite sample. While this issue may potentially be unsolvable for general EMX problems, we are able to address the issue in the case of a subfamily

of EMX problems, known as *union-bounded*. In this case, it suffices to convert the bandit learner into a weak *monotone compression scheme* (see below for the definition). We achieve the latter by effectively suppressing all information from these precise reward values $f^\star(x_h)$ in the execution of the bandit learner, arguing that its sequence of queried arms can still be used to construct such a compression scheme. This is by far the most technically involved portion of the argument, requiring a careful breakdown of cases and events in the (randomized) executions of the bandit learner, to eventually identify a strict subset of samples that will be (non-randomly) mapped to a function in $\mathcal{H}$ whose support includes the entire data set. With this monotone compression scheme in hand, we can then construct an EMX learner. Altogether, this establishes the equivalence of EMX learnability of a union-bounded class and bandit learnability of the constructed corresponding bandit problem. Noting that the undecidability of EMX learnability was indeed established for a union-bounded class $\mathcal{H}$, the above equivalence therefore extends the undecidability of EMX learnability to the bandit setting.

Undecidability of No-Regret Learnability: Although Theorem 1 is expressed in the PAC / optimization variant of bandit learning, we also prove the following equivalence to no-regret learnability.

Theorem 2 *Any $(\mathcal{X}, \mathcal{F})$ is learnable in the bandit setting* if and only if *it is no-regret learnable in the bandit setting.*

Thus, Theorem 1 also establishes undecidability of no-regret learnability in the bandit setting.

Corollary 1 *For $\mathcal{X} = \mathbb{R}$, there is a bandit problem $(\mathcal{X}, \mathcal{F})$ such that, whether or not $(\mathcal{X}, \mathcal{F})$ is no-regret learnable is independent of the ZFC axioms.*

1.2. Results for Binary-valued Rewards

While Theorem 1 indicates that a fully general theory of bandit learnability is essentially impossible (within ZFC), it leaves open the possibility of interesting special cases of families of bandit learning problems having general characterizations of learnability for function classes in the family. Indeed, as discussed above, Theorem 1 might be interpreted as revealing that there cannot be a fully general characterization of bandit learnability which is both simple and explicit (i.e., easy to interpret or evaluate), since any characterization would necessarily be sensitive to nuances of set-theoretic axioms. Faced with this situation, we suggest that the aim of the literature on bandit learnability, and phrasing of results therein, should pivot toward identifying and understanding interesting precisely-defined families of bandit learning problems for which simple and explicit characterizations of bandit learnability are possible.

To complement the above negative result, we develop one (incredibly simple) illustrative example of such a family, for which a simple complete characterization of bandit learnability *is* possible: namely, the case of *binary-valued* reward functions, that is, the family of all function classes $\mathcal{F}$ where every $f \in \mathcal{F}$ satisfies $\mathsf{image}(f) \subseteq \{0, 1\}$. We refer to such $(\mathcal{X}, \mathcal{F})$ as a *binary-valued bandit problem*. For this special case, we characterize which classes are learnable, and moreover, do so by defining a simple dimension which characterizes learnability in the bandit setting with binary-valued rewards. More specifically, we consider deterministic and randomized learners separately, and characterize the optimal query complexity for both cases. Interestingly, this also reveals a separation between the two: i.e., there are binary-valued bandit problems that are learnable by randomized learners but not learnable by deterministic learners. This is noteworthy, since there are some bandit learning algorithms in the literature that are deterministic (e.g., UCB).

Deterministic learners: To characterize the optimal query complexity of deterministic learning, we consider the following definition.

Definition 2 *Define the zero-teaching dimension of $\mathcal{F}$, denoted $\tau^0_{\mathcal{F}}$, as the smallest $t \in \mathbb{N}$ such that there exist $x_1, \ldots, x_t \in \mathcal{X}$ with*

$$\min_{f \in \mathcal{F} \setminus \{\mathbf{0}\}} \max_{1 \le i \le t} f(x_i) = 1,$$

where $\mathbf{0}$ is the all-zero function (which may or may not be in $\mathcal{F}$). If no such finite t exists, define $\tau^0_{\mathcal{F}} = \infty$.

The name zero-teaching dimension stems from a relation to the literature on teaching complexity (Goldman and Kearns, 1995), where $\tau^0_{\mathcal{F}}$ can equivalently be defined as the teaching dimension of $\mathbf{0}$ with respect to the class $\mathcal{F} \cup \{\mathbf{0}\}$. In words, $\tau^0_{\mathcal{F}}$ is the smallest number of points such that one of them must have $f^\star$ value 1, if there exists any point in $\mathcal{X}$ of value 1. We have the following result.

Theorem 3 *Any binary-valued bandit problem $(\mathcal{X}, \mathcal{F})$ is learnable by a deterministic algorithm if and only if $\tau^0_{\mathcal{F}} < \infty$. Moreover, the optimal query complexity $M(\epsilon)$ achievable by deterministic algorithms satisfies, $\forall \epsilon \in (0,1)$, $M(\epsilon) = \tau^0_{\mathcal{F}} - 1$.*

Indeed, it is rather obvious that the optimal query complexity for deterministic learners is is $\Theta(\tau^0_{\mathcal{F}})$.

Randomized learners: On the other hand, to understand learnability of binary-valued bandits by *randomized* learners, we consider the following definition.[4]

Definition 3 *Define the* maximin volume *of $\mathcal{F}$, denoted by $\tilde{\sigma}_{\mathcal{F}}$, as*

$$\tilde{\sigma}_{\mathcal{F}} = \sup_{P} \inf_{f \in \mathcal{F} \setminus \{\mathbf{0}\}} P(x : f(x) = 1),$$

where P ranges over all probability measures on $\mathcal{X}$.

We show that this simple quantity determines learnability in binary-valued bandit problems.

Theorem 4 *Any binary-valued bandit problem $(\mathcal{X}, \mathcal{F})$ is learnable if and only if $\tilde{\sigma}_{\mathcal{F}} > 0$. Moreover, the optimal query complexity $M(\epsilon)$ satisfies*

$$\frac{1-\epsilon}{\tilde{\sigma}_{\mathcal{F}}} - 1 \le M(\epsilon) \le \left\lceil \frac{1}{\tilde{\sigma}_{\mathcal{F}}} \ln\left(\frac{1}{\epsilon}\right) \right\rceil - 1.$$

While the above quantity gives a simple characterization of binary-valued bandit learnability, it turns out a related more-involved quantity provides an *exact* quantitative characterization of the query complexity. Specifically, consider the following definition.

4. To be formal in specifying what kind of randomization is allowed, we suppose there is a σ-algebra defined on $\mathcal{X}$, and for any given reward function the randomized learner's sequence of queries and return value should be jointly measurable under the product σ-algebra. Since the σ-algebra also informs which probability measures are valid in Definitions 3 and 4, the results are valid for any choice of this σ-algebra.

Definition 4 *Define the* randomized zero-teaching dimension *of $\mathcal{F}$, denoted $\tilde{\tau}^0_{\mathcal{F}}(\epsilon)$, as the smallest $t \in \mathbb{N}$ such that, there exists a sequence $x_1, \ldots, x_t$ of $\mathcal{X}$-valued random variables such that, $\forall f \in \mathcal{F} \setminus \{\mathbf{0}\}$,*

$$\mathbb{P}(\exists i \in \{1, \ldots, t\} : f(x_i) = 1) \geq 1 - \epsilon.$$

If no such finite t exists, define $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) = \infty$.

These two quantities are related via the following basic lemma.

Lemma 5 *If $\tilde{\sigma}_{\mathcal{F}} > 0$, then*

$$\frac{1-\epsilon}{\tilde{\sigma}_{\mathcal{F}}} \leq \tilde{\tau}^0_{\mathcal{F}}(\epsilon) \leq \left\lceil \frac{1}{\tilde{\sigma}_{\mathcal{F}}} \ln\left(\frac{1}{\epsilon}\right) \right\rceil.$$

Moreover, $\tilde{\sigma}_{\mathcal{F}} = 0$ if and only if $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) = \infty$.

As with the result for deterministic learners above, the fact that $\tilde{\tau}^0_{\mathcal{F}}(\epsilon)$ characterizes the optimal query complexity is rather obvious, given the simplicity of the binary-valued bandit scenario. Nevertheless, the result sheds light on the existence of special subfamilies of bandit problems which seem to avoid the negative results discussed in the preceding sections. We have the following result.

Theorem 5 *Any binary-valued bandit problem $(\mathcal{X}, \mathcal{F})$ is learnable if and only if $\tau^0_{\mathcal{F}}(\epsilon) < \infty$ for all $\epsilon > 0$. Moreover, for all $\epsilon \in (0, 1)$, the optimal query complexity $M(\epsilon)$ satisfies*

$$M(\epsilon) = \tilde{\tau}^0_{\mathcal{F}}(\epsilon) - 1.$$

Gaps between deterministic and randomized learnability: Interestingly, the optimal query complexity of randomized learners can be vastly smaller than that of deterministic learners. Indeed, there are function classes that are not even learnable by deterministic learners in the bandit setting, but which are learnable with a modest query complexity by randomized learners.

Example 1 *As a simple example of this, consider $\mathcal{X} = [0, 1]$ (equipped with the σ-algebra of Lebesgue-measurable sets), and the class $\mathcal{F}$ of all $f : [0, 1] \to \{0, 1\}$ having $|\{x : f(x) = 0\}| < \infty$. Note that $\tau^0_{\mathcal{F}} = \infty$, since for any finite sequence $x_1, \ldots, x_t$, there is a function $\mathbb{1}_{[0,1]\setminus\{x_1,\ldots,x_t\}} \in \mathcal{F}$ which is 0 on $x_1, \ldots, x_t$ (and 1 everywhere else). On the other hand, $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) = 1$ for any $\epsilon \in (0, 1)$, since taking $x_1 \sim \text{Uniform}(0, 1)$ yields $\mathbb{P}(f(x_1) = 1) = 1$ for any $f \in \mathcal{F}$.*

1.3. Results for General Real-valued Rewards

We can also extend the above results to provide general upper bounds on the query complexity of bandit learning for real-valued reward functions. In this case, in light of Theorem 1, we may not hope for a concise upper bound that is always optimal, since any optimal characterization of query complexity would need some aspect that varies with different additions to ZFC, and hence we would expect its definition to be somewhat involved. Nonetheless, we propose general analyses, which provide upper bounds on the query complexity of bandit learning, for deterministic and randomized learners, respectively, and which we instantiate for several concrete function classes below.

Deterministic learners: We begin with a complexity measure for deterministic learners. For any $c \in [0,1]$, define

$$\mathcal{F}_c = \left\{ f \in \mathcal{F} : \sup_{x \in \mathcal{X}} f(x) \geq c \right\}.$$

We propose the following definition, representing an extension of the zero-teaching dimension of $\mathcal{F}$ to real-valued functions, based on the *level sets* of the functions..

Definition 6 *For any $\epsilon \in (0,1)$, define the* level-set teaching dimension $\tau_{\mathcal{F}}(\epsilon)$ *as the smallest $t \in \mathbb{N}$ such that, for any $c \in [0,1]$ with $\mathcal{F}_c \neq \emptyset$, there exist $x_1, \ldots, x_t \in \mathcal{X}$ with*

$$\inf_{f \in \mathcal{F}_c} \max_{1 \leq i \leq t} f(x_i) \geq c - \epsilon.$$

If no such finite t exists, define $\tau_{\mathcal{F}}(\epsilon) = \infty$.

This quantity is directly inspired by applying the reasoning of the $\tau^0_{\mathcal{F}}$ complexity for binary-valued rewards, extended to the real-valued case by considering the *level sets* of the functions in the class. We have the following result.

Theorem 6 (Informal) *Any bandit problem $(\mathcal{X}, \mathcal{F})$ has query complexity $M(\epsilon)$ for deterministic learners satisfying $M(\epsilon) = O(\tau_{\mathcal{F}}(\epsilon)/\epsilon)$.*

While this upper bound is not always optimal (which is not surprising, in light of Theorem 1), we do instantiate the bound for various interesting examples below in Section 1.4.

Randomized learners: As we did for binary bandits, we can also extend the analysis to *randomized* learners, which can provide significantly stronger guarantees.

Definition 7 *For any $\epsilon \in (0,1)$, define the* maximin level-set volume *of $\mathcal{F}$, denoted by $\tilde{\sigma}_{\mathcal{F}}(\epsilon)$, as*

$$\tilde{\sigma}_{\mathcal{F}}(\epsilon) = \sup_{P} \inf_{c \in (0,1]} \inf_{f \in \mathcal{F}_c} P(x : f(x) \geq c - \epsilon),$$

where P ranges over all probability measures on $\mathcal{X}$.

For simplicity, if $\mathcal{F}_c = \emptyset$, we define $\inf_{f \in \mathcal{F}_c}$ to always evaluate to 1.

Theorem 7 *For any bandit problem $(\mathcal{X}, \mathcal{F})$, the optimal query complexity $M(\epsilon)$ satisfies*

$$M(\epsilon) \leq \frac{2}{\epsilon} \left\lceil \frac{1}{\tilde{\sigma}_{\mathcal{F}}(\epsilon/4)} \ln\left(\frac{2}{\epsilon}\right) \right\rceil.$$

As we did for binary bandits, we prove this by first relating $\tilde{\sigma}_{\mathcal{F}}(\epsilon)$ to a more-involved but more-directly relevant quantity, defined as follows.

Definition 8 *For any $\epsilon, \delta \in (0,1)$, define the* randomized level-set teaching dimension $\tilde{\tau}_{\mathcal{F}}(\epsilon, \delta)$ *as the smallest $t \in \mathbb{N}$ such that, for any $c \in [0,1]$ with $\mathcal{F}_c \neq \emptyset$, there exists a sequence $x_1, \ldots, x_t$ of $\mathcal{X}$-valued random variables with*

$$\inf_{f \in \mathcal{F}_c} \mathbb{P}\left(\max_{1 \leq i \leq t} f(x_i) \geq c - \epsilon \right) \geq 1 - \delta.$$

For any given c, the sequence $x_1, \ldots, x_t$ of random variables satisfying the above requirement is called a randomized $(c - \epsilon, \delta)$-level specifying set *for $\mathcal{F}_c$. If no such t exists, define $\tilde{\tau}_{\mathcal{F}}(\epsilon, \delta) = \infty$.*

The quantities $\tilde{\tau}_{\mathcal{F}}(\epsilon, \delta)$ and $\tilde{\sigma}_{\mathcal{F}}(\epsilon)$ are related via the following lemma.

Lemma 9 *For any $\epsilon, \delta \in (0,1)$,*

$$\frac{1-\delta}{\tilde{\sigma}_{\mathcal{F}}(\epsilon)} \leq \tilde{\tau}_{\mathcal{F}}(\epsilon, \delta) \leq \left\lceil \frac{1}{\tilde{\sigma}_{\mathcal{F}}(\epsilon)} \ln\left(\frac{1}{\delta}\right) \right\rceil.$$

Due to Lemma 9, to prove Theorem 7, it suffices to prove the following bound in terms of $\tilde{\tau}_{\mathcal{F}}(\epsilon, \delta)$.

Theorem 8 *For any bandit problem $(\mathcal{X}, \mathcal{F})$, the optimal query complexity $M(\epsilon)$ satisfies*

$$M(\epsilon) \leq \frac{2\tilde{\tau}_{\mathcal{F}}(\epsilon/4, \epsilon/2)}{\epsilon}.$$

1.4. Examples

To begin, we consider the well-studied case of $\mathcal{F}$ the set of *all functions* $\mathcal{X} \to [0,1]$. In this case, our proposed dimension $\tau_{\mathcal{F}}(\epsilon)$ captures the well-known fact that the optimal query complexity is linear in $|\mathcal{X}|$.

Example 2 *Consider any finite $\mathcal{X}$ and $\mathcal{F} = [0,1]^{\mathcal{X}}$, the set of* all functions $\mathcal{X} \to [0,1]$. *In this case, $\tau_{\mathcal{F}}(\epsilon) = |\mathcal{X}|$. To see this Note that, for any $\mathcal{X}'$ that is a strict subset of $\mathcal{X}$, there is a function $f \in \mathcal{F}$ equal $\mathbb{1}_{\mathcal{X} \setminus \mathcal{X}'}$: that is, f is 0 on $\mathcal{X}'$ and 1 on $\mathcal{X} \setminus \mathcal{X}'$. Thus, for any $c \in (\epsilon, 1]$, this function f is in $\mathcal{F}_c$ (its maximum value is $1 \geq c$), and yet its maximum value on $\mathcal{X}'$ is $0 < c - \epsilon$. Since the requirement in Definition 6 is always satisfied for $t = |\mathcal{X}|$ (by definition of $\mathcal{F}_c$), this implies $\tau_{\mathcal{F}}(\epsilon) = |\mathcal{X}|$.*

Next we consider the (also well-studied) case of $\mathcal{F}$ the set of *linear* functions

Example 3 *Consider $\mathcal{X} = \mathbb{S}^d$, the origin-centered unit sphere in $\mathbb{R}^d$, for some $d \in \mathbb{N}$, and $\mathcal{F} = \{x \mapsto w^\top x : w \in \mathbb{S}^d\}$. In this case, $\tau_{\mathcal{F}}(\epsilon) = \epsilon^{1-d}$.*

Every $f \in \mathcal{F}$ has $\sup_x f(x) = 1$. So $\mathcal{F}_c = \mathcal{F}$. Consider then the most-constraining case for the requirement in Definition 6: $c = 1$. Any $x_1, \ldots, x_t$ satisfying this requirement must be an ϵ-cover of the angles, hence has size at least ϵ^{1-d}.

Interestingly, this is an example where $\tau_{\mathcal{F}}(\epsilon)$ does not *recover the optimal query complexity. In particular, it is known that $O(d)$ query complexity is achievable (Bubeck and Cesa-Bianchi, 2012). Indeed, this is not hard to see. Any linearly independent $x_1, \ldots, x_d$ will uniquely identify the reward function, so that we can output the precise maximizing arm.*

We remark that, at the expense of a somewhat more-involved definition of the complexity measure, we can define a more-advanced variant of $\tau_{\mathcal{F}}(\epsilon)$ that captures this example as well. Specifically, rather than requiring the existence of a fixed *sequence $x_1, \ldots, x_t$ in Definition 6, we could instead define a set of* branching *sequences: $x_{()}, x_{(y_1)}, x_{(y_1,y_2)}, \ldots, x_{(y_1,\ldots,y_{t-1})}$, such that for every $f \in \mathcal{F}_c$, inductively letting $y_1 = f(x_{()})$ and $y_i = f(x_{(y_1,\ldots,y_{i-1})})$ for $i \geq 2$, we have $\max\{f(x_{()}), f(x_{(y_1)}), \ldots, f(x_{(y_1,\ldots,y_{t-1})})\} \geq c - \epsilon$. Defining $\tilde{\tau}_{\mathcal{F}}(\epsilon)$ as the smallest t for which such a set of branching sequences exists, for every $c \in [0,1]$, we can replace $\tau_{\mathcal{F}}(\epsilon)$ with $\tilde{\tau}_{\mathcal{F}}(\epsilon)$ in Theorem 6 and it will remain valid. Unlike $\tau_{\mathcal{F}}(\epsilon)$, this complexity $\tilde{\tau}_{\mathcal{F}}(\epsilon)$ captures the fact that the bandit learning algorithm may select its exploration points adaptively: i.e., that some information can be extracted even from points that do not yield a new best-arm-so-far, to inform which arm to*

try next. Indeed, this modified definition of the complexity measure is (almost tautologically) an optimal characterization of query complexity for any bandit problem $(\mathcal{X}, \mathcal{F})$.

Moreover, for the linear bandit problem, it is clear that $\tilde{\tau}_{\mathcal{F}}(\epsilon) \leq d + 1$*, so that this complexity measure captures the optimal query complexity of linear bandits.*

Example 4 *Let* $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ *be the unit Euclidean ball in* $\mathbb{R}^d$ *for some* $d \in \mathbb{N}$*, and fix any* $L \geq 1$ *and* $\alpha \in (0, 1]$*, and define the class of* Hölder smooth *functions:*

$$\mathcal{F} = \left\{ f : \sup_{x, x' \in \mathcal{X}} \frac{|f(x) - f(x')|}{\|x - x'\|^\alpha} \leq L \right\}.$$

In this case, $\tau_{\mathcal{F}}(\epsilon) = \left\lceil \left(\frac{L}{\epsilon}\right)^{d/\alpha} \right\rceil$*. For* $f \in \mathcal{F}_c$*, it suffices to get a point* x' *within distance* $(\epsilon/L)^{1/\alpha}$ *of the point* x *having* $f(x) \geq c$ *to guarantee* $f(x') \geq c - \epsilon$*. Thus, choosing an* $(\epsilon/L)^{1/\alpha}$*-cover of* $\mathcal{X}$ *suffices to satisfy the criterion in Definition 6. The stated expression on* $\tau_{\mathcal{F}}(\epsilon)$ *is the size of such a cover. On the other hand, any smaller set of points cannot be an* $(\epsilon/L)^{1/\alpha}$ *cover, so that for such a smaller set* $x_1, \ldots, x_t$ *there exists a point* $x \in \mathcal{X}$ *with none of* $x_1, \ldots, x_t$ *within distance* $(\epsilon/L)^{1/\alpha}$*, and we can find a function* $f \in \mathcal{F}_c$ *with* $f(x) = c$ *and every other point* x' *defined as* $\max\{c - L\|x' - x\|^\alpha, 0\}$*, in which case none of* $x_1, \ldots, x_t$ *will have* $f(x_i) \geq c - \epsilon$*.*

We also remark that this value of $\tau_{\mathcal{F}}(\epsilon)$ *matches the known query complexity of this problem (e.g., Kleinberg, Slivkins, and Upfal, 2008).*

1.5. Relation to the Disagreement-Based Approach

The analysis of Hashimoto, Yadlowsky, and Duchi (2018) expresses a result in terms of the disagreement coefficient, a quantity originally introduced in (Hanneke, 2007). Specifically, for $\mathcal{X}$ a finite set and any function class $\mathcal{F}$, they let $\mathcal{H} = \{h_{f,c} : f \in \mathcal{F}, c \in [0, 1]\}$, where $h_{f,c}(x) = \mathbb{1}[f(x) > c]$ is a *superlevel set* of f.[5] Let P_X be a uniform distribution over $\mathcal{X}$, and let[6]

$$\theta = \sup_{h \in \mathcal{H}} \sup_{r \geq 1/|\mathcal{X}|} \frac{P_X(\mathrm{DIS}(\mathrm{B}(h, r)))}{r},$$

where $\mathrm{B}(h, r) = \{h' \in \mathcal{H} : P_X(x : h'(x) \neq h(x)) \leq r\}$ is the r-ball centered at h, and $\mathrm{DIS}(\mathrm{B}(h, r)) = \{x : \exists h' \in \mathrm{B}(h, r), h'(x) \neq h(x)\}$ is the region of disagreement of the r-ball. Additionally, let V denote the VC dimension of $\mathcal{H}$ (Vapnik and Chervonenkis, 1974). They propose an algorithm which maintains weights $p^{(t)}(x)$ (summing to 1) over the arms $x \in \mathcal{X}$ in rounds $t = 1, \ldots, T$. On each round it queries n arms. Their conclusion, after T rounds, is expressed as a lower bound on $p^{(T)}(x^*)$, where $x^* = \mathrm{argmax}_{x \in \mathcal{X}} f^\star(x)$. Specifically, in their Theorem 3, they state such a lower bound in terms of T, under a condition that n be of a given sufficient size, and the bound is guaranteed to be valid with probability at least $1 - \delta$. While such a result is essentially of a different type than the PAC-style query complexity bounds studied here (they are more closely analogous to "Exact" or "PExact" learning guarantees), they have an immediate implication for the type of query complexity studied here: namely, we can consider the implication for the number of queries sufficient for their algorithm to guarantee $p^{(T)}(x^*) > \frac{1}{2}$, so that choosing

5. They use sublevel sets, but this is clearly equivalent.
6. They considered a supremum over all $r > 0$, but for finite $\mathcal{X}$ their results remain valid with merely $r \geq 1/|\mathcal{X}|$, which never increases the value of θ, yet makes it never larger than $|\mathcal{X}|$.

$\hat{x} = \text{argmax}_x\, p^{(T)}(x)$ will achieve $f^\star(\hat{x}) = \sup_x f^\star(x)$ with probability at least $1-\delta$. Setting $\delta = \epsilon$, this implies $\mathbb{E}[f^\star(\hat{x})] \geq \sup_x f^\star(x) - \epsilon$. After T rounds, the total number of queries by their algorithm is nT. Setting their lower bound on $p^{(T)}(x^*)$ to be greater than $\frac{1}{2}$ and minimizing the value of nT over the various parameters in their Theorem 3 subject to this constraint, we arrive at a guarantee on the query complexity of their algorithm:

$$O\left(\theta\left(V\log(\theta) + \log\left(\frac{1}{\epsilon}\right) + \log\log(|\mathcal{X}|)\right)\log(|\mathcal{X}|)\right).$$

We can recover a nearly-identical result as an (often loose) upper bound on the query complexity in our Theorem 6. We first note a relation between $\tau_{\mathcal{F}}(\epsilon)$ and θ. To start, notice that $\tau_{\mathcal{F}}(\epsilon)$ is upper bounded by $\text{TD}(\mathcal{H})$, where $\text{TD}(\mathcal{H})$ is the *teaching dimension* of the superlevel sets $\mathcal{H}$ (Goldman and Kearns, 1995). To see this, note that every $f \in \mathcal{F}$ has $h_{f,1}(x) = 0$ everywhere, since f is bounded by 1. Thus, $\mathcal{H}$ contains the everywhere-zero function $\mathbf{0}$. Thus, for any $c \in [0,1]$, the teaching dimension of $\{\mathbf{0}\} \cup \{h_{f,c-\epsilon} : f \in \mathcal{F}\}$ is at most $\text{TD}(\mathcal{H})$ (by monotonicity of TD). Let $x_1, \ldots, x_t$ be a minimal set such that any function in $\{h_{f,c-\epsilon} : f \in \mathcal{F}\}$ with value 0 on all of $x_1, \ldots, x_t$ must be equal 0 on all of $\mathcal{X}$: note that

$$t \leq \text{TD}(\{\mathbf{0}\} \cup \{h_{f,c-\epsilon} : f \in \mathcal{F}\}) \leq \text{TD}(\mathcal{H}).$$

Then every $f' \in \mathcal{F}$ for which $h_{f',c-\epsilon}$ is *not* everywhere 0 must have some x_i with $h_{f',c-\epsilon}(x_i) = 1$: that is, $f'(x_i) \geq c - \epsilon$. Hence, $x_1, \ldots, x_t$ satisfy the requirement in Definition 6. Since this is true for *every* $c \in [0,1]$, we conclude that $\tau_{\mathcal{F}}(\epsilon) \leq \text{TD}(\mathcal{H})$.

We conclude by recalling a relation between $\text{TD}(\mathcal{H})$ and θ from Theorem 5 of Wiener, Hanneke, and El-Yaniv (2015), which implies $\text{TD}(\mathcal{H}) = O(\theta(V\log(\theta) + \log\log(|\mathcal{X}|))\log(|\mathcal{X}|))$. For completeness, we include a brief proof here. Let $h^* \in \mathcal{H}$ be any function with teaching dimension $\text{TD}(h^*, \mathcal{H}) = \text{TD}(\mathcal{H})$. Consider randomly sampling with replacement from $\text{Uniform}(\mathcal{X})$ to produce a sequence of random samples $\tilde{x}_1, \tilde{x}_2, \ldots$. For $T = O(|\mathcal{X}|\log(|\mathcal{X}|))$, with probability at least $1/2$, we will have at least one copy of every x in $\mathcal{X}$ within $\tilde{x}_1, \ldots, \tilde{x}_T$. Let $V_0 = \mathcal{H}$ and for each $t \in \{1, \ldots, T\}$, let $V_t = \{h \in \mathcal{H} : \forall i \leq t, h(\tilde{x}_i) = h^*(\tilde{x}_i)\}$ and $Q_t = \mathbb{1}[\tilde{x}_t \in \text{DIS}(V_{t-1})]$. This sequence of Q_t values describes the behavior of the well-studied CAL active learning algorithm (Cohn, Atlas, and Ladner, 1994). A well-known bound on the sum of Q_t values, established by Hanneke (2011, 2014), shows that, with probability at least $3/4$,

$$\sum_{t=1}^{T} Q_t = O(\theta(V\log(\theta) + \log\log(|\mathcal{X}|))\log(|\mathcal{X}|)).$$

Moreover, since every t with $\tilde{x}_t \notin \text{DIS}(V_{t-1})$ has full agreement with $h^*(\tilde{x}_t)$ within V_{t-1}, we inductively have that equivalently $V_t = \{h \in \mathcal{H} : \forall i \leq t \text{ with } Q_i = 1, h(\tilde{x}_i) = h^*(\tilde{x}_i)\}$. Thus, with probability at least $1/4$ (by the union bound), the set V_T is the set of functions in $\mathcal{H}$ that agree with h^* on a set S of size $O(\theta(V\log(\theta) + \log\log(|\mathcal{X}|))\log(|\mathcal{X}|))$ (namely, the $\tilde{x}_i$ points having $Q_i = 1$), *and* every point x in $\mathcal{X}$ appears at least once in the sequence $\tilde{x}_1, \ldots, \tilde{x}_T$. The latter implies $V_T = \{h^*\}$. Thus, S is a teaching set for h^* with respect to $\mathcal{H}$, hence has size at least $\text{TD}(\mathcal{H})$ (by our choice of h^*). Altogether, we have

$$\tau_{\mathcal{F}}(\epsilon) = O(\theta(V\log(\theta) + \log\log(|\mathcal{X}|))\log(|\mathcal{X}|)).$$

This indicates that our analysis is essentially no worse than that of Hashimoto, Yadlowsky, and Duchi (2018) (up to log factors). On the other hand, one can give simple examples where the result in our Theorem 6 is significantly smaller than the result of Hashimoto et al. (2018).

Example 5 *Fix any $x_0 \in \mathcal{X}$, and consider the set $\mathcal{F}$ of functions that have $f(x_0) = 1$, and otherwise can take* any *values on $\mathcal{X} \setminus \{x_0\}$. Then $\tau_{\mathcal{F}}(\epsilon) = 1$, since x_0 always satisfies the criterion in Definition 6. On the other hand, $\theta = |\mathcal{X}| - 1$ and $V = |\mathcal{X}| - 1$, so that the bound of Hashimoto, Yadlowsky, and Duchi (2018) is $\Omega(|\mathcal{X}|^2 \log(|\mathcal{X}|))$. Thus, our analysis based on $\tau_{\mathcal{F}}(\epsilon)$ can offer an arbitrarily strong improvement over the result of Hashimoto, Yadlowsky, and Duchi (2018).*

As an additional remark, we can also extend Example 1 to reveal a large gap between the result of Hashimoto, Yadlowsky, and Duchi (2018) and the optimal query complexity from Theorems 4 and 5 for binary-valued bandits. Specifically, consider the following example.

Example 6 *Consider any finite set $\mathcal{X}$ and let $\mathcal{F}$ be the set of all function $f : \mathcal{X} \to \{0, 1\}$ having $\sum_{x \in \mathcal{X}} f(x) \geq (1/2)|\mathcal{X}|$. For this example, recalling the notation from Section 1.5, for $h = \mathbf{1}$ the constant-1 function (which is an element of $\mathcal{H}$ in this example), any $r \in (0, 1/2)$ has $\mathcal{F} \subseteq \mathrm{B}(h, r)$, and hence $\mathrm{DIS}(\mathrm{B}(h, r)) = \mathcal{X}$, so that $\theta = |\mathcal{X}|$. Since the VC dimension of $\mathcal{H}$ is $(1/2)|\mathcal{X}|$ here, the result of Hashimoto, Yadlowsky, and Duchi (2018) only provides a query complexity $\Omega(|\mathcal{X}|^2 \log(|\mathcal{X}|))$. On the other hand, letting $x \sim \mathrm{Uniform}(\mathcal{X})$, any $f \in \mathcal{F}$ has $\mathbb{P}(f(x) = 1) \geq \frac{1}{2}$, so that $\tilde{\sigma}_{\mathcal{F}} \geq \frac{1}{2}$. Together with Theorem 4, we see that $\mathcal{F}$ is in fact learnable with a modest query complexity $O\big(\log\big(\frac{1}{\epsilon}\big)\big)$ that is independent of $|\mathcal{X}|$.*

1.6. Comparison to the Decision-Estimation Coefficient Approach

A recent impressively thorough preprint of Foster, Kakade, Qian, and Rakhlin (2021a) attempts to approach the problem of providing a general characterization of bandit learnability (and indeed, even broader families of decision problems, including contextual bandits and reinforcement learning), in terms of a quantity they term the *Decision-Estimation Coefficient* (DEC). The definition of the DEC is in fact partly analogous to the definition of the *disagreement coefficient* (see Section 1.5)[7], though with important modifications making it more suitable for reward-maximization decision problems.

For the case of bandit learning, Foster, Kakade, Qian, and Rakhlin (2021a) formulate a general framework, which allows for *noisy* bandit problems. Specifically, they specify a bandit problem, for a space of arms $\mathcal{X}$, via a *model class* $\mathcal{M}$: a set of functions $M : \mathcal{X} \to \Delta(\mathbb{R})$, where $\Delta(\mathbb{R})$ denotes the set of probability measures on $\mathbb{R}$. Their underlying assumption is that there exists $M^\star \in \mathcal{M}$ representing the *true* reward distribution: that is, a bandit learner that pulls an arm $x_t \in \mathcal{X}$ on round t receives reward r_t sampled from the distribution $M^\star(x_t)$ (where r_t is conditionally independent of all past rewards, given x_t). The objective remains to obtain a near-maximum *expected* reward of the returned arm $\hat{x}$ (where this expectation is over both the randomness of the algorithm and the randomness of the reward that would be received upon pulling arm $\hat{x}$). Such a model class $\mathcal{M}$ naturally induces a function class $\mathcal{F} = \{f^M : M \in \mathcal{M}\}$, where $f^M(x) = \mathbb{E}_{r \sim M(x)}[r]$, so that the objective may be stated as achieving $\mathbb{E}[f^{M^\star}(\hat{x})] \geq \sup_x f^{M^\star}(x) - \epsilon$. Foster et al. (2021a) are largely focused on achieving low *regret* under such model classes, though their results also translate

7. Indeed, Foster, Kakade, Qian, and Rakhlin (2021a) prove formal relations between these quantities.

naturally to the reward-maximization setting as we have chosen to focus in the present work (see Section 4).

They propose a generic strategy they term *Estimation-to-Decision* (E2D), which iteratively prunes the model class essentially based on estimates of the distance of each model to $M^\star$: that is, they prune a model M from the class if a uniform confidence bound on its (Hellinger) distance to $M^\star$ reveals it is far from $M^\star$ (in expectation, under a distribution on $\mathcal{X}$ constructed in the algorithm). They then provide both an upper bound on the regret of E2D and a lower bound on the minimax optimal regret, in terms of (respectively) two variants of the Decision-Estimation Coefficient dec_γ. Without getting into too many fine details of these variations (which are not particularly consequential for the purpose of our comparison here), the essential definition is as follows. For $\gamma > 0$, and any model $\bar{M}$,

$$\mathrm{dec}_\gamma(\mathcal{M}, \bar{M}) = \inf_{p \in \Delta(\mathcal{X})} \sup_{M \in \mathcal{M}} \mathbb{E}_{x \sim p}\left[\sup_{x'} f^M(x') - f^M(x) - \gamma D^2_{\mathrm{H}}(M(x), \bar{M}(x))\right],$$

where $D^2_{\mathrm{H}}(P, Q) = \int \left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2$ is the Hellinger distance between distributions (supposing both are absolutely continuous with respect to some reference measure).

We summarize their upper and lower bounds, in an oversimplified form, for the purpose of comparison; the interested reader is referred to the original source for the formal statements. Their lower bound on the minimax regret, up to time T, is of a form roughly

$$\sup_{\gamma > 0} \min\left\{\sup_{\bar{M} \in \mathcal{M}} \mathrm{dec}_\gamma(\mathcal{M}, \bar{M})T, \gamma\right\}. \tag{1}$$

Complementing this, their upper bound on the regret of E2D is roughly of the form

$$\inf_{\gamma > 0} \max\left\{\sup_{\bar{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma(\mathcal{M}, \bar{M})T, \gamma \cdot \mathrm{est}(\mathcal{M}, T)\right\}, \tag{2}$$

where $\mathrm{co}(\mathcal{M})$ is the set of *mixtures* of models in $\mathcal{M}$, and $\mathrm{est}(\mathcal{M}, T) := \inf_{\epsilon > 0} \log \mathcal{N}(\mathcal{M}, \epsilon) + \epsilon^2 T$ is a measure of the complexity of uniform estimation of distances between models, where $\mathcal{N}(\mathcal{M}, \epsilon)$ is a bound on the ϵ^2-covering number of $\mathcal{M}$ under *uniform* Hellinger distance: $\sup_x D^2_{\mathrm{H}}(M(x), M'(x))$. Their theory offers some refinements of this upper bound as well, which do not significantly change our comparison below.

While these results are quite general, and certainly cover many interesting scenarios not addressed in the present work (e.g., stochastic rewards), we note here that they are not particularly well-suited to handling the case of *deterministic* rewards, which is the main focus of the present work. In particular, they do not capture the basic results of Section 1.2 on binary bandits.

To see this, in the case of the upper bound (2), note that for any class $\mathcal{F}$ of binary bandits, the corresponding model class $\mathcal{M} = \{M_f : f \in \mathcal{F}\}$ consists of models M_f which, for any arm x, simply have a single point mass at $f(x)$. Thus, since any distinct $M_f, M_{f'} \in \mathcal{M}$ have some x with $f(x) \neq f'(x)$, we may observe that $\sup_x D^2_{\mathrm{H}}(M(x), M'(x)) = 1$. Therefore, any infinite class $\mathcal{F}$ of binary bandits have a corresponding model class $\mathcal{M}$ with $\mathcal{N}(\mathcal{M}, \epsilon) = \infty$ for all $\epsilon > 0$, and hence $\mathrm{est}(\mathcal{M}, T) = \infty$. This means the upper bound (2) is always vacuously infinite for any infinite binary bandit class $\mathcal{F}$. A simple example of such a class, for which our analysis in Section 1.2 yields

finite query complexity (indeed, zero), is the class $\mathcal{F}$ of all binary functions f on $\mathcal{X} = \mathbb{N}$ satisfying $f(1) = 1$; this class has $\tau^0_{\mathcal{F}} = 1$, by taking $x_1 = 1$.

In the case of the lower bound (1), as a simple illustrative example of a large gap, consider $\mathcal{X} = \mathbb{N}$ and $\mathcal{F} = \{\mathbb{1}_{\{x\}} : x \in \mathcal{X}\}$: the class of *singletons*, which are 1 on exactly one arm. The corresponding model class $\mathcal{M}$ simply takes $M(x)$ as a point-mass at $f^M(x)$: i.e., $M(x)$ deterministically produces reward $f^M(x)$, where f^M may be any function in $\mathcal{F}$. It is a rather trivial matter to argue that $\tilde{\sigma}_{\mathcal{F}} = 0$ for this problem, and hence Theorem 4 implies this class is *not* learnable (a fact which is indeed rather obvious). For a non-learnable class, the optimal regret should grow as $\Theta(T)$. As noted above, $\mathcal{N}(\mathcal{M}, \epsilon) = \infty$ for this class, so that their upper bound in (2) is infinite. On the other hand, for the lower bound (1), for any $\bar{M} \in \mathcal{M}$, choosing p to be supported entirely on the point x with $f^{\bar{M}}(x) = 1$, any $M \in \mathcal{M}$ distinct from $\bar{M}$ has $\sup_{x'} f^M(x') - f^M(x) = 1$ and $D^2_{\mathrm{H}}(M(x), \bar{M}(x)) = 1$, so that $\mathrm{dec}_\gamma(\mathcal{M}, \bar{M}) \leq 1 - \gamma$. Thus, the regret lower bound in (1) is at most $\sup_{\gamma > 0} \min\{(1-\gamma)T, \gamma\} < 1$, far from the optimal regret $\Theta(T)$. In particular, the lower bound (1) does not imply the non-learnability of this model class.

In a very recent follow-up work, which was developed and published concurrently with (and independently from) the present work, Foster, Golowich, and Han (2023) propose a refinement of the DEC, which is able to improve some of the above gaps. They also propose a variant which directly addresses the near-maximization problem, rather than needing to convert from regret analysis. In particular, these new results are effectively able to replace $\sup_{\bar{M} \in \mathcal{M}}$ with $\sup_{\bar{M} \in \mathrm{co}(\mathcal{M})}$ in (1), which indeed recovers non-learnability of singletons. However, as they discuss, even with these refinements, there exist model classes $\mathcal{M}$ for which there remain arbitrarily large gaps between upper and lower bounds, due to the appearance of a notion of estimation complexity in the upper bound (analogous to $\mathrm{est}(\mathcal{M}, T)$ above). As discussed above, our Theorem 1 suggests that such gaps are inevitable and unavoidable for any simple and explicit theory stated in such generality. Aside from this, it is also desirable to have a clean, direct, and sharp analysis, for such a natural (albeit simple) special case as the binary bandit setting, as presented in our Section 1.2.

1.7. Comparisons to Recent Works on Learnability Under ZFC

In addition to Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a) (which we discuss at length below), other recent works have also studied limitations of ZFC in regard to learnability. Hanneke, Kontorovich, Sabato, and Weiss (2021) study the problem of *universal consistency* in metric spaces: that is, supervised learning algorithms for classification that guarantee their expected risk converges to the best possible (the Bayes risk) for all distributions. They propose a learning algorithm specified in terms of a metric associated with an abstract instance space $\mathcal{X}$, and show the algorithm is guaranteed to be universally consistent under any metric space for which there exists a universally consistent learning algorithm. They further provide a description of precisely which metric spaces admit such learning algorithms (namely, *essentially separable* metric spaces). They then argue that, under ZFC, it is impossible to prove the existence of metric spaces failing to satisfy this description: that is, it is impossible to prove the existence of metric spaces that do not admit universally consistent learners. We note that this is a very different kind of result compared to the present work, since it does not concern learnability under a *particular* space $\mathcal{X}$, but rather whether there exist metric spaces $\mathcal{X}$ where universal learning fails. In contrast, the undecidability result in our Theorem 1 is based on an explicit specification of a space $\mathcal{X}$ and function class $\mathcal{F}$ for which bandit learnability is independent of ZFC.

Another recent work, of Caro (2021), provides a number of claims about learnability statements whose truth values are undecidable (in either the computational or logical senses). However, we must note that the nature of those results is fundamentally different from the type of result given by our Theorem 1. The main point to note is that we define the function class $\mathcal{F}$ as a fixed *set*, explicitly defined in the proof of Theorem 1, of which we have full knowledge when determining learnability. In contrast, what is revealed by the work of Caro (2021) appears to be an interesting discussion of how undecidability can arise from various restrictions to the type of *access* to the function class a decider (algorithmic or human) may have. In particular, Caro (2021) considers a scenario in which we do not know a priori *what the function class is* that we are tasked with deciding learnability of, but rather (effectively) are permitted a kind of black-box query access to the class: plugging in parameter values θ and instances x, and receiving as output the value $f_\theta(x)$ of some function f_θ in the function class $\mathcal{F}$. To be clear, in that setting, the decider does not have any knowledge of which functions f are indexed by which parameter values θ, or any other structural information about the function class: only black-box queries to the map $(\theta, x) \to f_\theta(x)$. Indeed, Caro (2021) shows that, given only this black-box generic access, it is even impossible to decide whether the function class is infinite or instead contains only *one function* (redundantly indexed by all possible parameter values).[8]

In contrast, our Theorem 1 presents a single explicitly defined function class, and we may design any learning algorithm with full knowledge of the class, analyzing its behavior under various target functions. As revealed by our proof, the undecidability of learnability (Theorem 1) arises due to the influence of set-theoretic axioms on what kinds of learning algorithms *can exist* for this particular class. This is analogous to the result of Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a), and indeed our proof proceeds via relating the bandit setting to the setting studied by Ben-David et al. (2019a).

2. Outline of the Proof of Undecidability

The main technical innovation in this work is constructing an equivalence between members of a subfamily of bandit problems and a subfamily of EMX learning problems. This requires a rather sophisticated construction, and must address a number of challenges.

In EMX learning (Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff, 2019a), there is a domain $\mathcal{W}$ and a set $\mathcal{H}$ of binary-valued functions $h : \mathcal{W} \to \{0, 1\}$. There is an unknown distribution P on $\mathcal{W}$, and the learner is given as input a data set of n i.i.d. samples from P. Its goal is to select an element $\hat{h} \in \mathcal{H}$ with $P(x : \hat{h}(x) = 1) \geq \sup_{h \in \mathcal{H}} P(x : h(x) = 1) - \epsilon$.

Note that, at first, the relation between EMX learning and bandit learning seems completely non-obvious. In an EMX learning problem, a learner has access to samples from a distribution P, whereas bandit problems have no notion of data or distribution. Moreover, bandit problems are interactive, allowing algorithms that alter their behavior based on past observations, whereas EMX learners are given an algorithm-independent data set input.

The main idea underlying our construction is that, for a given EMX learning problem $(\mathcal{W}, \mathcal{H})$, we construct a bandit problem $(\mathcal{X}, \mathcal{F})$ where the reward functions in $\mathcal{F}$ are based on the *distribu-*

8. The result does appear to raise a very intriguing question regarding what it means to know "what the function class is". However, we believe in the case of the class $\mathcal{F}$ in our Theorem 1, at least we can be confident that the class $\mathcal{F}$ can be understood to a far more sophisticated extent than would be permitted by the kind of black-box queries permitted in the work of Caro (2021).

tion P in the EMX problem: different distributions P in the EMX problem correspond to different reward functions in the bandit problem. This idea becomes natural if we also set up a correspondence between *arms* in $\mathcal{X}$ and *functions* h in $\mathcal{H}$, so that for a given P and h, for the reward function f corresponding to P, there is an arm x_h whose reward $f(x_h)$ is (essentially) $P(x : h(x) = 1)$. Thus, a near-optimal choice of arm x_h in the bandit problem corresponds to a near-optimal choice of function h in the EMX problem, and vice versa.

While the above is a natural approach, it leaves out an important aspect of these problems: namely, the learning algorithm goes about finding a near-optimal arm or function in completely different ways in the two types of problems. Specifically, in the EMX problem, the learner is given an i.i.d. data set as input, whereas in the bandit problem, the learner sequentially selects arms to observe the rewards. In addressing these differences lies most of the technical challenge in establishing the equivalence.

The high-level summary of our approach to the first difference (the i.i.d. data) is to create a set of arms $\mathcal{X}_{\mathcal{W}}$ within $\mathcal{X}$ where, by appropriate randomization, the reward values in this region effectively simulate i.i.d. samples from P (here we restrict to $\mathcal{W} = [0, 1]$). To be clear, the rewards remain deterministic functions of $\mathcal{X}$ according to a function in $\mathcal{F}$; the randomization is in either a (simulated) adversary's selection of reward function or the learner's selection of arms, depending which direction of the equivalence we aim to establish. We accompany this with an increase of rewards for the above arms x_h to $(1/3)(2 + P(x : h(x) = 1))$, with all arms in $\mathcal{X}_{\mathcal{W}}$ having rewards at most $1/3$, to ensure the near-optimal arms in the bandit problem still correspond to near-optimal functions in the EMX problem.

Our solution to the second difference (the sequential nature of the bandit problem, and ability of the bandit learner to observe rewards, particularly for arms x_h as above) is more technical. Focusing on the sub-family of EMX problems that are *union bounded*, we argue that if the corresponding constructed bandit problem is learnable, then the bandit algorithm can be used to construct a weak *monotone compression scheme* for the EMX problem (i.e., a compression scheme of size $n - 1$ that reconstructs to an element of $\mathcal{H}$ that is a superset of the data), which Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a) have shown implies that the EMX problem is learnable. The intuitive idea is that the only case where the ability to get rewards for arms x_h is helpful is when the selected arm x_h corresponds to an h capturing more of the support of the distribution P than is already known from the observed rewards from $\mathcal{X}_{\mathcal{W}}$ (representing the i.i.d. samples). The ability of an algorithm to select such an informative arm x_h is thus basically equivalent to an ability to *guess* a never-before-seen point in the support of P. This ability is the essence of what makes EMX learning possible. Following this intuition, we define a reconstruction function for a compression scheme which accumulates all of the points from the supports of all of the functions h for which the algorithm pulls arm x_h, while suppressing any information in the reward values other than what the algorithm has already observed from the arms in $\mathcal{X}_{\mathcal{W}}$ it has pulled. In converting this into a compression scheme, the idea is to use samples from a given data set as values of the rewards for the arms in $\mathcal{X}_{\mathcal{W}}$, in which case the above positive outcomes would yield a correct "guess" at identifying one of the samples the algorithm did not actually observe as a reward. To complete the construction, since a compression scheme is, by definition, a *deterministic* function, we must convert this randomized procedure into a deterministic one; for this, we select values in $\mathcal{W}$ to be included in a reconstruction set which would be at least somewhat *likely* to be included in the randomly-constructed set. By a careful analysis of cases and events in the execution of this algorithm in a simulated scenario involving a random subset of the data, we are able to argue that

(1) in any data set of a particular size $m+1$, there exists a subset of a smaller size M such that the above reconstruction function, applied to the subset, would output a set which includes at least one other element of the $m+1$ points, and (2) this fact enables us to construct a compression scheme which, given the $m+1$ samples, identifies a subset of size m from which it will reconstruct an element of h with the full set of $m+1$ samples in its support. The definition and analysis of this compression scheme represents the most technically involved portion of the proof.

Putting all the above pieces together yields the equivalence between this family of EMX problems and a corresponding family of bandit problems. In particular, since there is an EMX problem within this family whose learnability is known to be undecidable (Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff, 2019a), this establishes that learnability of the corresponding bandit problem is also undecidable, which establishes Theorem 1.

2.1. Outline of the Paper

The rest of the paper is organized as follows. In Section 3 we present the proof of Theorem 1, by establishing an equivalence between a subfamily of EMX learning problems and a subfamily of bandit learning problems. In Section 4 we argue that no-regret learnability in the bandit setting is equivalent to learnability of a near-optimal arm. As a consequence, this establishes that no-regret learnability is also sometimes undecidable in the bandit setting.

In contrast, Section 5 considers the special case of *binary-valued* reward functions. Section 6 extends the idea underlying this analysis to real-valued functions. This yields an upper bound on the query complexity. In contrast to the binary-valued case, this upper bound is sub-optimal in some cases, which might not be surprising in light of the undecidability result in Theorem 1. We conclude with several important open problems in Section 7.

3. Independence from ZFC

In this section, we establish a correspondence between a family of EMX learning problems and a family of bandit learning problems, so that each such bandit problem is learnable if and only if its corresponding EMX problem is learnable. The undecidability of bandit learnability for an explicitly constructed class (Theorem 1) will follow from this correspondence.

3.1. Constructing Bandit Problems from EMX Problems

We focus on the case $\mathcal{W} = (0, 1/3)$.[9] As in (Ben-David et al., 2019a), we take the power set σ-algebra. This choice is not particularly important for the problem specification, since the EMX learning problem restricts to countably-supported probability measures anyway, and the class $\mathcal{H}$ of interest only considers finitely-supported sets, which would thus be measurable under any reasonable σ-algebra. However, as was true in the result of (Ben-David et al., 2019a), it *is* an important choice for the undecidability theorem, since it admits certain learning rules which would not be measurable under more-restrictive σ-algebras.

Before stating our formal construction, for completeness, we discuss the technical meaning of what a "learning algorithm" is, in the context of this construction. Specifically, for the EMX problem, a learning algorithm is a function $\hat{h} : \mathcal{W}^* \to \{0, 1\}$, with the constraint that, for any

9. The original space studied by Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a) was based on $\mathcal{W} = \mathbb{R}$. However, the example can be mapped to $(0, 1/3)$ without loss; this choice serves to simplify our construction.

$n \in \mathbb{N}$ and $w_1, \ldots, w_n \in \mathcal{W}$, $\hat{h}(w_1, \ldots, w_n, \cdot) \in \mathcal{H}$; there are no further restrictions on this function (recalling that we take the power set σ-algebra on $\mathcal{W}$). Clearly we can equivalently interpret this as a function mapping $w_1, \ldots, w_n$ to an element of $\mathcal{H}$.

For the bandit problem, there are two natural formulations, depending on whether the algorithm receives as input a *budget* of how many queries it can make, or whether the algorithm itself determines a stopping criterion. For the purpose of studying the (target-independent) sample complexity, there is no significant difference between these (since we can always choose the budget to be the value of the sample complexity itself), so for simplicity we suppose the algorithm receives as input a query budget. In this case, a learning algorithm may be regarded as a (possibly random) sequence of functions $x_t^{M,\epsilon} : (\mathcal{X} \times [0,1])^{t-1} \to \mathcal{X}$, $t \in \{1, \ldots, M+1\}$, where M is the budget. For a given $M \in \mathbb{N} \cup \{0\}$ and $\epsilon \in (0,1)$, the algorithm's sequence of arm pulls are $x_1^{M,\epsilon} = x_1^{M,\epsilon}()$, $x_2^{M,\epsilon} = x_2^{M,\epsilon}(x_1^{M,\epsilon}, f^\star(x_1^{M,\epsilon}))$, $x_3^{M,\epsilon} = x_3^{M,\epsilon}(x_1^{M,\epsilon}, f^\star(x_1^{M,\epsilon}), x_2^{M,\epsilon}, f^\star(x_2^{M,\epsilon}))$, and so on, up to $x_{M+1}^{M,\epsilon}$, which is interpreted as the arm *returned* by the algorithm, which we will typically abbreviate with $\hat{x}$, when M and ϵ are clear from the context. We note that, for simplicity, we suppose the bandit learner always queries up to its budget M number of arms; this is without loss of generality, since any algorithm that may terminate early can be represented here by an algorithm that simply makes additional queries up to the budget M and ignores these additional return values, returning the same $\hat{x}$ as the early-stopped algorithm.

Let $(\mathcal{W}, \mathcal{H})$ be any EMX learning problem which is *union bounded*: that is, for every $h_1, h_2 \in \mathcal{H}$, $\exists h_3 \in \mathcal{H}$ with $h_1 \cup h_2 \subseteq h_3$. We first construct the set $\mathcal{X}$ of arms for the corresponding bandit problem. There is a countable subset $\mathcal{X}_\mathcal{W}$, and another subset $\mathcal{X}_\mathcal{H} = \{x_h : h \in \mathcal{H}\}$, where every x_h is distinct, and where $\mathcal{X}_\mathcal{W}$ and $\mathcal{X}_\mathcal{H}$ are disjoint. We define $\mathcal{X} = \mathcal{X}_\mathcal{W} \cup \mathcal{X}_\mathcal{H}$.

Next, we describe the set $\mathcal{F}$ of reward functions. Let Π be the set of all countably-supported probability measures P on $\mathcal{W}$. For each $P \in \Pi$, there will be a subset $\mathcal{F}_P$ of reward functions, such that we define $\mathcal{F} = \bigcup_{P \in \Pi} \mathcal{F}_P$. Thus, to complete the construction, it remains only to define the subsets $\mathcal{F}_P$.

Before getting into the details, the high level idea of the construction is that, for each reward function f in $\mathcal{F}_P$, the arms x_h in $\mathcal{X}_\mathcal{H}$ return values $f(x_h) = (1/3)(P(h) + 2)$ for each $h \in \mathcal{H}$, where we use the notation $P(h) = \mathbb{E}_{X \sim P}[h(X)]$ for brevity. Thus, finding a good arm among $\mathcal{X}_\mathcal{H}$ has a correspondence to finding a near-maximal $P(h)$ value in the EMX problem; the choice of $(1/3)(P(h) + 2)$ rather than $P(h)$ ensures that the near-optimal arms in the bandit problem are all in $\mathcal{X}_\mathcal{H}$, rather than $\mathcal{X}_\mathcal{W}$ (which, as we explain below, will always have reward values at most $1/3$), so that (without loss of generality) we can focus on bandit learners that always choose $\hat{x} \in \mathcal{X}_\mathcal{H}$ (so that there is a corresponding output $\hat{h}$ for the corresponding EMX problem).

However, there are additional challenges in relating the EMX and bandit problems. Even given the above correspondence of optimal outputs, the EMX problem differs from the bandit problem in two important ways. First, it is not possible to exactly evaluate $P(h)$ in the EMX problem; it can only be estimated from data. In contrast, a bandit learner can query any x_h and observe $f(x_h) = (1/3)(P(h) + 2)$, effectively allowing it to evaluate $P(h)$ for any $h \in \mathcal{H}$. This poses a challenge in converting a bandit learner to an EMX learner, which we must address in the proof. It appears there is no general way to resolve this for general EMX problems (e.g., even the ability to estimate $P(h)$ from samples is not good enough to resolve this discrepancy). However, we note that the classes $\mathcal{H}$ of interest in the work of Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a) fall into a special subfamily, known as *union-bounded* classes, in which case Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a) showed that EMX learnability is equivalent to the

existence of a weak *monotone compression scheme*: namely, existence of a finite k and a function $\rho : \mathcal{W}^k \to \mathcal{H}$ such that, given any sequence S of $k+1$ points contained in some $h \in \mathcal{H}$ (i.e., $h(w) = 1$ for all $w \in S$), there exists a subsequence S' of S, of size k, such that $\rho(S')$ contains S. Based on this feature of union-bounded classes, we show that for the corresponding bandit problem (as outlined here), if the bandit problem is learnable, the bandit learner can be converted into a weak monotone compression scheme (of the above type), hence establishing learnability of the EMX problem.

A second difference from EMX learning also poses a challenge: namely, the EMX learner has access to i.i.d. data sampled from P. In contrast, the bandit learning problem, as studied in its simplest form in this work, has deterministic rewards. To address this, we propose to use the set of arms $\mathcal{X}_{\mathcal{W}}$ to effectively simulate having access to i.i.d. data. This construction is rather delicate, since we need to make it so that the conversion from EMX learner to bandit learner can simulate i.i.d. samples by choosing random elements from $\mathcal{X}_{\mathcal{W}}$ (which is accomplished by setting the rewards to be a map representing a random variable on $\mathcal{W}$), but on the other hand, for the conversion from bandit learner to EMX learner, we must ensure that the bandit learner cannot rely on the structure of rewards in $\mathcal{X}_{\mathcal{W}}$ beyond what samples would give, so that reward values can be simulated from the samples available in the EMX problem. In other words, we want that the region $\mathcal{X}_{\mathcal{W}}$ is *only* useful as a data source, and has no helpful relation to the target reward function $f^\star$ beyond this. To achieve the latter guarantee, we define $\mathcal{F}_P$ as a family of reward functions, allowing us to assign different reward values to different arms in $\mathcal{X}_{\mathcal{W}}$, and in the proof we effectively employ a *prior* distribution over (a finite subset of) $\mathcal{F}_P$ so that, for whatever queries the bandit learner would choose in $\mathcal{X}_{\mathcal{W}}$, answering according to the value of a sample from the EMX problem has roughly the same distribution as the reward value for a random reward function sampled from this prior, thus allowing the EMX samples to be used for answering queries in this region.

We now turn to the formal construction. As mentioned, $\mathcal{F} = \bigcup\{\mathcal{F}_P : P \in \Pi\}$, so that we will focus on describing the $\mathcal{F}_P$ sets. We describe the behavior of the functions in $\mathcal{F}_P$ in two parts. We first describe a function $f_{P,\mathcal{H}}$ on $\mathcal{X}_{\mathcal{H}}$, and then a set $\mathcal{F}_{P,\mathcal{W}}$ of functions on $\mathcal{X}_{\mathcal{W}}$. The final set $\mathcal{F}_P$ will then be formed by combination:

$$\mathcal{F}_P = \{x \mapsto f_{\mathcal{W}}(x)\mathbb{1}[x \in \mathcal{X}_{\mathcal{W}}] + f_{P,\mathcal{H}}(x)\mathbb{1}[x \in \mathcal{X}_{\mathcal{H}}] : f_{\mathcal{W}} \in \mathcal{F}_{P,\mathcal{W}}\}.$$

We begin with the function $f_{P,\mathcal{H}} : \mathcal{X}_{\mathcal{H}} \to [0,1]$, the simpler of the two parts, specified as follows:

$$\forall h \in \mathcal{H}, f_{P,\mathcal{H}}(x_h) = (1/3)(P(h) + 2).$$

The set $\mathcal{F}_{P,\mathcal{W}}$ is quite a bit more involved. Let $\mathcal{X}_{\mathcal{W}}^{(1)} = \{z_1^{(1)}, z_2^{(1)}, \ldots\}$ and $\mathcal{X}_{\mathcal{W}}^{(2)} = \{z_1^{(2)}, z_2^{(2)}, \ldots\}$ be disjoint countably infinite sets such that $\mathcal{X}_{\mathcal{W}} = \mathcal{X}_{\mathcal{W}}^{(1)} \cup \mathcal{X}_{\mathcal{W}}^{(2)}$. Let Σ_P denote the set of all sequences $\mathbf{w} = \{w_i\}_{i=1}^{\infty}$ satisfying the property: denoting by $\hat{P}_T^{\mathbf{w}}$ the empirical measure, specified by $\hat{P}_T^{\mathbf{w}}(\{w\}) = \frac{1}{T}\sum_{i=1}^{T}\mathbb{1}[w_i = w]$, it holds that

$$\lim_{T\to\infty}\sup_{w\in\mathcal{W}}\left|P(\{w\}) - \hat{P}_T^{\mathbf{w}}(\{w\})\right| = 0.$$

In particular, since P is countably supported, there indeed exist such sequences w_i, and moreover, this also implies the above property guarantees that

$$\lim_{T\to\infty}\left\|P - \hat{P}_T^{\mathbf{w}}\right\| = 0,$$

where $\|\cdot\|$ here denotes the total variation distance. For any such sequence $\mathbf{w} \in \Sigma_P$, let $\mathcal{T}_i(\mathbf{w})$ denote the set of all $T \in \mathbb{N}$ with $T \geq 3$ such that

$$\left\| P - \hat{P}_T^{\mathbf{w}} \right\| \leq \frac{1}{i}.$$

The set $\mathcal{F}_{P,\mathcal{W}}$ is defined as the set of all functions $f_P^{\mathbf{w},\mathbf{T}} : \mathcal{X}_{\mathcal{W}} \to [0,1]$, $\mathbf{w} = \{w_i\}_{i=1}^{\infty} \in \Sigma_P$, $\mathbf{T} = \{T_i\}_{i=1}^{\infty}$ with $T_i \in \mathcal{T}_i(\mathbf{w})$, defined as follows. For any $i \in \mathbb{N}$,

$$f_P^{\mathbf{w},\mathbf{T}}(z_i^{(1)}) = w_i$$

and

$$f_P^{\mathbf{w},\mathbf{T}}(z_i^{(2)}) = \frac{1}{T_i}.$$

In particular, note that $f_P^{\mathbf{w},\mathbf{T}}(x) \leq 1/3$ for every $x \in \mathcal{X}_{\mathcal{W}}$.

In the context of the proof below, the purpose of the points $z_i^{(1)}$ is for their reward values to represent the samples in the EMX problem. However, there is a difficulty in using them to simulate i.i.d. samples from P, to feed into an EMX learner, since the distribution P may be spread arbitrarily thinly on its support, requiring us to use a simulated distribution involving a large, but unknown, support size, in order to approximate sampling from P.[10] To resolve this, we allow the rewards on the $z_i^{(2)}$ points to effectively communicate how thinly this distribution is spread (or rather, how many w_i values to include in the support of a uniform distribution) enabling us to sample from a distribution that approximates P arbitrarily well.

This completes the construction of the set $\mathcal{F}$. We next state the formal claim, from which our Theorem 1 will follow.

Theorem 9 *For $\mathcal{W} = (0, 1/3)$ and any union-bounded $\mathcal{H}$ of finite sets, the bandit problem $(\mathcal{X}, \mathcal{F})$ constructed above is learnable if and only if the EMX problem $(\mathcal{W}, \mathcal{H})$ is learnable.*

Before stating the proof, we first need the following lemma from Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a,b).

Lemma 10 *(Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff, 2019b, Corollary 4) Any given union-bounded EMX problem $(\mathcal{W}, \mathcal{H})$ is EMX learnable if and only if there exists an $m \in \mathbb{N}$ for which there exists an $(m+1) \to m$ monotone compression scheme for $\mathcal{H}$: that is, a function $\eta : \mathcal{W}^m \to \mathcal{H}$ such that $\forall h \in \mathcal{H}$ and $x_1, \ldots, x_{m+1} \in h$, $\exists i_1, \ldots, i_m \in \{1, \ldots, m+1\}$ such that $\{x_1, \ldots, x_{m+1}\} \subseteq \eta(x_{i_1}, \ldots, x_{i_m})$.*

We are now ready for the proof of Theorem 9.

Proof of Theorem 9 We prove the "if" direction of the theorem for *any* EMX problem $(\mathcal{W}, \mathcal{H})$ as follows. Fix any $\epsilon \in (0, 2/3)$. Suppose $\mathcal{A}_{\text{emx}}$ is a learning algorithm for the EMX problem $(\mathcal{W}, \mathcal{H})$, guaranteed to return $\hat{h}$ with $\mathbb{E}P(\hat{h}) \geq \sup_{h \in \mathcal{H}} P(h) - 3\epsilon/2$ based on $M(3\epsilon/2)$ samples

10. It is tempting to simply sample from a uniform distribution on a continuous interval, and define the reward function as mapping uniform samples to P-distributed random variables. However, to the best of our knowledge, there is no way to implement this measurably, since the EMX learner these random variables would be composed with is only required to be measurable under the product σ-algebra.

from P, for any $P \in \Pi$: i.e., with sample complexity $M(3\epsilon/2)$. Let $m = M(3\epsilon/2)$. We construct a bandit learning algorithm based on $\mathcal{A}_{\text{emx}}$ as follows. Suppose $f^\star \in \mathcal{F}$ is the target reward function. Let $i_\epsilon = \lceil 2m/(3\epsilon) \rceil$. First pull the arm $z_{i_\epsilon}^{(2)}$ and let $T = 1/f^\star(z_{i_\epsilon}^{(2)})$. Sample $W_1, \dots, W_m$ i.i.d. $\text{Uniform}(\{z_1^{(1)}, \dots, z_T^{(1)}\})$ and pull these arms to observe rewards $X_i = f^\star(W_i)$, $i = 1, \dots, m$. Let $\hat{h} = \mathcal{A}_{\text{emx}}(X_1, \dots, X_m)$. Output the arm $\hat{x} = x_{\hat{h}}$ as the return value of the bandit learner.

We argue that this achieves ϵ-optimal expected reward, as follows. Let $P \in \Pi$ be the distribution for which $f^\star \in \mathcal{F}_P$, and let $\mathbf{w} = \{w_i\}_{i=1}^\infty \in \Sigma_P$ and $\mathbf{T} = \{T_i\}_{i=1}^\infty$ with $T_i \in \mathcal{T}_i(\mathbf{w})$, for which $f^\star(x) = f_P^{\mathbf{w},\mathbf{T}}(x)\mathbb{1}[x \in \mathcal{X}_\mathcal{W}] + f_{P,\mathcal{H}}(x)\mathbb{1}[x \in \mathcal{X}_\mathcal{H}]$. In particular, by definition, we have

$$\|P - \hat{P}_T^{\mathbf{w}}\| \leq \frac{\epsilon}{2m}.$$

Since each X_i has distribution $\hat{P}_T^{\mathbf{w}}$, and they are sampled independently, we have, for the distribution $\mathbb{P}_{(X_1,\dots,X_m)} = \left(\hat{P}_T^{\mathbf{w}}\right)^m$ of $(X_1, \dots, X_m)$,

$$\left\|\mathbb{P}_{(X_1,\dots,X_m)} - P^m\right\| \leq \frac{3\epsilon}{2}.$$

Thus, for $(X_1', \dots, X_m') \sim P^m$,

$$\mathbb{E}\Big[P(\hat{h})\Big] = \mathbb{E}[\mathcal{A}_{\text{emx}}(X_1, \dots, X_m)] \geq \mathbb{E}\big[\mathcal{A}_{\text{emx}}(X_1', \dots, X_m')\big] - \frac{3\epsilon}{2},$$

and by the guarantee of the EMX learner,

$$\mathbb{E}\big[\mathcal{A}_{\text{emx}}(X_1', \dots, X_m')\big] \geq \sup_{h \in \mathcal{H}} P(h) - \frac{3\epsilon}{2}.$$

Altogether,

$$\mathbb{E}[f^\star(\hat{x})] = \mathbb{E}\Big[(1/3)(P(\hat{h}) + 2)\Big] \geq (1/3)\left(\sup_{h \in \mathcal{H}} P(h) - 3\epsilon + 2\right) = \sup_x f^\star(x) - \epsilon.$$

Since the bandit algorithm pulls exactly $m + 1 = M(3\epsilon/2) + 1$ arms, we conclude that the sample complexity of learning $(\mathcal{X}, \mathcal{F})$ in the bandit setting is at most $M(3\epsilon/2) + 1$, and hence $(\mathcal{X}, \mathcal{F})$ is learnable in the bandit setting.

Second, we address the "only if" direction, holding for any union-bounded set $\mathcal{H}$ of finite sets, for $\mathcal{W} = (0, 1/3)$. This direction is quite a bit more involved, since a bandit algorithm might query at the arms x_h, which is a feature not directly present in the EMX problem. However, we will make use of Lemma 10 from Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a,b), establishing equivalence of union-closed EMX learnability to monotone compressibility. In particular, it is for this reason that we restrict $(\mathcal{W}, \mathcal{H})$ to be a union-bounded EMX learning problem. Suppose $\mathcal{A}_\text{b}$ is a learning algorithm for the Bandit problem $(\mathcal{X}, \mathcal{F})$, with $\epsilon = 1/18$, and query complexity $M = M(1/18) \in \mathbb{N}$. Without loss of generality, suppose $\mathcal{A}_\text{b}$ always returns an arm $x_{\hat{h}}$ in $\mathcal{X}_\mathcal{H}$ (since *all* arms in $\mathcal{X}_\mathcal{H}$ have higher rewards than *all* arms in $\mathcal{X}_\mathcal{W}$). Additionally, without loss of generality, let us suppose $\mathcal{A}_\text{b}$ is well-behaved even when the rewards it receives are not consistent with any $f \in \mathcal{F}$: that is, it still pulls M arms and returns some arm $x_{\hat{h}} \in \mathcal{X}_\mathcal{H}$.

We construct a monotone compression scheme η defined as follows. Let $m = \lceil 3M/2 \rceil$. Fix any sequence $S' = \{x_1, \dots, x_M\}$ of M distinct points all contained in some set in $\mathcal{H}$. We aim to specify the set returned by $\eta(S')$.

Consider an execution of $\mathcal{A}_{\text{b}}$ where we specify the returned rewards as follows. Initialize k to 0. As the algorithm executes, if it ever pulls any arm $z_i^{(2)} \in \mathcal{X}_{\mathcal{W}}^{(2)}$, it always receives a reward $\frac{1}{m+1}$. On the other hand, when at some time in its execution, the algorithm pulls an arm $z_i^{(1)} \in \mathcal{X}_{\mathcal{W}}^{(1)}$, then if it has previously pulled any arm $z_j^{(1)}$ with i and j equivalent mod $m+1$, then the algorithm receives reward equal to the reward it received when it pulled arm $z_j^{(1)}$; otherwise, if it has never pulled such an arm previously, then it receives as its reward the value x_{k+1}; after such a round, we increment $k \leftarrow k+1$. Finally, if at some round in its execution, the algorithm pulls an arm x_h in $\mathcal{X}_{\mathcal{H}}$, it receives as the reward the value

$$\frac{1}{3}\left(\frac{|h \cap \{x_1, \dots, x_k\}|}{m+1} + 2\right).$$

In this case, we do *not* increment k after such a round.

After the algorithm halts (after M total arm pulls), it returns some arm $x_{\hat{h}_{S'}}$. Define $H_{S'}$ as a subset of $\mathcal{H}$, with elements $\hat{h}_{S'}$ along with all $h \in \mathcal{H}$ for which the algorithm pulled arm x_h during this entire execution. Note that, if $\mathcal{A}_{\text{b}}$ is a randomized algorithm, then $H_{S'}$ may also be random. Let $\hat{\mathcal{W}}(S')$ be the set of all $w \in \mathcal{W}$ such that

$$\mathbb{P}(w \in \bigcup H_{S'}) \geq \frac{1}{4}\frac{1}{M+1}\frac{1}{m+1}.$$

Define $\eta(S')$ as any $h_{S'} \in \mathcal{H}$ with $h_{S'} \supseteq S' \cup \hat{\mathcal{W}}(S')$ (the existence of which is argued below).

To show $\eta(S')$ is well-defined, we need to argue that such an $h_{S'}$ exists. For this, note that each $w \in \hat{\mathcal{W}}(S')$ is contained in some $h \in \mathcal{H}$ (by definition of $H_{S'}$), and S' is contained in some element of $\mathcal{H}$ (by assumption); thus, to show such an $h_{S'}$ will be guaranteed to exist by the union-bounded property of $\mathcal{H}$, it suffices to argue that $\hat{\mathcal{W}}(S')$ is finite. To show the latter, note that $H_{S'}$ is a finite set (of size at most $M+1$), and each $h \in H_{S'}$ is an element of $\mathcal{H}$, and hence is itself a finite subset of $\mathcal{W}$. Thus, $\bigcup H_{S'}$ is always a finite set. Let $\hat{K} = |\bigcup H_{S'}|$, the cardinality of the set. Since $\hat{K}$ is always finite, there exists a finite $K^* \in \mathbb{N}$ such that, with probability at least $1 - \frac{1}{8}\frac{1}{M+1}\frac{1}{m+1}$, $\hat{K} \leq K^*$. Thus, for each $w \in \hat{\mathcal{W}}(S')$,

$$\mathbb{P}(w \in \bigcup H_{S'} \text{ and } \hat{K} \leq K^*) \geq \frac{1}{4}\frac{1}{M+1}\frac{1}{m+1} - \frac{1}{8}\frac{1}{M+1}\frac{1}{m+1} = \frac{1}{8}\frac{1}{M+1}\frac{1}{m+1}.$$

Enumerating $\bigcup H_{S'} =: \{w'_1, \dots, w'_{\hat{K}}\}$ with $w'_1 < w'_2 < \cdots < w'_{\hat{K}}$, by the union bound, for each $w \in \hat{\mathcal{W}}(S')$, there exists $k \in \{1, \dots, K^*\}$ such that

$$\mathbb{P}(w'_k = w) \geq \frac{1}{K^*}\frac{1}{8}\frac{1}{M+1}\frac{1}{m+1}.$$

For each $k \in \{1, \dots, K^*\}$, there can be at most $K^* 8(M+1)(m+1)$ distinct values w for which this holds (the corresponding events $\{w'_k = w\}$ being mutually exclusive). Therefore,

$$\left|\hat{\mathcal{W}}(S')\right| \leq (K^*)^2 8(M+1)(m+1) < \infty,$$

so that, letting $h_w \in \mathcal{H}$ be any set with $w \in h_w$, the finite union $\bigcup\{h_w : w \in \hat{\mathcal{W}}(S')\}$ contains $\hat{\mathcal{W}}(S')$ and is itself contained in some set $h \in \mathcal{H}$ by the union-bounded property, and hence (since S' is also contained in an element of $\mathcal{H}$ by assumption) $S' \cup \hat{\mathcal{W}}(S')$ is contained in some $h_{S'} \in \mathcal{H}$, so that our definition of $\eta(S')$ above is valid.

Since we aim to argue η provides an $(m+1) \to m$ monotone compression scheme, we need to extend the definition of η to allow arguments of size m, rather than M (recalling $M < m$). Toward this end, for any finite $T > M$ and sequence $S' = \{x'_1, \ldots, x'_T\}$ of distinct elements all contained in some set in $\mathcal{H}$, define $\eta(S')$ as any $h_{S'} \in \mathcal{H}$ with $h_{S'} \supseteq \bigcup\{\eta(x'_{i_1}, \ldots, x'_{i_M}) : i_1, \ldots, i_M \in \{1, \ldots, T\} \text{ distinct}\}$. Again, such an $h_{S'}$ is guaranteed to exist by the union-bounded property of $\mathcal{H}$.

We next argue that η is indeed a $(m+1) \to m$ monotone compression scheme for $\mathcal{H}$. Let $h \in \mathcal{H}$ and $S = \{x_1, \ldots, x_{m+1}\} \subseteq h$. For each $i \in \{1, \ldots, m+1\}$, let $S_i = \{x_j : j \neq i\}$. We will argue that $\exists i \in \{1, \ldots, m+1\}$ with $\eta(S_i) \supseteq S$, so that η is a valid $(m+1) \to m$ monotone compression scheme for $\mathcal{H}$. Note that it will suffice to find even *one* sequence S' of length M in S for which $h_{S'}$ contains even *one* of the other points in S, since letting x_i be this point, by definition of η we would have $x_i \in \eta(S_i)$, and since $S_i \subseteq \eta(S_i)$ (which follows from the definition of $\eta(S')$ for sequences S' of length M), it follows that $S \subseteq \eta(S_i)$. Thus, we will show there exists a sequence $S' = \{x_{i_1}, \ldots, x_{i_M}\}$ for which $\exists i \notin \{i_1, \ldots, i_M\}$ with $x_i \in \eta(S')$. As a trivial case, if S contains any duplicated elements, then since $S' \subseteq \eta(S')$ by definition, we trivially have $S \subseteq \eta(S_i)$ where x_i is any one of the duplicated values. To focus on the nontrivial case, for the remainder of the proof let us suppose all elements of the sequence S are distinct.

We will construct a distribution over (a finite subset of) the reward functions in $\mathcal{F}_P$, were $P = \text{Uniform}(S)$. Let $S_\sigma = \{\hat{x}_1, \ldots, \hat{x}_{m+1}\}$ be a uniform random permutation of the sequence of elements of S (independent of the randomness of $\mathcal{A}_b$). Let $S' = \{\hat{x}_1, \ldots, \hat{x}_M\}$. Define a (randomly selected) reward function $f^\star \in \mathcal{F}_P$ as follows. The sequence $\mathbf{w} = \{w_i\}_{i=1}^\infty$ is defined by $w_i = \hat{x}_j$ where $j \in \{1, \ldots, m+1\}$ is equivalent to i mod $m+1$: that is, $\mathbf{w}$ repeats copies of the sequence in S_σ infinitely. Define the sequence $\mathbf{T} = \{T_i\}_{i=1}^\infty$ as $T_i = \frac{1}{m+1}$ for all i. Note that $f_P^{\mathbf{w},\mathbf{T}} \in \mathcal{F}_{P,\mathcal{W}}$. Finally, define the target reward function as $f^\star(x) = f_P^{\mathbf{w},\mathbf{T}}(x)\mathbb{1}[x \in \mathcal{X}_\mathcal{W}] + f_{P,\mathcal{H}}(x)\mathbb{1}[x \in \mathcal{X}_\mathcal{H}]$.

Note that for any possible value of $f^\star$ under this distribution, since $f^\star \in \mathcal{F}_P$, if a returned arm $x_{\hat{h}}$ has reward $(1/3)(P(\hat{h}) + 2) \geq \max_{h \in \mathcal{H}}(1/3)(P(h) + 2) - 1/9 = (1/3)((2/3) + 2)$, it must have $|\hat{h} \cap S| \geq M + 1$ (by our choice of $m = \lceil 3M/2 \rceil$). In particular, when this occurs, it must be that $\hat{h} \cap S \setminus S' \neq \emptyset$: that is, $\hat{h}$ contains at least one point in S and not in S'.

Now for the key observation. Note that if we were to execute $\mathcal{A}_b$ under reward function $f^\star$ above, the sequence of rewards it receives is a random sequence (both due to randomness of $\mathcal{A}_b$ and randomness of the permutation S_σ). In particular, since S_σ is a uniform random permutation, this sequence of rewards is distributionally equivalent to the sequence of rewards that would be received in an execution which receives as rewards for pulling arms in $\mathcal{X}_\mathcal{W}^{(1)}$ the same reward values as in the definition of $\eta(S')$ above (i.e., the reward received when pulling arm $z_i^{(1)}$ is $\hat{x}_j$ where j is equivalent to i mod $m+1$), though (unlike in $\eta(S')$) still receiving reward $f^\star(x_h)$ upon pulling arms x_h in $\mathcal{X}_\mathcal{H}$. Without loss, let us suppose this is precisely how the rewards are defined in the execution of $\mathcal{A}_b$ under the target reward function $f^\star$ (this merely amounts to a coupling among the random variables S' and the sequence of query choices of the algorithm, which, since S' is a uniform random subset, does not change the distribution of rewards received, arms pulled, or the value $x_{\hat{h}}$ returned).

In particular, with this coupling in place, the execution of $\mathcal{A}_b$ under the target reward function $f^\star$, and the execution of $\mathcal{A}_b$ in the definition of $\eta(S')$, are identical (supposing we also couple the internal randomness in these two executions: that is, they execute with the same internal random bits) up until the first round in which $\mathcal{A}_b$ pulls an arm $x_h \in \mathcal{X}_\mathcal{H}$ with $h \cap S$ containing an element of S *not* previously received as a reward from a past pull of an arm in $\mathcal{X}_\mathcal{W}^{(1)}$: that is, in the execution of $\eta(S')$, the first round in which it pulls an x_h with $h \cap S \setminus \{\hat{x}_1, \ldots, \hat{x}_k\} \neq \emptyset$ (for the k value at that round, in the definition of $\eta(S')$). In particular, on the event that $\mathcal{A}_b$ never pulls such an arm x_h, we have that $x_{\hat{h}_{S'}} = x_{\hat{h}}$.

This equivalence is the purpose we have in mind when defining η. Indeed, we may note that if $f^\star(x_{\hat{h}}) \geq \sup_x f^\star(x) - 1/9$ (so that $\hat{h} \cap S \setminus S' \neq \emptyset$), then *either* $\hat{h}_{S'} = \hat{h}$, so that $h_{S'} \cap S \setminus S' \neq \emptyset$ and hence $\bigcup H_{S'} \cap S \setminus S' \neq \emptyset$, *or* the execution of $\mathcal{A}_b$ in the definition of $\eta(S')$ pulls some x_h with $h \cap S \setminus \{\hat{x}_1, \ldots, \hat{x}_k\} \neq \emptyset$, so that $\bigcup H_{S'}$ contains this value x_i which is in $S \setminus \{\hat{x}_1, \ldots, \hat{x}_k\}$. The point, intuitively, is that in either case the algorithm $\mathcal{A}_b$ has succeeded in *guessing* some point in S it has never seen before. The remainder of the proof argues that this ability to guess the value of a never-before-seen point from S results in some S_i having $x_i \in \eta(S_i)$.

Noting that $\sup_x f^\star(x) = 1$, Markov's inequality and our choice of $\epsilon = 1/18$ imply

$$\mathbb{P}\left(f^\star(x_{\hat{h}}) \geq \sup_x f^\star(x) - 1/9\right) = 1 - \mathbb{P}\left(1 - f^\star(x_{\hat{h}}) > \frac{1}{9}\right)$$
$$\geq 1 - 9\left(1 - \mathbb{E}[f^\star(x_{\hat{h}})]\right) \geq 1 - 9\epsilon = \frac{1}{2}.$$

Thus,

$$\mathbb{P}\left(P(\hat{h}) \geq \frac{2}{3}\right) \geq \frac{1}{2},$$

which, by the law of total probability, implies

$$\mathbb{E}\left[\mathbb{P}\left(P(\hat{h}) \geq \frac{2}{3}\middle| S'\right)\right] \geq \frac{1}{2}.$$

This implies that, with non-zero probability over the draw of S', it holds that

$$\mathbb{P}\left(P(\hat{h}) \geq \frac{2}{3}\middle| S'\right) \geq \frac{1}{2}.$$

By the arguments above, on this event, we have

$$\mathbb{P}\left(\hat{h} \cap S \setminus S' \neq \emptyset\middle| S'\right) \geq \frac{1}{2}.$$

Therefore, there exists some non-random choice of a sequence $S' = \{\hat{x}_1, \ldots, \hat{x}_M\}$ in S such that, if the algorithm receives rewards for arms pulled in $\mathcal{X}_\mathcal{W}^{(1)}$ as described in the definition of $\eta(S')$ (though still with $f^\star(x_h)$ values for arms x_h from $\mathcal{X}_\mathcal{H}$) then

$$\mathbb{P}\left(\hat{h} \cap S \setminus S' \neq \emptyset\right) \geq \frac{1}{2}. \tag{3}$$

Consider the execution of $\mathcal{A}_\mathfrak{b}$ in the above scenario (again, with rewards for arms in $\mathcal{X}_\mathcal{W}^{(1)}$ as described in the definition of $\eta(S')$), and denote by $\hat{h}^*$ the $\hat{h}$ corresponding to the arm $x_{\hat{h}}$ that would

be returned. Also consider the execution of $\mathcal{A}_{\text{b}}$ in the definition of $\eta(S')$: that is, in addition to the rewards in $\mathcal{X}_{\mathcal{W}}^{(1)}$ defined as in the execution leading to (3), we also use the modified the reward values for arms x_h pulled in $\mathcal{X}_{\mathcal{H}}$, so that the algorithm receives reward $(1/3)\left(\frac{|h\cap\{\hat{x}_1,\ldots,\hat{x}_k\}|}{m+1}+2\right)$ for the value k at that round in the execution described in the definition of $\eta(S')$. Denote by $\tilde{h}$ the $\hat{h}$ corresponding to the arm $x_{\hat{h}}$ that would be returned in this latter case.

There are two cases to consider. In Case 1, in the execution of $\mathcal{A}_{\text{b}}$ leading to $\hat{h}^*$, at every time t, either it pulls an arm in $\mathcal{X}_{\mathcal{W}}$, or it pulls an arm $x_h \in \mathcal{X}_{\mathcal{H}}$ for which $h \cap S \subseteq \{\hat{x}_1,\ldots,\hat{x}_k\}$, for the value of k at that round in the execution. As mentioned, the rewards it receives (from both the arms in $\mathcal{X}_{\mathcal{W}}$ and the arms in $\mathcal{X}_{\mathcal{H}}$) are identical to the rewards it would receive in the execution in the definition of $\eta(S')$ (again, supposing the same random bits are shared by the two executions, so that their choices remain identical as long as their received rewards remain identical). Thus, in this first case, we have $\tilde{h} = \hat{h}^*$.

On the other hand, consider a second case, Case 2, where the execution of $\mathcal{A}_{\text{b}}$ leading to $\hat{h}^*$ includes at least one time t in which it pulls an arm $x_{h^\flat}$ in $\mathcal{X}_{\mathcal{H}}$ for which $h^\flat \cap S$ contains at least one element $\tilde{x}$ of S not contained in $\{\hat{x}_1,\ldots,\hat{x}_k\}$, for the value of k in that round of the execution. Let us consider this $h^\flat$ and $\hat{x}_1,\ldots,\hat{x}_k$ for the *first* time t this happens; let us name the corresponding values of t and k as $\hat{t}$ and $\hat{k}$, respectively. Since, up until that time $\hat{t}$, the arms pulled by $\mathcal{A}_{\text{b}}$ and the rewards received in the execution leading to $\hat{h}^*$ and the execution leading to $\tilde{h}$ will be identical, naturally the execution in the definition of $\eta(S')$ will also pull this arm $x_{h^\flat}$ at time $\hat{t}$.

We have chosen S' based on (3) so that, with probability at least $\frac{1}{2}$, either Case 1 occurs and $\tilde{h} \cap S \setminus S' \neq \emptyset$ (call this Event 1), or Case 2 occurs and hence $h^\flat \cap S \setminus \{\hat{x}_1,\ldots,\hat{x}_{\hat{k}}\} \neq \emptyset$ (call this Event 2). By the pigeonhole principle, at least one of these two events occurs with probability at least $\frac{1}{4}$.

Consider first the scenario where Event 1 has probability at least $\frac{1}{4}$. Since there are $m+1-M = \lceil M/2\rceil + 1$ elements in $S \setminus S'$, by the pigeonhole principle there must exist some $x_i \in S \setminus S'$ for which, with probability at least $\frac{1}{4}\frac{1}{\lceil M/2\rceil+1}$, it holds that $x_i \in \tilde{h}$. In particular, in this scenario, since $\frac{1}{4}\frac{1}{\lceil M/2\rceil+1} \geq \frac{1}{4}\frac{1}{M+1}\frac{1}{m+1}$, we have $x_i \in \eta(S')$, and therefore $x_i \in \eta(S_i)$, so that $\eta(S_i) \supseteq S$.

Finally, consider the scenario where Event 2 has probability at least $\frac{1}{4}$: that is, with probability at least $\frac{1}{4}$, Case 2 occurs and hence $h^\flat \cap S \setminus \{\hat{x}_1,\ldots,\hat{x}_{\hat{k}}\} \neq \emptyset$. At time $\hat{t}$ in the execution, the algorithm has received, as the set of distinct values of rewards for the arms it has pulled in $\mathcal{X}_{\mathcal{W}}^{(1)}$, precisely the set $\{\hat{x}_1,\ldots,\hat{x}_{\hat{k}}\}$: the first $\hat{k}$ values in the sequence $S' = \{\hat{x}_1,\ldots,\hat{x}_M\}$. Note that $\hat{k}$ is a random variable (which has a well-defined value, albeit still random, whenever Case 2 occurs). By the pigeonhole principle, there exists a value $k^* \in \{0,1,\ldots,M\}$ such that, with probability at least $\frac{1}{4}\frac{1}{M+1}$, Case 2 occurs and $\hat{k} = k^*$. In the event Case 2 occurs, define $\hat{i}$ as the smallest $i \in \{1,\ldots,M\}$ such that $x_i \in h^\flat \cap S \setminus \{\hat{x}_1,\ldots,\hat{x}_{\hat{k}}\}$. By the pigeonhole principle, there exists a value $i^* \in \{1,\ldots,m+1\}$ such that, with probability at least $\frac{1}{4}\frac{1}{M+1}\frac{1}{m+1}$, Case 2 occurs while $\hat{k} = k^*$ and $\hat{i} = i^*$. In particular, on this event, $x_{i^*} \notin \{\hat{x}_1,\ldots,\hat{x}_{k^*}\}$.

Now define a sequence $S'' = \{\hat{x}_1,\ldots,\hat{x}_{k^*},\check{x}_{k^*+1},\ldots,\check{x}_M\}$, where each $\check{w}_i$ can be chosen as any elements of $S \setminus (\{\hat{x}_1,\ldots,\hat{x}_{k^*}\} \cup \{x_{i^*}\})$. In particular, this implies the elements of S'' are all contained in S_{i^*}. Now consider the execution of $\mathcal{A}_{\text{b}}$ in the definition of $\eta(S'')$ (again supposing shared internal random bits with the execution in the definition of $\eta(S')$). In particular, since the first k^* elements of S'' are the same as in S', the above argument implies that with probability at

least $\frac{1}{4}\frac{1}{M+1}\frac{1}{m+1}$, the execution of $\mathcal{A}_\mathrm{b}$ in the definition of $\eta(S'')$ will pull an arm $x_{h^\flat} \in \mathcal{X}_\mathcal{H}$ with $x_{i^*} \in h^\flat$. In particular, this implies $x_{i^*} \in \eta(S'')$, and therefore $x_{i^*} \in \eta(S_{i^*})$, so that $\eta(S_{i^*}) \supseteq S$.

Hence, in either scenario (i.e., the scenario where Event 1 has probability at lesat $\frac{1}{4}$ or the scenario where Event 2 has probability at least $\frac{1}{4}$), we have shown there exists $i \in \{1, \ldots, m+1\}$ such that $\eta(S_i) \supseteq S$. Since this argument holds for *any* sequence S of length $m+1$, we conclude that η is a valid $(m+1) \to m$ monotone compression scheme for $\mathcal{H}$, and hence, by Lemma 10, the EMX problem $(\mathcal{W}, \mathcal{H})$ is learnable. This completes the proof. ■

In particular, note that Theorem 1 now follows immediately from Theorem 9 and the independence from ZFC of EMX learnability on $\mathcal{W}$, as established by Ben-David, Hrubes, Moran, Shpilka, and Yehudayoff (2019a). Specifically, they show the following result.

Lemma 11 (Ben-David et al., 2019a) *For $\mathcal{W} = (0, 1/3)$ and $\mathcal{H}$ the class of all finite subsets of $\mathcal{W}$, EMX learnability of $(\mathcal{W}, \mathcal{H})$ is independent of the ZFC axioms. Specifically, $(\mathcal{W}, \mathcal{H})$ is EMX learnable under the* Continuum Hypothesis *axiom (i.e., that the continuum is the smallest cardinality strictly greater than the integers), whereas $(\mathcal{W}, \mathcal{H})$ is not EMX learnable under the axiom that there are infinitely many distinct cardinalities between the integers and the continuum (each of these axioms are known to be independent of ZFC; Jech, 2003; Kunen, 1980).*

Proof of Theorem 1 For $\mathcal{W} = (0, 1/3)$, the class $\mathcal{H}$ of all finite subsets of $\mathcal{W}$ is union-bounded (indeed, it is closed under finite unions). Therefore, taking $(\mathcal{X}, \mathcal{F})$ to be the bandit problem corresponding to the EMX problem $(\mathcal{W}, \mathcal{H})$, as guaranteed to exist by Theorem 9, we have that $(\mathcal{X}, \mathcal{F})$ is learnable in the bandit setting if and only if $(\mathcal{W}, \mathcal{H})$ is learnable in the EMX setting. Since Lemma 11 implies that the latter is independent of ZFC, the conclusion of Theorem 1 follows. ■

Remark 12 *Bandit optimization is also known as zeroth-order optimization, since we do not have access to oracles for derivatives. We remark that having access to such an oracle would not improve the undecidability situation. Specifically, the above scenario can easily be extended to one where we have access to additional oracles for any number of derivatives, and yet the undecidability proof would remain valid.*

4. Undecidability of No-Regret Learnability

In this section, we prove Theorem 2, establishing equivalence of no-regret bandit learnability to PAC bandit learnability. In light of Theorem 1, this has the further implication that there is a bandit problem $(\mathcal{X}, \mathcal{F})$ whose no-regret learnability is independent of the ZFC axioms.

Theorem 2 (restated) Any $(\mathcal{X}, \mathcal{F})$ is learnable in the bandit setting if and only if it is no-regret learnable in the bandit setting.

Proof of Theorem 2 Let $(\mathcal{X}, \mathcal{F})$ be any problem learnable in the bandit setting. Suppose $\mathcal{A}$ is an algorithm for learning the bandit problem $(\mathcal{X}, \mathcal{F})$ (i.e., in the PAC/optimization setting), guaranteeing query complexity $M(\varepsilon)$. Let us suppose $\mathcal{A}$ is an *anytime* algorithm: that is, we may stop it after any a-priori chosen number n of queries, and guarantee the expected loss is at most ε as long as $n \geq M(\varepsilon)$, for any choice of ε. This is without loss of generality, since given any algorithm

that would terminate in at most $M(\varepsilon)$ rounds, for given ε, we may repeatedly halve the value of $\varepsilon = 2^{-1}, 2^{-2}, \ldots$ to produce a sequence of arms $\hat{x}_1, \hat{x}_2, \ldots$: that is, we run $\mathcal{A}$ for $\varepsilon = 2^{-1}$ until it terminates (after at most $M(2^{-1})$ queries) and returns an arm $\hat{x}_1$, then re-running it with $\varepsilon = 2^{-2}$ until it terminates (after at most $M(2^{-2})$ queries) and returns an arm $\hat{x}_2$, and so on. For any given n, we may then choose the $\varepsilon_n = 2^{-i}$ for i maximal such that $n \leq \sum_{j=1}^{i} M(2^{-j})$ and return $\hat{x}_i$, which guarantees $\mathbb{E}[f^\star(\hat{x}_i)] \geq \sup_x f^\star(x) - \varepsilon_n$, which is still a PAC learner since $\varepsilon_n \to 0$ due to the fact that $M(\varepsilon) < \infty$ for all $\varepsilon > 0$. The query complexity is increased to $M'(\varepsilon) = \sum_{j=1}^{i} M(2^{-j})$, where $i = \lceil \log_2(1/\varepsilon) \rceil$, but this is still a finite number for any ε, and hence the algorithm remains a bandit learner. We proceed with the assumption that such a conversion has already been applied, so that $\mathcal{A}$ is already an anytime algorithm, and $M(\varepsilon)$ is its query complexity.

Let $x_1, x_2, \ldots$ denote the sequence of arms pulled by $\mathcal{A}$, and let $r_1, r_2, \ldots$ denote the rewards it receives from each pull, respectively. If the algorithm were to be terminated after some n rounds, let $\hat{x}_n$ denote the arm it would return (including the case $n = 0$, where we may define $\hat{x}_0$ as some arbitrary fixed choice). Now consider a no-regret learner, based on the well-known "ε-greedy" strategy, defined as follows. We distinguish between *exploration* rounds, in which we pull the next x_n in the sequence of arms $\mathcal{A}$ would pull, and *exploitation* rounds, in which we pull an arm $\hat{x}_n$ for an appropriate choice of n. Suppose we are on round t, and we have done n "exploration rounds" so far, in which $\mathcal{A}$ chose arms $x_1, \ldots, x_n$ and received rewards $r_1, \ldots, r_n$. Now with (independent) probability p_t we let $\mathcal{A}$ pull its next arm x_{n+1} and receive reward r_{n+1} (an "exploration" round). Otherwise ("exploitation" round), on the probability $1 - p_t$ event, we pull the arm $\hat{x}_n$ that $\mathcal{A}$ would return if we were to halt it there. In either case, let $\tilde{x}_t$ denote the arm pulled by this algorithm on round t.

Intuitively, the more arms $\mathcal{A}$ gets to pull, the better the reward of its return arm $\hat{x}_n$. So the $1 - p_t$ probability option (exploitation) gets better if we let $\mathcal{A}$ explore more (the p_t probability option). By setting p_t to decrease gradually, as $\hat{x}_n$ gets better over time, we gradually exploit more and more. But we don't decrease p_t too quickly, so we still explore infinitely often as the number of rounds t grows.

Formally, let $\{p_t\}_{t \in \mathbb{N}}$ be any non-increasing sequence satisfying $p_t \to 0$ with $\sum_t p_t = \infty$. Let n_t denote the value of n above, after round t (i.e., if round t explores, then $n_t = n_{t-1} + 1$, and otherwise $n_t = n_{t-1}$). Let ε_n be a non-increasing sequence in $[0, 1]$ with $\varepsilon_n \to 0$ such that, for every $n \in \mathbb{N}$, $n \geq M(\varepsilon_n)$; such a sequence exists by the fact that $\mathcal{A}$ is a learning algorithm with finite query complexity $M(\varepsilon)$ for every $\varepsilon > 0$.

For any t, we have

$$\begin{aligned}\mathbb{E}[f^\star(\tilde{x}_t)|n_{t-1}] &= \mathbb{E}\big[p_t f^\star(x_{n_{t-1}+1}) + (1-p_t) f^\star(\hat{x}_{n_{t-1}})\big|n_{t-1}\big] \\ &\geq (1-p_t)\mathbb{E}\big[f^\star(\hat{x}_{n_{t-1}})\big|n_{t-1}\big] \geq (1-p_t)\left(\sup_x f^\star(x) - \varepsilon_{n_{t-1}}\right).\end{aligned}$$

Thus,

$$\mathbb{E}[f^\star(\tilde{x}_t)] \geq (1-p_t)\left(\sup_x f^\star(x) - \mathbb{E}[\varepsilon_{n_{t-1}}]\right) \geq \sup_x f^\star(x) - p_t - \mathbb{E}[\varepsilon_{n_{t-1}}].$$

Note that n_{t-1} is a sum of $t-1$ independent Bernoulli random variables, with $\mathbb{E}[n_{t-1}] = \sum_{t'=1}^{t-1} p_{t'} =: \bar{n}_{t-1}$. Thus, by a Chernoff bound, with probability at least $1 - e^{-\bar{n}_{t-1}/8}$, it holds that $n_{t-1} \geq \frac{1}{2}\bar{n}_{t-1}$. Thus, since ε_n is non-increasing and bounded by 1,

$$\mathbb{E}[\varepsilon_{n_{t-1}}] \leq \varepsilon_{(1/2)\bar{n}_{t-1}} + e^{-\bar{n}_{t-1}/8}.$$

By the choice of p_t satisfying $\sum_{t'=1}^{\infty} p_{t'} = \infty$, we have $\bar{n}_{t-1} \to \infty$ as $t \to \infty$. Thus, since $\varepsilon_n \to 0$ as $n \to \infty$, we have $\mathbb{E}[\varepsilon_{n_{t-1}}] \to 0$. Therefore, $\sum_{t=1}^{T} \mathbb{E}[\varepsilon_{n_{t-1}}] = o(T)$. Moreover, since $p_t \to 0$, we have $\sum_{t=1}^{T} p_t = o(T)$. Letting $\mathcal{R}(T) = \sum_{t=1}^{T} p_t + \sum_{t=1}^{T} \mathbb{E}[\varepsilon_{n_{t-1}}] = o(T)$, and noting that $\mathcal{R}(T)$ has no dependence on the particular target reward function $f^\star \in \mathcal{F}$, altogether we have that

$$T \sup_x f^\star(x) - \mathbb{E}\left[\sum_{t=1}^{T} f^\star(\tilde{x}_t)\right] \leq \mathcal{R}(T),$$

so that this algorithm is a no-regret learner, and hence $(\mathcal{X}, \mathcal{F})$ is no-regret learnable in the bandit setting.

For the other direction in the equivalence, if $(\mathcal{X}, \mathcal{F})$ is no-regret learnable in the bandit setting, and B is any no-regret learner for $(\mathcal{X}, \mathcal{F})$ with some regret bound $\mathcal{R}(T) = o(T)$, if we run B for T rounds, and then output an arm $\hat{x} \sim \mathrm{Uniform}(x_1, \ldots, x_T)$ chosen uniformly at random from those $x_1, \ldots, x_T$ that B has pulled on rounds $1, \ldots, T$, this will have

$$\mathbb{E}[f^\star(\hat{x})] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} f^\star(x_t)\right] \geq \sup_x f^\star(x) - \frac{1}{T}\mathcal{R}(T).$$

Since $\mathcal{R}(T) = o(T)$, for any $\varepsilon > 0$, there exists $T_\varepsilon \in \mathbb{N}$ such that any $T \geq T_\varepsilon$ satisfies $\frac{1}{T}\mathcal{R}(T) \leq \varepsilon$, and hence the above choice of $\hat{x}$ provides a bandit learner (in the PAC/optimization sense) with query complexity $M(\varepsilon) \leq T_\varepsilon$.

This completes the proof that any bandit problem $(\mathcal{X}, \mathcal{F})$ is PAC learnable if and only if it is no-regret learnable. ■

As discussed above, Theorems 2 and 1 together have Corollary 1 as an immediate implication: that is, there exists a bandit problem $(\mathcal{X}, \mathcal{F})$ such that, whether or not $(\mathcal{X}, \mathcal{F})$ is no-regret learnable is independent of the ZFC axioms.

5. The Optimal Query Complexity of Binary-Valued Bandits

While the above negative results, in some sense, indicate that it is not possible to provide a fully-general characterization of bandit learnability, it nevertheless leaves open the potential for theories of bandit learning that cover many important cases of learnable and non-learnable classes. Toward this end, we are interested in identifying abstract dimensions which capture certain interesting cases.

We begin with the simplest case here: namely, *binary-valued* bandits. We say $(\mathcal{X}, \mathcal{F})$ is a binary-valued bandit problem if every $f \in \mathcal{F}$ has image contained in $\{0, 1\}$: that is, $\{f(x) : x \in \mathcal{X}\} \subseteq \{0, 1\}$. Note that this case is particularly simple since achieving $f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon$ for some $\epsilon \in (0, 1)$ is equivalent to choosing $\hat{x}$ equal any x with $f^\star(x) = 1$ (if such an x exists).

In this section, we identify the optimal query complexity and optimal algorithms for learning binary-valued bandit problems. We treat separately the case of restricting to deterministic learners vs allowing randomized learners. We identify the optimal query complexity in both cases, and supply a learner matching this complexity. Interestingly, this also reveals a separation between deterministic and randomized learners, in terms of learnability.

5.1. Deterministic Learners

We begin with the case of deterministic learners. For convenience, we restate the definitions and results stated in Section 1.

Definition 2 (restated). *Define the zero-teaching dimension of $\mathcal{F}$, denoted $\tau^0_{\mathcal{F}}$, as the smallest t such that, there exist $x_1, \ldots, x_t \in \mathcal{X}$ with*

$$\min_{f \in \mathcal{F} \setminus \{\mathbf{0}\}} \max_{1 \leq i \leq t} f(x_i) = 1,$$

where $\mathbf{0}$ is the all-zero function (which may or may not be in $\mathcal{F}$).

We have the following result, stating that this quantity determines learnability of binary-valued bandit problems by *deterministic* learners. The result itself is rather obvious, given the simplicity of the binary-valued bandit scenario.

Theorem 3 (restated). *Any binary-valued bandit problem $(\mathcal{X}, \mathcal{F})$ is learnable by a deterministic algorithm if and only if $\tau^0_{\mathcal{F}} < \infty$. Moreover, the optimal query complexity $M(\epsilon)$ achievable by deterministic algorithms satisfies, $\forall \epsilon \in (0,1)$, $M(\epsilon) = \tau^0_{\mathcal{F}} - 1$.*

Proof The result is almost immediate from the definition of $\tau^0_{\mathcal{F}}$. For completeness, we present the argument in detail, beginning with a proof that $M(\epsilon) \leq \tau^0_{\mathcal{F}} - 1$. If $t = \tau^0_{\mathcal{F}} < \infty$, let $x_1, \ldots, x_t$ satisfy the criterion in Definition 2: that is,

$$\min_{f \in \mathcal{F} \setminus \{\mathbf{0}\}} \max_{1 \leq i \leq t} f(x_i) = 1.$$

We may use an algorithm that pulls the arms $x_1, \ldots, x_{t-1}$. If one of these arms x_i yields a reward of 1, the algorithm returns this arm x_i as its output $\hat{x}$. Otherwise it returns the arm x_t as its output $\hat{x}$. We are guaranteed that, for any target reward function $f^\star \in \mathcal{F}$, if $f^\star \neq \mathbf{0}$ then at least one of $x_1, \ldots, x_t$ has $f^\star(x_i) = 1$. In this case, if it is one of $x_1, \ldots, x_{t-1}$, the algorithm will return it as $\hat{x}$, and hence $f^\star(\hat{x}) = 1 = \sup_x f^\star(x)$. If $f^\star \neq \mathbf{0}$ and none of $x_1, \ldots, x_{t-1}$ has $f^\star(x_i) = 1$, then it must be that $f^\star(x_t) = 1$, and indeed the algorithm returns x_t as its output $\hat{x}$ in this case, so that again we have $f^\star(\hat{x}) = 1 = \sup_x f^\star(x)$. The only remaining case is the situation with $f^\star = \mathbf{0}$. In this case, the algorithm returns x_t as its $\hat{x}$, but in fact *any* choice of $\hat{x}$ would be optimal, since $f^\star(\hat{x}) = 0 = \sup_x f^\star(x)$ in the case of $f^\star = \mathbf{0}$. Thus, in every case, we guarantee $f^\star(\hat{x}) = \sup_x f^\star(x)$, and the algorithm pulls $t-1$ arms.

To conclude the proof, we prove the complementary lower bound: that is, $M(\epsilon) \geq \tau^0_{\mathcal{F}} - 1$. Let t be any finite natural number with $t < \tau^0_{\mathcal{F}}$. Let $\mathcal{A}$ be any deterministic learning algorithm that makes at most $t-1$ queries; for simplicity, we suppose the algorithm always makes $t-1$ queries (otherwise we can make additional queries that are subsequently ignored to bring it up to $t-1$ total). consider a hypothetical run of the algorithm in which *every* reward it receives has value 0, and let $x_1, \ldots, x_{t-1}$ denote the sequence of arms it would pull in this hypothetical run, and let x_t denote the arm $\hat{x}$ the algorithm would return after these $t-1$ queries. Since $t < \tau^0_{\mathcal{F}}$, by minimality of $\tau^0_{\mathcal{F}}$ there must exist some $f^\star \in \mathcal{F} \setminus \{\mathbf{0}\}$ such that $f^\star(x_1) = \cdots = f^\star(x_t) = 0$. For this $f^\star$ as the target reward function, by induction the algorithm will indeed make as its $t-1$ queries the arms $x_1, \ldots, x_{t-1}$ in sequence, since at any time $s \leq t-1$ it will indeed have received reward 0 for each of the arms $x_1, \ldots, x_{s-1}$ it has previously pulled, and thus it will indeed choose x_s as its next query; moreover, having pulled the arms $x_1, \ldots, x_{t-1}$ and received reward 0 each time, it will indeed return x_t as its output arm $\hat{x}$. Thus, since $f^\star(x_t) = 0$, and $f^\star \neq \mathbf{0}$, we see that $f^\star(\hat{x}) = 0 = \sup_x f^\star(x) - 1 < \sup_x f^\star(x) - \epsilon$. This shows that for any $t < \tau^0_{\mathcal{F}}$, we have $M(\epsilon) > t - 1$. Therefore, $M(\epsilon) \geq \tau^0_{\mathcal{F}} - 1$. ■

5.2. Randomized Learners

Before proceeding with the discussion of learnability, we first show the relation between $\tilde{\tau}^0_{\mathcal{F}}(\epsilon)$ and $\tilde{\sigma}_{\mathcal{F}}$.

Proof of Lemma 5 We begin with establishing the rightmost inequality. Suppose $\tilde{\sigma}_{\mathcal{F}} > 0$. Let t equal the quantity on the right hand side. Let $\delta \in (0,1)$ and let P_δ be any probability measure on $\mathcal{X}$ for which

$$\inf_{f \in \mathcal{F} \setminus \{\mathbf{0}\}} P_\delta(x : f(x) = 1) > (1-\delta)\tilde{\sigma}_{\mathcal{F}}.$$

Define $x_1, \ldots, x_t$ as i.i.d. P_δ-distributed random variables. For any $f \in \mathcal{F} \setminus \{\mathbf{0}\}$,

$$\mathbb{P}(\nexists i \in \{1,\ldots,t\} : f(x_i) = 1) = (1 - P_\delta(x : f(x) = 1))^t < (1 - (1-\delta)\tilde{\sigma}_{\mathcal{F}})^t.$$

Since $\tilde{\sigma}_{\mathcal{F}} > 0$, the strict inequality $1 - \tilde{\sigma}_{\mathcal{F}} < e^{-\tilde{\sigma}_{\mathcal{F}}}$ holds. We can therefore choose $\delta > 0$ sufficiently small to satisfy $1 - (1-\delta)\tilde{\sigma}_{\mathcal{F}} \leq e^{-\tilde{\sigma}_{\mathcal{F}}}$. With this choice of δ, the rightmost expression above is at most

$$e^{-\tilde{\sigma}_{\mathcal{F}} t} \leq \epsilon.$$

Thus, $t \geq \tilde{\tau}^0_{\mathcal{F}}(\epsilon)$.

Next we establish the leftmost inequality in the lemma statement. Suppose $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) < \infty$. Let $t = \tilde{\tau}^0_{\mathcal{F}}(\epsilon)$ and let $x_1, \ldots, x_t$ be a sequence of $\mathcal{X}$-valued random variables such that $\forall f \in \mathcal{F} \setminus \{\mathbf{0}\}$, $\mathbb{P}(\exists i \in \{1,\ldots,t\} : f(x_i) = 1) \geq 1 - \epsilon$. Let P be the uniform mixture of the marginal distributions of the x_i random variables: that is, for measurable subsets $A \subseteq \mathcal{X}$,

$$P(A) = \frac{1}{t}\sum_{i=1}^{t} \mathbb{P}(x_i \in A).$$

For any $f \in \mathcal{F} \setminus \{\mathbf{0}\}$, we have

$$P(x : f(x) = 1) = \frac{1}{t}\sum_{i=1}^{t} \mathbb{P}(f(x_i) = 1) \geq \frac{1}{t}\mathbb{P}(\exists i \in \{1,\ldots,t\} : f(x_i) = 1) \geq \frac{1-\epsilon}{t},$$

where the first inequality is due to the union bound. Thus,

$$\inf_{f \in \mathcal{F} \setminus \{\mathbf{0}\}} P(x : f(x) = 1) \geq \frac{1-\epsilon}{t}.$$

Moreover, by definition of $\tilde{\sigma}_{\mathcal{F}}$,

$$\tilde{\sigma}_{\mathcal{F}} \geq \inf_{f \in \mathcal{F} \setminus \{\mathbf{0}\}} P(x : f(x) = 1).$$

Together, we have $\tilde{\sigma}_{\mathcal{F}} \geq \frac{1-\epsilon}{t}$, or equivalently, $\frac{1-\epsilon}{\tilde{\sigma}_{\mathcal{F}}} \leq t$, which establishes the leftmost inequality in the lemma.

Finally, we address the case of $\tilde{\sigma}_{\mathcal{F}} = 0$ or $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) = \infty$. The proof of the rightmost inequality only required $\tilde{\sigma}_{\mathcal{F}} > 0$, and yields the implication that $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) < \infty$ in this case. On the other hand, the proof of the leftmost inequality only required $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) < \infty$, and yields the implication that

$\tilde{\sigma}_{\mathcal{F}} > 0$ in this case. Thus, these proofs together imply that $\tilde{\sigma}_{\mathcal{F}} > 0$ if and only if $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) < \infty$, or equivalently, $\tilde{\sigma}_{\mathcal{F}} = 0$ if and only if $\tilde{\tau}^0_{\mathcal{F}}(\epsilon) = \infty$. ■

We are now ready for the proof of Theorem 5.

Proof of Theorem 5 For the upper bound, the algorithm pulls arms $x_1, \ldots, x_{t-1}$. If one of them has reward 1, it returns $\hat{x}$ as that one. Otherwise it returns $\hat{x} = x_t$.

For the lower bound, take any learner and let $x_1, \ldots, x_{t-1}$ be the sequence of arms it would pull if every reward it receives is 0, and let x_t be its returned arm $\hat{x}$ again in the case that all of the rewards it receives are 0. If $t < \tilde{\tau}^0_{\mathcal{F}}(\epsilon)$, there exists a choice of $f^\star \in \mathcal{F} \setminus \{\mathbf{0}\}$ such that, with probability strictly greater than ϵ, every x_i has $f^\star(x_i) = 0$. In particular, in this case, the algorithm will indeed pull arms $x_1, \ldots, x_{t-1}$, and will indeed return $\hat{x} = x_t$, and moreover this implies $f^\star(\hat{x}) = 0$. Thus, $\mathbb{E}[f^\star(\hat{x})] < 1 - \epsilon = \sup_x f^\star(x) - \epsilon$. ■

To conclude this discussion, we note that Theorem 4 now follows immediately from Theorem 5 together with Lemma 5.

5.3. Separation Between Deterministic and Randomized Learners

As mentioned above, it is interesting to note that the optimal query complexity of randomized learners can be vastly smaller than that of deterministic learners, and indeed, Section 1.1 gives an example of a function class that is easily learnable by randomized algorithms, but which is not learnable by deterministic algorithms: namely, the class of indicators for sets of Lebesgue measure 1 on $[0, 1]$.

6. A General Learning Algorithm for Real-Valued Rewards

This section presents the extension of the technique above, for binary rewards, to the case of general $[0, 1]$-valued rewards. Unlike the binary-valued case, the technique presented here is not always optimal, and in a sense this is necessary, given Theorem 1.

As we did for the case of binary-valued rewards, we present results for deterministic and randomized learners separately.

6.1. A Deterministic Learner

The following algorithm extends the zero-teaching set idea to general level sets, together with a search for the appropriate level set cut-off.

Algorithm $\mathcal{A}_{\text{det}}(\mathcal{F}, Q, \epsilon)$:
0. $q \leftarrow 0$, $V \leftarrow \mathcal{F}$, $r_{\max} \leftarrow 0$, let $x_{\max} \in \mathcal{X}$ arbitrary
1. For $r^* = (1/2)\epsilon, \epsilon, (3/2)\epsilon, 2\epsilon, \ldots \left\lceil \frac{2-\epsilon}{\epsilon} \right\rceil \frac{\epsilon}{2}$
2. Let S_{r^*} be a minimal specifying set for level r^* wrt V
3. For each x in S_t
4. Pull arm x and get reward r; $q \leftarrow q + 1$
5. Let $V \leftarrow \{f \in V : f(x) = r\}$
6. If $r > r_{\max}$, set $(x_{\max}, r_{\max}) \leftarrow (x, r)$
7. If $q = Q$, Return $x_{\max}$ 8. Return $x_{\max}$

We have the following query complexity bound for this algorithm.

Theorem 10 *For any $f^\star \in \mathcal{F}$ and $\epsilon \in (0,1)$, for any $Q \geq \left\lceil \frac{2\tau_{\mathcal{F}}(\epsilon)}{\epsilon} \right\rceil$, $\mathcal{A}_{det}(\mathcal{F}, Q, \epsilon)$ makes at most Q queries and returns a point $\hat{x} \in \mathcal{X}$ with $f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon$.*

Proof First note that, for any nonempty $V \subseteq \mathcal{F}$, $\tau(V_c, c-\epsilon) \leq \tau(\mathcal{F}_c, c-\epsilon)$ (by monotonicity of the max function). In particular, on any given round t, we have

$$|S_t| \leq \tau(V_{\hat{r}_t^*}, \hat{r}_t^* - \epsilon) \leq \tau_{\mathcal{F}}(\epsilon).$$

From this, and the fact that the algorithm only has $\left\lceil \frac{2}{\epsilon} \right\rceil$ rounds, the stated value of Q suffices for the guarantee on $\hat{x}$. ■

The upper bound of Theorem 6 follows immediately from this.

6.2. A Randomized Learner

In this section we present a randomized learner for general bounded real-valued bandit problems. As was true of the binary-valued case in Section 5.2, the corresponding analysis offers significant improvements compared to the deterministic learner above.

We will show that the query complexity bound in Theorem 8 is achieved by the following algorithm.

Algorithm $\mathcal{A}_{\text{rand}}(\mathcal{F}, Q, \epsilon)$:
0. $q \leftarrow 0$, $V \leftarrow \mathcal{F}$, $r_{\max} \leftarrow 0$, let $x_{\max} \in \mathcal{X}$ arbitrary
1. For $r^* = (1/2)\epsilon, \epsilon, (3/2)\epsilon, 2\epsilon, \ldots \left\lceil \frac{2-\epsilon}{\epsilon} \right\rceil \frac{\epsilon}{2}$
2. Let S_{r^*} be a minimal randomized $(r^* - \frac{\epsilon}{4}, \frac{\epsilon}{2})$-level specifying set for V_{r^*}
3. For each x in S_{r^*}
4. Pull arm x and get reward r; $q \leftarrow q+1$
5. Let $V \leftarrow \{f \in V : f(x) = r\}$
6. If $r > r_{\max}$, set $(x_{\max}, r_{\max}) \leftarrow (x, r)$
7. If $q = Q$, Return $x_{\max}$
8. Return $x_{\max}$

We have the following query complexity bound for this algorithm.

Theorem 11 *For any $f^\star \in \mathcal{F}$ and $\epsilon \in (0,1)$, for any*

$$Q \geq \frac{2\tilde{\tau}_{\mathcal{F}}(\epsilon/4, \epsilon/2)}{\epsilon},$$

$\mathcal{A}_{rand}(\mathcal{F}, Q, \epsilon)$ makes at most Q queries and returns a point $\hat{x} \in \mathcal{X}$ with

$$\mathbb{E} f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon.$$

Proof First note that, for any nonempty $V \subseteq \mathcal{F}$, the minimal size of a randomized $(c-\epsilon,\delta)$-level specifying set for V_c is no larger than that for $\mathcal{F}_c$ (by monotonicity of the max function). In particular, on any given round, we have

$$|S_{r^*}| \leq \tilde{\tau}_{\mathcal{F}}(\epsilon/4, \epsilon/2).$$

Since there are at most $2/\epsilon$ rounds, the size of Q in the theorem suffices to complete all rounds. Also note that we always have $f^\star \in V$. Moreover, note that the largest r^* in the algorithm for which $\sup_x f^\star(x) \geq r^*$ has size at least $\sup_x f^\star(x) - \epsilon/2$. Thus, for that round of the algorithm, by definition of S_{r^*}, with probability at least $1-\epsilon/2$, S_{r^*} contains at least one arm x with $f^\star(x) \geq r^* - \epsilon/4 \geq \sup_x f^\star(x) - \epsilon/2$. In particular, on this event, this implies that the returned $\hat{x} = x_{\max}$ upon termination satisfies $f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon/2$. Thus, since this fails only with probability at most $\epsilon/2$, we have

$$\mathbb{E} f^\star(\hat{x}) \geq \sup_x f^\star(x) - \epsilon.$$

■

As with the binary case, there exist function classes that are not learnable deterministically, but which are quite learnable by simple randomized learners. The same example given there remains valid: i.e., $\mathcal{F}$ as indicators of all measurable subsets of $[0,1]$ having Lebesgue measure 1.

The above completes the proof of Theorem 8 stated in Section 1.1. To complete the proof of Theorem 7 from that section, it suffices to prove the relation between $\tilde{\tau}_{\mathcal{F}}(\epsilon,\delta)$ and $\tilde{\sigma}_{\mathcal{F}}(\epsilon)$ stated in Lemma 9.

Proof of Lemma 9 We begin with establishing the rightmost inequality. Suppose $\epsilon, \delta \in (0,1)$, and suppose $\tilde{\sigma}_{\mathcal{F}}(\epsilon) > 0$ (so that the right inequality is non-vacuous). Let t equal the quantity on the right hand side. Let $\gamma \in (0,1)$ and let P_γ be any probability measure on $\mathcal{X}$ for which

$$\inf_{c\in(0,1]} \inf_{f\in\mathcal{F}_c} P_\gamma(x : f(x) \geq c-\epsilon) > (1-\gamma)\tilde{\sigma}_{\mathcal{F}}(\epsilon).$$

Define $x_1, \ldots, x_t$ as i.i.d. P_γ-distributed random variables. For any $c \in [0,1]$ and $f \in \mathcal{F}_c$,

$$\mathbb{P}\left(\max_{1\leq i\leq t} f(x_i) < c-\epsilon\right) = \left(1 - P_\gamma(x : f(x) \geq c-\epsilon)\right)^t < \left(1-(1-\gamma)\tilde{\sigma}_{\mathcal{F}}(\epsilon)\right)^t.$$

Since $\tilde{\sigma}_{\mathcal{F}}(\epsilon) > 0$, the strict inequality $1-\tilde{\sigma}_{\mathcal{F}}(\epsilon) < e^{-\tilde{\sigma}_{\mathcal{F}}(\epsilon)}$ holds. We can therefore choose $\gamma > 0$ sufficiently small to satisfy $1-(1-\gamma)\tilde{\sigma}_{\mathcal{F}}(\epsilon) \leq e^{-\tilde{\sigma}_{\mathcal{F}}(\epsilon)}$. With this choice of γ, the rightmost expression above is at most

$$e^{-\tilde{\sigma}_{\mathcal{F}}(\epsilon)t} \leq \delta.$$

Thus, $t \geq \tilde{\tau}_{\mathcal{F}}(\epsilon,\delta)$.

Next we establish the leftmost inequality in the lemma statement. Suppose $\tilde{\tau}_{\mathcal{F}}(\epsilon,\delta) < \infty$ (so that the inequality is non-vacuous). Let $t = \tilde{\tau}_{\mathcal{F}}(\epsilon,\delta)$ and for some $c \in [0,1]$ let $x_1, \ldots, x_t$ be a sequence of $\mathcal{X}$-valued random variables such that $\forall f \in \mathcal{F}_c$, $\mathbb{P}\left(\max_{1\leq i\leq t} f(x_i) \geq c-\epsilon\right) \geq 1-\delta$. Let P be the uniform mixture of the marginal distributions of the x_i random variables: that is, for measurable subsets $A \subseteq \mathcal{X}$,

$$P(A) = \frac{1}{t}\sum_{i=1}^{t} \mathbb{P}(x_i \in A).$$

For any $f \in \mathcal{F}_c$, we have

$$P(x : f(x) \geq c - \epsilon) = \frac{1}{t}\sum_{i=1}^{t} \mathbb{P}(f(x_i) \geq c - \epsilon) \geq \frac{1}{t}\mathbb{P}\left(\max_{1 \leq i \leq t} f(x_i) \geq c - \epsilon\right) \geq \frac{1-\delta}{t},$$

where the first inequality is due to the union bound. Since this argument holds for any choice of c, we have

$$\inf_{c \in (0,1]} \inf_{f \in \mathcal{F}_c} P(x : f(x) \geq c - \epsilon) \geq \frac{1-\delta}{t}.$$

Moreover, by definition of $\tilde{\sigma}_{\mathcal{F}}(\epsilon)$,

$$\tilde{\sigma}_{\mathcal{F}}(\epsilon) \geq \inf_{c \in (0,1]} \inf_{f \in \mathcal{F}_c} P(x : f(x) \geq c - \epsilon).$$

Together, we have $\tilde{\sigma}_{\mathcal{F}}(\epsilon) \geq \frac{1-\delta}{t}$, or equivalently, $\frac{1-\delta}{\tilde{\sigma}_{\mathcal{F}}(\epsilon)} \leq t$, which establishes the leftmost inequality in the lemma. ■

7. Open Problems

We conclude this work by stating some important open problems.

Decidability of Binary Bandit Learnability: First, we have shown that general bandit learning problems can be undecidable within ZFC. However, we also provided a concise characterization of bandit learnability in the case of binary-valued bandits. This raises a natural question:

Is bandit learnability decidable within ZFC for all binary-valued bandit problems?

Decidability of Bandit Learnability with Noise: A second important direction for study is bandit learnability with *noise*. Namely, it is common to consider the rewards $r_t(x)$ from pulling each arm x (on round t) to be independent random variables (independent across multiple rounds pulling arm x as well). In this case, $f^\star(x)$ is the conditional mean: $f^\star(x) = \mathbb{E}[r_t(x)]$. In this case, $f^\star \in \mathcal{F}$ corresponds to a *well-specified model* assumption. Let us still suppose $r_t(x)$ takes values in a bounded range (say, $[0, 1]$ without loss of generality). The objective of bandit learning then is that, within $M(\varepsilon)$ queries (observing the $r_t(x_t)$ values for the query sequence x_t, chosen adaptively, observing the past $r_{t'}(x_{t'})$, $t' < t$, values when selecting x_t), the learner should return $\hat{x}$ such that $\mathbb{E}[f^\star(\hat{x})] \geq \sup_x f^\star(x) - \varepsilon$: i.e., it nearly optimizes the conditional mean reward. We say a given $(\mathcal{X}, \mathcal{F})$ is learnable in the bandit setting with noise under the well-specified model assumption if this is achievable for all $f^\star \in \mathcal{F}$ (with finite query complexity $M(\varepsilon)$, for all $\varepsilon > 0$, by some algorithm $\mathcal{A}$) for all t-invariant distributions for each $r_t(x) \in [0, 1]$ subject to $f^\star(x) = \mathbb{E}[r_t(x)]$ for all $x \in \mathcal{X}$.

Bandit learning with noise is potentially a harder problem than the noise-free bandit learning problems considered in the present work. For instance, for $\mathcal{X} = \mathbb{N} \cup \{0\}$, consider a countable function class $\mathcal{F} = \{f_i : i \in \mathbb{N}\}$ of functions $\mathbb{N} \cup \{0\} \to [0, 1]$ such that, for $x \in \mathbb{N}$, $f_i(x) = \mathbb{1}[x = i]$. Clearly, bandit learning would be impossible for such a function class, since there is no structure to help the search for this single arm taking the maximum value 1, while all other arms have reward 0. However, we can add extremely helpful structure by setting the value $f_i(0)$ at 0 to

be 2^{-i}. Thus, while querying $f^\star(0)$ never itself provides a large reward (relative to the max reward, which is 1), it does reveal the precise identity of $f^\star$: that is, it reveals the $i \in \mathbb{N}$ for which $f^\star = f_i$, and hence also reveals the location of an $x^* = i$ with $f^\star(x^*) = \sup_x f^\star(x) = 1$, so that the learner may return $\hat{x} = x^*$ after just one query (i.e., $M(\epsilon) = 1$ for all $\epsilon \geq 0$). On the other hand, if we allow noise, the learner could be denied this information. For instance, for $f^\star = f_i$, the distribution of $r_t(0)$ could be $\text{Bernoulli}(2^{-i})$, while the rest could be noise-free: $r_t(x) = f^\star(x)$ for $x \in \mathbb{N}$. While indeed $\mathbb{E}[r_t(0)] = 2^{-i} = f^\star(0)$ (so that the well-specified model assumption is satisfied), for any finite number m, there are infinitely many function $f \in \mathcal{F}$ which, with very high probability, would simply return $r_t(0) = 0$ for all $t \leq m$, so that the learner would be unable to distinguish among these functions, and hence would still have an infinite number of possible $x \in \mathbb{N}$ where the x^* maximizing $f^\star$ may be. This can be made formal via the probabilistic method (for any given m, choosing the target i at random from a sufficiently large segment of $\{i_m, \ldots, i'_m\}$ for i_m sufficiently large), so that the expected reward of the learner can be made arbitrarily small; we leave the details as an exercise.

While learnability of noisy bandit problems may be more restrictive than noise-free problems, it is also, in some sense, less delicate: that is, less dependent on precise function values in the structure of $\mathcal{F}$. This can be seen in the above example. Also for this reason, the type of construction giving rise to our undecidability proof in Theorem 1 fails when noise is allowed. In particular, the constructed $(\mathcal{X}, \mathcal{F})$ in Theorem 1 is not learnable with noise, and this fact is decidable within ZFC. This leads to the following natural question:

Is bandit learnability with noise under the well-specified model assumption always decidable within ZFC?

References

R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995.

S. Ben-David, P. Hrubes, S. Moran, A. Shpilka, and A. Yehudayoff. Learnability can be undecidable. *Nature Machine Intelligence*, 1:44–48, 2019a.

S. Ben-David, P. Hrubes, S. Moran, A. Shpilka, and A. Yehudayoff. On a learning problem that is independent of the set theory ZFC axioms. *arXiv:1711.05195*, 2019b.

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

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

M. C. Caro. Undecidability of learnability. *arXiv:2106.01382*, 2021.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv:2112.13487*, 2021a.

D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Proceedings of the* 34^{th} *Conference on Learning Theory*, 2021b.

D. J. Foster, N. Golowich, J. Qian, A. Rakhlin, and A. Sekhari. A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv:2211.14250*, 2022.

D. J. Foster, N. Golowich, and Y. Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *Proceedings of the* 36^{th} *Conference on Learning Theory*, 2023.

S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the* 24^{th} *International Conference on Machine Learning*, 2007.

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014.

S. Hanneke, A. Kontorovich, S. Sabato, and R. Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129–2150, 2021.

T. B. Hashimoto, S. Yadlowsky, and J. C. Duchi. Derivative free optimization via repeated classification. In *Proceedings of the* 21^{st} *International Conference on Artificial Intelligence and Statistics*, 2018.

T. Jech. *Set Theory: Third Millenium Edition, Revised and Expanded*. Springer, Berlin, 2003.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 18*, 2004.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the* 40^{th} *ACM Symposium on Theory of Computing*, 2008.

K. Kunen. *Set Theory: An Introduction to Independence Proofs*. Elsevier, Amsterdam, 1980.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

S. Minsker. Estimation of extreme values and associated level sets of a regression function via selective sampling. In *Proceedings of the* 26^{th} *Conference on Learning Theory*, 2013.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55:527–535, 1952.

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

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 16(4):713–745, 2015.