
Greedy Algorithm for Structured Bandits: A Sharp Characterization of Asymptotic Success / Failure

Aleksandrs Slivkins
Microsoft Research, New York City
slivkins@microsoft.com

Yunzong Xu
University of Illinois Urbana-Champaign
xyz@illinois.edu

Shiliang Zuo
University of Illinois Urbana-Champaign
szuo3@illinois.edu

Abstract

We study the greedy (exploitation-only) algorithm in bandit problems with a known reward structure. We allow arbitrary finite reward structures, while prior work focused on a few specific ones. We fully characterize when the greedy algorithm asymptotically succeeds or fails, in the sense of sublinear vs. linear regret as a function of time. Our characterization identifies a partial identifiability property of the problem instance as the necessary and sufficient condition for the asymptotic success. Notably, once this property holds, the problem becomes easy—*any* algorithm will succeed (in the same sense as above), provided it satisfies a mild non-degeneracy condition. Our characterization extends to contextual bandits and interactive decision-making with arbitrary feedback. Examples demonstrating broad applicability and extensions to infinite reward structures are provided.

1 Introduction

Online learning algorithms often face uncertainty about the counterfactual outcomes of their actions. To navigate this uncertainty, they balance two competing objectives: *exploration*, making potentially suboptimal decisions to acquire information, and *exploitation*, leveraging known information to maximize rewards. This trade-off is central to the study of *multi-armed bandits* [Slivkins, 2019, Lattimore and Szepesvári, 2020], a foundational framework in sequential decision-making.

While exploration is central to bandit research, it presents significant challenges in practice, esp. when an algorithm interacts with human users. First, exploration can be wasteful and risky for the current user, imposing a burden that may be considered unfair since its benefits primarily accrue to future users. Second, exploration adds complexity to algorithm design, and its adoption in large-scale applications requires substantial buy-in and engineering support compared to a system that only exploits [Agarwal et al., 2016, 2017]. Third, exploration may be incompatible with users’ incentives when actions are controlled by the users. *E.g.*, an online platform cannot *force* users to try and review new products; instead, users gravitate toward well-reviewed or familiar options [Kremer et al., 2014].¹

A natural alternative is the *greedy algorithm* (Greedy), which exploits known information at every step without any intentional exploration. This approach sidesteps the aforementioned challenges and often better aligns with user incentives. In particular, it models the natural dynamics in an online

1. Enforcing exploration in such settings is very challenging (Kremer et al. [2014] and follow-up work, see Slivkins [2023] for an overview). While exploration *can* be made incentive-compatible, doing so involves considerable performance and/or monetary costs and additional complexity. More importantly, it hinges upon substantial (even if standard) assumptions from economic theory.

platform where each user acts in self-interest, making decisions based on full observations of previous users' actions and outcomes, *e.g.*, purchases and product reviews [Acemoglu et al., 2022].

Despite its simplicity and practical appeal, Greedy is widely believed to perform poorly. This belief is deeply ingrained in the bandit literature, which overwhelmingly focuses on exploration as a necessary ingredient for minimizing regret. A key motivation for this focus comes from well-known failure cases in *unstructured* K -armed bandits. A classic example is as follows: “Suppose the reward of each arm follows an independent Bernoulli distribution with a fixed mean, and Greedy is initialized with a single sample per arm. If the best arm initially returns a 0 while another arm returns a 1, Greedy permanently excludes the best arm.”

However, beyond such examples, the broader picture remains murky, especially for the widely-studied *structured bandits* – bandit problems with a known reward structure (*e.g.*, linearity, Lipschitzness, convexity) – where observing some actions provides useful information about others. Formally, a reward structure restricts the possible *reward functions* that map arms to their mean rewards. Reward structures reduce the need for explicit exploration, making the bandit problem more tractable. For *some* of them, Greedy in fact succeeds, *e.g.*, two-armed bandits with expected rewards that sum up to a known value. The literature provides a few examples of failure for some specific (one-dimensional, linear) reward structures, and a few non-trivial examples of success (*e.g.*, for linear contextual bandits); see more on this in Related Work. Likewise, large-scale experiments yield mixed results: some settings confirm the need for exploration, but others indicate that Greedy performs well [Bietti et al., 2021]. This contrast raises a fundamental question: *When—and why—does Greedy fail or succeed?*

Our Contributions. We work towards the missing foundation for structured bandits: a general theory of Greedy. Our main result allows finite, but otherwise *arbitrary* reward structures. We provide a complete characterization of when Greedy asymptotically *fails* (incurs linear regret) vs when it *succeeds* (achieves sublinear regret). Our characterization applies to *every* problem instance, resolving it in the positive or negative direction, not (just) in the worst case over a particular reward structure. The negative results are of primary interest here, as they substantiate the common belief that Greedy performs poorly, and the positive results serve to make the characterization precise.

A key insight is identifying a new “partial identifiability” property of the problem instance, called *self-identifiability*, as a *necessary and sufficient* condition for the asymptotic success. Self-identifiability asserts that, given the reward structure, fixing the expected reward of a suboptimal arm uniquely identifies it as suboptimal. We prove that Greedy achieves sublinear regret under self-identifiability, and suffers from linear regret otherwise. The negative result is driven by the existence of a *decoy*: informally, an alternative reward model such that its optimal arm is suboptimal for the true model and both models coincide on this arm. We show that with some positive probability, Greedy gets permanently stuck on such a decoy, for an infinite time horizon. For the positive result, Greedy succeeds (only) because self-identifiability makes the problem instance intrinsically “easy”. In fact, this success is *not* due to any particular cleverness of Greedy: we show that *any* reasonable algorithm (satisfying a mild non-degeneracy condition) achieves sublinear regret under self-identifiability.

Our characterization allows for essentially an arbitrary interaction protocol between the algorithm and the environment (Section 5). Specifically, we handle the model of “decision-making with structured observations” (DMSO, Foster et al. [2021]), which allows for arbitrary auxiliary feedback after each round. This model subsumes contextual bandits, combinatorial semi-bandits, and bandits with graph-based feedback, as well as episodic reinforcement learning. Before moving to this full generality, our presentation focuses on contextual bandits, where we obtain quantitatively stronger guarantees (Section 4), and “vanilla” bandits as a paradigmatic case for building key intuition and cleaner definitions (Section 3).

We apply our machinery to several examples, both positive and negative (Appendix D). We demonstrate that most infinite structures of interest admit meaningful finite analogs via discretization. We find that Greedy fails in linear bandits, Lipschitz bandits and “polynomial bandits” (with arms in $\mathbb{R}$ and polynomial expected rewards), and does so for *almost all* problem instances. For linear *contextual* bandits, Greedy succeeds if the context set is “sufficiently diverse”, but may fail if it is “low-dimensional”. For *Lipschitz* contextual bandits, Greedy behaves very differently, failing for almost all instances. One informal takeaway is that Greedy fails as a common case for most/all bandit structures of interest, whereas for *contextual* bandits it can go either way, depending on the structure. The success of Greedy appears to require context diversity and a parametric reward structure.

The second main result of this work concerns *infinite* (e.g., continuous) reward structures (Section 6). While our earlier analysis gave a sharp “if and only if” characterization for finite reward structures, such a complete characterization is more challenging to obtain for infinite ones. To make progress, we provide a characterization parameterized by a notion of “margin” (separating instances for which the positive result applies from instances for which the negative result applies), with guarantees that deteriorate as the margin vanishes. Subject to this margin, our result handles arbitrary infinite reward structures. It applies to structured bandits with finite action sets, requiring stronger notions of self-identifiability and decoy existence (parameterized by the margin), as well as new analysis ideas.

Discussion. The distinction between linear and sublinear regret is a standard “first-order” notion of success vs failure in bandits. Our positive results attain logarithmic, instance-dependent regret rates, possibly with a large multiplicative constant determined by the reward structure and the instance. Our negative results establish a positive (but possibly very small) constant probability of a “failure event” where Greedy gets permanently stuck on a decoy, for an infinite time horizon. Optimizing these constants for an arbitrary reward structure appears difficult. However, we achieve much better constants for the partial characterization in Section 6.

The greedy algorithm is initialized with some warm-up data collected from the same problem instance (and it needs at least 1 warm-up sample to be well-defined). Our negative results require exactly one warm-up sample for each context-arm pair. All our positive results allow for an arbitrary amount of initial data. Thus, our characterization effectively defines “success” as sublinear regret for any amount of warm-up data, and “failure” as linear regret for *some* amount of warm-up data.

We assume that Greedy is given a *regression oracle*: a subroutine to perform (least-squares) regression given the reward structure. As in “bandits with regression oracles” (referenced below), we separate out computational issues, leveraging prior work on regression, and focus on the statistical guarantees.

Related Work. Bandit reward structures studied in prior work include linear and combinatorial structures [e.g., Awerbuch and Kleinberg, 2008, McMahan and Blum, 2004, György et al., 2007, Cesa-Bianchi and Lugosi, 2012], convexity [e.g., Kleinberg, 2004, Flaxman et al., 2005, Bubeck et al., 2017], and Lipschitzness [e.g., Kleinberg, 2004, Kleinberg et al., 2008, Bubeck et al., 2011], as well as some others. Each of these is a long line of work on its own, with extensions to contextual bandits [e.g., Li et al., 2010, Slivkins, 2014]. There’s also some work on bandits with *arbitrary* reward structures [Amin et al., 2011, Combes et al., 2017, Jun and Zhang, 2020, Degenne et al., 2020, Parys and Golrezaei, 2024], and particularly contextual bandits with regression oracles [e.g., Agarwal et al., 2012, Foster et al., 2018, Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2022]. For more background, see books Slivkins [2019], Lattimore and Szepesvári [2020], Foster and Rakhlin [2023].

For Greedy, positive results with near-optimal regret rates focus on linear contextual bandits with diverse/smoothed contexts [Kannan et al., 2018, Bastani et al., 2021, Raghavan et al., 2023, Kim and Oh, 2024]. Both context diversity and parametric reward structure are essential. Our positive results for the same setting are incomparable: more general in terms of context diversity assumptions, but weaker in terms of the regret bounds. (This line of prior work does not contain any negative results.) Greedy is also known to attain $o(T)$ regret in various scenarios with a very large number of near-optimal arms [Bayati et al., 2020, Jedor et al., 2021].²

Negative results for Greedy are derived for “non-structured” K -armed bandits: from trivial extensions of the single-sample-per-arm example mentioned above, to an exponentially stronger characterization of failure probability [Banihashem et al., 2023], to various “near-greedy” algorithms / behaviors, both “frequentist” and “Bayesian” (same paper). Negative results for non-trivial reward structures concern dynamic pricing with linear demands [Harrison et al., 2012, den Boer and Zwart, 2014] and dynamic control in a generalized linear model [Lai and Robbins, 1982, Keskin and Zeevi, 2018]. Banihashem et al. [2023] also obtain negative results for the Bayesian version of Greedy in Bayesian bandits, under a certain “full support” assumption on the prior.³

2 Preliminaries: structured contextual bandits (StructuredCB)

We have *action set* $\mathcal{A}$ and *context set* $\mathcal{X}$. In each round $t = 1, 2, \dots$, a context $x_t \in \mathcal{X}$ arrives, an algorithm chooses an action (arm) $a_t \in \mathcal{A}$, and a reward $r_t \in \mathbb{R}$ is realized. The context is drawn

2. E.g., for Bayesian bandits with $\gg \sqrt{T}$ arms, where the arms’ mean rewards are sampled uniformly.

3. Essentially, the prior covers all reward functions $\{\text{arms}\} \rightarrow [0, 1]$ with probability density at least $p > 0$.

independently from some fixed and known distribution over $\mathcal{X}$.⁴ The reward r_t is an independent draw from a unit-variance Gaussian with unknown mean $f^*(x_t, a_t) \in [0, 1]$.⁵ A *reward function* is a function $f : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$; in particular, f^* is the *true* reward function. The reward structure is given by a known class $\mathcal{F}$ of reward functions which contains f^* ; the assumption $f^* \in \mathcal{F}$ is known as *realizability*. To recap, the problem instance is a pair $(f^*, \mathcal{F})$, where $\mathcal{F}$ is known and f^* is not.

We focus on *finite* reward structures, *i.e.*, assume (unless specified otherwise) that $\mathcal{X}, \mathcal{A}, \mathcal{F}$ are all finite. While this does not hold for most reward structures from prior work, one can discretize them to ensure finiteness. Indeed, when reward functions can take infinitely many values, one could round each function value to the closest point in some finite subset $S \subset [0, 1]$, *e.g.*, all integer multiples of some $\varepsilon > 0$. Likewise, one could discretize contexts, arms, or function parameters, when they are represented as points in some metric space, *e.g.*, as real-valued vectors. Or, one could define finite reward structures *directly*, with similar discretizations built-in (see Appendix D for examples).

We are interested in expected regret $\mathbb{E}[R(t)]$ as a function of round t . Regret is standard: $R(t) := \sum_{s \in [t]} (r^*(x_s) - r_s)$, where $r^*(x) := \max_{a \in \mathcal{A}} f^*(x, a)$, best expected reward given context x .

The greedy algorithm (Greedy) is defined as follows. It is initialized with $T_0 \geq 1$ rounds of warm-up data, denoted $t \in [T_0]$.⁶ Each such round yields a context-arm pair $(x_t, a_t) \in \mathcal{X} \times \mathcal{A}$ chosen exogenously, and reward $r_t \in \mathbb{R}$ drawn independently from the resp. reward distribution: unit-variance Gaussian with mean $f^*(x_t, a_t)$. At each round $t > T_0$, Greedy computes a reward function via least-squares regression (implemented via a “regression oracle”, as per Section 1):

$$f_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{s \in [t]} (f(x_s, a_s) - r_s)^2. \quad (2.1)$$

Note that there are no ties in (2.1) with probability one over the random rewards. Once reward function f_t is chosen, the algorithm chooses the best arm for f_t and context x_t , *i.e.*,

$$a_t \in \operatorname{argmax}_{a \in \mathcal{A}} f_t(x_t, a). \quad (2.2)$$

For ease of presentation, we posit that $f(x, \cdot)$ has a unique maximizer, for each feasible function $f \in \mathcal{F}$ and each context $x \in \mathcal{X}$; call such f *best-arm-unique*. (Our results can be adapted to allow for reward functions with multiple best arms, see Appendix A.)

Notation. Let K be the number of arms; identify the action set as $\mathcal{A} = [K]$. The number of times a given arm a was chosen for a given context x before round t is denoted $N_t(x, a)$, and the corresponding average reward is $\bar{r}_t(x, a)$. Average reward over the warm-up stage is denoted $\bar{r}_{\text{warm}}(x, a) := \bar{r}_t(x, a)$ with $t = T_0 + 1$. We’ll work with an alternative loss function,

$$\text{MSE}_t(f) := \sum_{(x,a) \in \mathcal{X} \times \mathcal{A}} N_t(x, a) (\bar{r}_t(x, a) - f(x, a))^2. \quad (2.3)$$

Note that it is equivalent to (2.1) for minimization, in the sense that $f_t = \operatorname{argmin}_{f \in \mathcal{F}} \text{MSE}_t(f)$.

3 Characterization for structured bandits

Let us focus on the paradigmatic special case of multi-armed bandits, call it StructuredMAB. Formally, there is only one context, $|\mathcal{X}| = 1$. The context can be suppressed from the notation; *e.g.*, reward functions map arms to $[0, 1]$. An arm is called *optimal* for a given reward function f (or, by default, for $f = f^*$) if it maximizes expected reward $f(\cdot)$, and *suboptimal* otherwise.

We start with two key definitions. *Self-identifiability* (which drives the positive result) asserts that fixing the expected reward of any suboptimal arm identifies this arm as suboptimal.

Definition 1 (Self-identifiability). *Fix a problem instance $(f^*, \mathcal{F})$. A suboptimal arm a is called self-identifiable if fixing its expected reward $f^*(a)$ identifies this arm as suboptimal given $\mathcal{F}$, *i.e.*, if arm a is suboptimal for any reward function $f \in \mathcal{F}$ consistent with $f(a) = f^*(a)$. If all suboptimal arms have this property, then the problem instance is called self-identifiable.*

4. Whether the context distribution is known to the algorithm is inconsequential, since Greedy (particularly, the regression in Eq. (2.1)) does not use this knowledge. W.l.o.g., $\mathcal{X}$ is the support set of the context distribution.

5. Gaussian reward noise is a standard assumption in bandits (along with *e.g.*, 0-1 rewards), which we make for ease of presentation. Our positive results carry over to rewards with an arbitrary sub-Gaussian noise, without any modifications. Likewise, our negative results carry over to rewards $r_t \in [0, 1]$ with an arbitrary *near-uniform* distribution, *i.e.*, one specified by a p.d.f. on $[0, 1]$ which is bounded away from 0 by an absolute constant.

6. We also refer to the first T_0 rounds as *warm-up stage*, and the subsequent rounds as *main stage*.

A *decoy* (whose existence drives the negative result) is another reward function f_{dec} such that its optimal arm a_{dec} is suboptimal for f^* and both reward functions coincide on this arm.

Definition 2 (Decoy). *Let f^*, f_{dec} be two reward functions, with resp. optimal arms a^*, a_{dec} . Call f_{dec} a decoy for f^* (with a decoy arm a_{dec}) if it holds that $f_{\text{dec}}(a_{\text{dec}}) = f^*(a_{\text{dec}}) < f^*(a^*)$.*

We emphasize that self-identifiability and decoys are new notions, not reducible to structural notions from prior work, see Appendix B. It is easy to see that they are equivalent, in the following sense:

Claim 1. *An instance $(f^*, \mathcal{F})$ is self-identifiable if and only if f^* has no decoy in $\mathcal{F}$.*

In our characterization, the complexity of the problem instance $(f^*, \mathcal{F})$ enters via its *function-gap*,

$$\Gamma(f^*, \mathcal{F}) = \min_{\text{functions } f \in \mathcal{F}: f \neq f^*} \min_{\text{arms } a: f(a) \neq f^*(a)} |f^*(a) - f(a)|. \quad (3.1)$$

We may also write $\Gamma(f^*) = \Gamma(f^*, \mathcal{F})$ when the function class $\mathcal{F}$ is clear from context.

Theorem 1. *Fix a problem instance $(f^*, \mathcal{F})$ of StructuredMAB.*

- (a) *If the problem instance is self-identifiable, then Greedy (with any warm-up data) satisfies $\mathbb{E}[R(t)] \leq T_0 + (K/\Gamma(f^*))^2 \cdot O(\log t)$ for each round $t \in \mathbb{N}$.*
- (b) *Suppose the warm-up data consists of one sample for each arm. Assume f^* has a decoy $f_{\text{dec}} \in \mathcal{F}$, with decoy arm a_{dec} . Then with some probability $p_{\text{dec}} > 0$ it holds that Greedy chooses a_{dec} for all rounds $t \in (T_0, \infty)$. We can lower-bound p_{dec} by $e^{-O(K/\Gamma^2(f_{\text{dec}}))}$.*

Discussion. Thus, Greedy succeeds, in the sense of achieving sublinear regret for any warm-up data, if and only if the problem instance is self-identifiable. Else, Greedy fails for some warm-up data, incurring linear expected regret. Specifically, regret is $\mathbb{E}[R(t)] \geq (t - T_0) \cdot p_{\text{dec}} \cdot (f^*(a^*) - f^*(a_{\text{dec}}))$ for each round $t \in (T_0, \infty)$, where a^* is the best arm.

The correct perspective is that Greedy fails on every problem instance unless self-identifiability makes it intrinsically “easy”. Indeed, consider any bandit algorithm that avoids playing an arm once it is identified, with high confidence, as suboptimal and having a specific expected reward. This defines a mild yet fundamental non-degeneracy condition: a reasonable bandit algorithm should never take an action that provides neither new information (exploration) nor utility from existing information (exploitation), whether it prioritizes one or balances both. The class of algorithms satisfying this condition is broad—for instance, an algorithm may continue playing some arm a indefinitely as long as the reward structure *permits* this arm to be optimal. However, under self-identifiability, any algorithm satisfying this condition achieves sublinear regret (see Appendix C for details).

The failure probability p_{dec} could be quite low. When there are multiple decoys $f_{\text{dec}} \in \mathcal{F}$, we could pick one (in the analysis) which maximizes function-gap $\Gamma(f_{\text{dec}})$. We present a more efficient analysis under stronger assumptions (which also applies to infinite function classes), see Section 6.

Proof Sketch for Theorem 1(a). We show that a suboptimal arm a cannot be chosen more than $\tilde{O}(K/\Gamma^2(f^*))$ times throughout the main stage. Indeed, suppose a is chosen this many times by some round $t > T_0$. Then $\bar{r}_t(a)$, the empirical mean reward for a , is within $\Gamma(f^*)/2$ of its true mean $f^*(a)$ with high probability, by a standard concentration inequality. This uniquely identifies $f^*(a)$ by definition of the function-gap, which in turn identifies a as a suboptimal arm for any feasible reward function. Intuitively, this should imply that a cannot be chosen again. Making this implication formal is non-trivial, requiring an additional argument invoking $\text{MSE}_t(\cdot)$, as defined in (2.3).

First, we show that $\text{MSE}_t(f^*) \leq \tilde{O}(K)$ with high probability, using concentration. Next, we observe that any reward function f with $f(a) \neq f^*(a)$ will have a larger $\text{MSE}_t(\cdot)$, and therefore cannot be chosen in round t . It follows that $f_t(a) = f^*(a)$. Consequently, arm a is suboptimal for f_t (by self-identifiability), and hence cannot be chosen in round t .

Proof Sketch for Theorem 1(b). To show that Greedy gets permanently trapped on the decoy arm *despite reward randomness*, we define two carefully-constructed events. The first ensures that the warm-up data causes Greedy to misidentify f_{dec} as the true reward function for all non-decoy arms:

$$E_1 = \{ |\bar{r}_{\text{warm}}(a) - f_{\text{dec}}(a)| < \Gamma(f_{\text{dec}})/2 \text{ for each arm } a \neq a_{\text{dec}} \}. \quad (3.2)$$

This concerns the single warm-up sample per non-decoy arm. The second event ensures that the empirical mean of the decoy arm a_{dec} remains close to $f^*(a_{\text{dec}})$ for all rounds after the warm-up:

$$E_2 = \{ \forall t > T_0, |\bar{r}_t(a_{\text{dec}}) - f^*(a_{\text{dec}})| \leq \Gamma(f_{\text{dec}})/2 \}. \quad (3.3)$$

Under $E_1 \cap E_2$, Greedy always chooses the decoy arm. To lower-bound $\Pr[E_1 \cap E_2]$, note that E_2, E_1 are independent (as they concern, resp., a_{dec} and all other arms), analyze each event separately.

4 Characterization for structured contextual bandits (StructuredCB)

The ideas from Section 3 need non-trivial modifications. The naive reduction to bandits — treating each contexts-to-arms mapping as a “super-arm” in StructuredMAB — does *not* work because Greedy now observes contexts. Further, such reduction would replace the dependence on K in Theorem 1 with the number of mappings, i.e., $K^{|\mathcal{X}|}$, whereas we effectively replace it with $K \cdot |\mathcal{X}|$.

Some notation: mappings from contexts to arms are commonly called *policies*. Let Π denote the set of all policies. Expected reward of policy $\pi \in \Pi$ is $f^*(\pi) := \mathbb{E}_x[f^*(x, \pi(x))]$, where the expectation is over the fixed distribution of context arrivals. A policy π is called *optimal* for reward function f if it maximizes expected reward $f(\pi)$, and *suboptimal* otherwise. Let π^* be the optimal policy for f^* . Note that $\pi(x) \in \operatorname{argmax}_{a \in \mathcal{A}} f(x, a)$ for each context x , which is unique by assumption.

Greedy can be described in terms of policies: it chooses policy π_t in each round t , before seeing the context x_t , and then chooses arm $a_t = \pi_t(x_t)$. Here π_t is the optimal policy for the f_t from Eq. (2.1).

As in Section 3, the positive and negative results are driven by, resp., self-identifiability and the existence of a suitable “decoy”. Let’s extend these key definitions to contextual bandits.

Definition 3 (Self-identifiability). *Fix a problem instance $(f^*, \mathcal{F})$. A suboptimal policy $\pi \in \Pi$ is called self-identifiable if fixing its expected rewards $f^*(x, \pi(x))$ for all contexts $x \in \mathcal{X}$ identifies this policy as suboptimal given $\mathcal{F}$. Put differently: if this policy is suboptimal for any reward function $f \in \mathcal{F}$ such that $f(x, \pi(x)) = f^*(x, \pi(x))$ for all contexts x . If each suboptimal policy has this property, then the problem instance is called self-identifiable.*

Definition 4 (Decoy). *Let f^*, f_{dec} be two reward functions, with resp. optimal policies π^*, π_{dec} . Call f_{dec} a decoy for f^* (with a decoy policy π_{dec}) if it holds that $f_{\text{dec}}(\pi_{\text{dec}}) = f^*(\pi_{\text{dec}}) < f^*(\pi^*)$ and moreover $f_{\text{dec}}(x, \pi_{\text{dec}}(x)) = f^*(x, \pi_{\text{dec}}(x))$ for all contexts $x \in \mathcal{X}$.*

In words, the decoy and f^* completely coincide on the decoy policy, which is a suboptimal policy for f^* . The equivalence of these definitions holds word-by-word like in Claim 1.

The notion of function-gap is extended in a natural way:

$$\Gamma(f^*, \mathcal{F}) = \min_{\text{functions } f \in \mathcal{F}: f \neq f^*} \min_{(x,a) \in \mathcal{X} \times \mathcal{A}: f(x,a) \neq f^*(x,a)} |f^*(x,a) - f(x,a)|. \quad (4.1)$$

Our results are also parameterized by the distribution of context arrivals, particularly by the smallest arrival probability across all contexts, denoted p_0 . (W.l.o.g., $p_0 > 0$.)

Theorem 2. *Fix a problem instance $(f^*, \mathcal{F})$ of StructuredCB. Let $X = |\mathcal{X}|$.*

- (a) *If the problem instance is self-identifiable, then Greedy (with any warm-up data) satisfies $\mathbb{E}[R(t)] \leq T_0 + (|\mathcal{X}|K/\Gamma(f^*))^2/p_0 \cdot O(\log t)$ for each round $t \in \mathbb{N}$.*
- (b) *Suppose the warm-up data consists of one sample for each context-arm pair. Assume f^* has a decoy $f_{\text{dec}} \in \mathcal{F}$, with decoy policy π_{dec} . Then with some probability $p_{\text{dec}} > 0$, Greedy chooses π_{dec} in all rounds $t \in (T_0, \infty)$. We have $p_{\text{dec}} \geq X^{-O(KX/\Gamma^2(f_{\text{dec}}))}$.*

Remark 1. Greedy succeeds (i.e., achieves sublinear regret for any warm-up data) if and only if the problem instance is self-identifiable. Else, Greedy fails for some warm-up data, with linear regret:

$$\mathbb{E}[R(t)] \geq (t - T_0) \cdot p_{\text{dec}} \cdot (f^*(\pi^*) - f^*(\pi_{\text{dec}})) \text{ for each round } t \in (T_0, \infty). \quad (4.2)$$

New Proof Ideas. For Theorem 2(a), directly applying the proof techniques from the MAB case gives a regret bound linear in $|\Pi| = K^X$. Instead, we develop a non-trivial potential argument to achieve regret bound polynomial in KX . For Theorem 2(b), we give new definitions of events E_1, E_2 extending (3.2), (3.3) by carefully accounting for contexts, and refine the deviation analysis to remove the dependence on $|\Pi|$. Proof sketches and full proofs are in Appendix F.

5 Interactive decision-making with arbitrary feedback

We consider Decision-Making with Structured Observations (DMSO), a general framework for sequential decision-making with a known structure [Foster et al., 2021]. It allows for arbitrary feedback observed after each round, along with the reward.⁷ The root challenge is that this feedback is usually correlated with rewards. Greedy must account for these correlations, not just compute the best fit based on rewards. Our solution is to develop a natural variant of Greedy based on maximum-likelihood estimation. The analysis then becomes much more technical compared to StructuredCB, requiring us to track changes in log-likelihood and define the “model-gap” in terms of KL-divergence.

Preliminaries. DMSO is defined as follows. Instead of “arms” and “contexts”, we have two new primitives: a policy set Π and observation set $\mathcal{O}$. The interaction protocol is as follows: in each round $t = 1, 2, \dots$, the algorithm selects a policy $\pi_t \in \Pi$, receives a reward $r_t \in \mathcal{R} \subset \mathbb{R}$, and observes an observation $o_t \in \mathcal{O}$. A *model* is a mapping from Π to a distribution over $\mathbb{R} \times \mathcal{O}$. The reward-observation pair (r_t, o_t) is an independent sample from distribution $M^*(\pi_t)$, where M^* is the true model. The problem structure is represented as a (known) model class $\mathcal{M}$ which contains M^* . We assume that $\Pi, \mathcal{M}, \mathcal{R}, \mathcal{O}$ are all finite.⁸ To recap, the problem instance is a pair $(M^*, \mathcal{M})$, where $\mathcal{M}$ is known but M^* is not. This completes the definition of DMSO.

StructuredMAB is a simple special case of DMSO with one possible observation. StructuredCB is subsumed by interpreting the observations o_t as contexts and defining $M^*(\pi)$ accordingly, to account for the distribution of context arrivals, the reward distribution, and the reward function.⁹ The observations in DMSO can also include auxiliary feedback present in various bandit models studied in prior work. To wit: rewards of “atomic actions” in combinatorial semi-bandits [e.g., Györfy et al., 2007, Chen et al., 2013], per-product sales in multi-product dynamic pricing [e.g., Keskin and Zeevi, 2014, den Boer, 2014], and rewards of all “adjacent” arms in bandits with graph-based feedback [Alon et al., 2013, 2015]. Moreover, the observations can include MDP trajectories in episodic reinforcement learning [see Agarwal et al., 2020, for background]. DMSO subsumes all these scenarios, under the “realizability” assumption $M^* \in \mathcal{M}$.

We use some notation. Let $f(\pi|M)$ be the expected reward for choosing policy π under model M , and $f^*(\pi) := f(\pi|M^*)$. A policy is called *optimal* (under model M) if it maximizes $f(\cdot|M)$, and *suboptimal* otherwise. Let π^* be an optimal policy for M^* . The history $\mathcal{H}_t$ at round t consists of (π_s, r_s, o_s) tuples for all rounds $s < t$. $D \stackrel{d}{=} D'$ denotes that distributions D, D' are equal.

Modified Greedy. The modified greedy algorithm (GreedyMLE) uses maximum-likelihood estimation (MLE) to analyze reward-observation correlations. As before, the algorithm is initialized with $T_0 \geq 1$ rounds of warm-up data, denoted $t \in [T_0]$. Each round yields a tuple $(\pi_t, r_t, o_t) \in \Pi \times \mathbb{R} \times \mathcal{O}$, where the policy π_t is chosen exogenously, and the (r_t, o_t) pair is drawn independently from the corresponding distribution $M^*(\pi_t)$. At each round $t > T_0$, the algorithm determines

$$M_t \in \operatorname{argmax}_{M \in \mathcal{M}} \mathcal{L}(M | \mathcal{H}_t), \quad (5.1)$$

the model with the highest likelihood $\mathcal{L}(M|\mathcal{H}_t)$ given history $\mathcal{H}_t$ (with ties broken arbitrarily).¹⁰ Then the algorithm chooses the optimal policy given this model: $\pi_t \in \operatorname{argmax}_{\pi \in \Pi} f(\pi|M_t)$. For simplicity, we assume that the model class $\mathcal{M}$ guarantees no ties in this argmax. Here $\mathcal{L}(M|\mathcal{H}_t)$ is an algorithm-independent notion of likelihood: the probability of seeing the reward-observation pairs in history $\mathcal{H}_t$ under model M , if the policies in $\mathcal{H}_t$ were chosen in the resp. rounds. In a formula,

$$\mathcal{L}(M | \mathcal{H}_t) := \prod_{s \in [t-1]} \Pr_{M(\pi_s)}(r_s, o_s). \quad (5.2)$$

W.l.o.g. we can restrict Π to policies that are optimal for some model; in particular $|\Pi| \leq |\mathcal{M}|$.

Our characterization. We adapt the definitions of “self-identifiability” and “decoy” so that “two models coincide on a policy” means having the same distribution of reward-observation pairs.

Definition 5 (Self-identifiability). *Fix a problem instance $(M^*, \mathcal{M})$. A suboptimal policy π is called self-identifiable if fixing distribution $M^*(\pi)$ identifies this policy as suboptimal given $\mathcal{M}$. That is: if*

7. Bandit formulations with partial feedback that does *not* include the reward [known as *partial monitoring*, e.g., Bartók et al., 2014, Antos et al., 2013], are outside our scope.

8. Finiteness of $\mathcal{R}, \mathcal{O}$ is for ease of presentation. We can also handle infinite $\mathcal{R}, \mathcal{O}$ if all outcome distributions $M(\pi)$ have a well-defined density, and Assumption 1 is stated in terms of these densities.

9. Here we work with discrete rewards, whereas our treatment in Sections 3 and 4 assumes Gaussian rewards.

10. As in Section 2, the regression is implemented via a “regression oracle”; we focus on statistical guarantees.

policy π is suboptimal for any model $M \in \mathcal{M}$ with $M(\pi) \stackrel{d}{=} M^*(\pi)$. The problem instance is called self-identifiable if all suboptimal policies have this property.

Definition 6 (Decoy). Let M^*, M_{dec} be two models, with resp. optimal policies π^*, π_{dec} . Call M_{dec} a decoy for M^* (with a decoy policy π_{dec}) if $M_{\text{dec}}(\pi_{\text{dec}}) \stackrel{d}{=} M^*(\pi_{\text{dec}})$ (i.e., the two models completely coincide on π_{dec}) and moreover $\pi f^*(\pi_{\text{dec}}) < f^*(\pi^*)$ (i.e., π_{dec} is suboptimal for f^*).

Claim 2. A DMSO instance $(M^*, \mathcal{M})$ is self-identifiable if and only if M^* has no decoy in $\mathcal{M}$.

We define *model-gap*, a modification of function gap which tracks the difference in reward-observation distributions (expressed via KL-divergence, denoted D_{KL}). The model gap of model $M \in \mathcal{M}$ is

$$\Gamma(M, \mathcal{M}) := \min_{M' \in \mathcal{M}, \pi \in \Pi: M(\pi) \neq M'(\pi)} D_{\text{KL}}(M(\pi), M'(\pi)).$$

Our characterization needs an assumption on the ratios of probability masses: ¹¹

Assumption 1. The ratio $\Pr_{M(\pi)}(r, o) / \Pr_{M'(\pi)}(r, o)$ is upper-bounded by $B < \infty$, for any models $M, M' \in \mathcal{M}$, any policy $\pi \in \Pi$, and any outcome $(r, o) \in \mathcal{R} \times \mathcal{O}$.

Theorem 3. Fix an instance $(M^*, \mathcal{M})$ of DMSO with Assumption 1 and model-gap $\Gamma = \Gamma(M^*, \mathcal{M})$.

- (a) If the problem instance is self-identifiable, then GreedyMLE (with any warm-up data) satisfies $\mathbb{E}[R(t)] \leq T_0 + (|\Pi| \ln(B)/\Gamma)^2 \cdot O(\ln(|\mathcal{M}| \cdot t))$ for each round $t \in \mathbb{N}$.
- (b) Suppose the warm-up data consists of $N_0 := c_0 \cdot (\ln(B)/\Gamma)^2 \log |\mathcal{M}|$ samples for each policy, for an appropriately chosen absolute constant c_0 (for the total of $T_0 := N_0 |\Pi|$ samples). Assume M^* has a decoy $M_{\text{dec}} \in \mathcal{F}$, with decoy policy π_{dec} . Then with some probability $p_{\text{dec}} \geq B^{-O(N_0 |\Pi|)} > 0$, GreedyMLE chooses π_{dec} in all rounds $t \in (T_0, \infty)$.

GreedyMLE succeeds (i.e., achieves sublinear regret for any warm-up data) if and only if the problem instance is self-identifiable. Else, it fails for some warm-up data, with linear regret like in Eq. (4.2). We also provide a more efficient lower bound on p_{dec} in Theorem 3(b), replacing B with a term that only concerns two relevant models, M_{dec}, M^* (not all of $\mathcal{M}$). Letting D_∞ be the Renyi divergence,

$$p_{\text{dec}} \geq e^{-O(C_{\text{dec}} N_0 |\Pi|)}, \text{ where } C_{\text{dec}} = \max_{\pi \in \Pi} D_\infty(M_{\text{dec}}(\pi) \| M^*(\pi)) \leq \log B. \quad (5.3)$$

Proof Sketch for Theorem 3. We consider the likelihood of a particular model $M \in \mathcal{M}$ given the history at round $t \geq 2$, $\mathcal{L}_t(M) := \mathcal{L}(M | \mathcal{H}_t)$. We track the per-round change in log-likelihood:

$$\Delta \ell_t(M) := \log \mathcal{L}_{t+1}(M) - \log \mathcal{L}_t(M) = \log(\Pr_{M(\pi_t)}(r_t, o_t)). \quad (5.4)$$

Let $\mathcal{L}_1(\cdot) = 1$, so that (5.4) is also well-defined for round $t = 1$.

We argue that the likelihood of M^* grows *faster* than that of any other model $M \in \mathcal{M}$. Specifically, we focus on $\Phi_t(M) := \mathbb{E}[\Delta \ell_t(M^*) - \Delta \ell_t(M)]$. We claim that

$$(\forall t \in \mathbb{N}) \quad \text{If } M^*(\pi_t) \stackrel{d}{=} M(\pi_t) \text{ then } \Phi_t(M) = 0 \text{ else } \Phi_t(M) \geq \Gamma. \quad (5.5)$$

In more detail: if the two models completely coincide on policy π_t , then $\Delta \ell_t(M^*) = \Delta \ell_t(M)$, and otherwise we invoke the definition of the model-gap. We use (5.5) for both parts of the theorem. The proof of Eq. (5.5) is where we directly analyze regression and invoke the model-gap.

Part (a). Suppose GreedyMLE chooses some suboptimal policy π_t in some round $t > T_0$ of the main stage. By Eq. (5.5) and self-identifiability, it follows that $\Phi_t(M_t) \geq \Gamma$. (Indeed, by (5.5) the only alternative is $M^*(\pi_t) \stackrel{d}{=} M_t(\pi_t)$, and then self-identifiability implies that policy π_t is suboptimal for model M_t , contradiction.) Likewise, we obtain that $\Phi_t(M) \geq \Gamma$ for any model $M \in \mathcal{M}$ for which policy π_t is optimal; let $\mathcal{M}_{\text{opt}}(\pi)$ be the set of all models for which policy π is optimal.

We argue that suboptimal policies $\pi \in \Pi$ cannot be chosen “too often”. Indeed, fix one such policy π . Then with high probability (w.h.p.) the likelihood of any model $M \in \mathcal{M}_{\text{opt}}(\pi)$ falls below that of M^* , so this model cannot be chosen again. So, w.h.p. this *policy* cannot be chosen again. ¹²

11. Related (but incomparable) assumptions on mass/density ratios are common in the literature on online/offline RL, [e.g., Munos and Szepesvári, 2008, Xie and Jiang, 2021, Zhan et al., 2022, Amortila et al., 2024].

12. This last step takes a union bound over the models $M \in \mathcal{M}_{\text{opt}}(\pi)$, hence $\log |\mathcal{M}|$ in the regret bound.

Part (b). We define independent events E_1 and E_2 , resp., on the warm-up process and on all rounds when the decoy is chosen, so that $E_1 \cap E_2$ guarantees that GreedyMLE gets forever stuck on the decoy. While this high-level plan is the same as before, its implementation is far more challenging.

To side-step some technicalities, we separate out $N_0/2$ warm-up rounds in which the decoy policy π_{dec} is chosen. Specifically, w.l.o.g. we posit that π_{dec} is chosen in the last $N_0/2$ warm-up rounds, and let $\mathcal{H}_{\text{warm}} = \mathcal{H}_{T'_0+1}$, $T'_0 := T_0 - N_0/2$ be the history of the rest of the warm-up.

First, we consider the “ghost process” (ghost) for generating $\mathcal{H}_{\text{warm}}$: in each round $t \leq T'_0$, the chosen policy π_t stays the same, but the outcome (r_t, o_t) is generated according to the decoy model M_{dec} . Under ghost, each round raises the likelihood $\mathcal{L}_t(M_{\text{dec}})$ *more* compared to any other model $M \in \mathcal{M}$. Namely, write $\Delta \ell_t(M) = \Delta \ell_t(M \mid \mathcal{H}_t)$ explicitly as a function of history $\mathcal{H}_t$, and let

$$\Phi_t(M, M_{\text{dec}}) := \mathbb{E} [\Delta \ell_t(M_{\text{dec}} \mid \mathcal{H}_t) - \Delta \ell_t(M \mid \mathcal{H}_t)], \quad (5.6)$$

where $\mathcal{H}_t$ comes from ghost. Reusing Eq. (5.5) (with M_{dec} now replacing true model M^*), yields:

$$\text{If } M_{\text{dec}}(\pi_t) \stackrel{d}{=} M(\pi_t) \text{ then } \Phi_t(M, M_{\text{dec}}) = 0 \text{ else } \Phi_t(M, M_{\text{dec}}) \geq \Gamma. \quad (5.7)$$

For each model $M \in \mathcal{M}$ different from M_{dec} , there is a policy $\pi \in \Pi$ on which these two models differ. This policy appears N_0 times in the warm-up data, so by Eq. (5.7) we have $\sum_{t \in [T'_0]} \Phi_t(M, M_{\text{dec}}) \geq \Gamma \cdot N_0$. Consequently, letting $\mathcal{M}_{\text{other}} := \mathcal{M} \setminus \{M_{\text{dec}}\}$, event

$$E_1 = \{ \forall M \in \mathcal{M}_{\text{other}} \quad \mathcal{L}(M_{\text{dec}} \mid \mathcal{H}_{\text{warm}}) > \mathcal{L}(M \mid \mathcal{H}_{\text{warm}}) \}$$

happens w.h.p. when $\mathcal{H}_{\text{warm}}$ comes from ghost.¹³ Since ghost and $\mathcal{H}_{\text{warm}}$ have bounded Renyi divergence, we argue that with some positive probability, event E_1 happens under $\mathcal{H}_{\text{warm}}$.

Let’s analyze the rounds in which the decoy policy π_{dec} is chosen. Let $t(j)$ be the j -th such round, $j \in \mathbb{N}$. We’d like to argue that throughout all these rounds, the likelihood of the decoy model M_{dec} grows faster than that of any other model $M \in \mathcal{M}$. To this end, consider event

$$E_2 := \{ \forall j > N_0/2, \forall M \in \mathcal{M}_{\text{other}}, \quad \sum_{i \in [j]} \Psi_i(M) \geq 0 \},$$

where $\Psi_j(M) := \Delta \ell_{t(j)}(M_{\text{dec}}) - \Delta \ell_{t(j)}(M)$. Here, we restrict to $j > N_0/2$ to ensure that E_1, E_2 concern disjoint sets of events, and hence are independent. $E_1 \cap E_2$ implies that in each round $t > T_0$, $\mathcal{L}_t(M_{\text{dec}}) > \mathcal{L}_t(M)$ for any model $M \in \mathcal{M}_{\text{other}}$, and so GreedyMLE chooses the decoy policy.

Finally, we argue that E_2 happens with positive probability. W.l.o.g., the outcomes (r_t, o_t) in all rounds $t = t(j)$, $j \in \mathbb{N}$ are drawn in advance from an “outcome tape”.¹⁴ We leverage Eq. (5.5) once again. Indeed, $\Psi_j(M) = 0$ for every model $M \in \mathcal{M}$ that fully coincides with M_{dec} on the decoy policy π_{dec} , so we only need to worry about the models $M \in \mathcal{M}$ for which this is not the case. Then $\sum_{i \in [j]} \mathbb{E} [\Psi_i(M)] \geq j \cdot \Gamma$. We obtain $\sum_{i \in [j]} \mathbb{E} [\Psi_i(M)] \geq 0$ with positive-constant probability by appropriately applying concentration separately for each $j > N_0/2$ and taking a union bound.

6 Structured bandits with an Infinite Function Class

We obtain a partial characterization for StructuredMAB, which handles an arbitrary *infinite* function class $\mathcal{F}$ and yields better constants compared to Theorem 1. The success of Greedy requires a stronger notion of self-identifiability: *approximately* fixing the expected reward of a suboptimal arm identifies it as suboptimal. The failure of Greedy requires a stronger notion of a decoy function, which must lie in the “interior” of $\mathcal{F}$. The characterization is “partial” in the sense that it does not yield a full dichotomy. However, the boundary between success and failure is controlled by a tunable “margin” parameter $\varepsilon > 0$, which can be made arbitrarily small (and optimized based on the instance).

Definition 7. A problem instance $(f^*, \mathcal{F})$ is ε -self-identifiable, $\varepsilon \geq 0$, if any suboptimal arm a of f^* is suboptimal for any reward function $f \in \mathcal{F}$ with $|f(a) - f^*(a)| \leq \varepsilon$. An ε -interior of $\mathcal{F}$, $\text{int}(\mathcal{F}, \varepsilon)$ is the set of all functions $f \in \mathcal{F}$, such that any reward function f' with $\|f' - f\|_2 \leq \varepsilon$ is also in $\mathcal{F}$.

13. This argument invokes a concentration inequality, which in turn uses Assumption 1. Likewise, Assumption 1 is used for another application of concentration in the end of the proof sketch.

14. Its entries $j \in \mathbb{N}$ are drawn independently from $M^*(\pi_{\text{dec}})$, and $(r_{t(j)}, o_{t(j)})$ is defined as the j -entry.

For a “continuous” function class such as linear functions or Lipschitz functions, $\text{int}(\mathcal{F}, \varepsilon)$ typically includes all but an $O(\varepsilon)$ -fraction of $\mathcal{F}$. The choice of the ℓ_2 norm in the definition of ε -interior is not essential: any ℓ_p norm suffices. We provide the main theorem below; see proof in Appendix H.

Theorem 4. Fix a problem instance $(f^*, \mathcal{F})$ of StructuredMAB with an infinite function class $\mathcal{F}$ (but a finite action set $\mathcal{A}$). For any $\varepsilon > 0$ (which can be optimized based on f^*):

- (a) If the problem instance is ε -self-identifiable, then Greedy (with any warm-up data) satisfies $\mathbb{E}[R(t)] \leq T_0 + (K/\varepsilon)^2 \cdot O(\log t)$ for each round $t \in \mathbb{N}$.
- (b) Suppose the warm-up data consists of one sample for each arm. Assume f^* has a decoy $f_{\text{dec}} \in \text{int}(\mathcal{F}, \varepsilon)$, with decoy arm a_{dec} . Then with some probability $p_{\text{dec}} > 0$ it holds that Greedy chooses a_{dec} for all rounds $t \in (T_0, \infty)$. We can lower-bound p_{dec} by $e^{-O(K^2/\varepsilon^2)}$.

This result mirrors Theorem 1, with the function gap replaced by ε , allowing for instance-dependent optimization of ε and tighter bounds. The proof for part (a) carries over with simple modifications. In contrast, proving part (b) is considerably more subtle. In the infinite case, Greedy may not get stuck on a single reward function—it could almost surely switch among infinitely many. The key insight is that such fluctuations need not impact the arm choice: even as the predictor f_t changes, the greedy selection a_t may remain fixed. The proof exploits this decoupling, constructing events where the algorithm persistently selects a decoy arm, even as the greedy predictors continue to evolve.

Discussion: challenges. An “if and only if” characterization for arbitrary infinite function classes is very difficult. First, one can no longer rely on the function-gap being strictly positive, which is a cornerstone of our analysis in the finite case. Second, Greedy’s behavior can be highly unstable: the algorithm’s predictor f_t can fluctuate indefinitely within a continuous region of functions that are all similarly consistent with data yet induce very different greedy action choices. As a result, the intuitive logic of “getting stuck in a decoy and staying there forever” does not directly extend.

The partial characterization in Theorem 4 is our proposed route to address these challenges. The margin ε serves a dual purpose: it stands in for the now-absent function-gap, and it allows us to deal with the predictor’s instability (by showing that a_t can remain permanently fixed).

However, this is still insufficient for a complete characterization. For many natural function classes, our framework leaves a set of instances, typically of fraction $O(\varepsilon)$, uncharacterized. The boundary between success and failure instances in a general infinite space can be highly complex; success instances can be very close to failure instances, making a sharp separation difficult. Our ε -interior notion is designed to provide a robust buffer around this boundary, at the cost of leaving instances within that buffer “undecided.” A tight characterization in full generality would likely require a more fine-grained analysis exploiting additional structural properties of the function class $\mathcal{F}$.

7 Conclusions

We study Greedy in structured bandits and characterize its asymptotic success vs failure in terms of a simple partial-identifiability property of the problem structure. Our characterization holds for arbitrary finite structures and extends to bandits with contexts and/or auxiliary feedback. In particular, we find that Greedy succeeds only if the problem is intrinsically “easy” for any algorithm which satisfies a mild non-degeneracy condition. We also provide a partial characterization for StructuredMAB with infinite reward structures (and finite action sets).

Several examples, both positive and negative, instantiate our characterization for various well-studied reward structures. We find that failure tends to be a common case for *bandits*, whereas both failure and success are common for structured *contextual* bandits.

We identify three directions for further work. First, extend our characterization to infinite action sets and infinite function/model classes (ideally with a complete characterization, as discussed in Section 6). Second, consider *approximate* greedy algorithms, stemming either from approximate regression or from human behavioral biases. Such algorithms, representing myopic human behavior under behavioral biases, were studied in Banihashem et al. [2023], but only for *unstructured* multi-armed bandits. Third, while our “asymptotic” perspective enables a general characterization, *stronger regret guarantees* are desirable for particular reward structures (and are known for only a few).

References

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *25th Advances in Neural Information Processing Systems (NIPS)*, pages 2312–2320, 2011.
- Naoki Abe, Alan W. Biermann, and Philip M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.
- Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Learning From Reviews: The Selection Effect and the Speed of Learning. *Econometrica*, 2022. Working paper available since 2017.
- Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *15th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pages 19–26, 2012.
- Alekh Agarwal, Sarah Bird, Markus Cozowicz, Miro Dudik, Luong Hoang, John Langford, Li-hong Li, Dan Melamed, Gal Oshri, Siddhartha Sen, and Aleksandrs Slivkins. Multiworld testing: A system for experimentation, learning, and decision-making, 2016. A white paper, available at <https://github.com/Microsoft/mwt-ds/raw/master/images/MWT-WhitePaper.pdf>.
- Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt, 2017. Technical report at arxiv.org/abs/1606.03966.
- Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms, 2020. Book draft, circulated since 2019. Available at <https://rltheorybook.github.io>.
- Rajeev Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33(6):1926–1951, 1995.
- Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *27th Advances in Neural Information Processing Systems (NIPS)*, pages 1610–1618, 2013.
- Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *28th Conf. on Learning Theory (COLT)*, pages 23–35, 2015.
- Kareem Amin, Michael Kearns, and Umar Syed. Bandits, query learning, and the haystack dimension. In *24th Conf. on Learning Theory (COLT)*, 2011.
- Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024.
- András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. *Theor. Comput. Sci.*, 473:77–99, 2013.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. of Machine Learning Research (JMLR)*, 3:397–422, 2002. Preliminary version in *41st IEEE FOCS*, 2000.
- Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved Rates for the Stochastic Continuum-Armed Bandit Problem. In *20th Conf. on Learning Theory (COLT)*, pages 454–468, 2007.
- Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *J. of Computer and System Sciences*, 74(1):97–114, February 2008. Preliminary version in *36th ACM STOC*, 2004.
- Kiarash Banihashem, MohammadTaghi Hajiaghayi, Suho Shin, and Aleksandrs Slivkins. Bandit social learning: Exploration under myopic behavior, 2023. Working paper at <https://arxiv.org/abs/2302.07425>. Preliminary version in *NeurIPS 2023*, titled “Bandit Social Learning under Myopic Behavior”.

- Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.
- Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021. Working paper available on arxiv.org since 2017.
- Mohsen Bayati, Nima Hamidi, Ramesh Johari, and Khashayar Khosravi. Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. In *33rd Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *J. of Machine Learning Research (JMLR)*, 22:133:1–133:49, 2021.
- Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1): 1–122, 2012. Published with Now Publishers (Boston, MA, USA). Also available at <https://arxiv.org/abs/1204.5721>.
- Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. Online Optimization in X-Armed Bandits. *J. of Machine Learning Research (JMLR)*, 12:1587–1627, 2011. Preliminary version in *NIPS 2008*.
- Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *49th ACM Symp. on Theory of Computing (STOC)*, pages 72–85. ACM, 2017.
- Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *J. Comput. Syst. Sci.*, 78(5): 1404–1422, 2012. Preliminary version in *COLT 2009*.
- Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *20th Intl. Conf. on Machine Learning (ICML)*, pages 151–159, 2013.
- Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual Bandits with Linear Payoff Functions. In *14th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2011.
- Richard Combes, Stefan Magureanu, and Alexandre Proutière. Minimal exploration in structured stochastic bandits. In *30th Advances in Neural Information Processing Systems (NIPS)*, pages 1763–1771, 2017.
- Varsha Dani, Thomas P. Hayes, and Sham Kakade. Stochastic Linear Optimization under Bandit Feedback. In *21th Conf. on Learning Theory (COLT)*, pages 355–366, 2008.
- Rémy Degenne, Han Shao, and Wouter M. Koolen. Structure adaptive algorithms for stochastic bandits. In *37th Intl. Conf. on Machine Learning (ICML)*, volume 119, pages 2443–2452, 2020.
- Arnoud V. den Boer. Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of Operations Research*, 39(3):863–888, 2014.
- Arnoud V. den Boer and Bert Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, 60(3):770–783, 2014.
- Abraham Flaxman, Adam Kalai, and H. Brendan McMahan. Online Convex Optimization in the Bandit Setting: Gradient Descent without a Gradient. In *16th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 385–394, 2005.
- Dylan J. Foster and Alexander Rakhlin. Beyond UCB: optimal and efficient contextual bandits with regression oracles. In *37th Intl. Conf. on Machine Learning (ICML)*, 2020.
- Dylan J. Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making, 2023. URL <https://arxiv.org/abs/2312.16730>.
- Dylan J. Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles. In *35th Intl. Conf. on Machine Learning (ICML)*, pages 1534–1543, 2018.

- Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.
- Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.
- András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *J. of Machine Learning Research (JMLR)*, 8:2369–2403, 2007.
- J. Michael Harrison, N. Bora Keskin, and Assaf Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586, 2012.
- Baihe Huang, Kaixuan Huang, Sham M. Kakade, Jason D. Lee, Qi Lei, Runzhe Wang, and Jiaqi Yang. Optimal gradient-based algorithms for non-concave bandit optimization. In *34th Advances in Neural Information Processing Systems (NeurIPS)*, pages 29101–29115, 2021.
- Matthieu Jedor, Jonathan Louëdec, and Vianney Perchet. Be greedy in multi-armed bandits, 2021. Working paper, available on <https://arxiv.org/abs/2101.01086>.
- Kwang-Sung Jun and Chicheng Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. In *33rd Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. *Advances in neural information processing systems*, 31, 2018.
- Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2231–2241, 2018.
- N. Bora Keskin and Assaf Zeevi. On incomplete learning and certainty-equivalence control. *Oper. Res.*, 66(4):1136–1167, 2018.
- N. Bora Keskin and Assaf J. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014.
- Seok-Jin Kim and Min-hwan Oh. Local anti-concentration class: Logarithmic regret for greedy linear contextual bandit. *Advances in Neural Information Processing Systems*, 37:77525–77592, 2024.
- Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems (NIPS)*, 2004.
- Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *40th ACM Symp. on Theory of Computing (STOC)*, pages 681–690, 2008.
- Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the “wisdom of the crowd”. *J. of Political Economy*, 122(5):988–1012, 2014. Preliminary version in *ACM EC 2013*.
- T.L. Lai and Herbert Robbins. Iterated least squares in multiperiod control. *Advances in Applied Mathematics*, 3(1):50–73, 1982.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.
- Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *19th Intl. World Wide Web Conf. (WWW)*, 2010.
- Tyler Lu, Dávid Pál, and Martin Pál. Showing Relevant Ads via Lipschitz Context Multi-Armed Bandits. In *14th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2010.
- H. Brendan McMahan and Avrim Blum. Online Geometric Optimization in the Bandit Setting Against an Adaptive Adversary. In *17th Conf. on Learning Theory (COLT)*, pages 109–123, 2004.
- Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

- Bart P. G. Van Parys and Negin Golrezaei. Optimal learning for structured bandits. *Management Science*, 70(6):3951–3998, 2024.
- Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. Greedy algorithm almost dominates in smoothed contextual bandits. *SIAM J. on Computing (SICOMP)*, 52(2):487–524, 2023. Preliminary version at *COLT 2018*.
- Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *26th Conf. on Learning Theory (COLT)*, volume 30, pages 3–24, 2013.
- David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operation Research*, 47(3):1904–1931, 2022.
- Aleksandrs Slivkins. Contextual bandits with similarity information. *J. of Machine Learning Research (JMLR)*, 15(1):2533–2568, 2014. Preliminary version in *COLT 2011*.
- Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019. Published with Now Publishers (Boston, MA, USA). Also available at <https://arxiv.org/abs/1904.07272>.
- Aleksandrs Slivkins. Exploration and persuasion. In Federico Echenique, Nicole Immorlica, and Vijay Vazirani, editors, *Online and Matching-Based Market Design*. Cambridge University Press, 2023. Also available at <https://arxiv.org/abs/2410.17086>.
- Andrew J Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR, 2023.
- Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.
- Qian Yu, Yining Wang, Baihe Huang, Qi Lei, and Jason D. Lee. Sample complexity for quadratic bandits: Hessian dependent bounds and optimal algorithms. In *36th Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.
- Geng Zhao, Banghua Zhu, Jiantao Jiao, and Michael Jordan. Online learning in stackelberg games with an omniscient follower. In *40th Intl. Conf. on Machine Learning (ICML)*, pages 42304–42316, 2023.

NeurIPS Paper Checklist

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

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

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

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

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

IMPORTANT, please:

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

1. Claims

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

Answer: [Yes]

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

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

2. Limitations

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

Answer: [Yes]

Justification: Our main result assumes a finite function class and comes with possibly large constants. We discuss these limitations in detail, and also partially address them in our second main result in Section 6.

Throughout, we assume *realizability*—the standard assumption in structured bandit problems that the true model lies within the model class. While this assumption may not hold exactly in practice, it is pervasive in the bandits literature and serves as a foundational premise for theoretical analysis. Accordingly, our results should be interpreted in the context of this widely adopted framework.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [Yes]

Justification: The paper provide full set of assumptions and complete proofs.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [NA]

Justification: The paper does not contain experimental results.

Guidelines:

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

5. Open access to data and code

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

Answer: [NA]

Justification: The paper does not contain experimental results, hence the question does not apply.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.

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

6. Experimental setting/details

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

Answer: [NA]

Justification: The paper does not contain experimental results.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [NA]

Justification: This question does not pertain since the paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [NA]

Justification: The paper does not contain experimental results.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: We confirm the research conform with the Code of Ethics.

Guidelines:

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

10. Broader impacts

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

Answer: [NA]

Justification: This is a theoretical paper on multi-armed bandits, focusing on foundational aspects of the field. We do not anticipate any direct societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [NA]

Justification: This question does not apply to the current submission.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [NA]

Justification: This question does not apply to the current submission.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

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

Answer: [NA]

Justification: This question does not apply to the current submission.

Guidelines:

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

14. Crowdsourcing and research with human subjects

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

Answer: [NA]

Justification: This question does not apply to the current submission.

Guidelines:

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

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

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

Answer: [NA]

Justification: This question does not apply to the current submission.

Guidelines:

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

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

TECHNICAL APPENDICES

Contents

A	StructuredCB with tie-breaking	23
B	Novelty of self-identifiability	24
C	Self-identifiability makes the problem easy	24
D	Examples	25
D.1	(Discretized) linear bandits	26
D.2	(Discretized) linear contextual bandits	26
D.3	(Discretized) Lipschitz Bandits	27
D.4	(Discretized) polynomial bandits	28
D.5	(Discretized) quadratic bandits	30
E	StructuredMAB characterization: Proof of Theorem 1	30
E.1	StructuredMAB Success: Proof of Theorem 1(a)	30
E.2	StructuredMAB Failure: Proof of Theorem 1(b)	31
F	StructuredCB characterization: Proof of Theorem 2	32
F.1	StructuredCB Success: Proof of Theorem 2(a)	33
F.2	StructuredCB Failure: Proof of Theorem 2(b)	35
G	DMSO characterization: Proof of Theorem 3	35
G.1	DMSO Success: Proof of Theorem 3(a)	36
G.2	DMSO Failure: Proof of Theorem 3(b)	36
H	StructuredMAB with an Infinite Function Class: Proof of Theorem 4	38
H.1	Success: Proof of Theorem 4(a)	38
H.2	Failure: Proof of Theorem 4(b)	38

A StructuredCB with tie-breaking

Let us outline how to adjust our definitions and results to account for ties in Eq. (2.2). We assume that the ties are broken at random, with some minimal probability $q_0 > 0$ on every optimal arm (i.e., every arm in $\operatorname{argmax}_{a \in \mathcal{A}} f_t(x_t, a)$). More formally, Greedy breaks ties in Eq. (2.2) according to an independent draw from some distribution D_t over the optimal arms with minimal probability at least q_0 . Subject to this assumption, the tie-breaking distributions D_t can be arbitrary, both within a given round and from one round to another.

The positive results (Definition 3 and Theorem 2(a)) carry over word-by-word, both the statements and the proofs. The negative results (Definition 4 and Theorem 2(b)) change slightly. Essentially, whenever we invoke *the* optimal arm for decoy f_{dec} , we need to change this to *all* optimal arms for f_{dec} .

Definition 8 (decoy). *Let f^* be a reward functions, with optimal policy π^* . Another reward function f_{dec} is called a decoy for f^* if any optimal policy π_{dec} for f_{dec} satisfies $f_{\text{dec}}(\pi_{\text{dec}}) = f^*(\pi_{\text{dec}}) < f^*(\pi^*)$ and moreover $f_{\text{dec}}(x, \pi_{\text{dec}}(x)) = f^*(x, \pi_{\text{dec}}(x))$ for all contexts $x \in \mathcal{X}$.*

The equivalence of self-identifiability and not having a decoy holds as before, i.e., the statement of Claim 1 carries over word-by-word. Moreover, it is still the case that “self-identifiability makes the problem easy”: all of Appendix C carries over as written.

Theorem 5 (negative). *Fix a problem instance $(f^*, \mathcal{F})$ of StructuredCB. Suppose the warm-up data consists of one sample for each context-arm pair. Assume f^* has a decoy $f_{\text{dec}} \in \mathcal{F}$. Let Π_{dec} be the set of all policies that are optimal for f_{dec} . Then with some probability $p_{\text{dec}} > 0$, Greedy only chooses policies $\pi_t \in \Pi_{\text{dec}}$ in all rounds $t \in (T_0, \infty)$. We have $p_{\text{dec}} \geq X^{-O(KX/\Gamma^2(f_{\text{dec}}))}$, where $X = |\mathcal{X}|$.*

Under these modifications, Remark 1 applies word-by-word. In particular, existence of a decoy implies linear regret, where each round t with $\pi_t \in \Pi_{\text{dec}}$ increases regret by $f^*(\pi^*) - f^*(\pi_{\text{dec}})$.

Proof of Theorem 5. The proof of Theorem 2(b) mostly carries over, with the following minor modifications. Let $\mathcal{A}_{\text{dec}}^*(x) = \operatorname{argmax}_{a \in \mathcal{A}} f_{\text{dec}}(x, a)$ be the set of optimal arms for the decoy f_{dec} for a given context x . The two events E_1 and E_2 (as originally defined eq. (F.1) and eq. (F.2)) will be modified to be invoked on all decoy context-arm pairs.

$$\begin{aligned} E_1 &= \{ |\bar{r}_{\text{warm}}(x, a) - f_{\text{dec}}(x, a)| < \Gamma(f_{\text{dec}})/2 \text{ for each } x \in \mathcal{X} \text{ and arm } a \notin \mathcal{A}_{\text{dec}}^*(x) \}, \\ E_2 &= \{ |\bar{r}_t(x, \pi_{\text{dec}}(x)) - f^*(x, \pi_{\text{dec}}(x))| < \Gamma(f_{\text{dec}})/2 \text{ for each } x \in \mathcal{X}, a \in \mathcal{A}_{\text{dec}}^*(x), \text{ and round } t > T_0 \}. \end{aligned}$$

Analyzing the probability for event E_1 still follows from Lemma 20. Analyzing the probability for event E_2 follows from Lemma 21, but with the choice of σ will be chosen as $\sigma = \Theta(\Gamma(f_{\text{dec}})/\sqrt{\ln(|\mathcal{X}|K)})$, and we still have $\Pr[E_2] \geq 0.9$.

B Novelty of self-identifiability

We argue that self-identifiability is a novel notion. Specifically, we compare it to (i) knowing the optimal value, and (ii) Graves-Lai coefficient being 0.

First, one could ask if self-identifiability is equivalent to knowing the value of the best arm. However, the former does not imply the latter. Consider the simple example $\mathcal{F} = \{(3, 1), (2, 1)\}$. Both functions are self-identifiable in $\mathcal{F}$, but clearly the optimal value differs.

Second, consider the Graves-Lai coefficient [Graves and Lai, 1997, Wagenmaker and Foster, 2023]. Let us define it formally, for the sake of completeness. Consider DMSO, as defined in Section 5, with model class $\mathcal{M}$. Let

$$\Delta(\pi|M) = f(\pi_M|M) - f(\pi|M)$$

be the suboptimality gap for model M and policy π , where π_M is the optimal policy for M . Let $\mathcal{M}^{\text{alt}}$ be the set of models that disagree with M on the optimal policy:

$$\mathcal{M}^{\text{alt}}(M) := \{M' \in \mathcal{M} | \pi_M \neq \pi_{M'}\}.$$

Now, the Graves-Lai coefficient is defined as

$$\text{GLC}(\mathcal{M}, M) = \inf_{\eta \in \mathbb{R}_+^{\Pi}} \left\{ \sum_{\pi \in \Pi} \eta_{\pi} \Delta(\pi|M) \mid \forall M' \in \mathcal{M}^{\text{alt}}(M) : \sum_{\pi \in \Pi} \eta_{\pi} D_{\text{KL}}(M(\pi) \| M'(\pi)) \geq 1 \right\}.$$

Intuitively, the Graves-Lai coefficient measures the “verification” cost of verifying whether a given function f^* (or a given model M^* in the DMSO setting) is indeed the true model. The Graves-Lai coefficient being 0 implies that the learner can ascertain that f^* or M^* is indeed the true model by simply executing the set of optimal policies $\Pi(f^*)$ or $\Pi(M^*)$.

Now, one could ask if self-identifiability is equivalent to $\text{GLC}(\mathcal{M}, M) = 0$. We observe that this is not the case: the two notions are incomparable. For a counterexample, consider StructuredMAB with two arms and $\mathcal{F} = \{(2, 1), (0.5, 1)\}$. Problem instance $f^* = (0.5, 1)$ is self-identifiable, since revealing the sub-optimal arm as having reward 0.5 immediately rules out $(2, 1)$ as being the true model. But the $\text{GLC} > 0$, since to ascertain $(0.5, 1)$ as being the true model one necessarily has to choose the 1st arm and experiment. On the other hand, one can see $f^* = (2, 1)$ is not self-identifiable but has $\text{GLC} = 0$. In this example, Greedy succeeds when $\text{GLC} > 0$ (larger GLC suggests larger regret of the optimal algorithm in GLC-based theory) but fails when $\text{GLC} = 0$ (lower GLC suggests lower regret of the optimal algorithm in GLC-based theory)! Hence GLC *does not* capture the per-instance behavior of Greedy.

However, GLC has *some* connection to our machinery. Namely, if $\text{GLC}(\mathcal{F}, f) = 0$ for some reward function f , then f necessarily cannot be a decoy for any other reward function f^* . That said, $\text{GLC}(\mathcal{F}, f)$ provides no information about whether f itself admits a decoy. We believe that GLC precisely characterizes the asymptotic performance of the *optimal* algorithm [Graves and Lai, 1997, Wagenmaker and Foster, 2023], whereas self-identifiability precisely captures the asymptotic behavior of Greedy—a generally suboptimal algorithm.

C Self-identifiability makes the problem easy

Our characterization raises a natural question: does the success of Greedy under self-identifiability stem from the algorithm itself, from self-identifiability, or both? Put differently, when Greedy succeeds, does it make any non-trivial effort toward its success?

Surprisingly, our characterization provides a definitive negative answer: Greedy succeeds because self-identifiability makes the problem intrinsically “easy.” We prove that whenever self-identifiability holds, *any* reasonable algorithm (satisfying a mild non-degeneracy condition defined below) also achieves sublinear regret. This, in a sense, reveals the “triviality” of the greedy algorithm: it succeeds only when the problem is so easy that any reasonable algorithm would succeed.

To formalize this, we must clarify what we mean by “reasonable algorithms.” Clearly, we need to exclude certain degenerate cases, such as static algorithms that pick a single arm forever, neither exploring nor exploiting information. We argue that a reasonable algorithm should at least *care about* information—whether through exploration, exploitation, or both. In other words, a reasonable

algorithm should never select an action that serves *neither* any exploration purpose (i.e., bringing new information) *nor* any exploitation purpose (i.e., utilizing existing information). This principle naturally leads to *information-aware* algorithms formally defined below.

We work in the setting of StructuredCB, and explain how to specialize it to StructuredMAB.

Definition 9. Consider some round t in StructuredCB. We say policy π is δ -identified-and-suboptimal if there exists a suitable concentration event which happens with probability $1 - \delta$, such that under the concentration event, its mean rewards $f^*(x, \pi(x))$ for each context x are exactly identified given the current history, and moreover this identification reveals that the policy is suboptimal given the function class.

For StructuredMAB, this definition specializes to defining δ -identified-and-suboptimal arms.

Definition 10. An algorithm for StructuredCB (resp., StructuredMAB) is called δ -information-aware if at each round, it does not choose any policy (resp., arm) that is δ -identified-and-suboptimal.

Let us define the concentration events: $\mathcal{E}_{\text{MAB}}$ for StructuredMAB and $\mathcal{E}_{\text{CB}}$ for StructuredCB:

$$\mathcal{E}_{\text{MAB}} := \{ |\bar{r}_t(a) - f^*(a)| > \beta_t(N_t(a)) \quad \forall a \in \mathcal{A}, t \in \mathbb{N} \}, \quad (\text{C.1})$$

$$\begin{aligned} \mathcal{E}_{\text{CB}} := \{ |\bar{r}_t(x, a) - f^*(x, a)| < \beta_t(N_t(x, a)) \text{ and } N_t(x, a) > \Omega(N_t(\pi) p_0) \\ \text{with } a = \pi(x) \quad \forall x \in \mathcal{X}, \pi \in \Pi, t \in \mathbb{N} \}, \end{aligned} \quad (\text{C.2})$$

where $\beta_t(n) = \sqrt{\frac{2}{n} \log \left(\frac{10K|\mathcal{X}|t \cdot n^2}{3\delta} \right)}$ and $N_t(x)$ is the number of times context x has been observed before round t . Here, p_0 is the smallest context arrival probability, like in Section 4. Note that $\mathcal{E}_{\text{MAB}}$ is just a specialization of $\mathcal{E}_{\text{CB}}$.

Theorem 6. Consider StructuredCB with time horizon T . Any $1/T$ -information-aware algorithm ALG achieves a sublinear regret $\mathbb{E}[R(T)]$ under self-identifiability.

Proof. Assume $\mathcal{E}_{\text{CB}}$ holds. Fix any suboptimal policy π . We show π can only be chosen $o(T)$ times.

By the definition of $\beta_t(\cdot)$ in the event $\mathcal{E}_{\text{CB}}$, there must exist some parameter $T' = \tilde{\Theta}(1/\Gamma^2(f^*)) (= o(T))$, such that $\beta_t(T') < \Gamma(f^*)$. Then, if the suboptimal policy π is executed above the threshold $\Omega(T'/p_0)$, we have $N_t(x, a) > T'$, and consequently for any context x ,

$$|\bar{r}_t(x, \pi(x)) - f^*(x, \pi(x))| < \beta(T') < \Gamma(f^*).$$

Then recall for any function f and context-arm pair (x, a) , we have either $f(x, a) = f^*(x, a)$ or $|f(x, a) - f^*(x, a)| \geq \Gamma(f^*)$. This precisely means the policy π becomes identified, and by self-identifiability, any information-aware algorithm will not keep choosing π . Hence, the total regret of the information-aware algorithm is at most $O(T'|\Pi|)$, which is sublinear $o(T)$. $\square$

D Examples

Let us instantiate our characterization for several well-studied reward structures from bandits literature. We consider linear, Lipschitz, and (one-dimensional) polynomial structures, for bandits as well as contextual bandits. All reward structures in this section are discretized to ensure finiteness, as required for our complete characterization in Sections 3 to 5. (While our partial characterization in Section 6 handles infinite reward structures, a secondary goal of this section is to illustrate how common infinite reward structures can be meaningfully discretized so that the complete finite-structure results become directly applicable.) The discretization is consistent across different reward functions, in the sense that all functions take values in the same (discrete) set $\mathcal{R}$, with $|\mathcal{R}| \ll |\mathcal{F}|$. This prevents a trivial form of self-identifiability that could arise if each reward function f were discretized independently and inconsistently, resulting in some $f(a)$ values being unique and making f self-identifiable solely due to the discretization strategy specific to f .¹⁵

15. For example, consider an instance $(f^*, \mathcal{F})$ being *not* self-identifiable, with its decoy $f_{\text{dec}} \in \mathcal{F}$ satisfying $f^*(a_{\text{dec}}) = f_{\text{dec}}(a_{\text{dec}}) = 0.5$. Now, suppose we discretize $f^*(a_{\text{dec}})$ using discretization step 0.1 and discretize $f_{\text{dec}}(a_{\text{dec}})$ using discretization step 0.2. After this modification, $f^*(a_{\text{dec}})$ and $f_{\text{dec}}(a_{\text{dec}})$ would no longer be equal, and self-identifiability could occur.

On a high level, we prove that decoys exist for “almost all” instances of all bandit structures that we consider (*i.e.*, linear, Lipschitz, polynomial, and quadratic). Therefore, the common case in all these bandit problems is that Greedy fails.

For contextual bandits (CB), our findings are more nuanced. Linear CB satisfy identifiability when the context set is sufficiently diverse (which is consistent with prior work), but admit decoys (as a somewhat common case) when the context set is “low-dimensional”. In contrast, existence of decoys is the common case for Lipschitz CB. One interpretation is that self-identifiability requires both context diversity and a parametric reward structure which enables precise “global inferences” (*i.e.*, inferences about arms that are far away from those that have been sampled).

In what follows, we present each structure in a self-contained way, interpreting it as special case of our framework. Since our presentation focuses on best-arm-unique reward functions, our examples are focused similarly (except those for Linear CB). Throughout, let $[y, y']_\varepsilon$ be a uniform discretization of the $[y, y']$ interval with step $\varepsilon > 0$, namely: $[y, y']_\varepsilon := \{\varepsilon \cdot n \in [y, y'] : n \in \mathbb{N}\}$. Likewise, we define $(y, y')_\varepsilon := \{\varepsilon \cdot n \in (y, y') : n \in \mathbb{N}\}$.

D.1 (Discretized) linear bandits

Linear bandits is a well-studied variant of bandits [Auer, 2002, Abe et al., 2003, Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010].¹⁶ Formally, it is a special case of StructuredMAB defined as follows. Arms are real-valued vectors: $\mathcal{A} \subset \mathbb{R}^d$, where $d \in \mathbb{N}$ is the dimension. Reward functions are given by $f_\theta(a) = a \cdot \theta$ for all arms a , where $\theta \in \Theta \subset \mathbb{R}^d$. The parameter set Θ is known to the algorithm, so the function class is $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. The true reward function is $f^* = f_{\theta^*}$ for some $\theta^* \in \Theta$. (Fixing Θ , we interpret θ^* as a “problem instance”).

Linear bandits, as traditionally defined, let Θ be (continuously) infinite, *e.g.*, a unit ℓ_1 -ball, and sometimes consider an infinite (namely, convex) action set. Here, we consider a “discretized” version, whereby both Θ and $\mathcal{A}$ are finite. Specifically, $\Theta = ([-1, 1]_\varepsilon \setminus \{0\})^d$, *i.e.*, all parameter vectors in $[0, 1]^d$ with discretized non-zero coordinates. Action set $\mathcal{A}$ is an arbitrary finite subset of $[-1, 1]^d$ containing the hypercube $\{-1, 1\}^d$.¹⁷ Note that each reward function f_θ , $\theta \in \Theta$ has a unique best arm $a_\theta^* = \text{sign}(\theta) := (\text{sign}(\theta_i) : i \in [d]) \in \{-1, 1\}^d$.

We prove that linear bandits has a decoy for “almost all” problem instances.

Lemma 1. *Consider linear bandits with dimension $d \geq 2$, parameter set $\Theta = ([-1, 1]_\varepsilon \setminus \{0\})^d$, $\varepsilon \in (0, 1/4]$, and an arbitrary finite action set $\mathcal{A} \subset [-1, 1]^d$ containing the hypercube $\{-1, 1\}^d$. Consider an instance $\theta^* \in \Theta$ such that $\|\theta^*\|_1 - 2 \min_{i \in [d]} |\theta_i^*| \geq d\varepsilon$. Then θ^* has a decoy in Θ .*

Proof. Let $j \in [d]$ be a coordinate with the smallest $|\theta_j^*|$. Choose arm $a_{\text{dec}} \in \{-1, 1\}^d$ with $(a_{\text{dec}})_i = \text{sign}(\theta_i^*)$ for all coordinates $i \neq j$, and flipping the sign for $i = j$, $(a_{\text{dec}})_j = -\text{sign}(\theta_j^*)$. Note that $\langle \theta^* | a_{\text{dec}} \rangle = \|\theta^*\|_1 - 2 \min_{i \in [d]} |\theta_i^*| \in [d\varepsilon, d]$.

Now, for any given $\alpha \in [d\varepsilon, d]_\varepsilon$ and any sign vector $v \in \{-1, 1\}^d$, there is $\theta \in \Theta$ such that $\|\theta\|_1 = \alpha$ and its signs are aligned as $\text{sign}(\theta) = v$. Thus, there exists $\theta_{\text{dec}} \in \Theta$ such that $\|\theta_{\text{dec}}\|_1 = \langle \theta^* | a_{\text{dec}} \rangle$ and $\text{sign}(\theta_{\text{dec}}) = a_{\text{dec}}$. Note that a_{dec} is the best arm for θ_{dec} . Moreover, $\langle \theta_{\text{dec}} | a_{\text{dec}} \rangle = \|\theta_{\text{dec}}\|_1 = \langle \theta^* | a_{\text{dec}} \rangle < \|\theta^*\|_1 = \langle \theta^* | a^* \rangle$. So, θ_{dec} is a decoy for θ^* . $\square$

D.2 (Discretized) linear contextual bandits

Linear contextual bandits (CB) are studied since [Li et al., 2010, Chu et al., 2011, Abbasi-Yadkori et al., 2011]. Formally, it is a special case of StructuredCB defined as follows. Each context is a tuple $x = (x(a) \in \mathbb{R}^d : a \in \mathcal{A}) \in \mathcal{X} \subset \mathbb{R}^{d \times K}$, where $d \in \mathbb{N}$ is the dimension and $\mathcal{X}$ is the context set. Reward functions are given by $f_\theta(x, a) = x(a) \cdot \theta$ for all context-arm pairs, where $\theta \in \Theta \subset \mathbb{R}^d$ and Θ is a known parameter set. While Linear CB are traditionally defined with (continuously) infinite Θ and $\mathcal{X}$, we need both to be finite.

¹⁶. We consider *stochastic* linear bandits. A more general model of *adversarial* linear bandits is studied since Awerbuch and Kleinberg [2008], McMahan and Blum [2004], see Bubeck and Cesa-Bianchi [2012, Chapter 5] for a survey.

¹⁷. For ease of exposition, we relax the requirement that expected rewards must lie in $[0, 1]$.

Like in linear bandits, the function class is $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. The true reward function is $f^* = f_{\theta^*}$ for some $\theta^* \in \Theta$, which we interpret as a “problem instance”.

Remark 2. For this subsection, we do not assume best-arm-uniqueness, and instead rely on the version of our characterization that allows ties in (2.2), see Appendix A.

We show that self-identifiability holds when the context set is sufficiently diverse. Essentially, we posit that per-arm contexts $x(a)$ take values in some finite subset $S_a \subset \mathbb{R}^d$ independently across arms, and each S_a spans $\mathbb{R}^d$; no further assumptions are needed.

Lemma 2 (positive). *Consider linear CB with degree $d \geq 1$ and an arbitrary finite parameter set $\Theta \subset \mathbb{R}^d$. Suppose the context set is $\mathcal{X} = \prod_{a \in \mathcal{A}} S_a$, where $S_a \subset [-1, 1]^d$ are finite “per-arm” context sets such that each S_a spans $\mathbb{R}^d$. Then self-identifiability holds for all instances $\theta^* \in \Theta$.*

Proof. Fix some policy π . For a given context x , let $v(x) = x(\pi(x)) \in \mathbb{R}^d$ be the context vector produced by this policy. Let’s construct a set $\mathcal{X}_0 \subset \mathcal{X}$ of contexts such that $v(\mathcal{X}_0) := \{v(x) : x \in \mathcal{X}_0\}$, the corresponding set of context vectors, spans $\mathbb{R}^d$. Add vectors to $\mathcal{X}_0$ one by one. Suppose currently $v(\mathcal{X}_0)$ does not span $\mathbb{R}^d$. Then, for each arm $a \in \mathcal{A}$, the per-arm context set S_a is not contained in $\text{span}(v(\mathcal{X}_0))$; put differently, there exists a vector $v_a \in S_a \setminus \text{span}(v(\mathcal{X}_0)) \in \mathbb{R}^d$. Let $x = (x(a) = v_a : \forall a \in \mathcal{A}) \in \mathcal{X}$ be the corresponding context. It follows that $v(x) \notin \text{span}(v(\mathcal{X}_0))$. Thus, adding x to the set $\mathcal{X}_0$ increases $\text{span}(v(\mathcal{X}_0))$. Repeat this process till $v(\mathcal{X}_0)$ spans $\mathbb{R}^d$.

Thus, fixing expected rewards of policy π for all contexts in $\mathcal{X}_0$ gives a linear system of the form

$$v(x) \cdot \theta^* = \alpha(x) \quad \forall x \in \mathcal{X}_0,$$

for some known numbers $\alpha(x)$ and vectors $v(x)$, $x \in \mathcal{X}_0$. Since these vectors span $\mathbb{R}^d$, this linear system completely determines θ^* . $\square$

Remark 3. In particular, Lemma 2 holds when the context set is a (very) small perturbation of one particular context x . For a concrete formulation, let $S(a) = \{x(a) + \varepsilon e_i : i \in [d]\}$ for each arm a and any fixed $\varepsilon > 0$, where e_i , $i \in [d]$ is the coordinate- i unit vector. This is consistent with positive results for Greedy in Linear CB with smoothed contexts [Kannan et al., 2018, Bastani et al., 2021, Raghavan et al., 2023], where “nature” adds variance- σ^2 Gaussian noise to each per-arm context vector. (Greedy achieves optimal regret rates which degrade as σ increases, e.g., $\mathbb{E}[R(T)] \leq \tilde{O}(\sqrt{T}/\sigma)$.) We provide a qualitative explanation for this phenomenon.

On the other hand, decoys may exist when the context set $\mathcal{X}$ is degenerate. We consider $\mathcal{X} = \prod_{a \in \mathcal{A}} S_a$, like in Lemma 2, but now we posit that the per-arm sets S_a do not span $\mathbb{R}^d$, even jointly. We prove the existence of a decoy under some additional conditions.

Lemma 3 (negative). *Consider linear CB with parameter set $\Theta = [-1, 1]_\varepsilon^d$, for some degree $d \geq 2$ and discretization step $\varepsilon \in (0, 1/2]$ with $1/\varepsilon \in \mathbb{N}$. Suppose the context set is $\mathcal{X} = \prod_{a \in \mathcal{A}} S_a$, where $S_a \subset [-1, 1]^d$ are the “per-arm” context sets. Assume $\text{span}(S_1, \dots, S_{K-1}) \subset \mathbb{R}^{d-1}$ and $S_K = \{(0, 0, \dots, 0, 1)\}$. Then any instance $\theta^* \in \Theta$ with $\theta_d^* = 1$ and $\|\theta^*\|_1 < 2$ has a decoy in Θ .*

Proof. Consider vector $\theta_{\text{dec}} \in \Theta$ such that it coincides with θ^* on the first $d - 1$ components, and $(\theta_{\text{dec}})_d = -1$. We claim that θ_{dec} is a decoy for θ^* .

To prove this claim, fix context $x \in \mathcal{X}$. Let a^*, a_{dec} be some optimal arms for this context under θ^* and θ_{dec} , respectively. Then $a_{\text{dec}} \in [K - 1]$. (This is because the expected reward $x(a) \cdot \theta^*$ of arm a is greater than -1 when $a \in [K - 1]$, and exactly -1 when $a = K$.) Similarly, we show that $a^* = K$. It follows that $x(a_{\text{dec}}) \cdot \theta_{\text{dec}} = x(a_{\text{dec}}) \cdot \theta^*$, since θ_{dec} and θ^* coincide on the first $K - 1$ coordinates, and the last coordinate of $x(a_{\text{dec}})$ is 0. Moreover $x(a_{\text{dec}}) \cdot \theta^* < 1 = x(a^*) \cdot \theta^*$. Putting this together, $x(a_{\text{dec}}) \cdot \theta_{\text{dec}} = x(a_{\text{dec}}) \cdot \theta^* < x(a^*) \cdot \theta^*$, completing the proof. $\square$

D.3 (Discretized) Lipschitz Bandits

Lipschitz bandits is a special case of StructuredMAB in which all reward functions $f \in \mathcal{F}$ satisfy Lipschitz condition, $|f(a) - f(a')| \leq \mathcal{D}(a, a')$, for any two arms $a, a' \in \mathcal{A}$ and some known metric $\mathcal{D}$ on $\mathcal{A}$. Introduced in Kleinberg et al. [2008], Bubeck et al. [2011], Lipschitz bandits have been studied extensively since then, see Slivkins [2019, Ch. 4.4] for a survey. The paradigmatic case is

continuum-armed bandits [Agrawal, 1995, Kleinberg, 2004, Auer et al., 2007], where one has action set $\mathcal{A} \subset [0, 1]$ and metric $\mathcal{D}(a, a') = L \cdot |a - a'|$, for some $L > 0$.

Lipschitz bandits, as traditionally defined, allow *all* reward functions that satisfy the Lipschitz condition, and hence require an infinite function class $\mathcal{F}$. To ensure finiteness, we impose a finite action set $\mathcal{A}$ and constrain the set of possible reward values to a discretized subset $\mathcal{R} = [0, 1]_\varepsilon$. We allow all Lipschitz functions $\mathcal{A} \rightarrow \mathcal{R}$. Further, we restrict the metric $\mathcal{D}$ to take values in the same range $\mathcal{R}$. We call this problem *discretized Lipschitz bandits*.

We show that “almost any” best-arm-unique reward function has a best-arm-unique decoy.

Lemma 4. *Consider discretized Lipschitz bandits, with range $\mathcal{R} = [0, 1]_\varepsilon$ and metric $\mathcal{D}$. Let $\mathcal{F}$ be the set of all best-arm-unique Lipschitz reward functions $\mathcal{A} \rightarrow \mathcal{R}$. Consider a function $f \in \mathcal{F}$ such that $0 < f(a) < f(a^*)$ some arm a . Then f has a decoy $f_{\text{dec}} \in \mathcal{F}$ (with decoy arm a).*

Proof. Define reward function f_{dec} by $f_{\text{dec}}(a') = \min(0, f(a) - \mathcal{D}(a, a'))$ for all arms $a' \in \mathcal{A}$. So, f_{dec} takes values in $\mathcal{R}$ and is Lipschitz w.r.t. $\mathcal{D}$ (since $\mathcal{D}$ satisfies triangle inequality); hence $f_{\text{dec}} \in \mathcal{F}$. Also, f_{dec} has a unique best arm a (since $f(a) > 0$ and the distance between any two distinct points is positive). Note that $f_{\text{dec}}(a) = f(a) < f(a^*)$, so f_{dec} is a decoy. $\square$

This result extends seamlessly to *Lipschitz contextual bandits (CB)* [Lu et al., 2010, Slivkins, 2014], albeit with somewhat heavier notation. Formally, Lipschitz CB is a special case of StructuredCB which posits the Lipschitz condition for all context-arm pairs: for each reward function $f \in \mathcal{F}$,

$$|f(x, a) - f(x', a')| \leq \mathcal{D}((x, a), (x', a')) \quad \forall x, x' \in \mathcal{X}, a, a' \in \mathcal{A}, \quad (\text{D.1})$$

where $\mathcal{D}$ is some known metric on $\mathcal{X} \times \mathcal{A}$. As traditionally defined, Lipschitz CB allow all reward functions which satisfy (D.1). We define *discretized Lipschitz CB* same way as above: we posit finite $\mathcal{X}, \mathcal{A}$, restrict the range of the reward functions and the metric to range $\mathcal{R} = [0, 1]_\varepsilon$, and allow all functions $f : \mathcal{X} \times \mathcal{A} \rightarrow \mathcal{R}$ which satisfy (D.1). Again, we show that “almost any” best-arm-unique reward function has a best-arm-unique decoy.

Lemma 5. *Consider discretized Lipschitz CB, with range $\mathcal{R} = [0, 1]_\varepsilon$ and metric $\mathcal{D}$. Let $\mathcal{F}$ be the set of all best-arm-unique Lipschitz reward functions $\mathcal{X} \times \mathcal{A} \rightarrow \mathcal{R}$. Consider a best-arm-unique function $f \in \mathcal{F}$ such that for some policy π we have $0 < f(x, \pi(x)) < f(x, \pi^*(x))$ for each context x . Then f has a best-arm-unique decoy $f_{\text{dec}} \in \mathcal{F}$ (with decoy policy π).*

Proof. Define reward function f_{dec} by $f_{\text{dec}}(x, a) = \min(0, f(x, \pi(x)) - \mathcal{D}((x, \pi(x)), (x, a)))$ for all context-arm pairs (x, a) . Like in the proof of Lemma 4, we see that f_{dec} takes values in $\mathcal{R}$ and is Lipschitz w.r.t. $\mathcal{D}$, hence $f_{\text{dec}} \in \mathcal{F}$. And it has a unique best arm $\pi(x)$ for each context x . Finally, $f_{\text{dec}}(x, \pi(x)) = f(x, \pi(x)) < f(x, \pi^*(x))$, so f_{dec} is a decoy. $\square$

D.4 (Discretized) polynomial bandits

Polynomial bandits [Huang et al., 2021, Zhao et al., 2023] is a bandit problem with real-valued arms and polynomial expected rewards.¹⁸ We obtain a negative result for “almost all” instances of polynomial bandits, and a similar-but-cleaner result for the special case of “quadratic bandits”.

We define polynomial bandits as a special case of StructuredMAB with action set $\mathcal{A} \subset \mathbb{R}$ and reward functions f are degree- p polynomials, for some degree $p \in \mathbb{N}$. Denote reward functions as $f = f_\theta$, where $\theta = (\theta_0, \dots, \theta_p) \in \mathbb{R}^{p+1}$ is the parameter vector with $\theta_p \neq 0$, so that $f_\theta(a) = \sum_{q=0}^p \theta_q \cdot a^q$ for all arms a . The function set is $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, for some parameter set Θ . Typically one allows continuously many actions and parameters, *i.e.*, an infinite reward structure.

We consider *discretized* polynomial bandits, with finite $\mathcal{A}$ and Θ . The action space is $\mathcal{A} = [0, 1/2]_\varepsilon$, for some fixed discretization step $\varepsilon \in (0, \frac{1}{2p})$. The parameter set Θ needs to be discretized in a more complex way, in order to guarantee that the function class contains a decoy. Namely,

$$\Theta = \prod_{q=0}^p [-1/q, 1/q]_{\delta(q)}, \text{ where } \delta(q) = \varepsilon^{p+1-q}.$$

18. Huang et al. [2021], Zhao et al. [2023] considered a more general formulation of polynomial bandits, with multi-dimensional arms $a \in \mathbb{R}^d$. It was also one of the explicit special cases flagged in Parys and Golrezaei [2024].

We “bunch together” all polynomials with the same leading coefficient θ_p . Specifically, denote $\Theta_\gamma = \{\theta \in \Theta : \theta_p = \gamma\}$ and $\mathcal{F}_\gamma = \{f_\theta : \theta \in \Theta_\gamma\}$, for $\gamma \neq 0$.

We focus on reward functions f_θ such that

$$a_\theta^* = \arg \max_{a \in \mathcal{A}} f_\theta \text{ is unique} \quad \text{and} \quad f_\theta(a_\theta^*) > \sup_{a \in (\max \mathcal{A}, \infty)_\varepsilon} f_\theta(a);$$

call such f_θ *well-shaped*. In words, the “best feasible arm in $\mathcal{A}$ ” is unique, and dominates any larger discretized arm.¹⁹ (We do not attempt to characterize which polynomials are well-shaped.)

We prove that “almost any” well-shaped function $f_\theta \in \mathcal{F}_\gamma$ has a well-shaped decoy in $\mathcal{F}_\gamma$, for any non-zero γ in some (discretized) range.²⁰ Here, “almost all” is in the sense that every non-leading coefficient of θ must be bounded away from the boundary by 5ε , namely: $\theta_q \in [-1/q + 5\varepsilon, 1/q - 5\varepsilon]$ for all $q \neq p$. Let $\Theta_\gamma^{\text{bdd}}$ be the set of all such parameter vectors $\theta \in \Theta_\gamma$. Moreover, we consider θ such that the best arm satisfies $a_\theta^* > \varepsilon$.

Lemma 6. *Consider discretized polynomial bandits, as defined above, for some degree $p \geq 2$ and discretization step $\varepsilon \in (0, \frac{1}{2p})$. Fix some non-zero $\gamma \in [-1/p, 1/p]_\varepsilon$. Then any well-shaped reward function $f_\theta \in \mathcal{F}_\gamma$ with $\theta \in \Theta_\gamma^{\text{bdd}}$ and $a_\theta^* > \varepsilon$ has a well-shaped decoy in $\mathcal{F}_\gamma$.*

Proof. Fix one such function f_θ . Consider a function $f_{\text{dec}} : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$f_{\text{dec}}(a) \equiv f_\theta(a + \varepsilon) - (f_\theta(a_\theta^*) - f_\theta(a_\theta^* - \varepsilon)), \quad \forall a \in \mathbb{R}. \quad (\text{D.2})$$

In the rest of the proof we show that f_{dec} is a suitable decoy.

First, we observe that $f_{\text{dec}} = f_{\theta_{\text{dec}}}$, where $\theta_{\text{dec}} \in \mathbb{R}^{p+1}$ is given by $(\theta_{\text{dec}})_p = \theta_p$,

$$\begin{aligned} (\theta_{\text{dec}})_q &= \sum_{i=q}^p \theta_i \binom{i}{q} \varepsilon^{i-q}, \quad \forall q = \{p-1, \dots, 1\}, \text{ and} \\ (\theta_{\text{dec}})_0 &= \sum_{i=0}^p \theta_i \varepsilon^i - (f_\theta(a_\theta^*) - f_\theta(a_\theta^* - \varepsilon)). \end{aligned}$$

Second, we claim that $\theta_{\text{dec}} \in \Theta_\gamma$. Indeed, the above equations imply that all coefficients of θ_{dec} are suitably discretized: $(\theta_{\text{dec}})_q \in (-\infty, \infty)_{\delta(q)}$ for all $q \in \{0, \dots, p-1\}$. It remains to show that they are suitably bounded; this is where we use $\theta \in \Theta^{\text{bdd}}$. We argue this as follows:

- Since $|\theta_q| \leq 1/q$ for all $q = 0, \dots, p$ and $\sum_{i=1}^p 1/(i!) < e \leq 3$, a simple calculation shows that $|(\theta_{\text{dec}})_q - \theta_q| \leq 3\varepsilon$ for each $q \in \{p-1, \dots, 1\}$.
- Since $|\theta_q| \leq 1/q$ for all $q = 0, \dots, p$ and $a \in [0, 1/2]$, a simple calculation shows that $f_\theta(a)$ is 2-Lipchitz on $\mathcal{A}$, so $|f_\theta(a_\theta^*) - f_\theta(a_\theta^* - \varepsilon)| \leq 2\varepsilon$, and moreover $|(\theta_{\text{dec}})_0 - \theta_0| \leq 3\varepsilon + 2\varepsilon = 5\varepsilon$.

Claim proved.

Third, we prove that f_{dec} is well-shaped and is a decoy for f_θ . Indeed, Eq. (D.2) and $a_\theta^* > \varepsilon$, combined with the well-shaped condition (1) $a_\theta^* = \arg \max_{a \in \mathcal{A}} f_\theta$ being unique and (2) $f_\theta(a_\theta^*) < \sup_{a \in (\max \mathcal{A}, \infty)_\varepsilon} f_\theta(a)$, imply that (1) $a_{\text{dec}}^* = a_\theta^* - \varepsilon \in \mathcal{A}$ is the unique best arm under f_{dec} , i.e., $\arg \max_{a \in \mathcal{A}} f_{\text{dec}}(a)$ and (2) $f_{\text{dec}}(a_{\text{dec}}^*) < \sup_{a \in (\max \mathcal{A}, \infty)_\varepsilon} f_{\text{dec}}(a)$, which means that f_{dec} satisfies the well-shaped condition. Moreover, we have

$$f_{\text{dec}}(a_{\text{dec}}^*) = f_\theta(a_{\text{dec}}^*) < f_\theta(a_\theta^*),$$

where the equality holds by (D.2), and the inequality holds by the uniqueness of a_θ^* . $\square$

19. Being well-shaped is a mild condition. A sufficient condition is as follows: $\arg \max_{a \in (\infty, \infty)_\varepsilon} f_\theta$ is unique and lies in $(0, 1/2]$. Note that even if $\arg \max_{a \in \mathbb{R}} f_\theta$ is non-unique or falls outside $(0, 1/2]$, it is still possible that f_θ is well-shaped, since $\arg \max_{a \in \mathbb{R}} f_\theta$ is not necessarily in $(-\infty, \infty)_\varepsilon$.

20. As a corollary, if we consider the function set consisting of all “well-shaped reward functions in $\mathcal{F}_\gamma$ ”, then “almost any” function in this function set has a decoy in the same function set.

D.5 (Discretized) quadratic bandits

Quadratic bandits is a special case of polynomial bandits, as defined in Appendix D.4, with degree $p = 2$. Quadratic bandits (in a more general formulation, with multi-dimensional arms $a \in \mathbb{R}^d$) have been studied, as an explicit model, in Shamir [2013], Huang et al. [2021], Yu et al. [2023]. We obtain a similar negative guarantee as we do for polynomial bandits – “almost any” problem instance has a decoy – but in a cleaner formulation and a simpler proof.

Let’s use a more concrete notation: reward functions are $f_{\gamma,\mu,c}$ with

$$f_{(\gamma,\mu,c)}(a) = \gamma(a - \mu)^2 + c,$$

where the leading coefficient $\gamma < 0$ determines the shape (curvature) of the function and the other two parameters $\mu, c \in [0, 1]$ determine the location of the unique global maximum (i.e., (μ, c)).

Discretization is similar, but slightly different. The action space is $\mathcal{A} = [0, 1]_\varepsilon$, for some fixed discretization step $\varepsilon \in (0, 1/2]$. The parameter space Θ , i.e., the set of feasible (γ, μ, c) tuples, is defined as $\gamma \in [-1, -0.5]_\varepsilon$, $\mu \in [0, 1]_\varepsilon$ and $c \in [0, 1]_{\varepsilon^3}$. Note that $\mu \in \mathcal{A}$, so any function $f_{(\gamma,\mu,c)}$ has a unique optimizer at $a = \mu \in \mathcal{A}$.

We focus on function space $\mathcal{F}_\gamma := \{f_{(\gamma,\mu,c)} : (\gamma, \mu, c) \in \Theta\}$, grouping together all functions with the same leading coefficient γ . We prove that “almost any” function in $\mathcal{F}_\gamma$ has a decoy in $\mathcal{F}_\gamma$.

Lemma 7. *Consider discretized quadratic bandits, for some fixed discretization step $\varepsilon \in (0, 1/2]$. Fix any leading coefficient $\gamma \in [-1, -0.5]_\varepsilon$. Then for any reward function $f^* = f_{(\gamma,\mu,c)} \in \mathcal{F}_\gamma$, it has a decoy $f_{\text{dec}} \in \mathcal{F}_\gamma$, as long as μ, c are bounded away from 0: $\mu \geq \varepsilon$ and $c \geq |\gamma|\varepsilon^2$.*

Proof. Consider reward function $f_{\text{dec}} = f_{(\gamma, \mu-\varepsilon, c+\gamma\varepsilon^2)}$. Since $\mu \geq \varepsilon$ and $c \geq |\gamma|\varepsilon^2$, it follows that $f_{\text{dec}} \in \mathcal{F}_\gamma$. Let us prove that f_{dec} is a decoy for f^* . Note that $\mu - \varepsilon$ is a suboptimal action for f^* and is the optimal action for f_{dec} . Finally, it is easy to check that $f^*(\mu - \varepsilon) = \gamma\varepsilon^2 + c = f_{\text{dec}}(\mu - \varepsilon)$. $\square$

E StructuredMAB characterization: Proof of Theorem 1

E.1 StructuredMAB Success: Proof of Theorem 1(a)

Let us fix the time horizon t and show the bound on the expected regret $\mathbb{E}[R(t)]$. Recall that $\bar{r}_t(a)$ as the empirical mean for arm a and that $N_t(a)$ is the number of times arm a pulled up to round t . Also recall the greedy algorithm is minimizing the following loss function each round:

$$\text{MSE}_t(f) := \sum_{a \in [K]} N_t(a) (\bar{r}_t(a) - f(a))^2.$$

Lemma 8. *Define $\beta(n) = \sqrt{\frac{2}{n} \log \frac{\pi^2 K n^2}{3\delta}}$. With probability $1 - \delta$:*

$$\forall a, \tau, |\bar{r}_\tau(a) - f^*(a)| < \beta(N_\tau(a)).$$

Proof. This lemma is a standard Hoeffding plus union bound, this exact form has appeared in Jun et al. [2018]. $\square$

In the following we shall always assume the event in the previous lemma holds and choose $\delta = 1/t$.

Lemma 9. *Assume the event in Lemma 8, then we have the upper bound on MSE_τ for each round $\tau \in [t]$.*

$$\text{MSE}_\tau(f^*) \leq K \cdot O(\log t).$$

Proof. Note that under the event from the previous lemma, we have for each arm:

$$\begin{aligned} N_\tau(a) (\bar{r}_\tau(a) - f^*(a))^2 &\leq N_\tau(a) \cdot \beta^2(N_\tau(a)) \\ &\leq O(\log t). \end{aligned}$$

Then, summing over all arms completes the proof. $\square$

Lemma 10. Assume the event in Lemma 8. The number of times any suboptimal arm is chosen cannot exceed T' rounds, where T' is some parameter with $T' = (K/\Gamma(f^*)^2) \cdot O(\log t)$.

Proof. We prove this by contradiction. Consider any round τ during which some suboptimal arm a is chosen above this threshold T' . The reward for arm a is going to get concentrated within $O(\Gamma(f^*))$ to $f^*(a)$, in particular:

$$|\bar{r}_\tau(a) - f^*(a)| < \Gamma(f^*)/2.$$

Take any reward vector f' such that $f'(a) \neq f(a)$. By the definition of class-gap, we have:

$$|f'(a) - f^*(a)| \geq \Gamma(f^*),$$

hence

$$|\bar{r}_\tau(a) - f^*(a)| \geq \Gamma(f^*)/2.$$

Then the cumulative loss

$$\text{MSE}_\tau(f') \geq T' \cdot (\Gamma(f^*)/2)^2 = K \cdot \Omega(\log t).$$

Therefore any f' with $f'(a) \neq f(a)$ cannot possibly be minimizing $\text{MSE}_\tau(\cdot)$. That is to say, the reward vector f_τ minimizing $\text{MSE}_\tau(\cdot)$ must have $f_\tau(a) = f^*(a)$. Then, by self-identifiability, we precisely know that arm a is also a suboptimal arm for the reward vector f_τ . Hence, we obtain a contradiction, and arm a cannot possibly be chosen this round. $\square$

We complete the proof of Theorem 1(a) as follows. The regret incurred during the warmup data is at most T_0 . Fix any round $t > T_0$. After the warmup data, we know with probability $1 - 1/t$, any suboptimal arm can be pulled at most $(K/\Gamma(f^*)^2) \cdot O(\log t)$ times after the warmup data, and the regret is $(K/\Gamma(f^*)^2) \cdot O(\log t)$. With the remaining probability $1/t$, the regret is at most $O(t)$. Hence, the theorem follows.

E.2 StructuredMAB Failure: Proof of Theorem 1(b)

Recall the two events are defined as

$$E_1 = \{ |\bar{r}_{\text{warm}}(a) - f_{\text{dec}}(a)| < \Gamma(f_{\text{dec}})/2 \text{ for each arm } a \neq a_{\text{dec}} \}.$$

$$E_2 = \{ \forall t > T_0, |\bar{r}_t(a_{\text{dec}}) - f^*(a_{\text{dec}})| \leq \Gamma(f_{\text{dec}})/2 \}.$$

Lemma 11. Assume event E_1 and E_2 holds, then greedy algorithm only choose the decoy arm a_{dec} .

Proof. The proof is by induction. Assume by round t , the algorithm have only choose the decoy arm a_{dec} . Note that assuming event E_1 and E_2 holds, for any reward vector $f \neq f_{\text{dec}}$, we will have

$$|\bar{r}_t(a) - f_{\text{dec}}(a)| \leq |\bar{r}_t(a) - f(a)|,$$

with at least one inequality strict for one arm. Hence f_{dec} must (still) be the $\text{MSE}_t(\cdot)$ minimizer, and a_{dec} will be chosen in the next round. $\square$

Lemma 12. Event E_1 happens with probability at least $\left[\frac{\Gamma(f_{\text{dec}})}{\sqrt{2\pi\sigma^2}} \exp(-2/\sigma^2) \right]^{K-1}$.

Proof. The random variable $\bar{r}_{\text{warm}}(a)$ is a gaussian variable with mean $f^*(a)$ and variance σ^2 . It has a distribution density at x with the following form

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x - f^*(a))^2/(2\sigma^2)).$$

For any x in the interval $[f_{\text{dec}}(a) - \Gamma(f_{\text{dec}})/2, f_{\text{dec}}(a) + \Gamma(f_{\text{dec}})/2]$, by boundedness of mean reward, we have

$$|x - f^*(a)| \leq 2.$$

Then, the density of x at any point on the interval $[f_{\text{dec}}(a) - \Gamma(f_{\text{dec}})/2, f_{\text{dec}}(a) + \Gamma(f_{\text{dec}})/2]$ is at least

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-2/\sigma^2).$$

Therefore, for any arm a , we have the following.

$$\Pr[|\bar{r}_{\text{warm}}(a) - f_{\text{dec}}(a)| \leq \Gamma(f_{\text{dec}})/2] \geq \frac{\Gamma(f_{\text{dec}})}{\sqrt{2\pi\sigma^2}} \exp(-2/\sigma^2).$$

Since the arms are independent, it follows that event E_1 happens with probability

$$\left[\frac{\Gamma(f_{\text{dec}})}{\sqrt{2\pi\sigma^2}} \exp(-2/\sigma^2) \right]^{K-1}. \quad \square$$

Lemma 13. For some appropriately chosen $\sigma = \Theta(\Gamma(f_{\text{dec}}))$, we have event E_2 happens with probability at least

$$\Pr\{E_2\} \geq 0.9.$$

Proof. Denote the bad event

$$E_3 = \{\exists t > T_0, |\bar{r}_t - f_{\text{dec}}(a_{\text{dec}})| > \Gamma(f_{\text{dec}})/2\},$$

which is the complement of E_2 . We will obtain an upper bound on E_3 , therefore a lower bound on E_2 . Note that event E_2 (and E_3) is only about the decoy arm a_{dec} , and recall that $f^*(a_{\text{dec}}) = f_{\text{dec}}(a_{\text{dec}})$.

By union bound,

$$\begin{aligned} \Pr[E_3] &\leq \sum_{t=1}^T \Pr[|\bar{r}_t - f_{\text{dec}}(a_{\text{dec}})| > \Gamma(f^*)/2] \\ &\leq 2 \sum_{t=1}^T \exp(-t\Gamma(f_{\text{dec}})^2/\sigma^2) \\ &\leq 2 \exp(-\Gamma(f_{\text{dec}})^2/\sigma^2) / (1 - \exp(-\Gamma(f_{\text{dec}})^2/\sigma^2)). \end{aligned}$$

Here, the second inequality is by a standard Hoeffding bound, and the last inequality is by noting that we are summing a geometric sequence.

Then, we can choose some suitable σ with $\sigma = \Theta(\Gamma(f_{\text{dec}}))$ ensures $\Pr[E_3] < 0.1$ and that $\Pr[E_2] > 0.9$. $\square$

Lemma 14. For some appropriately chosen $\sigma = \Theta(\Gamma(f_{\text{dec}}))$, we have the following lower bound:

$$\Pr[E_1 \cap E_2] \geq [\Omega(\exp(-2/\sigma^2))]^{K-1}$$

Proof. Note that event E_1 and E_2 are independent, then, the probability of $E_1 \cap E_2$ can be obtained from the previous two lemmas. $\square$

Theorem 1(b) directly follows from the above lemmas.

F StructuredCB characterization: Proof of Theorem 2

We start with a proof sketch, and proceed with full proofs.

Part (a). Directly applying the proof technique from the MAB case results in a regret bound that is linear in $|\Pi| = K^X$. Instead, we apply a potential argument and achieve regret bound that is polynomial in KX . First, by a standard concentration inequality, we upper-bound the loss for f^* as $\text{MSE}_t(f^*) \leq \tilde{O}(KX)$ with high probability. Then, we use self-identifiability to argue that if in some round t of the main stage some suboptimal policy π is chosen, there must exist some context-arm pair $(x, \pi(x))$ that is “under-explored”: appeared less than $\tilde{O}(XK/\Gamma^2(f^*))$ times. This step carefully harnesses the structure of the contextual bandit problem. Finally, we introduce a well-designed potential function (see Lemma 18) that tracks the progress of learning over time. This

function increases whenever a suboptimal policy is executed on an under-explored context-action pair, allowing us to bound the total number of times any suboptimal policy is executed. A key challenge is that while the second step guarantees the existence of an “under-explored” context-arm pair, it does not ensure that the context actually appears when the associated policy is chosen. We address this using a supermartingale argument and the fact that each context arrives with probability at least p_0 in each round. Combining these steps, we upper-bound the expected number of times Greedy selects a suboptimal policy, and we bound the final expected regret via the regret decomposition lemma.

Part (b). As in the MAB case, we define event E_1 to ensure that the warm-up data misidentifies f_{dec} as the true reward function, and event E_2 that the empirical rewards of the decoy policy are tightly concentrated. The definitions are modified to account for contexts:

$$E_1 = \left\{ \left| \bar{r}_{\text{warm}}(x, a) - f^\dagger(x, a) \right| < \Gamma(f^\dagger)/2 \text{ for each } x \in \mathcal{X} \text{ and arm } a \neq \pi^\dagger(x) \right\}, \quad (\text{F.1})$$

$$E_2 = \left\{ \left| \bar{r}_t(x, \pi^\dagger(x)) - f^*(x, \pi^\dagger(x)) \right| < \Gamma(f^\dagger)/2 \text{ for each } x \in \mathcal{X} \text{ and round } t > T_0 \right\}. \quad (\text{F.2})$$

A *decoy* context-arm pair (x, a) is one with $a = \pi_{\text{dec}}(x)$. E_1 concerns the single warm-up sample for each non-decoy context-arm pair. E_2 asserts that the empirical rewards are concentrated for all decoy context-arm pairs (and all rounds throughout the main stage). The two events are independent, as they concern non-overlapping sets of context-arm pairs. Greedy always chooses the decoy arm when E_1, E_2 happen. To lower-bound $\Pr[E_1 \cap E_2]$, invoke independence, analyze each event separately.

F.1 StructuredCB Success: Proof of Theorem 2(a)

Recall $N_t(x, a)$ as the number of times that context x appears and arm a was chosen up until round t . Also recall the greedy algorithm is finding the function f that minimize the following function each round: $\text{MSE}_t(f) = \sum_{x,a} N_t(x, a) (\bar{r}_t(x, a) - f(x, a))^2$.

Let us fix any $t \in \mathbb{N}$. We will show the upper bound on the expected regret as stated in the theorem.

Lemma 15. Fix any $\delta \in (0, 1)$. Let $\beta(n) = \sqrt{\frac{2}{n} \log \frac{\pi^2 X K n^2}{3\delta}}$. Then with probability at least $1 - \delta$,

$$\forall x, a, s, |\bar{r}_s(x, a) - f(x, a)| \leq \beta(N_s(x, a)).$$

Proof. The proof is similar to that of Lemma 8 in the previous section, which is a Hoeffding-style concentration bound with a union bound. We can simply treat each context-arm pair (x, a) as an arm, and this directly yields the result. $\square$

In the following, we shall assume the event in the previous lemma holds, and choose $\delta = 1/t$.

Lemma 16. Assume the event in Lemma 15 holds. For any round $s \in [t]$, the cumulative loss at the true underlying function, $\text{MSE}_s(f^*)$, can be upper bounded as $|\mathcal{X}|K \cdot O(\log t)$.

Proof. We observe that

$$\begin{aligned} \text{MSE}_s(f^*) &= \sum_{x \in \mathcal{X}, a \in [K]} N_s(x, a) (\bar{r}_s(x, a) - f^*(x, a))^2 \\ &\leq \sum_{x,a} O(\log(|\mathcal{X}|Kt)) \\ &\leq |\mathcal{X}|K \cdot O(\log(|\mathcal{X}|Kt)). \end{aligned} \quad \square$$

Lemma 17. Assume the event in Lemma 15 holds. Fix any round s . Let T' be some suitably chosen parameter and $T' = |\mathcal{X}|K/\Gamma(f^*)^2 \cdot O(\log(|\mathcal{X}|Kt))$. Suppose Greedy executes some suboptimal policy π in round s . Then there exists context x , such that $N_s(x, \pi(x)) < T'$.

Proof. We prove this by contradiction. Suppose that a suboptimal policy π is executed at round s , and further suppose that for all context x , we have $N_s(x, \pi(x)) \geq T'$.

By the previous lemma, we have $\forall x$,

$$|\bar{r}_t(x, \pi(x)) - f^*(x, \pi(x))| < \beta(T_0) < \Gamma(f^*)/2.$$

Consider any function f such that:

$$\exists x : f(x; \pi(x)) \neq f^*(x; \pi(x)). \quad (\text{F.3})$$

By the definition of the class gap,

$$|f(x, a) - f^*(x, \pi(x))| \geq \Gamma(f^*),$$

and then

$$|f(x, a) - \bar{r}_t(x, \pi(x))| \geq \Gamma(f^*)/2.$$

Then, the term $\text{MSE}_s(f)$ can be lower bounded:

$$\text{MSE}_s(f) \geq T' \cdot (\Gamma(f^*)/2)^2 \geq |\mathcal{X}|K \cdot O(\log(|\mathcal{X}|Kt)).$$

Hence, any function satisfying Eq. (F.3) cannot possibly minimize $\text{MSE}_s(\cdot)$. In other words, the function minimizing the loss at this step f_t must satisfy

$$f_t(x; \pi(x)) = f^*(x; \pi(x)).$$

Finally, the self-identifiability condition precisely tells us the policy π must be suboptimal for f_t and hence cannot be executed at round t . We obtain a contradiction, and the lemma is proven. $\square$

Lemma 18. *Conditional on the event in Lemma 15, the expected total number of times of suboptimal policy execution is no larger than $|\mathcal{X}|KT'/p_0 = (|\mathcal{X}|K/\Gamma(f^*))^2/p_0 \cdot O(\log t)$.*

Proof. Define the potential function as

$$M_s = \sum_{x,a} \min(N_s(x, a), T').$$

Consider any round s that a suboptimal policy π is executed, by the previous lemma, there exists a context arm pair $(x, \pi(x))$ such that $N_s(x, \pi(x)) < T'$. With probability at least p_0 , such a context x will arrive, and M_s will increase by 1. Therefore, whenever a suboptimal policy is executed, with probability at least p_0 , we will have $M_{t+1} = M_t + 1$.

Let us use the indicator variable I_s to denote whether a suboptimal policy is executed in round s . Then M_s forms a supermartingale:

$$\mathbb{E}[M_s | M_{s-1}] \geq M_{s-1} + p_0 I_s.$$

Since we have that deterministically $M_t < |\mathcal{X}|KT'$, we know that the total number of times of suboptimal policy execution $N_t = \sum_{s=1}^t I_s$ satisfies

$$\mathbb{E}[N_t] = \mathbb{E}\left[\sum_{s=1}^t I_s\right] \leq \mathbb{E}\left[\sum_{s=1}^t (\mathbb{E}[M_s | M_{s-1}] - M_{s-1})\right]/p_0 < |\mathcal{X}|KT'/p_0.$$

Hence, the total number of suboptimal policies pull is upper bounded as desired. $\square$

Proof of Theorem 2(a). The regret incurred in the warmup phase is at most T_0 . With probability $1 - 1/t$, the number of suboptimal policy pulls can be bounded as in the lemma above. With the remaining $1/t$ probability the regret is at most $O(t)$. Finally, by the regret decomposition lemma (Lemma 4.5 in Lattimore and Szepesvári [2020]), we have

$$\begin{aligned} \mathbb{E}[R(t)] &\leq T_0 + |\mathcal{X}|KT'/p_0 + O(1) \\ &\leq T_0 + (|\mathcal{X}|K/\Gamma(f^*))^2/p_0 \cdot O(\log t) \end{aligned}$$

F.2 StructuredCB Failure: Proof of Theorem 2(b)

Recall the two events

$$\begin{aligned} E_1 &= \{ |\bar{r}_{\text{warm}}(x, a) - f_{\text{dec}}(x, a)| < \Gamma(f_{\text{dec}})/2 \text{ for each } x \in \mathcal{X} \text{ and arm } a \neq \pi_{\text{dec}}(x) \}, \\ E_2 &= \{ |\bar{r}_t(x, \pi_{\text{dec}}(x)) - f^*(x, \pi_{\text{dec}}(x))| < \Gamma(f_{\text{dec}})/2 \text{ for each } x \in \mathcal{X} \text{ and round } t > T_0 \}. \end{aligned}$$

Lemma 19. Assume event E_1 and E_2 holds. Then the greedy algorithm only executes the decoy policy π_{dec} .

Proof. We prove this by induction. Assume up until round t the greedy algorithm only executes π_{dec} . Consider any other function $f \neq f_{\text{dec}}$. Then we must have

$$\forall x, a |\bar{r}_t(x, a) - f(x, a)| \geq |\bar{r}_t(x, a) - f_{\text{dec}}(x, a)|.$$

And the inequality is strict for at least one (x, a) pair. Hence f_{dec} is (still) the reward function minimizing MSE in round t , and the policy π_{dec} will be executed. $\square$

Lemma 20. Event E_1 happens with probability at least $\Omega(\frac{\Gamma(f_{\text{dec}}) \exp(-2/\sigma^2)}{\sigma})^{|\mathcal{X}|K}$.

Proof. The proof is similar to the counterpart in multi-arm bandits. Note that $\bar{r}_{\text{warm}}(x, a)$ is gaussian distributed with variance σ^2 . We can obtain a lower bound by directly examining the distribution density of a gaussian. $\square$

Lemma 21. For some suitable chosen $\sigma = \Theta(\Gamma(f_{\text{dec}})/\sqrt{\ln(X)})$, event E_2 happens with probability 0.9.

Proof. Similar to the proof for multi-arm bandits, define the event

$$E_3 = \{ \exists t, x, |\bar{r}_t(x, \pi_{\text{dec}}(x)) - f_{\text{dec}}(x, \pi_{\text{dec}}(x))| \geq \Gamma(f_{\text{dec}})/2 \}$$

which is the complement of event E_2 . By a union bound,

$$\begin{aligned} \Pr[E_3] &= |\mathcal{X}| \sum_{t=1}^{\infty} \exp(-t\Gamma^2(f_{\text{dec}})/\sigma^2) \\ &\leq |\mathcal{X}| \exp(-\Gamma^2(f_{\text{dec}})/\sigma^2) / (1 - \exp(-\Gamma^2(f_{\text{dec}})/\sigma^2)) \end{aligned}$$

Choosing some suitable $\sigma = \Theta(\Gamma(f_{\text{dec}})/\sqrt{\ln X})$ ensures $\Pr[E_3] < 0.1$, and consequently $\Pr[E_2] > 0.9$. $\square$

Lemma 22. We have the following lower bound:

$$\Pr[E_1 \cap E_2] \geq \left[\log(|\mathcal{X}|) \exp(-2(\log |\mathcal{X}|)^2 / \Gamma(f_{\text{dec}})) \right]^{|\mathcal{X}|K}.$$

Proof. Note that event E_1 and E_2 are independent, hence the lemma follows by the previous two lemmas. $\square$

Theorem 2(b) now follows from the above lemmas.

G DMSO characterization: Proof of Theorem 3

Recall that $\Delta \ell_t(M)$ is the change in log-likelihood for model M in round t , as per (5.4). Note that

$$\Delta \ell_t(M) - \Delta \ell_t(M') = \log \left(\Pr_{M(\pi_t)}(r_t, o_t) / \Pr_{M'(\pi_t)}(r_t, o_t) \right) \in [-\log B, \log B].$$

The equality is by (5.4), and the inequality is by Assumption 1 (and this is how this assumption is invoked in our analysis). We use the notation $\sigma_0 = \log |B|$ in what follows.

G.1 DMSO Success: Proof of Theorem 3(a)

The below lemma bounds the number of times any suboptimal policy can be executed.

Lemma 23. *Let π° be any suboptimal policy. Fix $\delta \in \left(0, \frac{1}{|\mathcal{M}|}\right)$. With probability at least $1 - |\mathcal{M}|\delta$, the policy π° can be executed for at most $O\left(\sigma_0^2 \Gamma^{-2} \ln(1/\delta)\right)$ rounds.*

Proof. Let $\mathcal{M}^*(\pi^\circ)$ be the class of models whose optimal policy is π° . We show that after π° has been executed for T' rounds, any model in $\mathcal{M}^*(\pi^\circ)$ cannot be the MLE maximizer with probability $1 - |\mathcal{M}|\delta$. Let $Y_t(M)$ be the difference in log-likelihood of M^* and M in the t -th round:

$$Y_t(M) = \Delta \ell_t(M^*) - \Delta \ell_t(M).$$

Note that $Y_t(M)$ is a random variable where randomness comes from random realizations of reward-outcome pairs. $Y_t(M)$ can exhibit two types of behaviors:

1. $Y_t(M) = 0$, corresponding to the case where $M(\pi_t) \stackrel{d}{=} M^*(\pi_t)$ (i.e., models M and M^* coincide under π_t)
2. $Y_t(M)$ is a random sub-gaussian variable with variance $\leq \sigma_0^2$ and that $\mathbb{E}[Y_t] \geq \Gamma$.

Consider rounds s during which the policy π° is executed. Since π° is a suboptimal policy, during these rounds, we know that $Y_t(M)$ is of the second type for any M in $\mathcal{M}^*(\pi^\circ)$. That is to say, it is a subgaussian random variable with variance upper bounded by $O(\sigma_0^2)$, and that further

$$\mathbb{E}[Y_t(M)] = D_{\text{KL}}(M^*(\pi^\circ), \mathcal{M}^*(\pi^\circ)).$$

Since we have assumed π° is suboptimal for the true model M^* , we know that,

$$\mathbb{E}[Y_t(M)] \geq \Gamma.$$

Let $Z_t(M) = \sum_{\tau=1}^t Y_\tau(M)$.

$$\begin{aligned} \Pr[N_t(\pi^\circ) > T'] &\leq \Pr[\exists s, \pi \in \mathcal{M}^*(\pi^\circ), \text{ s.t. } Z_s(M(\pi)) \leq 0, N_s(\pi^\circ) = T'] \\ &\leq |\mathcal{M}|\delta. \end{aligned}$$

Here in the last line we choose $T' = O\left(\sigma_0^2 \Gamma^{-2} \ln(1/\delta)\right)$, completing the proof. $\square$

We complete the proof as follows. By a union bound, with probability $1 - |\mathcal{M}||\Pi|\delta$, the total number of rounds all suboptimal policy can be chosen is upper bounded by

$$O\left(|\Pi| \sigma_0^2 \Gamma^{-2} \ln(1/\delta)\right).$$

Choose $\delta = 1/(t|\Pi||\mathcal{M}|)$ and that $\log(1/\delta) = O(|\Pi|t)$, then with probability $1 - 1/t$, the total number of suboptimal policies executions can be upper bounded by

$$\frac{|\Pi|\sigma_0^2}{\Gamma^2} \cdot O(\log(|\Pi|t)).$$

G.2 DMSO Failure: Proof of Theorem 3(b)

In the subsequent discussion, we define $Q(E)$ as the probability of some event E occurring under the assumption that the data is generated by the decoy model M_{dec} (a hypothetical or ghost process). Similarly, we denote $P(E)$ as the probability of event E occurring under the assumption that the data is generated by M^* (the true process).

Recall that the two events are defined as follows.

$$E_1 = \{\forall M \in \mathcal{M}_{\text{other}} \quad \mathcal{L}(M_{\text{dec}} | \mathcal{H}_{\text{warm}}) > \mathcal{L}(M | \mathcal{H}_{\text{warm}})\}$$

and the event

$$E_2 := \left\{ \forall j > N_0/2, \forall M \in \mathcal{M}_{\text{other}}, \quad \sum_{i \in [j]} \Psi_i(M) \geq 0 \right\}.$$

Where we defined $\Psi_j(M) := \Delta \ell_{t(j)}(M_{\text{dec}}) - \Delta \ell_{t(j)}(M)$.

We first begin with the following concentration result. This result is stated in a general manner and not specific to our problem.

Lemma 24. Let $X_1, X_2, \dots$ be a sequence of random variables with $\mathbb{E}[X_i] > \Gamma$, and each is subgaussian with variance σ^2 . Then there exists some T' with $T' = \Theta(\sigma^2/\Gamma^2)$, such that with probability $1 - \delta$, for any $t > T' \cdot O(\log 1/\delta)$,

$$\sum_{\tau=1}^t X_\tau > 0.$$

Proof. This lemma is by a standard concentration with union bound. We perform the following bounds

$$\begin{aligned} \Pr \left[\exists t > T', \sum_{s=1}^t X_s > 0 \right] &\leq \sum_{t=T'}^{\infty} \Pr \left[\sum_{s=1}^t X_s > 0 \right] \\ &\leq \sum_{t=T'}^{\infty} \exp(-t\sigma^2/\Gamma^2) \\ &\leq \delta. \end{aligned}$$

Here the last line is by choosing a suitable $T' = \Theta(\sigma^2/\Gamma^2)$ and noticing we are summing a geometric sequence. $\square$

The below lemmas lower bound the probability that the likelihood of M_{dec} will be the unique highest after the warmup data (assuming under ghost process Q).

Lemma 25. We have a lower bound on E_1 under the ghost process:

$$Q(E_1) \geq 0.9.$$

Proof. Fix a model $M \neq M_{\text{dec}}$. There must exist at least one policy π that discriminates M and M_{dec} , in other words the distribution $M(\pi)$ and $M_{\text{dec}}(\pi)$ are different. Then, the expected change in log-likelihood of M_{dec} is at least $\Gamma(M_{\text{dec}})$ greater than that of M for each time a sample or policy π is observed:

$$\Phi_t(M, M_{\text{dec}}) \geq \Gamma.$$

where we defined $\Phi_t(M, M_{\text{dec}})$ as per Eq. (5.6),

$$\Phi_t(M, M_{\text{dec}}) := \mathbb{E} [\Delta \ell_t(M_{\text{dec}} | \mathcal{H}_t) - \Delta \ell_t(M | \mathcal{H}_t)].$$

Moreover, we know that in each round, either $\Phi_t(M, M_{\text{dec}}) = 0$, or $\Phi_t(M, M_{\text{dec}})$ is a subgaussian random variable with mean greater than Γ . Further, during rounds when π is sampled, the latter will happen. The policy π is sampled for $N_0 = c_0 \cdot (\sigma_0/\Gamma)^2 \log(|\Pi||M|)$ times in the warmup phase. Then, by a standard concentration inequality

$$\Pr[\ell_{\text{warm}}(M_{\text{dec}}) - \ell_{\text{warm}}(M) < 0] < 0.1/|\mathcal{M}|.$$

Now, we take a union bound over all models $|\mathcal{M}|$, and we obtain a lower bound for event E_1 . $\square$

What remains is the to show a lower bound for the event E_2 . We do so in the below lemma.

Lemma 26. Event E_2 happens with probability at least 0.9.

Note that event E_2 is only about when π_{dec} is sampled. Since $M_{\text{dec}}(\pi_{\text{dec}}) = M(\pi_{\text{dec}})$, the ghost process coincides with the true process.

Proof. Fix some model M . If the distribution for $M(\pi_{\text{dec}})$ and $M_{\text{dec}}(\pi_{\text{dec}})$ were the same, then Ψ_j would be 0 for any j . Hence we can assume $M(\pi_{\text{dec}})$ and $M_{\text{dec}}(\pi_{\text{dec}})$ are two different distributions. Then Ψ_j would be a subgaussian random variable with $\mathbb{E}[\Psi_j] > \Gamma(M_{\text{dec}})$. By the previous Lemma 24, the event E_2 holds specifically for model M with probability at least $1 - \delta/|M|$. Then, by a union bound, event E_2 holds with probability $1 - \delta$. $\square$

Lemma 27. If event E_1 and event E_2 holds, then only π_{dec} is executed.

Proof. The proof is by induction. Clearly after the warmup phase, the policy π_{dec} is executed. Now suppose up until round t the policy π_{dec} is executed, by event E_2 , the model M_{dec} remains the log-likelihood maximizer, and hence π_{dec} will still be chosen next round. $\square$

Proof of Theorem 3(b). Now let P be the true underlying process for which data is actually generated according to true model M^* . Recall $D_\infty(M_{\text{dec}}(\pi)|M^*(\pi)) \leq \log B$. Then on the warmup data consisting of $|\Pi|N_0$ samples, the density ratio of the ghost process and true process is bounded by $B^{|\Pi|N_0}$. Therefore, the probability of event E_1 can be bounded as follows.

$$P(E_1) \geq Q(E_1)/B^{|\Pi|N_0}.$$

And after warmup GreedyMLE only choose π_{dec} by event E_2 . Hence, the final probability lower bound of always choosing π_{dec} after the warmup is $\Omega(B^{-|\Pi|N_0})$.

H StructuredMAB with an Infinite Function Class: Proof of Theorem 4

H.1 Success: Proof of Theorem 4(a)

The proof of Theorem 1(a) carries over with a modified version of Lemma 10. Thus, it suffices to state and prove this modified lemma.

Lemma 28. *Let $\mathcal{F}$ be an infinite function class. Assume the event in Lemma 8. The number of times any suboptimal arm is chosen cannot exceed T' rounds, where T' is some parameter with $T' = (K/\varepsilon^2) \cdot O(\log t)$.*

Proof. We prove this by contradiction. Consider any round τ during which some suboptimal arm a is chosen above this threshold T' . The reward for arm a is going to get concentrated within $O(\varepsilon)$ to $f^*(a)$, in particular:

$$|\bar{r}_\tau(a) - f^*(a)| < \varepsilon/2.$$

Take any reward vector f' such that $|f'(a) - f(a)| > \varepsilon$. Then we have

$$|\bar{r}_\tau(a) - f(a)| \geq \varepsilon/2.$$

Then the cumulative loss

$$\text{MSE}_\tau(f') \geq T' \cdot (\varepsilon/2)^2 = K \cdot \Omega(\log t).$$

Therefore any f' with $|f'(a) - f(a)| \geq \varepsilon$ cannot possibly be minimizing $\text{MSE}_\tau(\cdot)$. That is to say, the reward vector f_τ minimizing $\text{MSE}_\tau(\cdot)$ must have $|f_\tau(a) - f^*(a)| < \varepsilon$. Then, by the strong notion of ε -self-identifiability, we precisely know that arm a is also a suboptimal arm for the reward vector f_τ . Hence, we obtain a contradiction, and arm a cannot possibly be chosen this round. $\square$

H.2 Failure: Proof of Theorem 4(b)

The proof of Theorem 4(b) follows the same structure as Theorem 1(b) (see Appendix E.2), with a key modification: although we still aim to show that Greedy becomes permanently stuck on a_{dec} with constant probability, the regression oracle may no longer return f_{dec} exactly—its output may fluctuate around f_{dec} due to reward noise and the continuity of $\mathcal{F}$. Our key insight is that, under suitable probabilistic events, these fluctuations do not change the greedy decision: the regression output may differ slightly from f_{dec} , but the resulting action remains a_{dec} .

Let us define the following events:

$$E_1 = \left\{ \left| \bar{r}_{\text{warm}}(a) - (f_{\text{dec}}(a) - \varepsilon/(2\sqrt{K})) \right| < \varepsilon/(4\sqrt{K}) \quad \text{for each arm } a \neq a_{\text{dec}} \right\},$$

$$E_2 = \left\{ \forall t > T_0, |\bar{r}_t(a_{\text{dec}}) - f^*(a_{\text{dec}})| \leq \varepsilon/(4\sqrt{K}) \right\}.$$

These events mirror (3.2) and (3.3), but with two changes: (1) we replace the confidence radius $\Gamma(f_{\text{dec}})/2$ by $\varepsilon/(4\sqrt{K})$, and (2) shift the baseline value of $f_{\text{dec}}(a)$ by $-\varepsilon/(2\sqrt{K})$ in E_1 .

We begin with three probabilistic lemmas.

Lemma 29. Event E_1 happens with probability at least $\left[\frac{\varepsilon}{2\sqrt{2\pi\sigma^2 K}} \exp(-2/\sigma^2) \right]^{K-1}$.

Proof. The random variable $\bar{r}_{\text{warm}}(a)$ is a gaussian variable with mean $f^*(a)$ and variance σ^2 . It has a distribution density at x with the following form

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x - f^*(a))^2/(2\sigma^2)).$$

For any x in the interval $[f_{\text{dec}}(a) - 3\varepsilon/(4\sqrt{K}), f_{\text{dec}}(a) - \varepsilon/(4\sqrt{K})]$, by boundedness of mean reward, we have

$$|x - f^*(a)| \leq 2.$$

Then, the density of x at any point on the interval $[f_{\text{dec}}(a) - 3\varepsilon/(4\sqrt{K}), f_{\text{dec}}(a) - \varepsilon/(4\sqrt{K})]$ is at least

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-2/\sigma^2).$$

Therefore, for any arm a , we have the following.

$$\Pr \left[\left| \bar{r}_{\text{warm}}(a) - (f_{\text{dec}}(a) - \varepsilon/(2\sqrt{K})) \right| < \varepsilon/(4\sqrt{K}) \right] \geq \frac{\varepsilon}{2\sqrt{2\pi\sigma^2 K}} \exp(-2/\sigma^2).$$

Since the arms are independent, it follows that event E_1 happens with probability

$$\left[\frac{\varepsilon}{2\sqrt{2\pi\sigma^2 K}} \exp(-2/\sigma^2) \right]^{K-1}. \quad \square$$

Lemma 30. For some appropriately chosen $\sigma = \Theta(\varepsilon/\sqrt{K})$, we have event E_2 happens with probability at least

$$\Pr\{E_2\} \geq 0.9.$$

Proof. Denote the bad event

$$E_3 = \left\{ \exists t > T_0, |\bar{r}_t - f_{\text{dec}}(a_{\text{dec}})| > \varepsilon/(4\sqrt{K}) \right\},$$

which is the complement of E_2 . We will obtain an upper bound on E_3 , therefore a lower bound on E_2 . Note that event E_2 (and E_3) is only about the decoy arm a_{dec} , and recall that $f^*(a_{\text{dec}}) = f_{\text{dec}}(a_{\text{dec}})$ by the definition of a decoy.

By union bound,

$$\begin{aligned} \Pr[E_3] &\leq \sum_{t=1}^T \Pr \left[|\bar{r}_t - f_{\text{dec}}(a_{\text{dec}})| > \varepsilon/(4\sqrt{K}) \right] \\ &\leq 2 \sum_{t=1}^T \exp(-t\varepsilon^2/(4\sigma^2 K)) \\ &\leq 2 \frac{\exp(-\varepsilon^2/(4\sigma^2 K))}{(1 - \exp(-\varepsilon^2/(4\sigma^2 K)))}. \end{aligned}$$

Here, the second inequality is by a standard Hoeffding bound, and the last inequality is by noting that we are summing a geometric sequence.

Then, we can choose some suitable σ with $\sigma = \Theta(\varepsilon/\sqrt{K})$ ensures $\Pr[E_3] < 0.1$ and that $\Pr[E_2] > 0.9$. $\square$

Lemma 31. For some appropriately chosen $\sigma = \Theta(\varepsilon/\sqrt{K})$, we have the following lower bound:

$$\Pr[E_1 \cap E_2] \geq [\Omega(\exp(-2/\sigma^2))]^{K-1}$$

Proof. Since event E_1 concerns all $a \neq a_{\text{dec}}$ and event E_2 concerns a_{dec} , we know that events E_1 and E_2 are independent. As a result,

$$\begin{aligned}\Pr[E_1 \cap E_2] &= \Pr[E_1] \Pr[E_2] \\ &\geq 0.9 \Pr[E_1] \\ &= \Omega \left(\left[\frac{\varepsilon}{2\sqrt{2\pi(\varepsilon^2/K)K}} \exp(-2/\sigma^2) \right]^{K-1} \right) \\ &= [\Omega(\exp(-2/\sigma^2))]^{K-1},\end{aligned}$$

where we utilize the previous two lemmas. $\square$

Having obtained the previous three probabilistic lemmas, we now prove a crucial lemma which is an extension of Lemma 11 to the infinite $\mathcal{F}$ setting. At this point, the key insight introduced in Section 6 plays a central role in the proof.

Lemma 32. *Assume event E_1 and E_2 holds, then greedy algorithm only choose the decoy arm a_{dec} .*

Proof. The proof is by induction. Assume by round t , the algorithm have only choose the decoy arm a_{dec} . Note that assuming event E_1 and E_2 holds. Consider the reward function f_t^{emp} given by the empirical means: $f_t^{\text{emp}}(a) = \bar{r}_t(a)$ for all $a \in \mathcal{A}$. By the induction assumption, $\bar{r}_t(a) = \bar{r}_{\text{warm}}(a)$ for each arm $a \neq a_{\text{dec}}$. Hence, by the definition of E_1 and E_2 , we have

$$\|f_t^{\text{emp}} - f_{\text{dec}}\|_2 \leq \sqrt{\sum_{a \in \mathcal{A}} \left(\frac{3\varepsilon}{4\sqrt{K}} \right)^2} = \frac{3\varepsilon}{4}.$$

Since f_{dec} is an ε -interior with respect to $\mathcal{F}$, we have $f_t^{\text{emp}} \in \mathcal{F}$.

Clearly, $f_t^{\text{emp}} \in \mathcal{F}$ is the unique minimizer of $\text{MSE}_t(\cdot)$. To see this, for any reward vector $f \neq f_t^{\text{emp}}$, we will have

$$|\bar{r}_t(a) - f_t^{\text{emp}}| = 0 \leq |\bar{r}_t(a) - f(a)|,$$

with at least one inequality strict for one arm. Hence the regression oracle will choose $f_t = f_t^{\text{emp}}$.

Although f_t^{emp} is not the same as f_{dec} , its optimal action is a_{dec} when E_1 and E_2 happen. This is because

$$\begin{aligned}f_t^{\text{emp}}(a_{\text{dec}}) &\geq f^*(a_{\text{dec}}) - \frac{\varepsilon}{4\sqrt{K}} \\ &= f_{\text{dec}}(a_{\text{dec}}) - \frac{\varepsilon}{4\sqrt{K}} \\ &= (f_{\text{dec}}(a_{\text{dec}}) - \frac{\varepsilon}{2\sqrt{K}}) + \frac{\varepsilon}{4\sqrt{K}} \\ &> f_t^{\text{emp}}(a)\end{aligned}$$

for all $a \neq a_{\text{dec}}$, where the first inequality follows from the definition of E_2 , the first equality follows from the definition of a decoy, and the last inequality follows from the definitions of E_1 and f_t^{emp} . $\square$

Theorem 4(b) directly follows from the above lemmas.